SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I11P118

Volume 11 Issue 11, 213-227, November 2024
© 2024 Seventh Sense Research Group®

Original Article

Optimizing Breast Cancer Recurrence Forecasting Using ANOVA Feature Selection and GRU Models

Arathi Chandran R I[1], V. Mary Amala Bai[2]

[1]*Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, India*
[2]*Department of Information Technology, Noorul Islam Centre for Higher Education, Kumaracoil, India*

[1]*Corresponding Author : Arathi.Chandran.R.I@outlook.com*

Received: 20 September 2024 Revised: 19 October 2024 Accepted: 17 November 2024 Published: 03 December 2024

Abstract - *The challenge of breast cancer recurrence remains a critical concern, prompting the need for effective predictive models that improve patient outcomes. This study introduces a novel prediction model, addressing common issues like complex model structures, high-dimensional data, and class imbalance. The model combines a Gated Recurrent Unit (GRU) with Analysis of Variance (ANOVA)-based feature selection to boost accuracy and reliability. Using the Wisconsin Breast Cancer (WBC) dataset, the study applies preprocessing techniques to enhance data quality. ANOVA is employed to select relevant features, which are input into the GRU model. The GRU's multi-layer architecture successfully identifies complex patterns in the data. The model achieves impressive results, with a mean accuracy of 96.49%, precision of 97.04%, recall of 96.67%, and an F1-score of 96.67%. The confusion matrix and ROC curve analyses also validate the model's performance in predicting recurrence. This GRU-ANOVA approach is promising to improve breast cancer recurrence predictions, offering critical insights for clinical decision-making and patient care.*

Keywords *- Breast cancer recurrence prediction, Gated Recurrent Unit, Analysis of Variance (ANOVA), Feature optimization, Wisconsin Breast Cancer (WBC) dataset .*

1. Introduction

Breast cancer is a malignant tumor formed when cells in breast tissue proliferate uncontrolled, evading the standard regulatory processes governing cell division and apoptosis. While the exact cause of breast cancer remains unclear, several well-known risk factors have been identified that can help predict the likelihood of developing the disease. These include age, family history, and genetic predisposition, which are critical in assessing a woman's risk of breast cancer [1].

Additionally, specific diagnostic markers are correlated with more severe variants of breast cancer and an increased likelihood of recurrence. Factors such as larger tumor size, reduced hormone receptor expression for estrogen and progesterone, lymph node involvement, and higher histologic grade are key indicators of a more hostile form of the disease. Identifying the most significant prognostic markers can provide oncologists with valuable insights into the potential for breast cancer recurrence, hence facilitating improved treatment options [2]. While recurrence can happen at any time, it is most common within the first five years after initial treatment, and the likelihood of recurrence is closely tied to these prognostic markers.

For many patients, the terrifying possibility of a breast cancer recurrence is genuine; it signifies the cancer cells' return after the first therapy. Cancer can recur in two main ways: locally, where it grows back in the same spot where the initial tumor was, or distantly, where it travels to other parts of the body. A significant management issue for breast cancer is the possibility of this recurrence occurring months or even years after the initial course of treatment. While distant recurrence, sometimes referred to as metastatic breast cancer, denotes the spread of cancer cells outside of the breast and adjacent lymph nodes, local recurrence implies that some cancer cells may have escaped early therapy [3]. Both types of recurrence can have serious consequences for patients, requiring additional care and perhaps affecting their life span and overall expectancy.

After therapy, a local recurrence usually occurs in the vicinity of the initial tumor site and frequently requires surgery, particularly after lumpectomy. Breast cancer can return even after a mastectomy, especially if there is significant involvement of lymph nodes. Surgical intervention, radiation therapy, chemotherapy, and hormone therapy are possible treatments for local recurrence [4]. Even though it often spreads to organs, including the brain, liver, lungs, or bones, distant metastasis of breast cancer is still classified as breast cancer. Regular follow-up meetings with healthcare experts, imaging tests (MRIs, CT scans, or mammograms), and blood tests to look for any oddities or changes are usually part of the monitoring process for signs of recurrence [5]. In order to enhance outcomes and survival rates for patients with breast cancer, early diagnosis of recurrence is essential for immediate intervention and care. Predicting the recurrence of breast cancer is critical to early detection, individualized treatment programs, better patient outcomes, effective resource management, and patient empowerment. Better treatment outcomes and prompt action are made possible by early identification. Patient care is optimized through customized treatment programs based on risk assessments [6]. By assisting in proactive treatment, predictive models lessen the negative effects on a patient's quality of life. Care for individuals at elevated risk is given priority in the efficient use of resources. Making informed decisions promotes patient empowerment. All things considered, recurrence prediction is critical to improving the role of patients in medical care, utilization of resources, and results [7]. To address this, the proposed research introduces an innovative model that combines ANOVA feature selection and GRU-based deep learning for accurate recurrence forecasting. This approach not only optimizes feature selection but also enhances prediction accuracy and reliability. The main aspects of the proposed research are outlined below:

- Introduces a new model designed specifically for predicting breast cancer recurrence with better accuracy and reliability.
- Presents a novel method for feature optimization that improves the precision of breast cancer recurrence detection.
- Provides a comprehensive comparison of the proposed method against current methodologies. This comparative analysis highlights the strengths and advantages of the new model, demonstrating its superior performance and potential benefits over existing approaches in breast cancer recurrence prediction.

The paper proceeds as follows: Section 2 reviews existing methods relevant to the current study. The proposed model is presented in Section 3. Section 4 showcases the experimental findings and subsequent discussion. Finally, Section 5 outlines the conclusions drawn from the study.

2. Related Works

Hussein et al. [8] examined BRCA1 oncoprotein expression in invasive ductal carcinoma of the breast, highlighting its relevance for prognosis and therapy. Given breast cancer's complexity and hormone dependence, accurate assessment of BRCA1 along with Estrogen Receptor (ER), Progesterone Receptor (PR), and Human epidermal growth factor receptor 2 (Her2/neu) was crucial. The study analyzed 83 paraffin-embedded samples from patients diagnosed between January 1, 2010, and March 13, 2012, using immunohistochemistry with the Ventana Benchmark system. Results showed BRCA1 expression in 20.5% of cases, with higher levels linked to advanced tumor grades and stages. Although negative BRCA1 expression generally correlated with negative ER, PR, and Her2/neu statuses, no significant associations were found with these markers or patient age. The findings underscored BRCA1's potential as a prognostic marker for aggressive tumors. Liu et al. [9] utilized the Shapley Additive Explanations (SHAP) method to develop a clinical decision assistance tool addressing model opacity concerns. Their analysis of data from 1,629 patients identified key variables affecting recurrence, leading to a highly accurate prediction model with AUC scores of 0.96 for Random Forest and 0.97 for Extra Trees. This transparency in decision-making enhanced trust in the model's recommendations within clinical settings. González-Castro et al. [10] focused on enhancing 5-year recurrence estimates by integrating structured and unstructured data from 823 breast cancer patients. They created three feature sets—organized, unorganized, and mixed—and evaluated five ML methods, with XGBoost performing best (accuracy = 0.900, recall = 0.907, F1-score = 0.897, AUROC = 0.807). Their findings revealed that structured data provided the most accurate results, with unstructured data performing slightly worse and blended data being less effective.

Howard et al. [11] presented a deep learning model that combined computerized histology and clinical threat variables to predict recurrence risk. Their model outperformed a traditional clinical nomogram, achieving an AUROC of 0.83 in a sample cohort compared to 0.76 in an external validation cohort ($p = 0.0005$), demonstrating its superior predictive power. Zeng et al. [12] used unstructured Electronic Health Record (EHR) data to build ML models predicting breast cancer recurrence post-surgery. Analyzing data from 1,841 patients with histopathological reports and medical records, they applied LSTM, XGBoost, and SVM algorithms. The LSTM model performed highest in both training (accuracy, F1 score) and testing cohorts, indicating its robust predictive capability.

Othman et al. [13] introduced a hybrid DL model combining copy number alteration, gene expression, and clinical data from the METABRIC dataset. Their model, utilizing CNN for feature extraction and GRU and LSTM for classification, attained an impressive accuracy of 98.0% with decision fusion, outperforming other methods and setting a new standard for predictive accuracy. Lulu Wang (2023) [14] examined advances in microwave imaging for breast cancer screening, highlighting its non-ionizing, non-invasive, and economical characteristics. The study emphasized the shortcomings of traditional imaging techniques such as X-ray mammography and ultrasound, suggesting microwave imaging as a more advantageous option. Microwave imaging exhibited improved accuracy and efficiency in tumor identification by integrating machine learning techniques.

