Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Breast cancer recurrence prediction with deep neural network and feature optimization

Arathi Chandran R I & V Mary Amala Bai

To cite this article: Arathi Chandran R I & V Mary Amala Bai (2024) Breast cancer recurrence prediction with deep neural network and feature optimization, Automatika, 65:1, 343-360, DOI: 10.1080/00051144.2023.2293280

To link to this article: https://doi.org/10.1080/00051144.2023.2293280

Published online: 08 Jan 2024.

AUTOMATIKA
2024, VOL. 65, NO. 1, 343–360
https://doi.org/10.1080/00051144.2023.2293280

Breast cancer recurrence prediction with deep neural network and feature optimization

Arathi Chandran R I[a] and V Mary Amala Bai[b]

[a]Department of Computer Applications, Noorul Islam Centre for Higher Education (NICHE), Kumaracoil, India; [b]Department of Information Technology, Noorul Islam Centre for Higher Education (NICHE), Kumaracoil, India

ABSTRACT

Breast cancer remains a pervasive global health concern, necessitating continuous efforts to attain effectiveness of recurrence prediction schemes. This work focuses on breast cancer recurrence prediction using two advanced architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), integrated with feature selection techniques utilizing Logistic Regression (LR) and Analysis of Variance (ANOVA). The well-known Wisconsin cancer registry dataset, which contains vital diagnostic data from breast mass fine-needle aspiration biopsies, was employed in this study. The mean values of accuracy, precision, recall and F1-score for the proposed LR-CNN-LSTM model were calculated as 98.24%, 99.14%, 98.30% and 98.14% respectively. The mean values of accuracy, precision, recall and F1-score for the proposed ANOVA-GRU model were calculated as 96.49%, 97.04%, 96.67% and 96.67% respectively. The comparison with traditional methods showcases the superiority of our proposed approach. Moreover, the insights gained from feature selection contribute to a deeper understanding of the critical factors influencing breast cancer recurrence. The combination of LSTM and GRU models with feature selection methods not only enhances prediction accuracy but also provides valuable insights for medical practitioners. This research holds the potential to aid in early diagnosis and personalized treatment strategies.

ARTICLE HISTORY
Received 26 September 2023
Accepted 1 December 2023

KEYWORDS
Breast cancer; recurrence; prediction; deep learning; LSTM; GRU; ANOVA; logistic regression; classification

1. Introduction

The alarming prevalence of breast cancer necessitates continuous advancements in diagnostic and prognostic methodologies to improve patient outcomes and quality of life [1]. In this context, predictive modelling and the integration of Deep Learning (DL) offer promising avenues to address the complex problems in breast cancer recurrence. Breast cancer recurrence remains a critical concern for patients and healthcare providers. Recurrence occurs when cancer cells reappear after initial treatment, often leading to more aggressive forms of the disease and challenging treatment scenarios [2]. Accurate prediction of breast cancer recurrence is crucial for guiding treatment decisions, enabling early intervention, and enhancing overall survival rates. The risk of distant recurrence after years of occurrence is depicted in Figure 1.

Traditional clinical risk assessment tools have limitations in handling the multifaceted nature of breast cancer recurrence, and there is a pressing need for more sophisticated predictive models that can leverage the wealth of available patient data [3]. Deep learning has been a popular approach for predictive modelling in several medical fields, including oncology, in recent years. By revealing hidden patterns and linkages in intricate datasets, these technologies have the potential to increase the precision and dependability of breast cancer recurrence prediction [4]. Among the diverse ML techniques available, Recurrent Neural Networks (RNNs) have demonstrated remarkable capabilities in modelling dynamic biological processes, such as cancer progression [5].

This work presents a comprehensive investigation into the application of RNNs, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), for breast cancer recurrence prediction. In conjunction with RNNs, we explore the integration of feature selection techniques utilizing Logistic Regression (LR) and Analysis of Variance (ANOVA) to identify the most influential diagnostic features associated with breast cancer recurrence. The Wisconsin Breast Cancer (WBC) dataset serves as our primary data source, containing a wealth of pertinent information extracted from fine-needle aspiration biopsies of breast mass. The motivation behind this research lies in the potential to revolutionize breast cancer recurrence prediction by harnessing the power of deep learning and statistical methods. We aim to address several critical research

CONTACT Arathi Chandran R I arathichandranri.ca@outlook.com Department of Computer Applications, Noorul Islam Centre for Higher Education (NICHE), Kumaracoil, Tamil Nadu, India

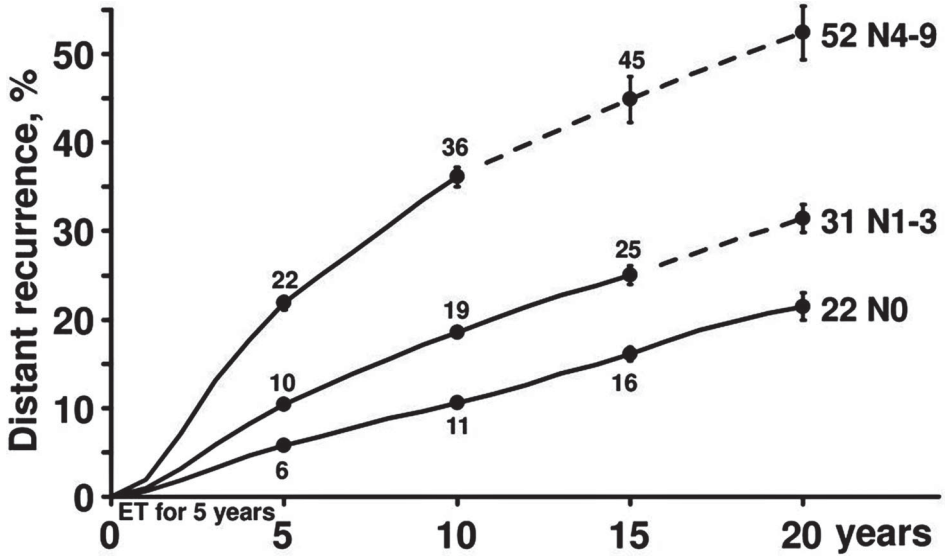

Figure 1. Chance of recurrence over years.

questions: Can LSTM and GRU, as advanced RNN architectures, improve the accuracy of breast cancer recurrence prediction compared to traditional machine learning algorithms? How can feature selection techniques, including LR and ANOVA, enhance the interpretability of predictive models and contribute to a better understanding of the factors influencing breast cancer recurrence.

Initially preprocess the WBC dataset, ensuring data quality and suitability for modelling. Next, we employ LR-based and ANOVA-based feature selection methods to identify the most relevant features associated with breast cancer recurrence. Subsequently, we develop predictive models using LSTM and GRU on the selected features. Finally, we rigorously evaluate the performance of these models against traditional machine learning algorithms, drawing valuable comparisons and conclusions. The significance of this research lies in its potential to provide healthcare professionals with more accurate tools for breast cancer recurrence prediction, enabling early intervention and tailored treatment strategies. Moreover, the insights gained from feature selection can contribute to a deeper understanding of the complex interplay of factors in breast cancer recurrence [6]. Ultimately, this work aims to enhance patient outcomes, reduce the burden of breast cancer recurrence, and advance the field of predictive modelling in oncology.

2. Literature review

Advances in treatment have improved survival rates, the prediction of breast cancer recurrence remains a complex and crucial challenge [7]. Accurate recurrence prediction is essential for tailoring treatment strategies, improving patient outcomes, and reducing the burden of recurrent breast cancer. In recent years, ML and DL techniques have exhibited the potential to enhance the accuracy of recurrence prediction models [8]. Breast cancer recurrence occurs when cancer cells reappear after initial treatment, often in a more aggressive form [9]. Accurate prediction is crucial for guiding treatment decisions, enabling early intervention, and improving overall survival rates. Traditional clinical risk assessment tools, while valuable, may have limitations in capturing the complex interplay of factors influencing recurrence [10]. ML approaches have shown promise in addressing these challenges. Various studies have explored the application of ML techniques, such as KNN and SVM, for breast cancer recurrence prediction [11].

