Breast Cancer Diagnosis and Prognosis Using Triple Hybrid Deep Learning Approach

Sanitha P S[1], Jayakrishnan B[2]

[1]PG Student, Dept. of Computer Science and Engineering, Mangalam college of Engineering, Kerala Technological University, Kottayam, Kerala, India
[2]Assistant Professor, Dept. of Computer Science and Engineering., Mangalam College of Engineering, Kerala Technological University, Kottayam, Kerala, India

Abstract

Breast cancer is one of the malignancies that affects women. Breast cancer is a condition that is brought on by abnormal breast cells that multiply and form tumours. If left untreated, tumours have the capacity to grow throughout the body and become fatal. Early initiation and thorough completion of treatment is associated with better outcomes and greater patient tolerance for breast cancer patients. These days, early detection of breast cancer is quite helpful and will help the women who battle the illness. The earliest detection of breast cancer can be successfully achieved with the use of machine learning-based approaches. Breast cancer can be diagnosed with great accuracy using a number of machine learning techniques, including CNN, RF, SVM, NB, KNN, AB, and others. Thus, I am introducing a triple hybrid deep learning method for breast cancer diagnosis and prognosis. This is the CNN, GRU, and LSTM combination.

Keywords: Machine Learning, Triple hybrid deep learning, CNN, GRU, LSTM

1. INTRODUCTION

One of the most serious cancers that affect women is breast carcinoma. It ranks among the most frequent causes of death due to cancer in females. A biopsy is a medical procedure that involves taking a sample of breast cells for testing purposes. Other recommended tests to detect breast cancer include breast ultrasounds, breast inspections performed by a doctor, mammograms, and breast magnetic resonance imaging (MRI). Of these testing methods, the biopsy is the one that can conclusively identify breast cancer. Following the testing of all these medical procedures, radiologists look at the results, talk with the doctors about them, and determine if the cells are malignant or not. Studies show that radiologists diagnose cancer incorrectly between 10% and 30% of the time when doing screening studies. Errors resulting from misinterpreting breast cancer symptoms account for 52% of errors, whereas failure to detect abnormal scan symptoms accounts for 43% of errors [4]. Due to this mistake rate, many benign tumours require biopsies, which puts the patient through needless expenditure and discomfort. Mistakes resulting from improper classification of mammography have a significant financial cost. This is because false negative results from screening mammography are a serious problem because early detection can drastically reduce treatment costs, delays, and effectiveness. On the other hand, if the illness is identified early, patients can avoid pointless treatments.

Thus, without the assistance of radiologists or specialists, researchers developed a Computer-Aided Diagnosis (CAD) that can swiftly and accurately identify tumours. A common application of artificial intelligence is machine learning, which uses data and expertise to forecast illness. To improve performance, it makes use of several strategies, such as statistical, a probabilistic approach and optimisation. Numerous studies have been conducted about the automated detection of breast carcinoma. The World Health Organisation reports that BC is the most common malignancy in women to receive a diagnosis. According to a survey, the most skilled physicians and specialists can identify breast cancer with accuracy of 79%, whereas machine learning approaches have a 91% accuracy rate and can diagnose diseases far more accurately than traditional methods [2]. As a result, machine learning has a big chance to lower the death rate. Many machine learning techniques, such as AdaBoost, Support Vector Machine, Naive Bayes, Knearest neighbours, Random Forest, etc., are employed in the detection of breast cancer. Nonetheless, this study included eleven machine learning classifiers to find the most suitable model for forecasting cancer of breasts.

Using various machine learning algorithms yields better results for breast cancer detection and prediction. However, when using various machine learning algorithms, one factor that affects the system's performance is the choice of data sets and how well they are selected. Thus, one of the primary issues with data sets lacking feature selection is that they also contain a large amount of unwanted data. I utilised the chi square feature selection strategy to solve this problem, and I also used a triple hybrid deep learning approach for greater accuracy and outcome.