Rajasekaran and Ram [15] proposed a hybrid LSTM-XGBoost model combined with Linear Discriminant Analysis

(LDA) for feature extraction. This system, utilizing hyperparameter tuning and cross-validation, demonstrated higher accuracy and efficiency in breast cancer prediction than existing approaches. Su et al. [16] developed the Breast Cancer Recurrence Network (BCR-Net), which forecasts the risk of recurrence using histopathology slides. BCR-Net achieved 68.9% and 71.1% accuracy for low and high-risk predictions on H&E slides, with an overall AUC of 0.775 for H&E and 0.811 for Ki67 slides, demonstrating its effectiveness in risk stratification with less computational overhead. Yao et al. [17] created a multi-modal DL model integrating clinical data, gene expression, and H&E-stained histopathology images. Their model, which divided tumor areas into image blocks and encoded them into 1D feature vectors, combined visual attributes with clinical and gene expression data. This approach yielded an AUC of 0.75, outperforming models that used only H&E images or clinical data.

Existing research on breast cancer recurrence prediction highlights several limitations that need addressing to enhance model performance. One significant challenge is the need for more robust and precise predictive models. Although numerous researchers have utilized Machine Learning (ML) and Deep Learning (DL) approaches using various data modalities—such as clinical information, histopathological images, and gene expression data—these models often fall short regarding prediction accuracy, efficiency and scalability. Current models frequently focus on specific types of data, which can limit their overall effectiveness. A more comprehensive approach integrating multiple data types could yield more reliable and applicable predictions for real-world scenarios. Moreover, many of these models have yet to be validated on larger, more diverse datasets, which is crucial for assessing their generalizability and efficacy in actual clinical settings. Addressing these limitations could lead to the development of more powerful tools for the early detection and management of breast cancer metastasis and recurrence. By bridging these research gaps, we can improve the precision and applicability of predictive models, ultimately enhancing patient outcomes.

3. Materials and Methods

Predicting breast cancer recurrence poses significant challenges owing to the difficulty of advanced models, the high dimensionality of datasets, correlated features, and class imbalance. To address these issues, we propose a novel approach integrating a Gated Recurrent Unit (GRU) model with feature optimization based on Analysis of Variance (ANOVA). This combination aims to improve the reliability and accuracy of recurrence predictions for breast cancer. The proposed model leverages the GRU’s ability to handle sequential data and capture temporal dependencies, which is crucial for analyzing patterns in patient data over time. By integrating ANOVA-based feature optimization, we address the problem of high dimensionality and correlated features, ensuring that the model focuses on the most relevant features for predicting recurrence. This approach improves model performance and mitigates the effects of class imbalance, leading to more reliable and accurate predictions. Figure 1 provides the block diagram of the suggested system.

3.1. Dataset

The research utilized the publicly accessible Wisconsin Breast Cancer (WBC) dataset, which has been meticulously prepared for analysis. To ensure the dataset’s suitability, it underwent rigorous filtering and formatting using various methodologies [18]. The WBC dataset, sourced from the WBC repository, comprises 569 cases with 34 additional features and a class attribute labeled “outcome,” denoting “R” for recurrent cases and “N” for non-recurring ones. Among these cases, 47 were recurrent, while the remaining 522 were non-recurrent. The emphasis was on persons diagnosed with invasive breast cancer who had not manifested distant metastases. Features were derived from Fine Needle Aspiration (FNA) images of breast masses.

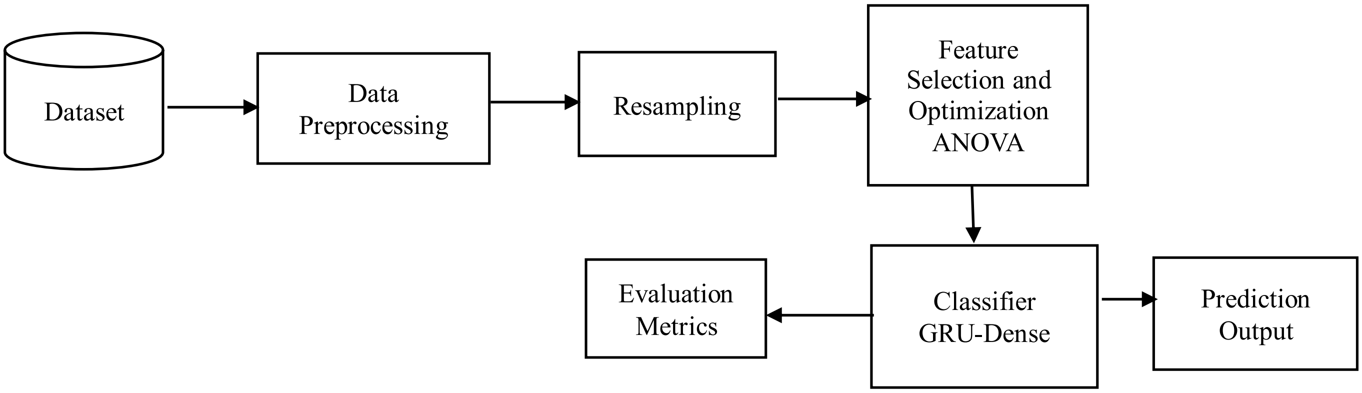

Fig. 1 Block diagram of the proposed model

3.2. Data Preprocessing

Data preprocessing is crucial for ensuring high-quality input for analysis by addressing errors, missing values, and inconsistencies in raw data. It involves several steps, including format standardization, discrepancy resolution, and dataset integration. Normalization is applied to ensure that the data is on a comparable scale, which adjusts parameters to fall within the range of 0 to 1. This ensures that each feature has a maximum value of 1 and a minimum value of 0, thereby standardizing the data. In addition, text-labeled data is converted to numerical formats using the Label Encoder technique, which assigns numerical values to categorical labels. For instance, in this study, “Recurrence” is encoded as 1 and “No-recurrence” as 0. This transformation, explained by Equation (1), is integral to preparing the dataset for analysis by ensuring consistency and comparability.

$$\left|x_j\right| = \frac{x_j}{\sqrt{x_j^2+y_j^2+z_i^2}} \qquad (1)$$

Where $\left|x_j\right|$ is the normalized value of the variable x_j along the x-axis for the jth data point

After preprocessing, we analyzed a set of thirty unique attributes, each revealing specific relationships. Figure 2 presents a histogram showing the distribution of selected geometric features such as shape, area, and perimeter. These geometric properties are instrumental in defining the dimensions and structure of cancer-affected tissues. These characteristics are used in image analysis to quantify and characterize object attributes within an image. Extracting these geometric properties for mammogram analysis is crucial since they yield significant insights into the geometric structures of cells. These features are vital for training the proposed GRU-ANOVA model, as they serve as key indicators of tissue shape and help enhance the model’s accuracy in detecting and analyzing tissue abnormalities.