DL techniques, particularly RNNs, have emerged as powerful tools for modelling sequential data [12]. LSTM and GRU networks, have demonstrated the ability to identify dependencies for modelling dynamic biological processes, such as cancer progression [13]. LSTM, with its memory cells and gating mechanisms, has been widely used in various medical domains, including cardiology and genomics, for modelling sequential data [14]. GRU, a simpler RNN variant, offers advantages in terms of training efficiency while maintaining competitive performance in sequence modelling tasks [15]. The success of predictive models often depends on the quality and relevance of input features. Feature selection methods play a crucial role in identifying the most informative attributes while reducing dimensionality and enhancing model interpretability [16]. Regression-based feature selection techniques, such as Lasso regression, assign coefficients to features, highlighting their importance in predictive models [17]. ANOVA evaluates the significance of

different variables and their impact on recurrence risk. Recent research has explored the integration of LSTM and GRU networks with feature selection techniques to improve breast cancer recurrence prediction. This approach combines the capabilities of deep learning in modelling sequential data with the benefits of feature selection in enhancing model interpretability.

Smith et al. [18] demonstrated the effectiveness of LSTM-based models in capturing temporal patterns in patient data, leading to significant improvements in recurrence prediction accuracy compared to traditional methods [19]. The incorporation of regression-based feature selection allowed for the identification of critical clinical and genomic factors contributing to recurrence risk. Similarly, Garcia et al. [20] extended this approach by incorporating GRU models into the predictive framework, highlighting the advantages of GRU's simplified architecture and efficient training. While the integration of LSTM and GRU networks with feature selection methods shows promise, several challenges and avenues for future research exist. One challenge is the need for larger and more diverse datasets to ensure the generalizability of predictive models. Model interpretability remains a critical concern in clinical applications. Addressing this challenge may involve developing techniques to visualize and explain the decisions made by LSTM and GRU models. Furthermore, ethical considerations surrounding the use of AI in healthcare, including breast cancer prediction, warrant careful attention [22]. Ensuring transparency, fairness, and privacy in model development and deployment is essential.

The integration of LSTM and GRU networks with feature selection techniques holds promise for improving breast cancer recurrence prediction. These approaches offer the potential to enhance prediction accuracy while providing valuable insights into the factors contributing to recurrence risk. As researchers continue to explore these methodologies and address the associated challenges, the future holds promise for more precise and interpretable predictive models that can positively impact the lives of breast cancer patients.

3. Dataset

This study employed the publicly available WBC dataset, which was downloaded without any restrictions. Subsequently, we conducted a thorough dataset pre-processing phase [23]. During this phase, we applied alternative techniques to structure and prepare the dataset for analysis. Data pre-processing is a crucial step in filtering and formatting the data in a way that makes it suitable for analysis. Real-world datasets often exist in various formats, and it is essential to adapt them for comprehensible utilization. Data pre-processing serves as a reliable method for addressing these challenges by transforming the dataset into a usable format for standard operations.

The data for this study was sourced from the WBC repository, comprising 569 cases. It includes a single class attribute, "outcome," with two primary values denoted as "R" for recurring and "N" for non-recurring, alongside 34 additional attributes. Among the cases, there were 47 instances of recurrence and 151 instances of non-recurrence. Each entry in this dataset contains follow-up information related to breast cancer cases, specifically focusing on patients diagnosed with invasive breast cancer without remote metastasis. The initial 30 attributes are derived from breast lump images, which was obtained through Fine Needle Aspiration (FNA). To facilitate the utilization of categorical features in our analysis, we employed a label encoder. This tool effectively transforms categorical feature levels into numerical values. In this study, label values ranging from 0 to 1 were assigned using the Label Encoder. Specifically, "Recurrence" was encoded as 1, while "No-recurrence" was encoded as 0.

Data normalization is a technique employed to rescale one or more parameters within a range of 0–1. This process ensures that the maximum value of each attribute becomes 1, while the minimum value becomes 0. Normalization is particularly beneficial when the researcher lacks prior knowledge about the data distribution. Following the application of the Label Encoder technique, which converts text-labelled datasets into numerical datasets, the entire dataset is transformed into a numeric format. The process of normalizing numeric datasets is elucidated through Equation 1.

$$|x_i| = \frac{x_i}{\sqrt{x_i^2 + y_i^2 + z_i^2}} \tag{1}$$

After completing the data pre-processing phase, we derived a set of 30 unique features, each of which exhibits specific interconnections. The distribution of these geometric features is visually represented in Figure 2 through a histogram. These geometric features, such as perimeter, area, and shape, play a vital role in characterizing the structure and dimensions of cancer-affected tissues. In the domain of image analysis, geometric features are commonly employed to describe and quantify the characteristics of objects within an image. Extracting these features from mammograms holds particular significance as they furnish valuable insights into the geometric shapes of cells. These geometric attributes serve as essential indicators of tissue morphology and are consequently crucial for our proposed DL models' training process.

In image analysis, structural features are instrumental in capturing the spatial arrangements of pixels within an object, thereby providing insights into its texture, patterns, and shape. Figure 3 serves as a visual representation of the distribution of these structural

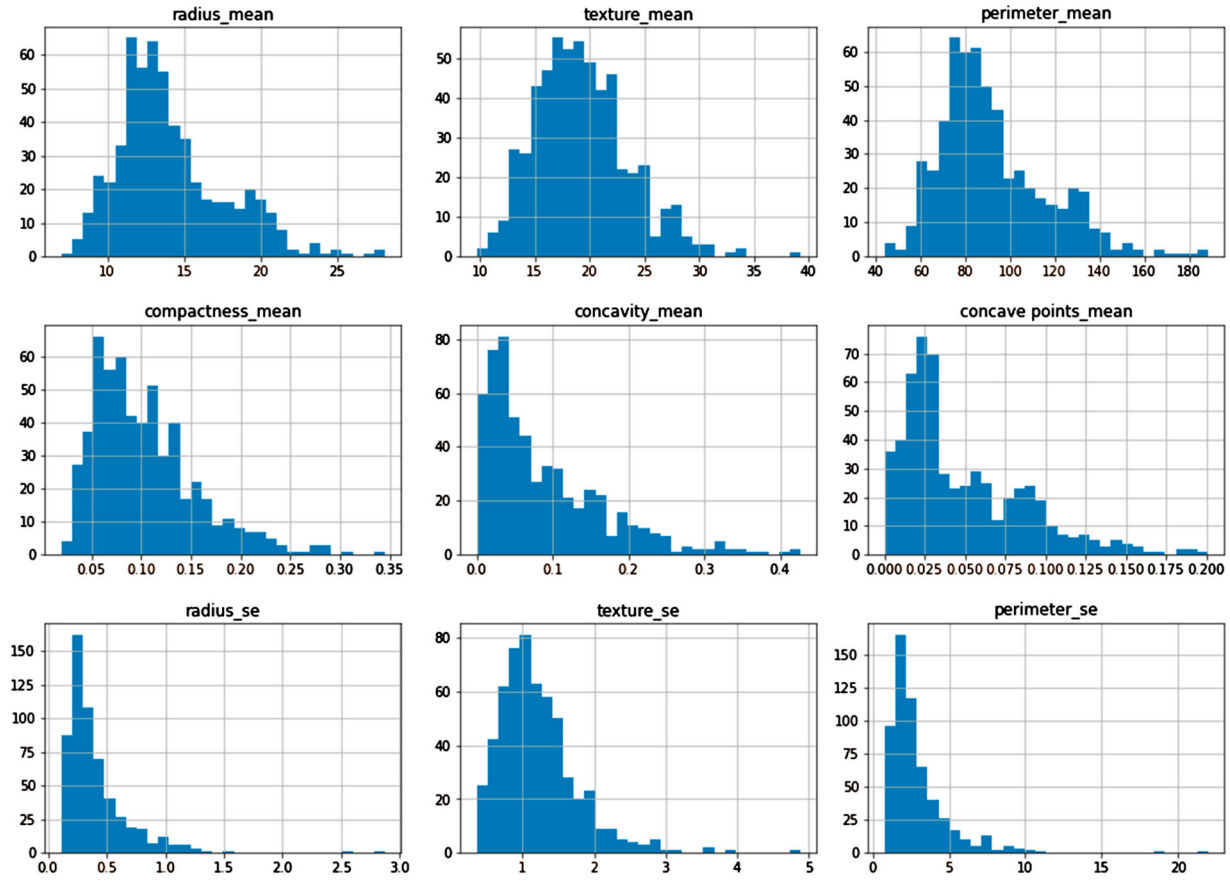

Figure 2. Geometrical feature distribution.