2. LITERATURE REVIEW

A combination of classifiers can be used, according to Usman Naseem et al. [1], to automatically identify breast cancer and determine its prognosis. This employs the four-machine learning-based classifiers NB, DT, SVM and LR to diagnose breast cancer early. This study used a set of machine learning algorithms to forecast and diagnose breast cancer. A thorough evaluation of the efficacy of various ensemble machine learning and machine learning-based classifiers is also provided here. Here, various sampling techniques were also assessed as a way to address the issue of class imbalance in datasets. It is also shown that the suggested approach to breast cancer detection performs better than a number of cutting-edge techniques. and analysis is carried out both with and without sampling approaches; these are the research contributions of this work. This paper's research gap is that feature selection is one of the most crucial factors based on the dataset. They make no mention of feature selection algorithms in this paper.

In [2], This study uses eleven algorithms and a range of machine learning approaches, such as cross-validation, hyperparameter adjusting, and feature optimisation techniques including PCA and feature scaling. The classifiers are RF, SVM, NCC, NB, MP, LR, DT, AB, KNN, GB, and VC. This work's research contribution is that they also developed a website that computes the result accurately using actual time inputs. The research gap in this work is that they do not consider self-developed data sets.

The dataset's characteristics, which comprised continuous, binary, normal, and associated data variables, were used by Yufan Feng et al. [3] to choose which machine learning methods to apply. The machine learning techniques of decision trees, random forests, logistic regression, multilayer perceptrons, and naive bayes for bagging were employed to generate prediction models. The research contribution of this work is the development of machine learning techniques to forecast survival of specific breast cancer in patients with MpBC, utilising a collection of data of 160 individuals with the disease with clinical, pathological, and biological features. Two different approaches to variable selection were used: While one calculates the direction and strength of the linear relationship among the two variables, the other evaluates qualities by looking at the gain ratio in regard to the class. In this method, refrain from including unrelated variables. or excessive. Lack of accurate therapy information is the research gap, and the small dataset raises the possibility of overfitting and skewed model evaluation. The inherent high expense of obtaining information pertaining to human subjects and wet lab research makes it difficult to access larger datasets. The study suggests a novel model in [4] which integrates SVM with the Improved AlexNet architecture to detect breast cancer. Because of their capacity to integrate, the Support Vector Machine (SVM) classifier and deep learning-based feature extraction are used in combination. The model's reliance on a single training and validation dataset, the use of three distinct optimizers, and the adoption of three distinct mammography picture sizes constitute the research gap in this instance. This study's objective is to automate the detection and classification of breast cancer in radiographs using a unique technique called Breast Net-SVM.

An already trained CNN-based DNN, AlexNet for processing thermogram-specific sub-networks, and a basic NN for processing clinical data are the components of the methodology Dennies Tsietso et al. [5] suggested. AlexNet trains more quickly because of its shallow architecture, particularly when working with multiple input images. This enhances classification performance by enabling the network to discover relationships between various thermograms. These are the contributions made by research towards the integration of transfer learning in order to lower computing costs and increase efficiency. Creation of a multi-input network that uses three breast images and clinical data to accurately classify patients as healthy or sick. Introducing an automated ROI extraction technique for effective symmetry analysis.

3. METHODOLOGY

The primary goal of this work is to create an automated machine learning system for breast cancer detection and prognosis. It will do this by utilising a triple hybrid deep learning architecture and a graphical user interface (GUI) that is easy to use and is written in Python Tkinter. Using a triple hybrid deep learning approach, breast cancer can be automatically detected and predicted based on this work. I can now detect and anticipate breast cancer with greater accuracy through this. In this instance, the issue with the existing methodology is likewise addressed through the application of the chi square feature selection method [1].