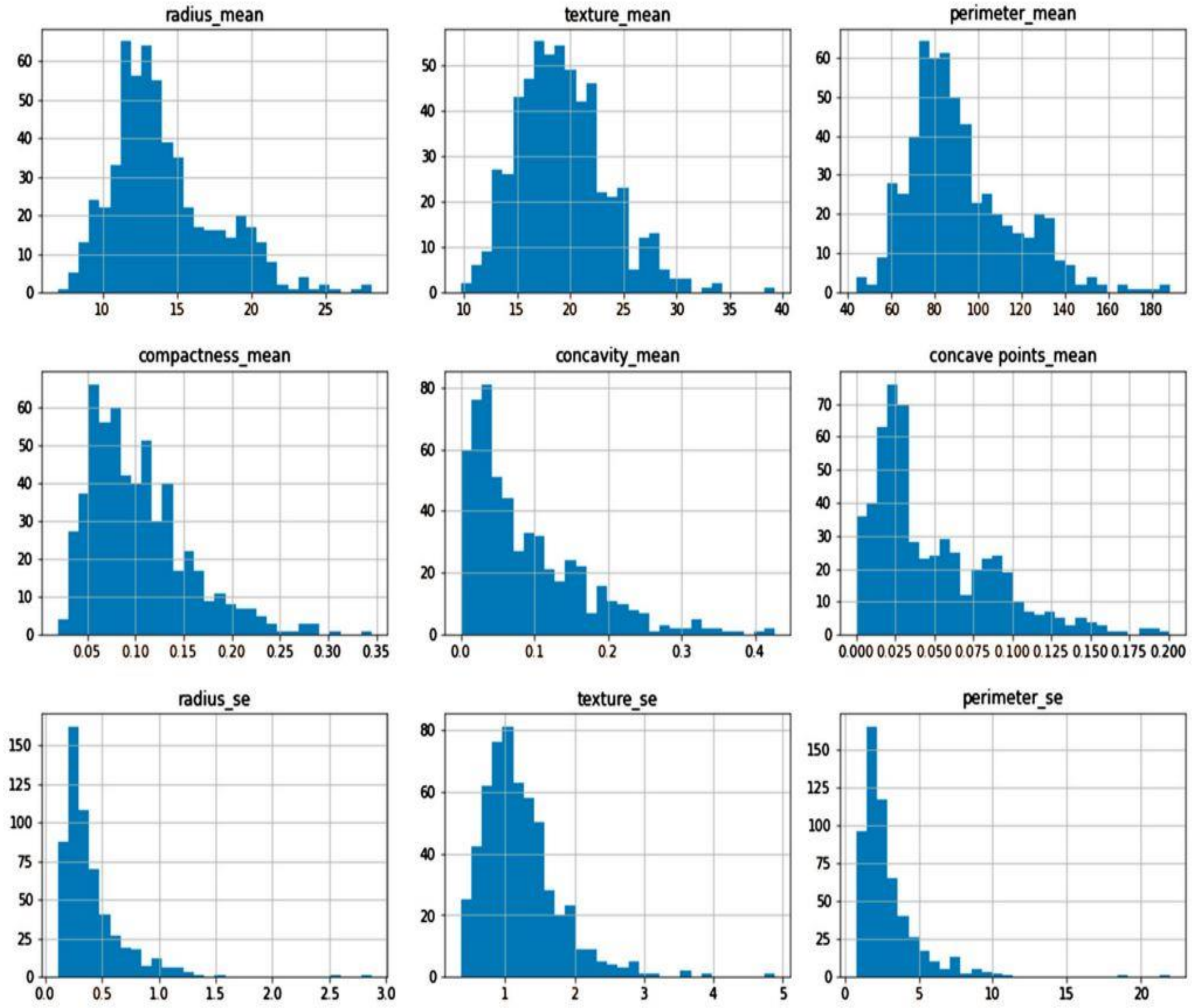

Fig. 2 Distribution of certain geometric features

Structural characteristics in image analysis delineate the spatial configurations of pixels within an object, providing details about its texture, patterns, and shape. As illustrated in Figure 3, these features depict how these characteristics are distributed across the dataset. By encoding these attributes mathematically, we can examine the relationships between different parts of the object, allowing for a detailed analysis of its overall structure and composition. This approach enhances our understanding of the object's visual elements and facilitates a deeper exploration of its internal connections and patterns.

Gabor filters are instrumental in detecting additional structural properties, such as texture pattern direction and frequency, by analyzing fluctuations in pixel brightness within the image. The filters capture detailed texture information by examining the binary patterns present in each pixel's neighborhood. In parallel, shape context descriptors provide a way to characterize object shapes by comparing contour points against a reference shape. This method helps in defining the spatial distribution and geometric properties of objects. Texture-based features, which record changes in pixel intensities across the spatial domain, are crucial for encoding objects' texture and surface attributes within an image. Figure 4 shows how these features are distributed and how well they capture important texture information. Various techniques, such as frequency analysis and transformation-based methods, are employed to compute these features, each offering unique insights into the image's structural and textural properties.