features within our dataset. These features are numerically encoded, facilitating the analysis of the relationships between different components within an object. Texture information is encapsulated through the examination of binary patterns within a circular neighbourhood surrounding each pixel. Additionally, our analysis encompasses other structural attributes, including the utilization of Gabor filters to ascertain the orientation and frequency of texture patterns. Furthermore, shape context descriptors play a significant role in characterizing object shapes by comparing the distribution of their contour points with those of a reference shape.

Texture-based features belong to a category of attributes that elucidate the spatial fluctuations in pixel intensities within an image. The depiction of the distribution of these specific features can be observed in Figure 4. Their primary function is to encapsulate valuable details regarding the texture or surface qualities exhibited by an object within the image. The computation of texture-based features is accomplished through diverse methodologies, including frequency analysis and transform-based analysis. The dataset consists of the records of 569 patients. The correlation between data indicates the ability of the dataset to predict breast cancer recurrence in an efficient way. The feature having perfect correlation (1) is dropped and the remaining features are considered. The correlation analysis of the given dataset is depicted in Figure 5. Highest correlation is obtained for "diagnosis" and lowest correlation is obtained for "*smoothness_se*". The correlation of remaining features falls in between these two features.

Feature correlation pertains to the degree of association or similarity between two or more attributes within a dataset. This aspect holds significant importance in both feature selection and the functioning of deep learning algorithms. This is because highly correlated features can have adverse effects on a model's performance and accuracy. When two features exhibit a strong correlation, it signifies that they convey comparable information and may potentially introduce redundancy, ultimately leading to overfitting. Consequently, it becomes imperative to assess and detect correlations among features in order to identify the most pertinent and mutually independent attributes for a given task.

In a heat map, data is presented in a matrix format where rows and columns symbolize distinct categories, with colours denoting the intensity of the values. Typically, the colour intensity adheres to a gradient scale, with darker hues signifying higher values and lighter shades representing lower values. Heat maps prove invaluable for detecting correlations, clusters, or outliers within the dataset. They offer insights that may be challenging to extract from alternative visualization techniques. To illustrate the correlation among features

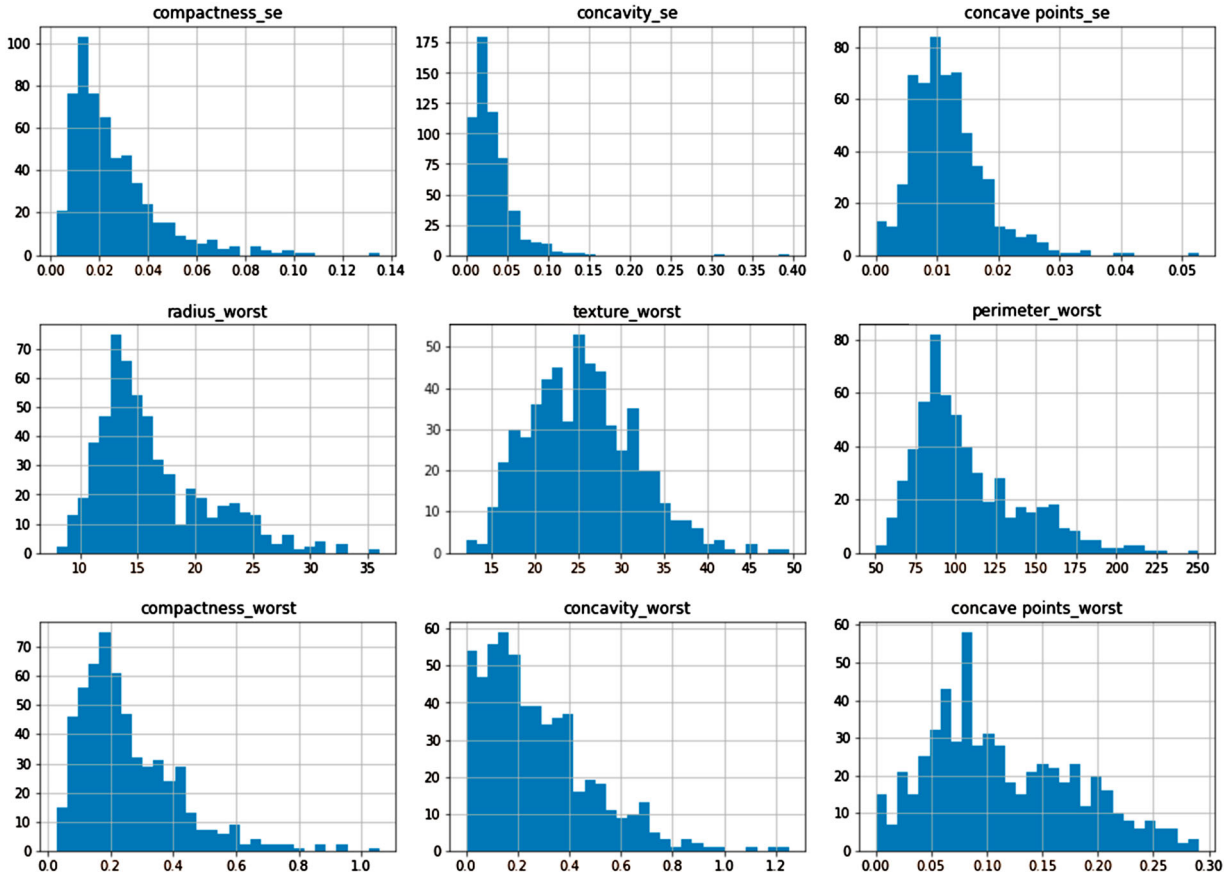

Figure 3. Structural feature distribution.