- **Data set**

The "Breast cancer Wisconsin Dataset," which is accessible to the public, is used to automatically diagnose and predict breast cancer. Data like id, diagnosis, radius_mean, texture_mean, area_mean, perimeter_mean, smoothness_mean, and so on are contained in it.

System Flow Diagram: The figure 1.a,1. b illustrates the suggested system flow diagram.

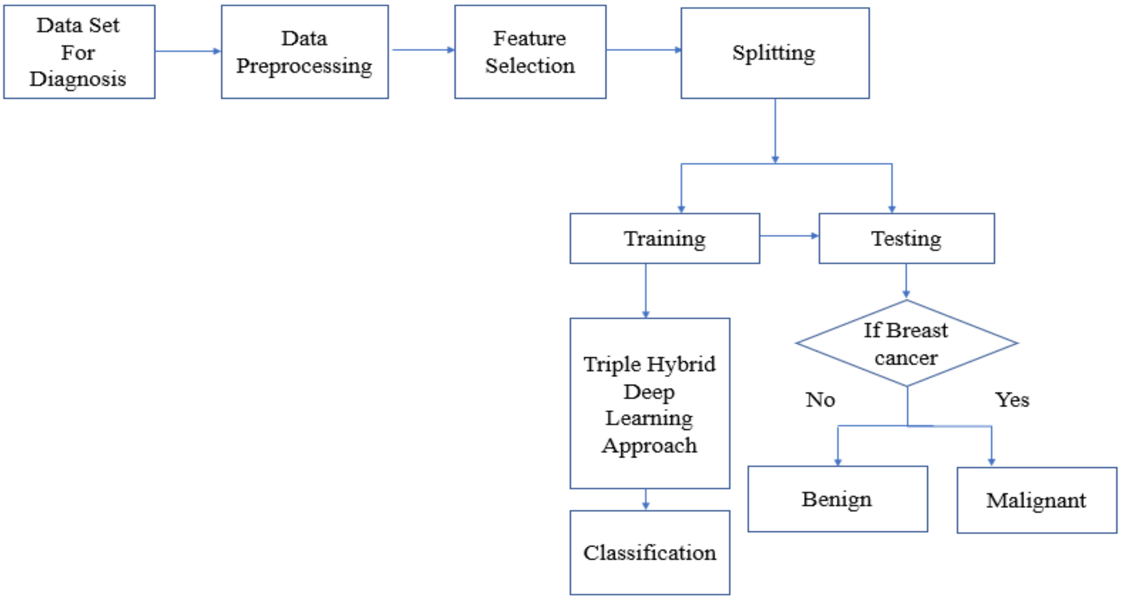

Fig.1.a: Proposed Model Flow Diagram for Diagnosis

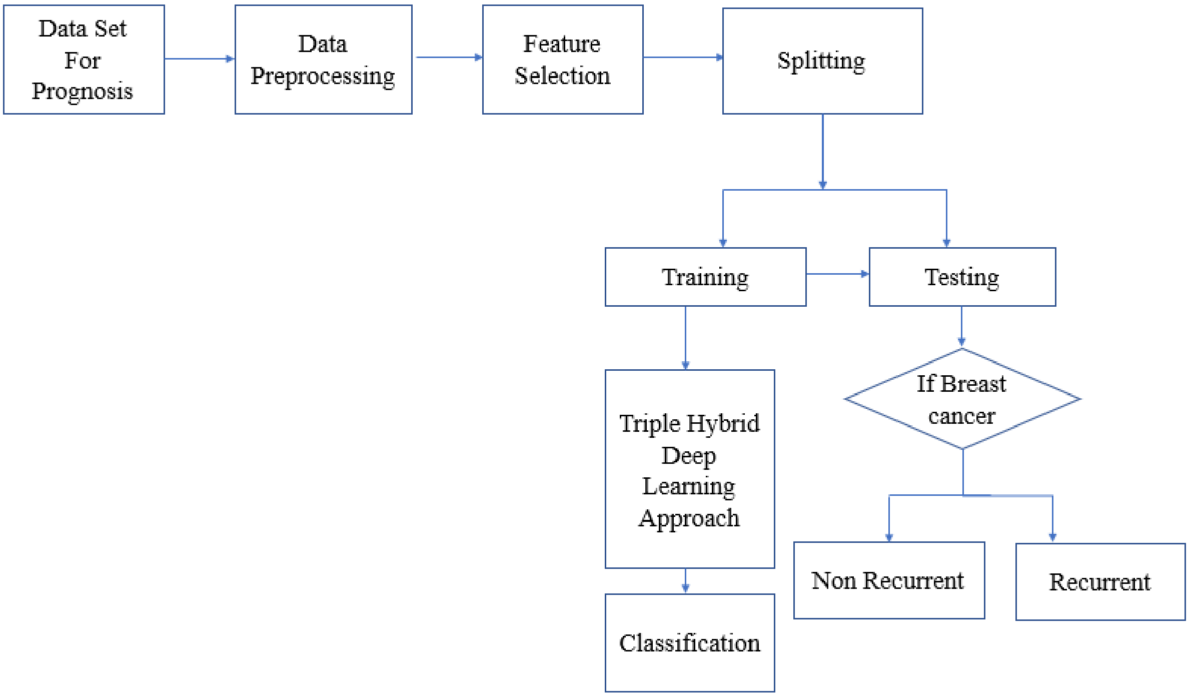

Fig.1.b: Proposed Model Flow Diagram for Prognosis

- **Data pre-processing**

Real-world data frequently contains mistakes, missing values, and may be in an unsuitable format that makes it impossible to apply machine learning models directly. The data pre-processing process, which also increases the accuracy and efficiency of the machine learning model, is required to clean the data and get it ready for a machine learning model.

The Data Preprocessing module is responsible for preparing the data for the machine learning model. This involves several important steps, including loading data from files into a suitable format (e.g., a pandas

Data Frame), filling in missing values through imputation or removal, and removing extraneous columns that don't increase the model's efficiency. This module will also standardize or normalize the data to increase model accuracy. Pandas functions to alter data and ensure that it is clean and ready for further processing will be written for each preprocessing task.

- **Feature Selection**

Only a small portion of the dataset's variables can be used to build the machine learning model; the remaining attributes are either unnecessary or redundant. Excessive and redundant attribute additions to the dataset may have a detrimental effect on the accuracy and general performance of the model. It is therefore crucial to identify and pick the best qualities from the data and to eliminate any unnecessary or excessive information. In this process, machine learning feature selection is helpful.

The Feature Selection module uses the Chi-square technique to extract the most relevant characteristics from the data. This method enhances the performance of the model by identifying the most important predictors of the target variable. This module will utilize the Chi-square test to prioritise features using chi2 and SelectKBest from scikit-learn. A function named select_features(data) will be implemented to automate this selection process and ensure that only the most relevant features are used for model training. The chi-square feature selection method is used in this investigation. The Chi-Square test is a crucial statistical method for analysing correlations in category data. Its uses are diverse and help researchers understand how different aspects relate to one another.

In the chi square feature selection procedure, select a k value. Based on this k value, select the attributes that are most valuable from the data. Next, to enhance output prediction, balance the data set. Complete the process in 570 data records to facilitate diagnosis. On 199 of the 570 total data points for the prognosis, carry out the complete process.

- **Train-Test Splitting**

The Train-Test Splitting module will divide the preprocessed data into training and testing sets. For training, most data will typically be split 80-20, with a portion left aside to evaluate the model's performance. This section ensures that the model can be effectively trained and assessed on untested data to gauge its capacity for generalization. The scikit-learn train_test_split technique and a random seed will be used to ensure split repeatability.