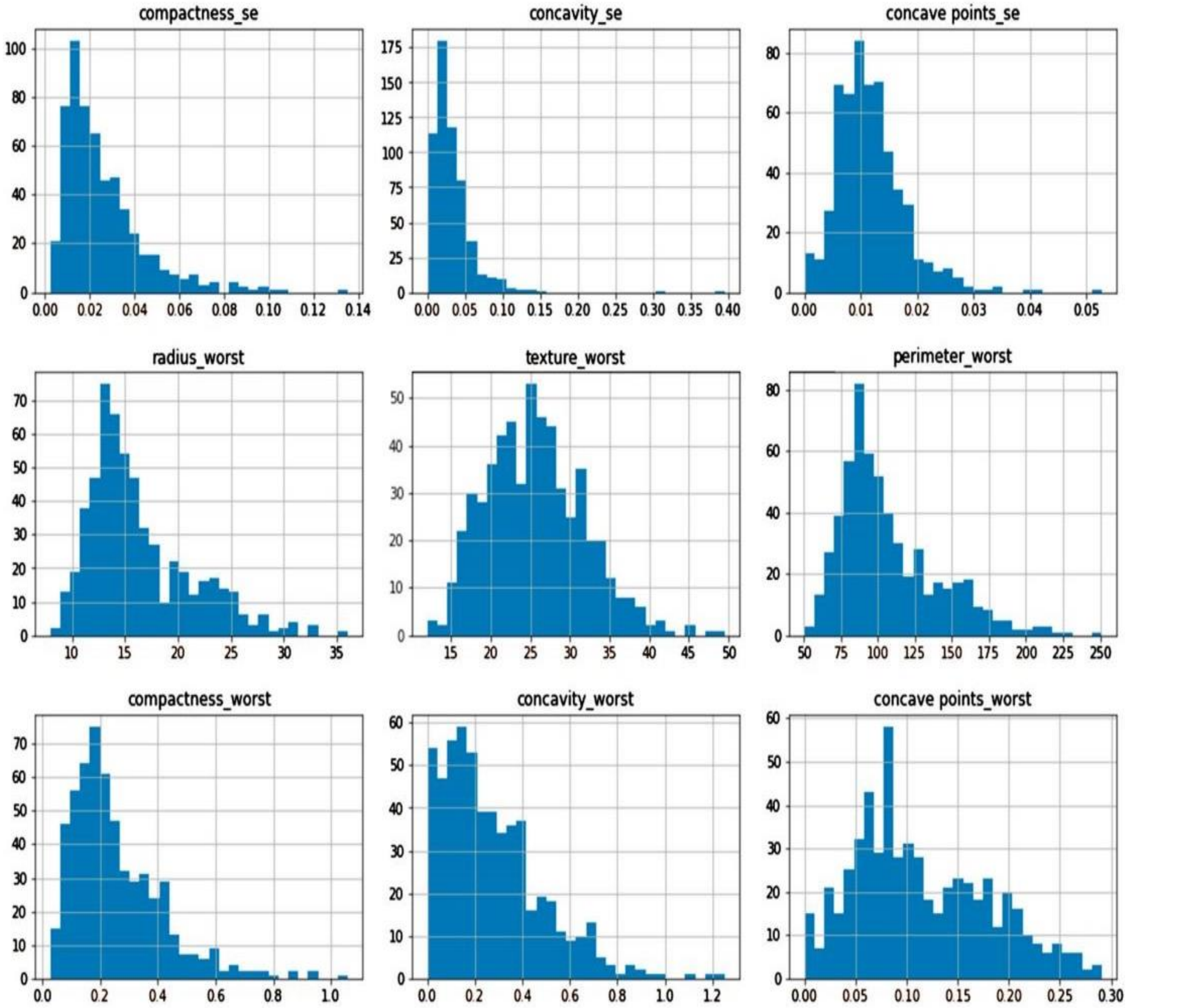

Fig. 3 Distribution of certain structural features

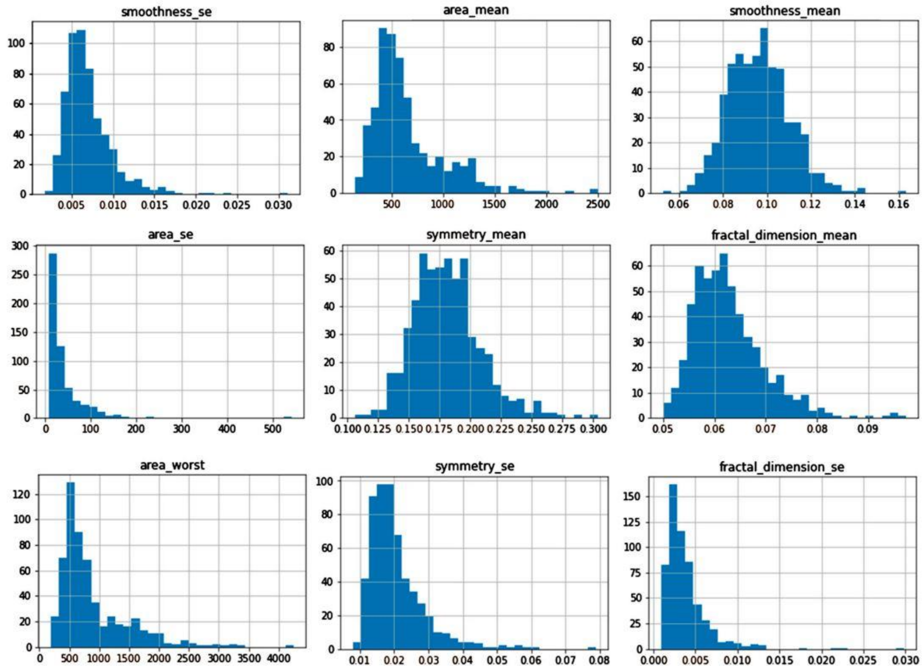

Fig. 4 Distribution of certain texture features

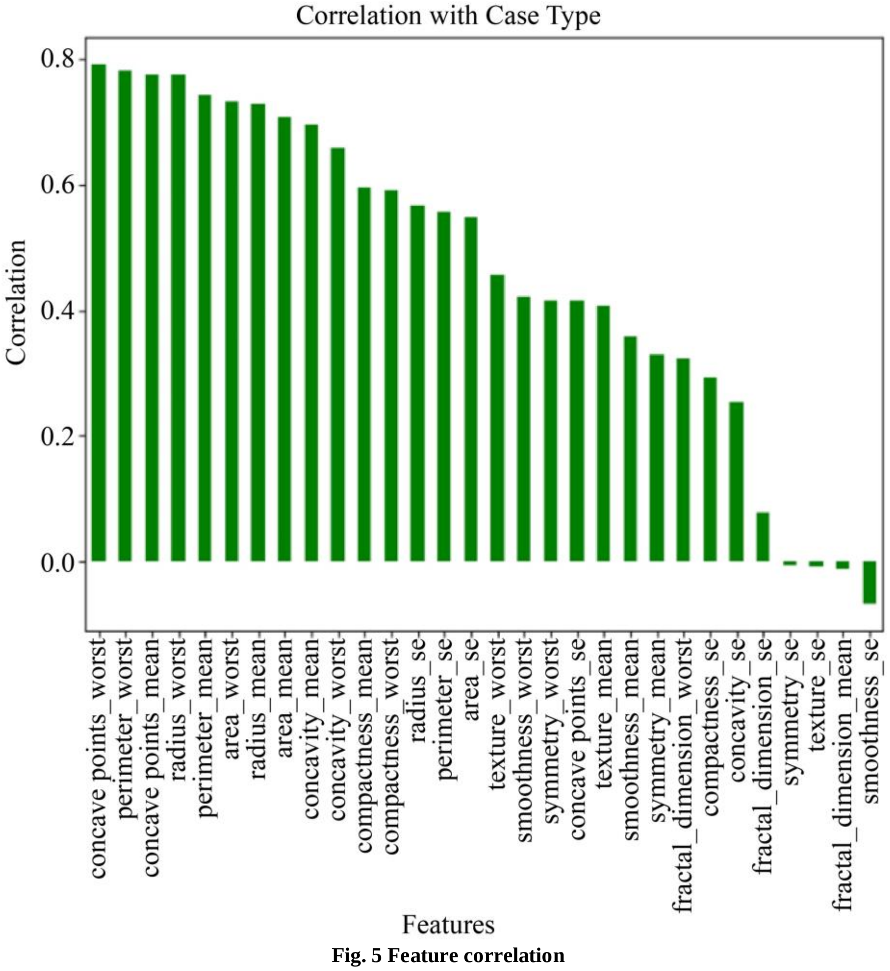

Fig. 5 Feature correlation

The correlation study demonstrates that the dataset, comprising records from 569 patients, is valuable for predicting breast cancer recurrence. Removing features with perfect correlation coefficients reduces redundancy and potential overfitting in our model. Feature correlation, which assesses the degree of association between features, is critical for feature selection and the effectiveness of DL algorithms. High correlations can negatively impact model accuracy and performance by introducing redundancy. Thus, identifying and evaluating feature correlations helps pinpoint the most pertinent and independent features for the task at hand. According to the correlation analysis illustrated in Figure 5, “diagnosis” exhibits the highest correlation, while “smoothness_se” shows the lowest, reflecting the varied significance of these features in the predictive model.

A heat map is an effective visualization tool that represents data in a matrix format, where colors signify the intensity of values. A gradient scale is typically used, with deeper hues indicating higher values and lighter shades representing lower ones. This visual format is particularly useful for identifying patterns, connections, clusters, or outliers within a dataset, which might be challenging to detect through other visualization methods.

For instance, as illustrated in Figure 6, a heat map can vividly display the correlations between features in a dataset, allowing for quick and intuitive insights into the relationships and strengths of those correlations.

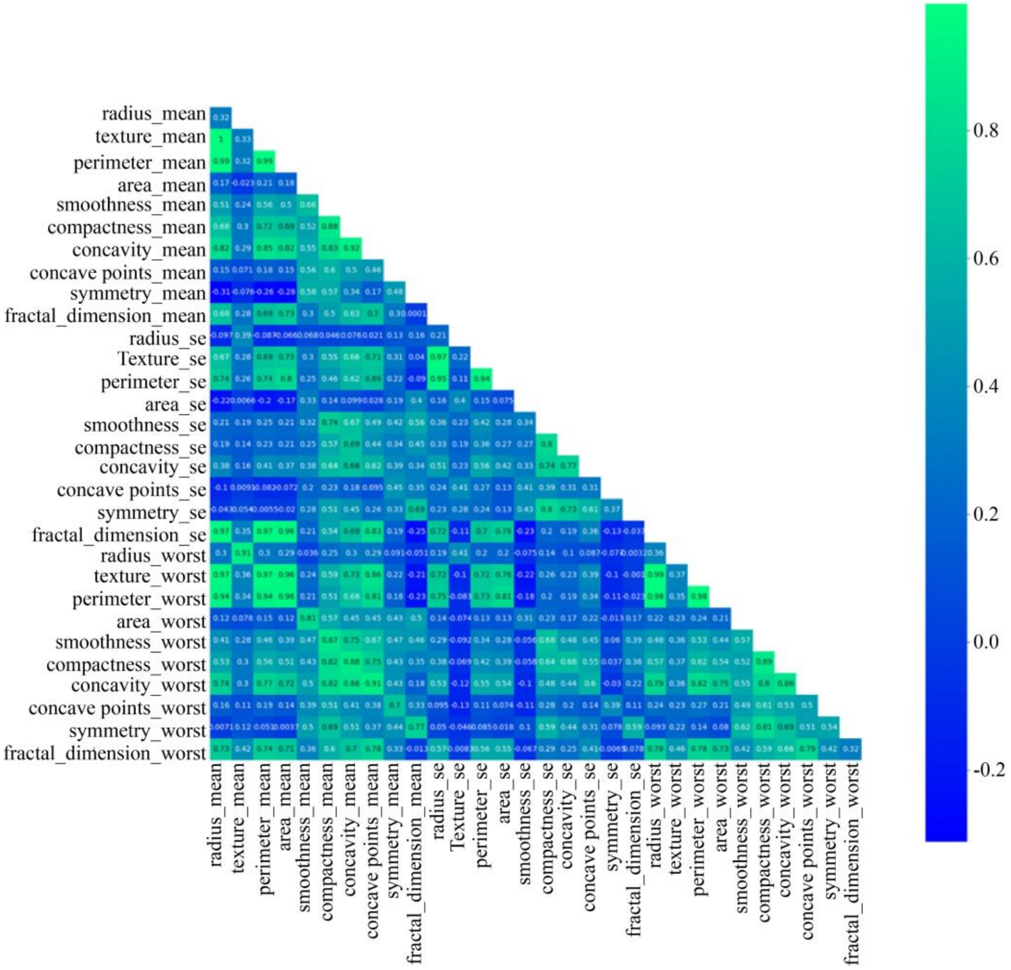

Fig. 6 Heat map of the dataset

3.3. Feature Selection

The Analysis of Variance (ANOVA) model plays a crucial role in identifying the variance among individual participant features and uncovering related features, as illustrated in Figure 7. By employing ANOVA for feature selection, the dataset is systematically ranked based on the F-statistic values assigned to each feature set. This ranking process simplifies the task of determining the most relevant subset of characteristics. Consequently, ANOVA facilitates a more effective evaluation of which features are most significant, thus streamlining the selection process and enhancing the overall quality of the dataset used for further analysis.