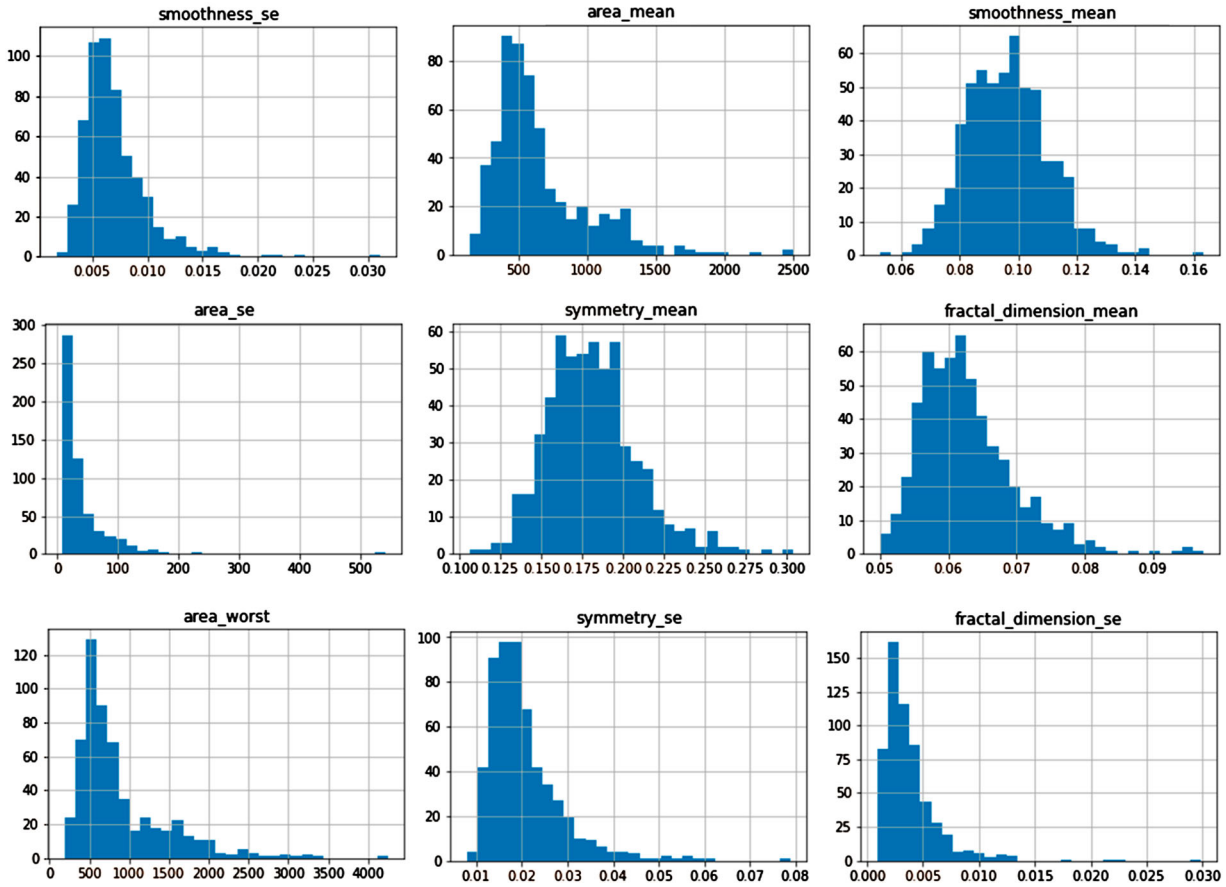

Figure 4. Texture feature distribution.

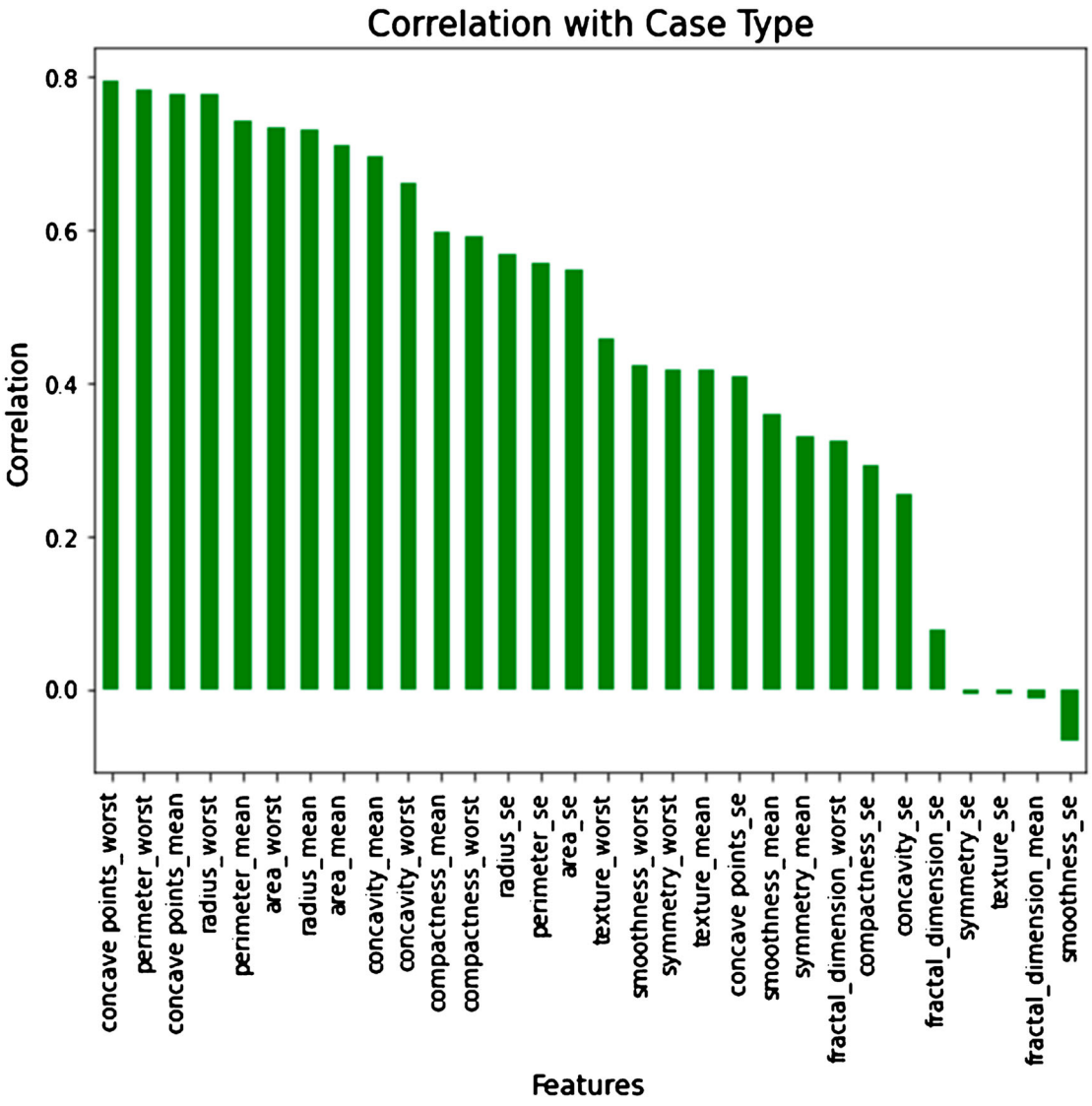

Figure 5. Correlation of features.

within the dataset, a heat map, as illustrated in Figure 6, is utilized.

4. Methodology

The primary aim of the proposed classifiers is to predict the likelihood of breast cancer occurrence in patients. Leveraging DL techniques, these classifiers automatically discern features from raw data through supervised learning, employing an end-to-end training process. The approach adopts two models: the ANOVA-GRU model and the LR-CNN-LSTM framework, both tailored for these predictions. Incorporating dropout layers after GRU layers, except for the dense layers, is essential to prevent overfitting in this model's design. It's imperative to fine-tune hyperparameters during model development to maximize its efficiency. This section outlines the model specifications and its associated hyperparameters. Hyperparameter adjustments play a crucial role in enabling the model to achieve optimal predictive performance. The deep learning models are fed with optimized features and subsequently trained to determine the likelihood of breast cancer recurrence.

4.1. Classification using CNN-LSTM and logistic regression

The proposed LR – Convolutional Neural Network (LR-CNN-LSTM) model consist of two parts. The LSTM memory model specializes in learning discrete participant features, classifying them into two categories: dense and sparse features. Among these, only the dense features are retained and utilized for training the CNN-LSTM classifier. On the other hand, the sparse features are disregarded. The CNN-LSTM-based generalization model is designed to learn shared features across participants. This approach serves to mitigate overall classification bias, ultimately enhancing the overall efficiency. The architecture of the LR-CNN-LSTM model is visually presented in Figure 7.