- **Triple Hybrid Deep Learning Approach**

Hybrid deep learning methodologies are more productive than machine learning techniques. This triple hybrid deep learning system uses convolutional neural networks (CNN), gated recurrent units (GRU), and long short-term memory (LSTM) to identify breast cancer autonomously. This mix of deep learning networks performs better than using a single deep learning model. Using a deep network trained on the training set, features are extracted, and the data is re-expressed using the retrieved features as input to a non-deep learning method trained to perform the final classification. This is the concept of the hybrid learning approach.

This article discusses a triple hybrid deep learning architecture with a focus on sequential data analysis. The first layer, called Convolutional Layer (Conv1D), has a 3-kernel size and 64 filters. Its purpose is to extract local patterns or features from the input sequences. The feature maps are then down sampled with a pool size of 2 using a MaxPooling Layer (MaxPooling1D) in order to preserve important information. Dropout layers are positioned following the max-pooling layer and periodically in between the recurrent layers in order to minimize overfitting by arbitrarily deactivating some input units during training. The architecture then alternates between layers of Long Short-Term Memory (LSTM) and Gated Recurrent

Units (GRU), both of which are configured to return sequences ('return sequences=True'), facilitating the propagation of output sequences to higher layers. LSTM and GRU units do a good job of capturing long-term dependencies and sequential patterns in data. Two thick layers follow the recurrent layers, learning complex patterns with the help of retrieved information. The final thick layer is suitable for binary classification applications since it employs a sigmoid activation function. Because it effectively models hierarchical representations of sequential data by integrating the advantages of convolutional, recurrent, and dense layers, this triple hybrid architecture is particularly well-suited for applications requiring the interpretation and prediction of sequential patterns. The core of the system is the Triple Hybrid Deep Learning Model, which integrates Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTM). Using CNNs for feature extraction, GRUs for sequential dependency management, and LSTMs for long-term dependency capture, this hybrid model leverages the benefits of each individual component. Built on top of TensorFlow/Keras, the model will have separate CNN, GRU, and LSTM layers that are combined to form a single architecture that can identify breast cancer and predict its prognosis.

- **Model Training**

The Model Training module is used to train the Triple Hybrid Deep Learning Model using the training set. The model will be put together with the appropriate optimizer, loss function, and assessment metrics before it is trained. The model.compile() and model.fit() functions from Keras will be used to set up and train the model, respectively. This module will have functions like train_model(model, X_train, y_train) to facilitate training and ensure that the model learns from the provided data in an effective way.

- **Performance Evaluation**

The Performance Evaluation module will assess the model's correctness and effectiveness once it has been trained using the testing data. The model will be used to evaluate the model's accuracy. The evaluation tools provided by scikit-learn and evaluate() will be employed to calculate additional metrics such as F1-score, precision, and recall. This thorough evaluation ensures that the model's performance is well understood and highlights both its advantages and potential improvement areas.

- **Model Saving**

The trained model's architecture and weights will be saved by the Model Saving module for later use. The Keras model.save() method will be used to save the trained model in the specified file directory. By ensuring that the model may be easily reloaded and utilised to generate predictions without the need for retraining, this feature improves the system's usability and efficacy.

- **Prediction**

In the Prediction module, users can enter new data for diagnosis and prognosis. This can be accomplished by loading the previously trained model, preprocessing the fresh data to conform to the training format, and then using model.predict() to generate predictions. The user will then see the expected results on the user interface (UI). A function named make_prediction(input_data) will manage the entire prediction pipeline, ensuring a seamless and user-friendly experience.

Breast Cancer Diagnosis: To identify breast cancer, a triple hybrid deep learning model is employed. This prediction shows the likelihood of cancer in the given sample.

Breast Cancer Prognosis: Breast cancer detection is accomplished by the use of the triple hybrid deep learning model. This forecast shows if the sample that was provided is recurring or not.