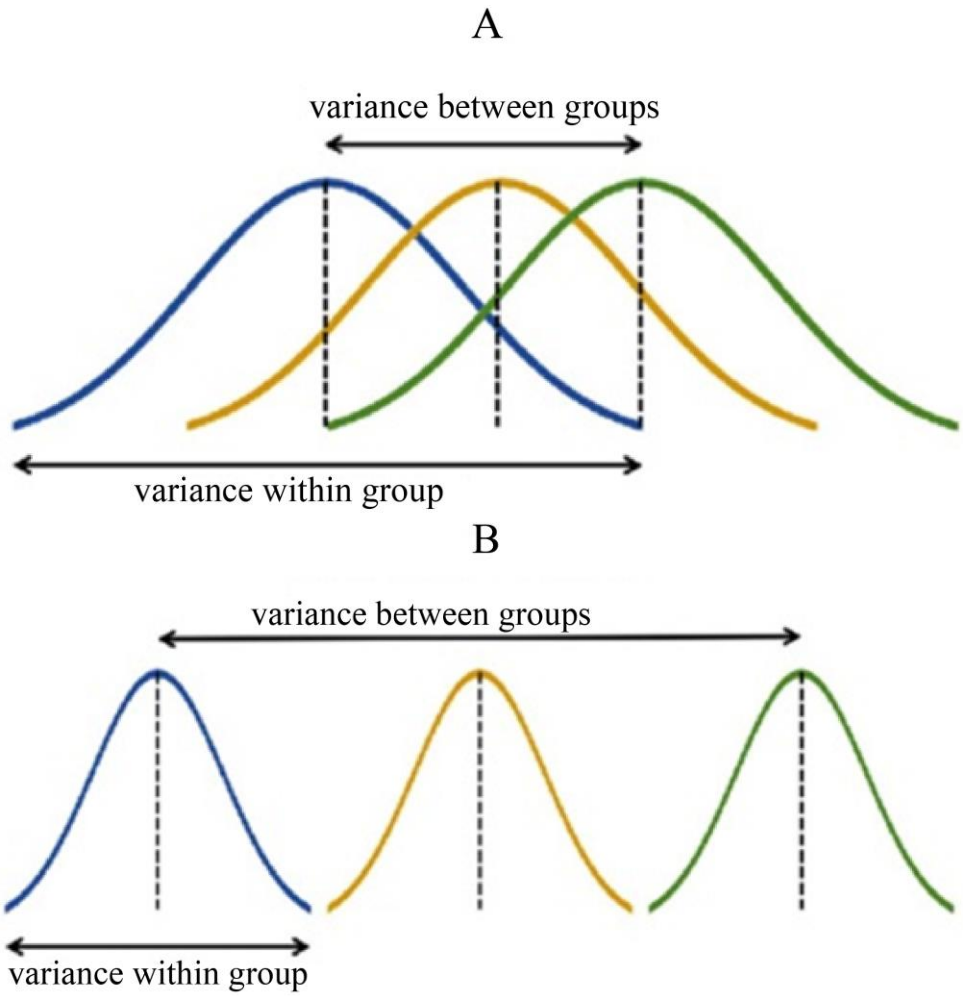

Fig. 7 ANOVA Operation

ANOVA is a statistical method that is highly effective for simultaneously analyzing the relationships between numerical and categorical variables. It employs the F-test to evaluate these associations. The equations used to compute the sum of squares that measure variability within the data are key to this analysis. For instance, the sum of squares within groups (SSW) is crucial for understanding how individual group variances contribute to the overall variability. Specifically, Equation (2) calculates SSW by assessing the deviations of data points within each group from their respective group means, providing insight into the extent of variability attributable to differences within groups rather than between them.

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_{\imath}\right)^2 \qquad (2)$$

Where X_{ij} refers to the j^{th} observation within the i^{th}group, while $\bar{X}_{\imath}$ denotes the mean of the i^{th}group. The total number of groups is represented by k, and n_i stands for the number of observations within the i^{th}group.

Degrees of Freedom (DF) are essential in determining test statistics and assessing the variability of a dataset. Specifically, the Degrees of Freedom Within the Group (DFW) are essential for understanding the distribution of data points and the precision of statistical estimates. The DFW is computed using Equation (3), which considers the quantity of observations and the number of parameters assessed within each group.

$$DFW = N - k \qquad (3)$$

The mean square value within the group (MSW) is represented by Equation (4).

$$MSW = \frac{SSW}{DFW} \qquad (4)$$

The sum of squares between groups (SSB) is illustrated by Equation (5).

$$SSB = \sum_{i=1}^{k} n_i(\bar{X}_{\imath} - \bar{X})^2 \qquad (5)$$

Where $\overline{X}_i$ is the mean of the i^{th}group, $\bar{X}$ is the overall mean.

The degree of freedom between the group (DFB) is given by Equation (6).

$$DFB = k - 1 \tag{6}$$

The mean square value between the group (MSB)is represented by Equation (7).

$$MSB = \frac{SSB}{DFB} \tag{7}$$

The F-statistic is a fundamental component of ANOVA analysis, playing a crucial role in determining whether there are significant differences between the means of various groups or treatments. By thoroughly testing the null hypothesis, which posits that all group means are equal, the F-statistic assesses the statistical significance of any observed differences between these groups. It allows researchers to quantify whether variations in group means are due to random chance or if they represent meaningful, systematic differences. The F-statistic is essential for identifying the significance of these variations and is mathematically represented by Equation (8).

$$F = \frac{MSB}{MSW} \tag{8}$$

3.4. Proposed GRU- ANOVA Model

A Gated Recurrent Unit (GRU) is a unique variant of Recurrent Neural Network (RNN) optimized for sequential data processing, particularly engineered to mitigate the vanishing gradient issue commonly encountered by conventional RNNs. By utilizing gating mechanisms, GRUs effectively manage the flow of information within the network, enabling them to capture long-range dependencies more efficiently.

Unlike other RNN variants, GRUs do not use a distinct cell state; instead, they rely solely on a hidden state to store and transmit information across time steps. At each timestamp (t_i), a new hidden state is generated by combining the input (x_{ti}) with the previous hidden state (h_{ti-1}), which is then passed forward to the next time step. The two main components of GRUs are the reset and update gates, which play crucial roles in determining what information should be forgotten or retained. Figure 8 illustrates the structure of the GRU cell.

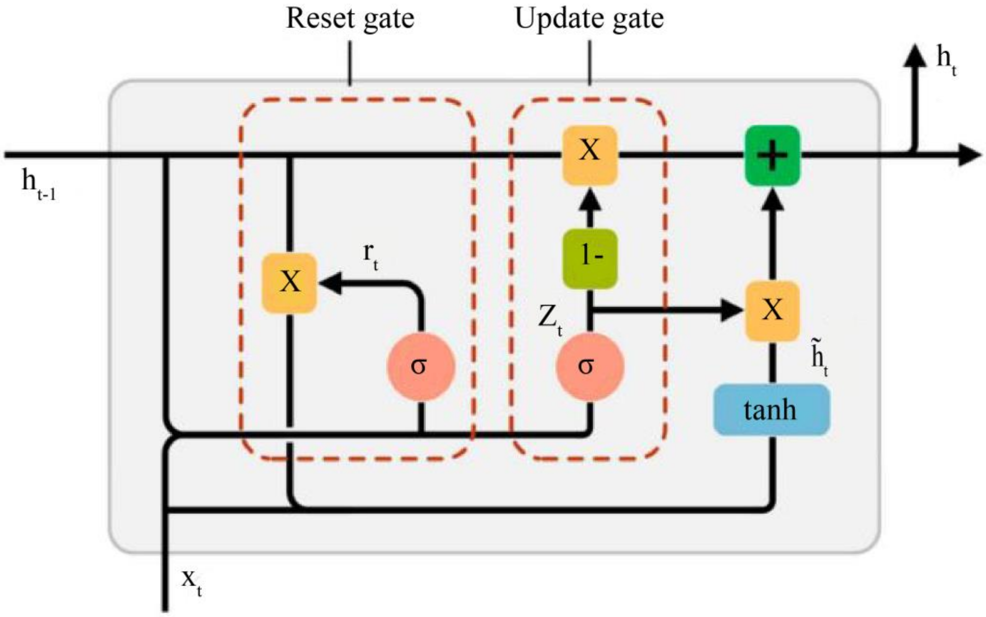

Fig. 8 GRU cell structure

The equations governing the operations of a GRU are as follows:

Reset Gate (r_{ti}),

$$r_{ti} = \sigma(W_r.\left[h_{ti-1,}x_{ti}\right] + b_r) \tag{9}$$

Update Gate (z_{ti}),

$$z_{ti} = \sigma(W_z.\left[h_{ti-1,}x_{ti}\right] + b_z) \tag{10}$$

Candidate Activation (h_{ti}),

$$h_{ti} = (1 - z_{ti}) * h_{ti-1} + z_{ti} * \overline{h_{ti}} \tag{11}$$

The update gate (z_{ti}) controls how much of the previous state is retained, while the reset gate (r_{ti}), determines how much of the past state is forgotten. Additionally, the candidate activation (h_{ti}), captures new data and integrates it into the hidden state. The proposed Modified GRU model consists of two parts: a generalization model based on GRU and a feature optimization model using ANOVA, as described in Algorithm

1. The GRU classifier is trained only on features exhibiting correlations, while irrelevant features are discarded, incorporating a convolutional neural network for enhanced performance. The ANOVA model plays a crucial role in identifying variations among individual participant features and revealing correlated attributes. The GRU-based generalization model then learns these shared traits. This integrated approach aims to reduce classification bias and improve task accuracy.