LR stands as one of the machine learning algorithms employed for addressing classification tasks. It serves

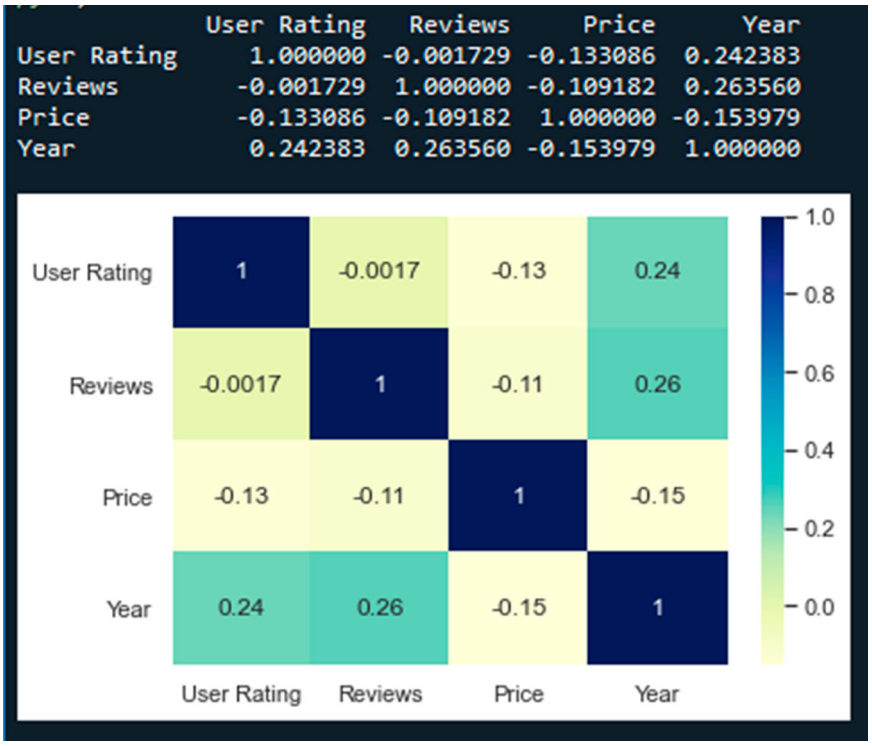

Figure 6. Correlation heatmap of the dataset.

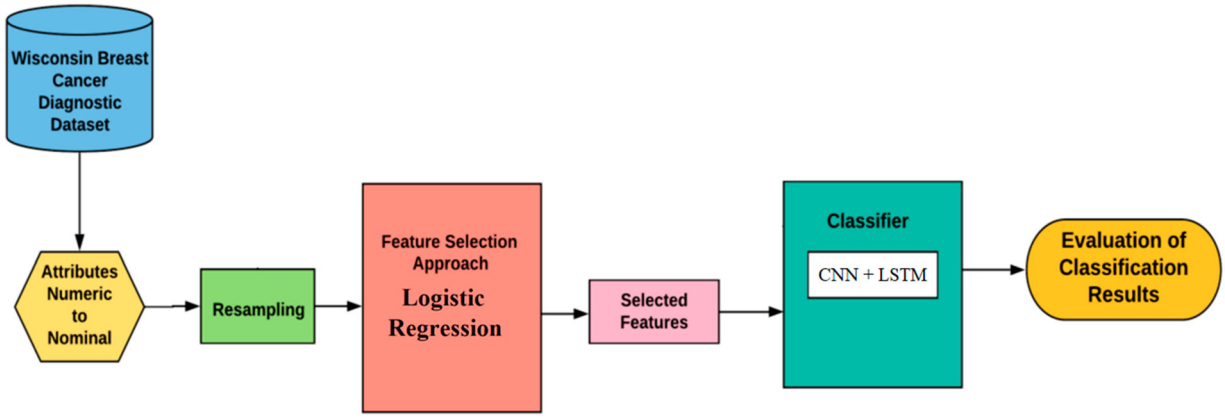

Figure 7. Architecture of LR-CNN-LSTM model.

the purpose of estimating the probability of an instance belonging to a specific class. Notably, LR is primarily utilized in the context of binary classification, where the objective is to classify instances into one of two possible classes. A graphical depiction of the LR operation is presented in Figure 8.

The model introduced in this study consists of two distinct components: a memory cell and a generalization scheme. The memory-based structure is conceptualized for the retrieval of data and it is articulated as:

$$f(x) = sig(xw^T + b), x \tag{2}$$

$$w^T = w_1, w_2, \ldots\ldots..w_{d1+d2+d\phi} \tag{3}$$

$$x = x_1, x_2, \ldots\ldots..x_{d1+d2+d\phi} \tag{4}$$

In this equation, regression coefficient is indicated by w^T, bias is denoted by b, X denotes independent variables and dense data is denoted by d1. Here d2 stands for the sparse feature data associated with the given data, The transformation of cross-product is defined as:

$$\phi(x) = \prod_{i=1}^{d} x_{i=1}^{c_i}, c_i \in \{0, 1\} \tag{5}$$

Within Equation (5), the variable c_i operates as a Boolean value, responsible for regulating features. The prediction function, as demonstrated in Equation (2), adopts the form of a sigmoid function. This function plays a pivotal role in controlling prediction values within the range of (0, 1) and facilitating classification for the provided dataset. The depiction of the segregation of dense and sparse features through LR is visually illustrated in Figure 9.

During the training phase, Stochastic Gradient Descent (SGD) served as the optimizer, and logistic loss was employed to quantify the loss. The logistic loss

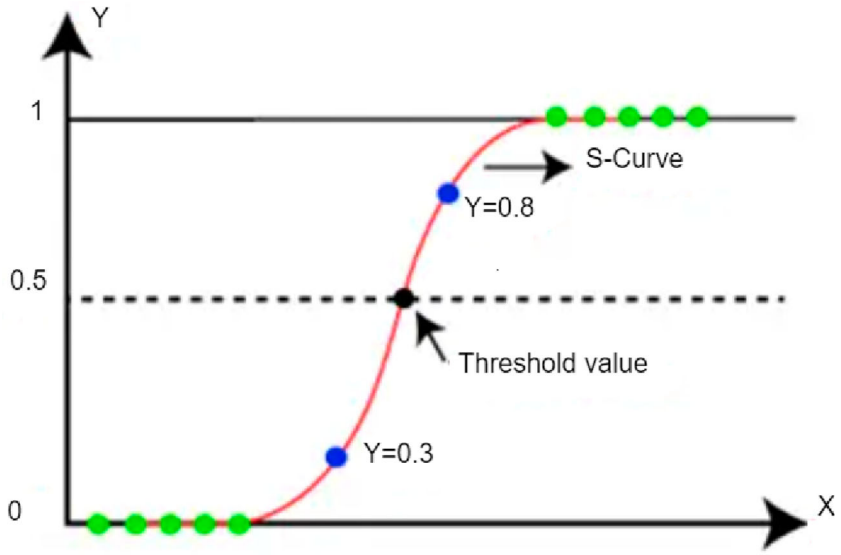

Figure 8. Logistic regression operation.

function for LR is defined as:

$$R(w) = \frac{1}{m}\sum_{i=1}^{m} y^1 \log(\sigma(b + wx^i)) \tag{6}$$

Consider a scenario involving "m" samples, where each sample "x" is characterized by a dimension of "m". Within this context, "σ" represents the sigmoid function, and "w" is the parameter vector. The "y^1" variable represents the estimate of i^{th} sample in question. In constructing the loss function for LR, it's noteworthy that when "y" equals 1, the latter equation becomes 0, and conversely, when "y" equals 0, the former is reduced to 0. Subsequently, the average of these individual loss values, computed across all "m" samples, yields the comprehensive loss function for LR. LSTM represents an advancement over traditional RNNs, introducing crucial improvements in handling sequential data. In LSTM, the current output is intricately linked to the previous state, allowing it to address some of the inherent limitations of conventional RNNs. One of the major challenges that LSTM tackles is mitigating issues related to long-term dependencies by incorporating gradient descent techniques. The recurrent hidden state h_t and the output y_t depend solely on the previous hidden state h_{t-1} and the current input x_t LSTM's architecture introduces mechanisms that enhance efficiency.

$$h_t = Ux_t + Wh_{t-1} \tag{7}$$

$$y_t = Vh_t \tag{8}$$

Bi-directional RNNs can encounter challenges during training, particularly when dealing with long-term dependencies in sequential data. The issues of vanishing and exploding gradients can become pronounced, rendering Bi-RNNs less suitable for scenarios involving extended dependencies. This represents a fundamental concern within recurrent networks. LSTMs have