- **Integration in GUI**

All of these modules will eventually be integrated into the main program, ensuring that all of the components communicate with each other. The main script will manage the preprocessing of the data, train-test splitting, feature selection, model training, evaluation, saving, and prediction. With the help of straightforward controls and immediate feedback, users will be able to execute tasks with ease thanks to the GUI's improved design. The program will contain robust error management and user feedback systems to increase system dependability. Additionally, specific instructions.

4. RESULTS & ANALYSIS

In this section, describe the experiment's findings. The Breast cancer Wisconsin (Diagnosis) and Breast cancer Wisconsin (Prognosis) databases are used in this investigation. Accuracy was used as an evaluation metric in this study. Every result that is given is an average. Accuracy is defined as follows: Accuracy is one factor that's used to classify tasks that are efficient. Accuracy will forecast the correct output. It is calculated as the ratio of the classifiers' total number of predictions to the number of correct predictions.
Accuracy:
Accuracy=Number of correct prediction / All samples
It is clear from this that the deep learning model performs better when it comes to of prognosis and diagnosis.

Table 1: Proposed Model. Average accuracy (%) is reported as evaluation metric.

Machine Learning Model	Diagnosis	Prognosis
Triple Hybrid Deep Learning approach (CNN+GRU+LSTM)	96.50	88.52

As a result, the triple hybrid deep learning strategy (CNN+GRU+LSTM) achieves 88.52% prognostic accuracy and 96.50% diagnosis accuracy. The quantity and calibre of datasets used in training define how accurate machine learning algorithms are. The outcome will be predicted using the dataset.
Precision: Dividing the total number of true positives by the total number of false positives and true positives yields the precision formula.
Recall: Recall is calculated by dividing the total number of true positives by the sum of true positives and false negatives.
F1 score=2 * (Precision * Recall) / (Precision + Recall)
Support
The number of real instances of the class in the given dataset is known as support.

Table 2: Diagnosis Classification Report

Classification	Precision	Recall	F1-Score	Support
Benign	0.96	0.96	0.97	72
Malignant	0.97	0.96	0.96	71

Table 3: Prognosis Classification Report

Classification	Precision	Recall	F1-Score	Support
Non-Recurrent	0.84	0.93	0.88	28

Recurrent	0.93	0.85	0.89	33

5. CONCLUSION & FUTURE SCOPES

Breast cancer is a common disease that mostly affects women worldwide. Early detection of this condition is critical to lowering death rates and improving patient outcomes. As per a survey, machine learning techniques are more accurate in identifying breast cancer than even the most proficient doctors and specialists. Early discovery of BC will greatly improve the prognosis and chances of recovery because it may encourage patients to get surgery immediately away. Therefore, it is crucial to design a system that enables the medical community to accurately and quickly diagnose breast cancer. Because machine learning (ML) is so good at modelling important feature recognition from complicated BC datasets, it is frequently utilised for pattern classification of breast cancer (BC). The triple hybrid deep learning strategy for early detection and prognosis of breast cancer is described in this research study.

FUTURE SCOPES:

1. Multi-Modal Data Inclusion: By adding other data modalities such as genetic data, mammography images, or patient demographics, the breast cancer diagnostic and prognosis system's feature space can be significantly increased, and its overall model performance can be enhanced. Several data kinds are incorporated into the model to enable a more comprehensive understanding of the ailment and the patient's attributes. These data types also aid in the model's ability to detect subtle patterns and correlations that may not be evident when depending solely on one data source. Then, data from these several sources can be effectively integrated using fusion techniques such as early or late fusion, which take advantage of the complementary characteristics of multiple modalities. By combining information from mammography pictures for tumour diagnosis, genetic data for susceptibility assessment, and patient demographics for risk factor comprehension, the program builds a more complete picture of every case. As a result, forecasts become more accurate and personalized. This integrative approach allows medical professionals to treat patients with breast cancer with specialized interventions and improves the prognostic and diagnostic capacities of the system.