Algorithm 1: GRU- ANOVA Classification

Input: WBC dataset
Output: Prediction of Breast Cancer Recurrence

```
1. Calculate the mean for within and between the groups in the dataset
2. Calculate the sum of squares for the dataset
3. Calculate the Degree of Freedom
4. Calculate mean of squares
5. Compute the F-Statistic
6. Optimization of the features in the dataset utilizing ANOVA
7. Feature extraction using GRU
8. Categorization of the feature using sigmoid function
9. Obtain the predicted output
```

In this novel approach, the features optimized through the use of the ANOVA model are subsequently fed into the GRU classifier for enhanced performance. Several network parameters across different layers are critical in determining the final classification outcome. Table 1 outlines all the layers that constitute the proposed GRU classification model, highlighting the importance of each layer in refining the model's predictive accuracy. By optimizing these parameters, the approach ensures more precise categorization, demonstrating the effectiveness of the integrated ANOVA and GRU architecture.

Table 1. Model summary

Layers Type	Output Shape	Parameters
GRU	30 × 50	7950
Dropout	30 × 50	0
GRU	30 × 50	15300
Dropout	30 × 50	0
GRU	30 × 50	15300
Dropout	30 × 50	0
GRU	30 × 50	15300
Dropout	30 × 50	0
Fully Connected Layer		151
Total		53901
Trainable		53901
Non – Trainable		0

The GRU model, initialized with 30 features for the WBC dataset, classifies the final output into two categories: Recurrence (1) and No-Recurrence (0), after processing through the convolutional and hidden layers. The model demonstrates optimal performance with 512 neurons and 200 epochs for both training and validation. The decision to use 200 epochs was made because, at this point, the GRU model reaches convergence, where both accuracy and loss metrics stabilize, yielding the best results. This ensures the model achieves high reliability and effectiveness in its predictions. Figure 9 shows the model architecture of the proposed GRU model.

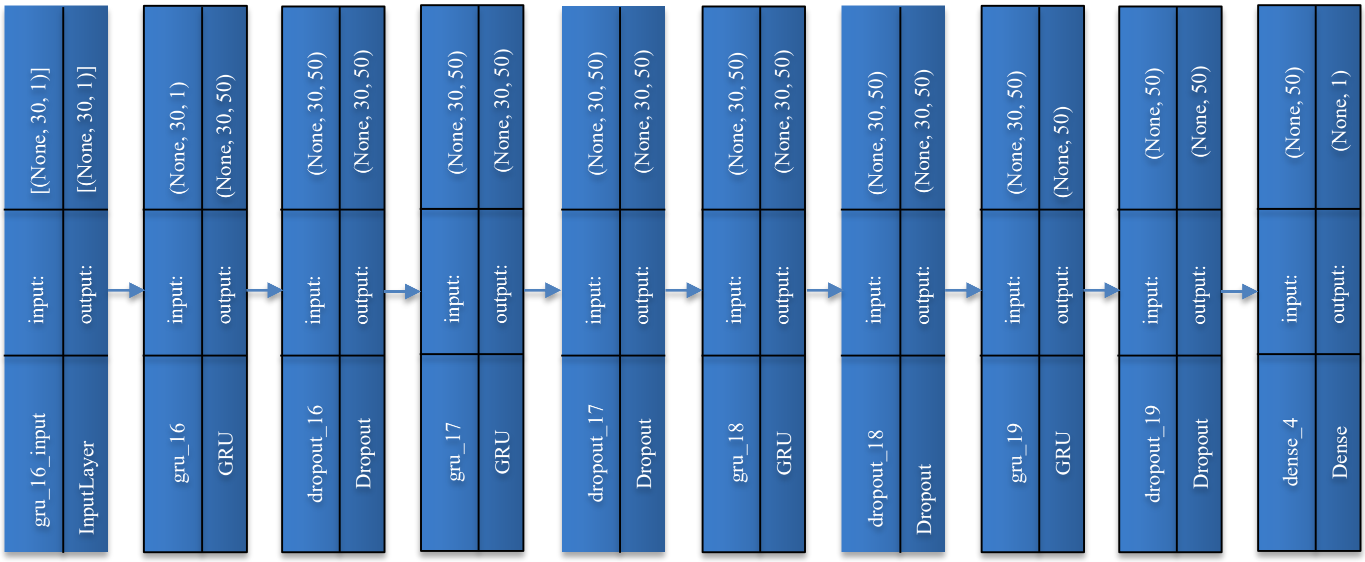

Fig. 9 Model architecture of proposed GRU model

3.5. Hardware and Software Setup

The research utilized a powerful computing setup, including an NVIDIA GeForce GTX 1080Ti GPU, an Intel Core i7 processor, 32GB of RAM, and the Keras library [20], integrated with TensorFlow and executed in Python. The user-friendly Keras interface, along with Google Colab's extensive computational capabilities, facilitated the design of models and ensured efficient training and implementation of complex neural network structures. Hyperparameters, essential settings that control the operation and behavior of a deep learning framework during the training process, are predetermined by the user, unlike model parameters, which are learned from the data itself, as shown in Table 2.

Table 2. Hyperparameter specifications

Hyperparameters	**Values**
Optimizer	Adam
Loss function	Binary Cross entropy
No. of epochs	200
Batch size	24
Activation Function	ReLU

4. Results and Discussion

The optimization process focuses on minimizing training loss, which measures how well the model fits the training data after each epoch. A lower loss value signifies better model performance. On the other hand, accuracy is typically inversely related to loss and indicates the percentage of correct predictions out of the total predictions made on the training data. Similar to training metrics, validation metrics assess the model's performance but use validation data that remains unseen during training. This helps ensure an unbiased evaluation of the model's capacity to generalize to novel data.

$$Log\ Loss = \frac{1}{N}\sum_{i=1}^{N} -(y_i * log(p_i) + (1 - y_i) * log(1 - p_i) \qquad (12)$$

Identifying pertinent markers that accurately classify the case type is essential for predicting breast cancer recurrence. The probability of belonging to class 0 is represented by $(1 - p_i)$, while p_i signifies the probability of belonging to class 1. In this classification framework, the first part of the formula becomes more significant when the observation is classified as class 1, while the second part becomes negligible. Conversely, when the true class is 0, the first part diminishes, and the second component gains importance. Identifying the right features for this prediction involves selecting attributes that strongly correlate with the target variable and effectively distinguishing between the different groups.

Feature selection plays a vital role in improving the accuracy and efficiency of machine learning models across various applications. In breast cancer recurrence prediction, overfitting can be a significant challenge. It occurs when the model achieves excellent performance metrics during training but fails to generalize to new, unseen data, as indicated by poor validation results. Monitoring training and validation losses and accuracy plots over multiple epochs is essential for evaluating the model's generalization ability. These graphical representations allow researchers to track the model's performance over time, helping to identify trends in accuracy improvement and assess the impact of feature optimization strategies in mitigating overfitting. For instance, analyzing accuracy trends, as shown in Figure 10, provides insight into the model's learning progress and effectiveness in predicting breast cancer recurrence.