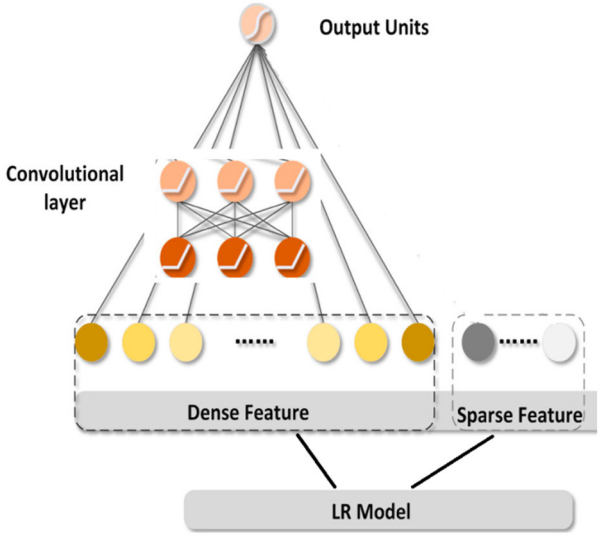

Figure 9. Separation of dense and sparse features.

demonstrated their effectiveness in managing the vanishing and exploding gradient problems efficiently, particularly in contexts with longer dependencies. Importantly, LSTMs incorporate three critical gates allowing for the effective capture and retention of important sequential patterns. This architectural enhancement is visually represented in Figure 10.

In this formulation, the LSTM architecture introduces several crucial components to enhance its ability to capture and manage sequential data effectively. The "ct" memory cell plays a central role in storing and maintaining information over time. Additionally, three gating mechanisms are employed: the "i" (input) gate, responsible for introducing new information into the memory cell; the "f" (forget) gate, which controls the clearing of cell memory; and the "o" (output) gate, regulating the exposure of memory content in producing the output. In this context, LSTM proves valuable in the

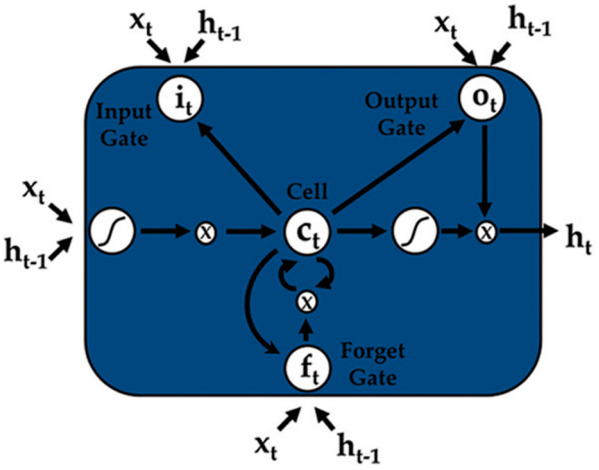

Figure 10. Proposed LSTM cell.

computation of an enhanced hidden state as follows:

$$i_t = \sigma[U_i h_{t-1} + V_i C_{t-1} + W_i x_i] \quad (9)$$

$$f_t = \sigma[U_f h_{t-1} + V_f C_{t-1} + W_f x_t] \quad (10)$$

$$O_t = \sigma[U_o h_{t-1} + V_o C_{t-1} + W_o x_t] \quad (11)$$

$$c_t = \tanh[U_c h_{t-1} + W_c x_t] \quad (12)$$

$$c_t = f_t^i \otimes c_{t-1} + i_t \otimes \tilde{c}_t \quad (13)$$

$$h_t = \tanh(c_t) \quad (14)$$

The gating mechanisms are facilitated by specific activation functions, with the sigmoid function denoted as $\sigma(.)$ and the hyperbolic tangent function as tanh(.). This sophisticated architecture and gating mechanism of LSTM effectively address the challenge of capturing longer dependencies in sequential data. The architecture and parameters of a LR-CNN-LSTM model can be tailored to the specific problem and dataset, providing flexibility and adaptability across various applications. Hyperparameter tuning and careful design are necessary to achieve optimal results for any regression task. The LR-CNN-LSTM algorithm used the classification of given dataset is described in Algorithm 1.

Algorithm 1: LR-CNN-LSTM Classification

Input: Wisconsin Breast Cancer Dataset
Output: Breast Cancer Recurrence Prediction
Step 1: Initialize the values in the Dataset.
Step 2: Compute the regression coefficient w^T
Step 3: Using w^T separate sparse and dense features.
Step 4: Discard the sparse features and select the dense features.
Step 5: Convert the data into a 1-D matrix.
Step 6: Perform convolution operation to extract the features.
Step 7: Utilize the LSTM to categorize the features.
Step 8: Obtain the predicted output.

In this novel approach, the features optimized using the LR model are given as the input to the CNN-LSTM classifier. During each layer, numerous network parameters influence the classification result. The layers in the proposed classifier are explained in Table 1.

Figure 11 shows the proposed CNN-LSTM model for WBC dataset that initialize with 30 features. After the processing in convolutional and hidden layers, final output is generated with two classes (Recurrence = 1, No-Recurrence = 0). In this model 512 neurons and 200 epochs were utilized for training and validation. 200 epochs were used because it is the best value at which the CNN-LSTM model converges and the loss and accuracy are stable providing best results.

4.2. Classification Using GRU and ANOVA

The proposed model, (GRU-ANOVA) model, is composed of two distinct components: the Gated Recurrent Unit (GRU)-based generalization model and the ANOVA-based feature optimization model. Within the ANOVA model, its function is to identify the variance within individual participant features and uncover any correlated features. Among these, only the correlated features are retained and utilized for training the GRU-DNN classifier, while any remaining features are disregarded. The GRU-based generalization model then proceeds to learn the correlated features shared among participants. This comprehensive approach aims to mitigate overall classification bias and enhance the overall accuracy of classification tasks. The architectural representation of the ANOVA-GRU model can be observed in Figure 12.

Through the application of ANOVA for feature selection, the dataset undergoes a ranking process based on F-statistic values assigned to each set of features. This ranking effectively identifies the optimal subset of features within the dataset. ANOVA, a statistical test, is particularly valuable when dealing with a combination of both numerical and categorical variables. It serves as a means to investigate the correlations between these features, with the F-test for ANOVA being instrumental in assessing these correlations. The calculation of the sum of squares is achieved through the following equations.

$$SS_w = \sum_{j=1}^{k} \sum_{j=1}^{l} X - \bar{X}_j \quad (15)$$

Table 1. Proposed CNN-LSTM model summary.

Layers	Type	Output shape	Parameters
Input Layer	Conv1D	30 × 256	1024
Batch Normalization	Batch Norm	30 × 256	1024
ReLU	Activation	30 × 256	0
Convolution Layer	Conv1D	30 × 128	98432
Batch Normalization	Batch Norm	30 × 128	1024
ReLU	Activation	30 × 128	0
Maxpooling	Pooling	15 × 128	0
LSTM	Recurrent	64	49408
Hidden Layer	Dense	128	930
Fully Connected Layer	Dense	1	129
Total			158849
Trainable			158081
Non-Trainable			768