2. Integration with Electronic Health Records (EHR): Integrating the breast cancer diagnosis and prognosis system with electronic health record (EHR) systems is one of the most significant ways to enhance healthcare procedures, facilitate seamless data sharing, and foster professional communication among healthcare providers. By seamlessly connecting the system with EHR systems, clinicians may access patient data, including medical histories, diagnostic results, and treatment plans, inside a single interface. Through this interface, the requirement for manual data entry is eliminated, which not only boosts productivity but also ensures that comprehensive patient data is available to back educated decisions. When the diagnosis system and EHR are synchronized in real-time, healthcare providers may also keep an eye on patient progress, update records, and share findings with interdisciplinary teams more effectively.

3. Personalized treatment recommendations and patient risk stratification: Provide patients access to elements that will enable them to be grouped based on their risk profiles and receive personalized treatment recommendations. This may mean creating customized treatment regimens based on the results of prediction models that estimate the likelihood of different outcomes (like response to specific treatments or the risk of recurrence). The system might become a more comprehensive tool for personalized treatment with the addition of this feature, which would ultimately improve patient outcomes and care.

4. Using Real-Time Data Analysis: Provide forecasts and comments immediately by utilising real-time data analysis technologies. This can be especially useful in therapeutic scenarios where quick decision-making is essential. For example, by integrating real-time data from continuous patient monitoring or diagnostic equipment, the system can continuously update its projections and provide doctors with the most up-to-date information. Real-time analysis can increase the system's responsiveness and effectiveness in busy medical environments.

References

1. Usman Naseem, Junaid Rashid, Liaqat Ali, Jungeun Kim, Qazi Emad Ul Haq, Mazhar Javed Awan, And Muhammad Imran "An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers" in IEEE Access, vol. 10, pp. 78242- 78252, 2022, doi: 10.1109/ACCESS.2022.3174599**.**
2. Khandaker Mohammad Mohi Uddin, Nitish Biswas , Sarreha Tasmin Rikta , Samrat Kumar Dey "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique" in Computer Methods and Programs in Biomedicine Update ELESVIER.
3. Yufan Feng, Natasha McGuire, Alexandra Walton, AP-MBC Consortium, Stephen Fox, Antonella Papa, Sunil R. Lakhani , Amy E. McCart Reed "Predicting breast cancer-specific survival in metaplastic breast cancer patients using machine learning algorithms" in Journal of Pathology Informatics ELESVIER.
4. Jawad Ahmad, Sheeraz Akram, Arfan Jaffar, Muhammad Rashid, And Sohail Masood Bhatti "Breast Cancer Detection Using Deep Learning: An Investigation Using the DDSM Dataset and a Customized AlexNet and Support Vector Machine" in IEEE Access DOI: 10.1109/ACCESS.2023.3311892.
5. Dennies Tsietso, (Student Member, IEEE), Abid Yahya, (Senior Member, IEEE), Ravi Samikannu, (Senior Member, IEEE), Muhammad Usman Tariq, (Member, IEEE), Muhammad Babar, (Member, IEEE), Basit Qureshi, (Senior Member, IEEE), AND ANIS KOUBAA, (Senior Member, IEEE) "Multi-Input Deep Learning Approach for Breast Cancer Screening Using Thermal Infrared Imaging and Clinical Data" in IEEE Access 2023.
6. https://www.sciencedirect.com/science/article/pii/S1877050923001229
7. https://www.scaler.com/topics/deep-learning/gru-network/
8. https://blog.insightdatascience.com/automating-breast-cancer-detection-with-deep-learning-d8b49da17950
9. https://www.researchgate.net/publication/325064884_Machine_Learning_with_Applications_in_Breast_Cancer_Diagnosis_and_Prognosis
10. Eshav Tambre, Khushi Dhake, Needhi Patil, Dnyaneshwar Pawa R, Pranav Mahadik "Breast Cancer Detection Using Deep Learning" in International Journal of Research Publication and Reviews 2023.