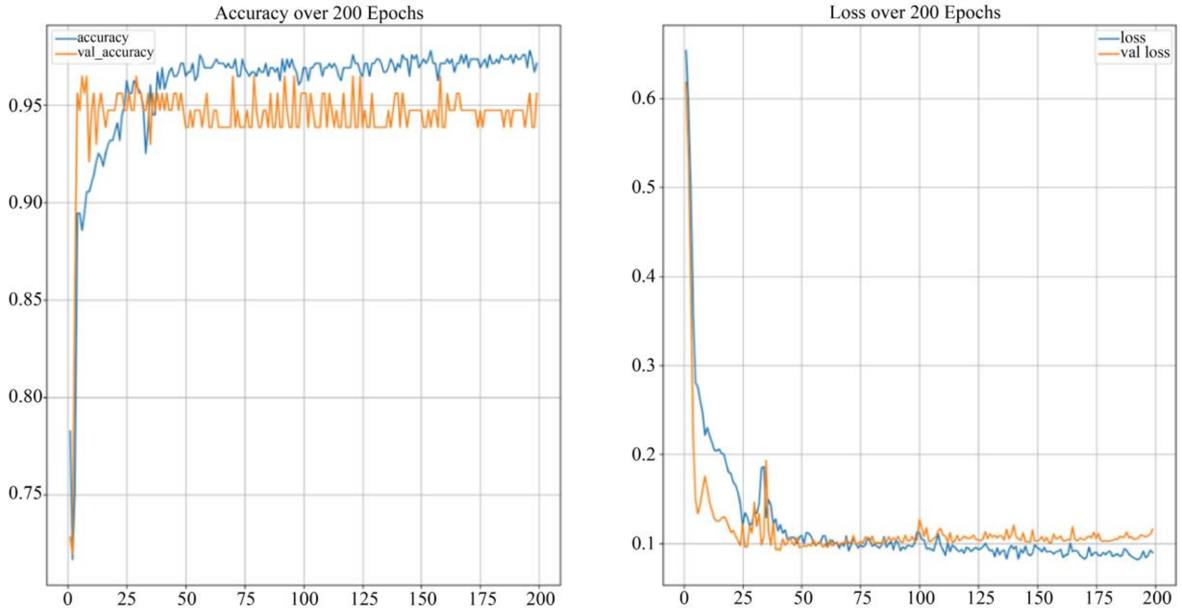

Fig. 10 Accuracy and loss plot of the proposed model

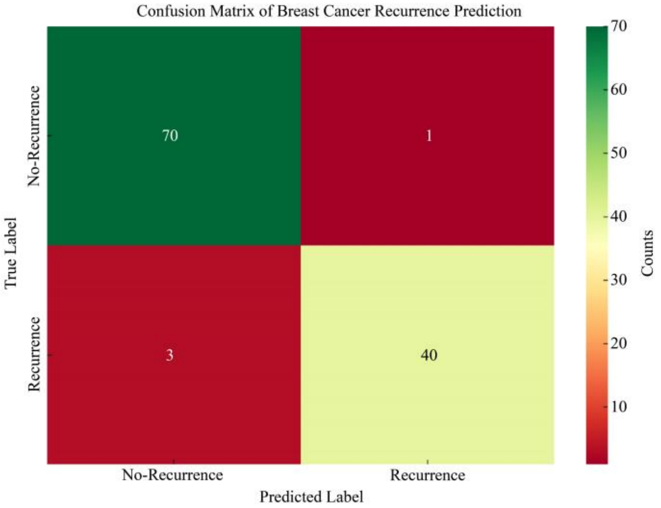

Fig. 11 Confusion matrix

When analyzing model performance, it's notable that metrics such as precision, recall, accuracy, and F1 score generally improve as the proportion of data allocated for training increases. This suggests that a larger training set enables the model to learn better and generalize more effectively. Conversely, these metrics tend to decline as the percentage of data designated for validation increases. The optimal results were observed when 30% of the data was used for training and 70% for validation. This finding was derived from employing a 10-fold cross-validation method, which helps ensure robust and reliable performance evaluation.

A confusion matrix is a performance assessment tool employed in classification tasks to gauge the accuracy of a classification model. It offers a comprehensive analysis of the model's predictions in relation to the actual results. Figure 11 displays the confusion matrix, which illustrates the performance of the GRU-ANOVA classification algorithm in distinguishing between patients with and without recurrence in the validation dataset comprising 114 patients.

The matrix reveals that the algorithm accurately identified 70 patients as not having a recurrence, with only one instance of misclassification. For patients with recurrence, 40 were correctly categorized, though three were incorrectly classified. These results underscore the effectiveness of the proposed strategy in distinctly differentiating the two categories.

Performance indicators obtained from the confusion matrix provide a comprehensive assessment of the proposed model's effectiveness. The system's performance is primarily assessed based on four parameters: accuracy, precision, recall, and F1-score. The measurements derived from the notions of False Positive (FP), False Negative (FN), True Negative (TN), and True Positive (TP), as delineated in Equation (13), Equation (14), Equation (15), and Equation (16), are crucial for evaluating the model's efficacy. Figure 12 shows the performance evaluation metrics obtained in the proposed research.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

$$F1-Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \tag{16}$$

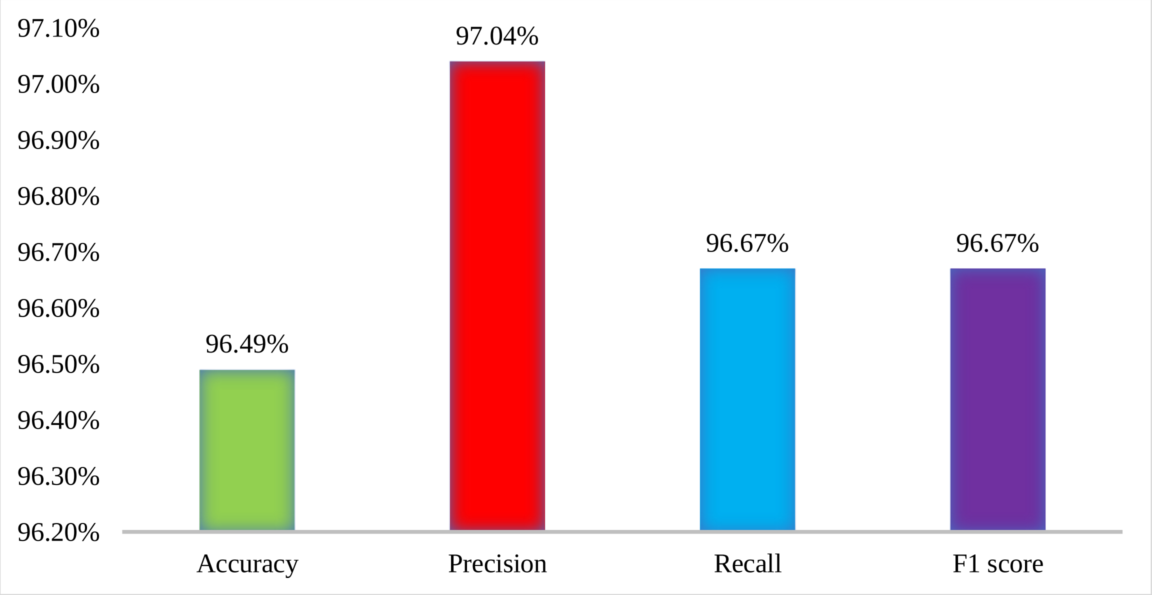

Fig. 12 Performance evaluation of the proposed model

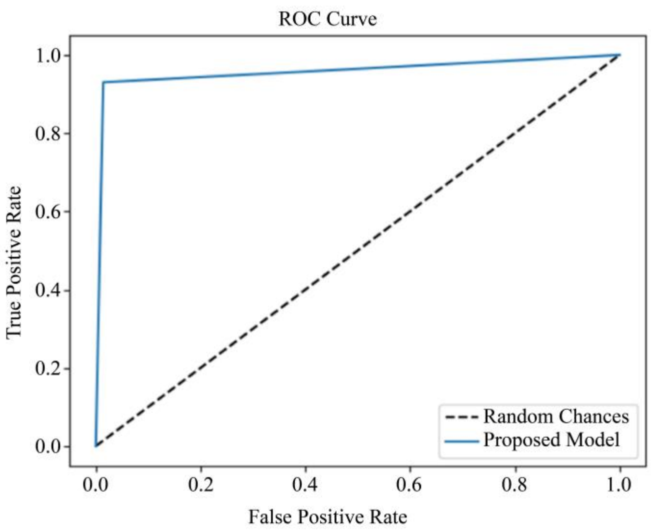

Fig. 13 ROC curve

The ROC curve is a crucial tool for evaluating a model's performance by illustrating how sensitivity and specificity are balanced as the decision threshold varies. Figure 13 displays the ROC curve and AUC for the proposed GRU-ANOVA Classifier, highlighting its effectiveness in distinguishing between classes. The GRU-ANOVA model achieved impressive metrics, including a mean accuracy of 96.49%, precision of 97.04%, recall of 96.67%, and F1-score of 96.67%. These results underscore the significant impact of integrating recurrence and additional layers into the model's architecture, contributing to its exceptional classification performance.

The proposed GRU-ANOVA Classifier outperforms existing methods across various performance metrics, as depicted in Table 3 and Figure 14. When compared to Support Vector Machines (SVM), which achieved an accuracy of 95.70% and an F1-score of 95.83%, the GRU-ANOVA model demonstrates superior accuracy (96.49%) and F1-score (96.67%). Although the Random Forest model exhibits a higher accuracy of 96.71%, the GRU-ANOVA Classifier excels in precision (97.04%) and recall (96.67%), surpassing the Random Forest's precision of 96.77% and recall of 95.14%. Notably, General CNN, with an accuracy of 85.83%, and Naïve Bayes, with an accuracy of 92.94%, lagging behind the GRU-ANOVA model, highlighting its robust performance compared to these methods. Furthermore, the GRU-ANOVA Classifier outperforms other models, such as KNN, Decision Tree, and Logistic Regression, which have lower accuracy and F1 scores. This superior performance underscores the effectiveness of incorporating recurrence and additional layers into the model's architecture.