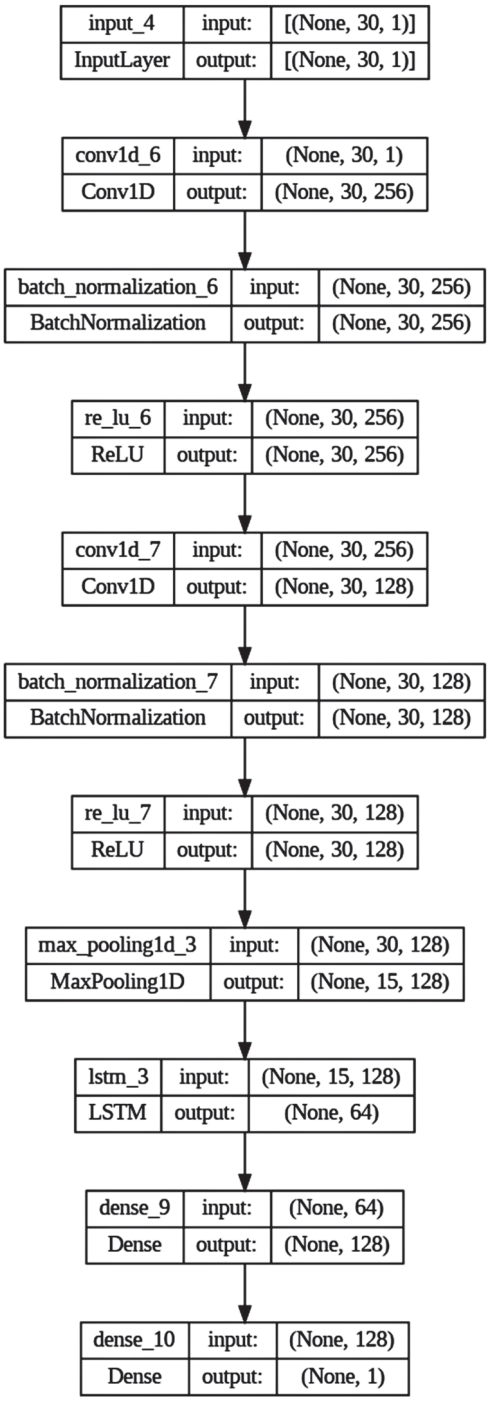

Figure 11. Proposed CNN-LSTM model for breast cancer recurrence prediction.

$$SS_b = \sum_{j=1}^{k} \sum_{j=1}^{l} \bar{X}_j - X \tag{16}$$

The degrees of freedom (*df*) are a crucial concept used to determine the variability in a dataset and to calculate test statistics. Degree of freedom are associated with different sources of variation in the ANOVA model is given by,

$$df_w = k - 1 \tag{17}$$

$$df_b = n - k \tag{18}$$

Mean Square (MS_W) is calculated by dividing SS_W within groups by *df* associated with group variability given in following equations.

$$MS_w = \frac{SS_w}{df_w} \tag{19}$$

$$MS_b = \frac{SS_b}{df_b} \tag{20}$$

The F-statistic, often referred to as the F-test or F-ratio, holds a pivotal position within the realm of ANOVA. This statistical measure serves as a critical tool for assessing whether significant variations exist in the means of two or more groups or treatments. The F-statistic, as defined in Equation 21, is employed to rigorously test the null hypothesis positing that the group means are equal. It essentially aids in the evaluation of whether any observed disparities between these groups carry statistical significance.

$$F = \frac{MS_b}{MS_w} \tag{21}$$

GRUs employ a gating mechanism to regulate the flow of information, distinguishing them from LSTM networks by not having a distinct cell state. Instead, GRUs utilize a hidden state (H_t). This simplified architecture offers the advantage of faster training. During every timestamp (t), input (X_t) is taken along with (H_{t-1}). Subsequently, H_t is generated, which is transferred on to the next timestamp in the sequence. Unlike LSTMs, which feature three gates, GRUs incorporate primarily two gates named reset and update. The GRU cell structure is visually depicted in Figure 13.

Update Gate (z) plays a pivotal role in deciding the extent to which prior knowledge should be carried forward into the future, serving as an analogous counterpart to the output gate, as outlined in Equation 21.

$$z_t = \sigma[W_z(h_{t-1}, x_t)] \tag{21}$$

Reset Gate (r) is responsible for determining the portion of previous knowledge to discard, bearing similarity to the combined function of the input gate and the forget gate in an LSTM recurrent unit.

$$r_t = \sigma[W_r(h_{t-1}, x_t)] \tag{22}$$

The current memory gate (h_t) in a GRU, although often overlooked in standard GRU discussions, is integrated as a means to present non-linearity and ensure zero-mean input, aiming to diminish the influence of prior information being propagated to future. The ANOVA-GRU algorithm used the classification of given dataset is described in Algorithm 2.

$$\tilde{h}_t = \tanh[W(h_{t-1} * r_t, x_t)] \tag{23}$$

$$h_t = \tilde{h}_t * z_t + h_{t-1} * (1 - z_t) \tag{24}$$

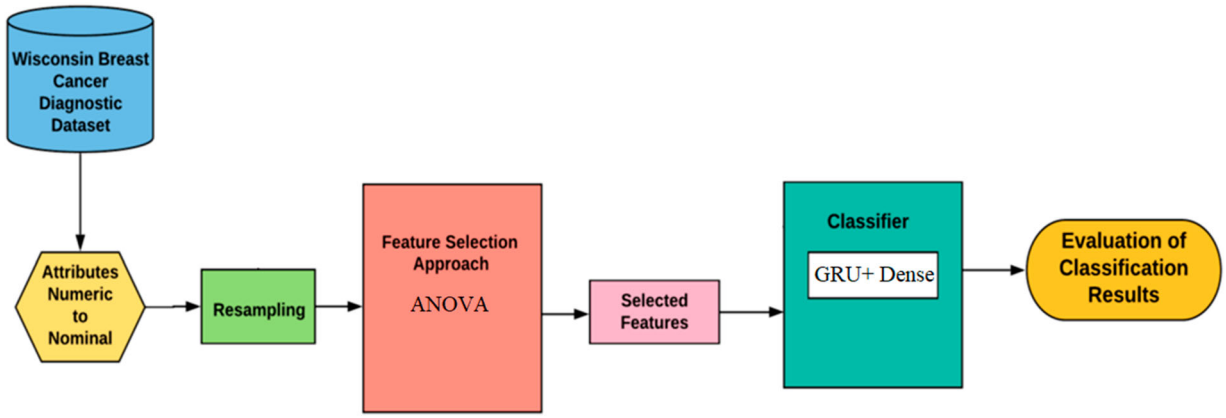

Figure 12. Architecture of ANOVA-GRU model.

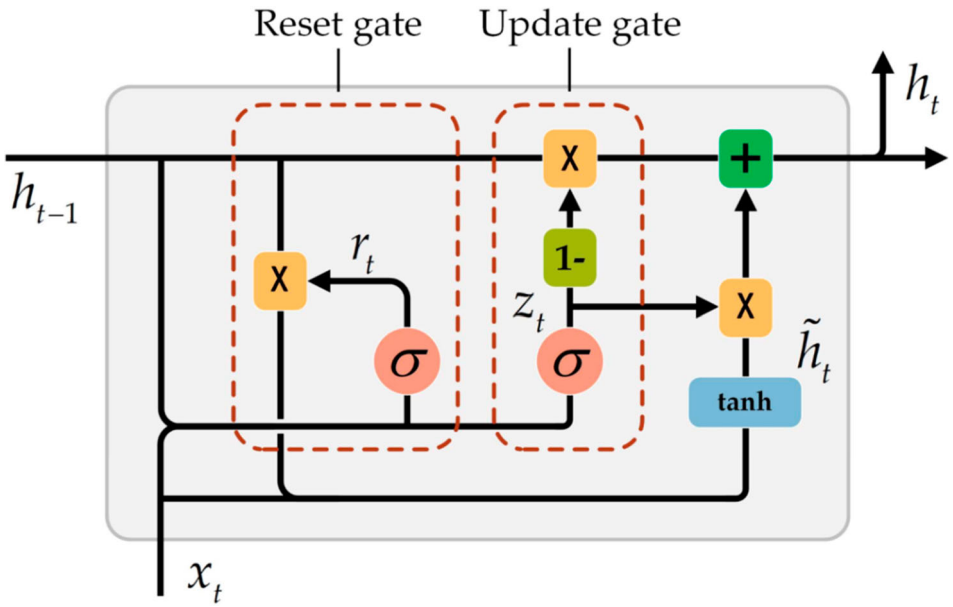

Figure 13. Structure of GRU cell.