Table 3. Performance comparison with existing methods

Methodology	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naïve Bayes [18]	92.94	93.85	90.89	92.63
Random Forest [19]	96.71	96.77	95.14	94.36
KNN [20]	95.03	95.24	94.18	95.52
SVM [21]	95.70	96.28	93.48	95.83
Decision Tree [22]	90.46	91.34	90.18	90.89
Logistic Regression [23]	95.74	96.83	95.32	95.10
General CNN [24]	85.83	87.34	83.13	86.38
GRU-ANOVA	**96.49**	**97.04**	**96.67**	**96.67**

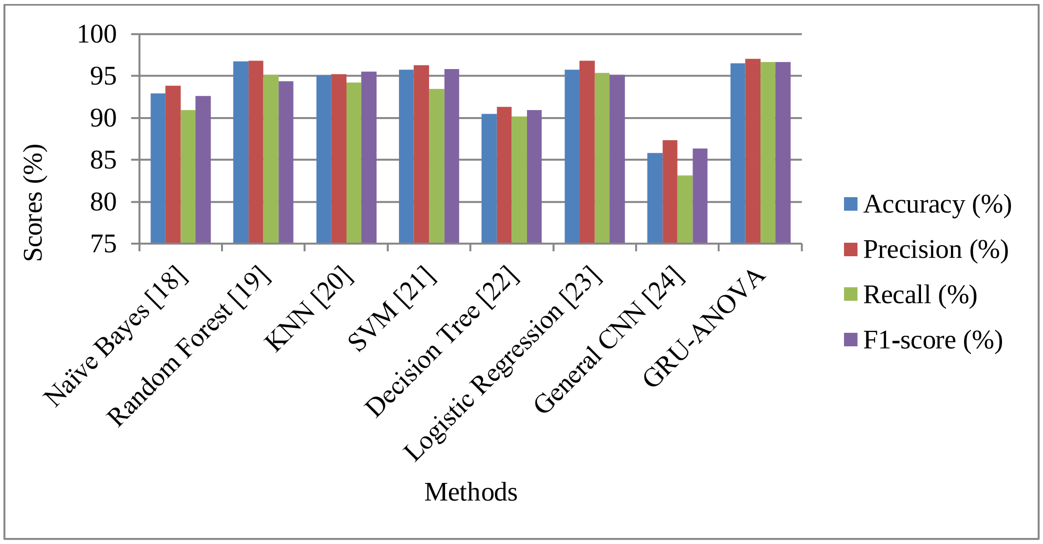

Fig. 14 Performance comparison of the proposed model with existing methods

5. Conclusion

Breast cancer recurrence is a significant challenge in patient care, as it signifies the return of cancer cells following initial treatment and can manifest either locally or distantly. This recurrence complicates patient management and affects prognosis. This study introduces an innovative approach to predicting breast cancer recurrence by utilizing a modified Gated Recurrent Unit (GRU) model combined with ANOVA-based feature optimization. The proposed model demonstrates impressive performance metrics, achieving a mean accuracy of 96.49%, precision of 97.04%, recall of 96.67%, and an F1-score of 96.67%. These results are supported by a thorough confusion matrix and ROC curve analysis, highlighting the model's capability to effectively distinguish between recurrence and non-recurrence cases. The model's robustness in managing complex datasets and addressing issues such as class imbalance underscores its potential as a valuable tool for early detection and personalized treatment planning in breast cancer management. However, further validation studies within clinical settings are recommended to fully assess its real-world applicability and impact on patient care.

Acknowledgements

The author expresses profound appreciation to the supervisor for providing guidance and unwavering support throughout the course of this study.

References

[1] O. Mohammad Mehdi Owrang, Ginger Schwarz, and Fariba Jafari Horestani, "Prediction of Breast Cancer Recurrence with Machine Learning," *Encyclopedia of Information Science and Technology*, Sixth Edition, pp. 1-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Amal Alzu'bi et al., "Predicting the Recurrence of Breast Cancer Using Machine Learning Algorithms," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13787-13800, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Shi-Jer Lou et al., "Machine Learning Algorithms to Predict Recurrence within 10 Years after Breast Cancer Surgery: A Prospective Cohort Study," *Cancers*, vol. 12, no. 12, pp. 1-16, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Noreen Fatima et al., "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and their Analysis," *IEEE Access*, vol. 8, pp. 150360-150376, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Björn Stenkvist et al., "Predicting Breast Cancer Recurrence," *Cancer*, vol. 50, no. 12, pp. 2884-2893, 1982. [CrossRef] [Google Scholar] [Publisher Link]

[6] Donald Courtney et al., "Breast Cancer Recurrence: Factors Impacting Occurrence and Survival," *Irish Journal of Medical Science*, vol. 191, pp. 2501-2510, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mallika Siva Donepudi et al., "Breast Cancer Statistics and Markers," *Journal of Cancer Research and Therapeutics*, vol. 10, no. 3, pp. 506-511, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mahmoud Hussein, Mohammed Elnahas, and Arabi Keshk, "A Framework for Predicting Breast Cancer Recurrence," *Expert Systems with Applications*, vol. 240, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ying Liu et al., "Clinical Decision Support Tool for Breast Cancer Recurrence Prediction Using SHAP Value in Cooperative Game Theory," *Heliyon*, vol. 10, no. 2, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Lorena González-Castro et al., "Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records," *Cancers*, vol. 15, no. 10, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Frederick M. Howard et al., "Integration of Clinical Features and Deep Learning on Pathology for the Prediction of Breast Cancer Recurrence Assays and Risk of Recurrence," *NPJ Breast Cancer*, vol. 9, no. 1, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Lixuan Zeng et al., "The Innovative Model Based on Artificial Intelligence Algorithms to Predict Recurrence Risk of Patients with Postoperative Breast Cancer," *Frontiers in Oncology*, vol. 13, pp. 1-12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nermin Abdelhakim Othman, Manal A. Abdel-Fattah, and Ahlam Talaat Ali, "A Hybrid Deep Learning Framework with Decision-Level Fusion for Breast Cancer Survival Prediction," *Big Data and Cognitive Computing*, vol. 7, no. 1, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Lulu Wang, "Microwave Imaging and Sensing Techniques for Breast Cancer Detection," *Micromachines*, vol. 14, no. 7, pp. 1-29, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] G. Rajasekaran, and C. Sunitha Ram, "Breast Cancer Prediction Based on Feature Extraction Using Hybrid Methodologies," *International Journal of Soft Computing and Engineering*, vol. 13, no. 2, pp. 20-28, 2023. [CrossRef] [Publisher Link]

[16] Ziyu Su et al., "BCR-Net: A Deep Learning Framework to Predict Breast Cancer Recurrence from Histopathology Images," *PLOS ONE*, vol. 18, no. 4, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Yuhua Yao et al., "ICSDA: A Multi-Modal Deep Learning Model to Predict Breast Cancer Recurrence and Metastasis Risk by Integrating Pathological, Clinical and Gene Expression Data," *Briefings in Bioinformatics*, vol. 23, no. 6, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] T.P. Latchoumi, T.P. Ezhilarasi, and K. Balamurugan, “Bio-Inspired Weighed Quantum Particle Swarm Optimization and Smooth Support Vector Machine Ensembles for Identification of Abnormalities in Medical Data,” *SN Applied Sciences*, vol. 1, no. 10, pp. 1-10, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jelmar Quist et al., “Random Forest Modelling of High-Dimensional Mixed-Type Data for Breast Cancer Classification,” *Cancers*, vol. 13, no. 5, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Vincent Peter C. Magboo, and Ma. Sheila A. Magboo, “Machine Learning Classifiers on Breast Cancer Recurrences,” *Procedia Computer Science*, vol. 192, pp. 2742-2752, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Leila Ghasem Ahmad et al., “Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence,” *Journal of Health & Medical Informatics*, vol. 4, no. 2, pp. 1-3, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[22] Jimin Guo et al., “Revealing Determinant Factors for Early Breast Cancer Recurrence by Decision Tree,” *Information Systems Frontiers*, vol. 19, pp. 1233-1241, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[23] Annemieke Witteveen et al., “Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence,” *Medical Decision Making*, vol. 38, no. 7, pp. 822-833, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[24] Meha Desai, and Manan Shah, “An Anatomization on Breast Cancer Detection and Diagnosis Employing Multi-Layer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN),” *Clinical eHealth*, vol. 4, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]