Algorithm 2: ANOVA-GRU Classification

Input: Wisconsin Breast Cancer Dataset
Output: Breast Cancer Recurrence Prediction
Step 1: Calculate all the means in the Dataset.
Step 2: Compute the Sum of Squares
Step 3: Compute the Degrees of Freedom (df).
Step 4: Compute the Mean of Squares.
Step 5: Calculate the F-Statistic.
Step 5: Using F-statistic value, optimize the features in the dataset.
Step 6: Perform GRU operation to extract the features.
Step 7: Utilize the sigmoid function to categorize the features.
Step 8: Obtain the predicted output.

In this novel approach, the features optimized using the ANOVA model are given as the input to the GRU classifier. During each layer, numerous network parameters influence the classification result. The layers involved in the proposed GRU classification model are explained in Table 2.

Figure 14 shows the proposed GRU model for WBC dataset that initialize with 30 features. After the processing in convolutional and hidden layers, final output is generated with two classes (Recurrence = 1, No-Recurrence = 0). In this model 512 neurons and 200 epochs were utilized for training and validation. 200 epochs were used because it is the best value at which the GRU model converges and the loss and accuracy are stable providing best results.

Table 2. Proposed GRU model summary.

Layers	Type	Output shape	Parameters
Input Layer (GRU)	Recurrent	30 × 50	7950
Dropout	Dropout	30 × 50	0
GRU	Recurrent	30 × 50	15300
Dropout	Dropout	30 × 50	0
GRU	Recurrent	30 × 50	15300
Dropout	Dropout	30 × 50	0
GRU	Recurrent	30 × 50	15300
Dropout	Dropout	30 × 50	0
Fully Connected Layer	Dense	1	51
Total			53901
Trainable			53901
Non-Trainable			0

5. Results and discussion

Training loss represents the amount of loss relative to the training data at the conclusion of each epoch, with

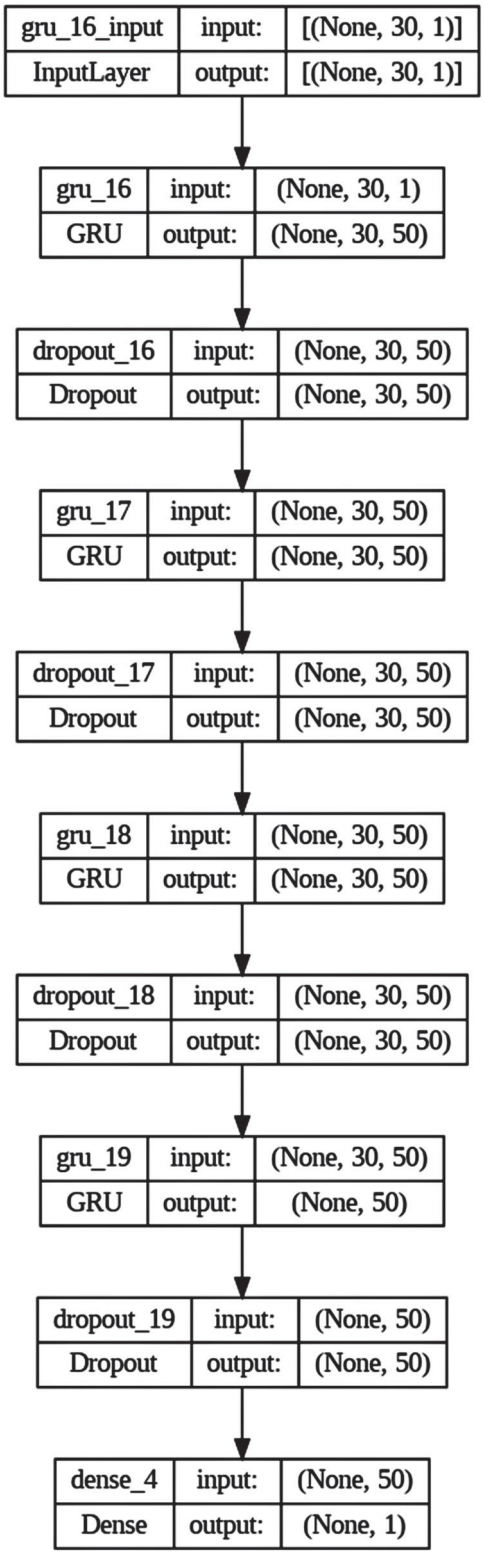

Figure 14. Proposed GRU model for Breast Cancer recurrence prediction.

the optimization process aiming to minimize it, hence lower values indicating better performance. Accuracy, on the other hand, is the ratio of correct predictions to all predictions made on the training data, typically inversely related to the loss. Validation metrics, like training metrics, provide similar measures but are computed using validation data, ensuring that they remain unseen by the model during training. The Binary Cross-Entropy (BCE) is employed to quantify the loss during each iteration, where "y" signifies the output label (1 for recurrence and 0 for no-recurrence), and "p(y)" represents the prediction probability for all "N" data points. The mathematical expression for BCE is presented in Equation 25.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i.\log\ [p(y_i)] + (1-y_i).\log\ [1-p(y_i)] \qquad (25)$$

The identification of pertinent features that aptly characterize the case type holds paramount importance in the realm of breast cancer recurrence prediction. Typically, these relevant features are chosen based on their correlation with the target variable and their capacity to differentiate between distinct categories. Once selected, these features serve as the foundation for constructing predictive models that excel in forecasting the outcomes of new cases. This process of pinpointing relevant features is a fundamental and pivotal stage in numerous data-driven applications, exerting a profound influence on the performance and precision of the ultimate model. In this formula, we observe that for all correctly predicted points (y = 1), the logarithm of the predicted probability (log(p(y))) is added to the loss to account for the log probability correction. Conversely, for incorrectly predicted points (y = 0), the sum of the logarithm of (1-p(y)) is computed to determine the log probability of the result being incorrect. Overfitting becomes a concern when the training loss and accuracy appear favourable, but the validation counterparts exhibit poor performance, indicating an inability to generalize to new data. To comprehensively evaluate the performance of breast cancer recurrence prediction, these metrics are typically combined into training and validation loss plots. These visual representations illustrate how the model's accuracy evolves across epochs or iterations during training. Throughout this training process, the model is exposed to labeled data, striving to discern underlying patterns and features associated with each label. Over multiple epochs, the model adapts its parameters and weights to better align with the data, enhancing its accuracy. The analysis of accuracy fluctuations over epochs can reveal the impact of novel feature optimization techniques in mitigating overfitting. In assessing the efficiency of the breast cancer recurrence prediction algorithm, training and validation accuracy are assessed across various epochs, as depicted in Figure 15 and Figure 16.

In the process of analyzing accuracy, the dataset is divided into two distinct categories: validation data and training data. Notably, metrics like precision, accuracy, recall, and F1-score tend to decrease as the percentage of validation data within the dataset

increases. Conversely, the performance metrics exhibit an upward trend as the percentage of training data increases. Throughout the experiment, various combinations of training and validation data are explored, with the optimal results achieved when using 30% for training and 70% for validation. Incorporating 30 feature maps and utilizing cross-validation, a diverse set of performance parameters is computed. The best performance is attained when allocating 30% of the data to validation and 70% to training. These results are derived from a 10-fold cross-validation approach, with the highest recorded accuracy reaching 97.63%. After rigorous experimentation with the dataset, the proposed models are fine-tuned, leading to improved classification outcomes. The choice of the algorithms is driven by its capability to enhance performance parameters, and this implementation of deep learning techniques stands out for its accuracy and computational efficiency.

The confusion matrix depicted in Figure 17 provides a clear idea about the performance of proposed LR-CNN-LSTM classification algorithm in discriminating various classes in the validation data. The total number of patients considered for validation is 114. In this validation dataset, 71 patients belong to class 0 (No-Recurrence) and 43 patients belong to class 1 (Recurrence). In the case of no-recurrence all 71 patients are correctly categorized. In the case of recurrence, 41 patients are correctly categorized and 2 patients are wrongly categorized. This indicates that, the proposed

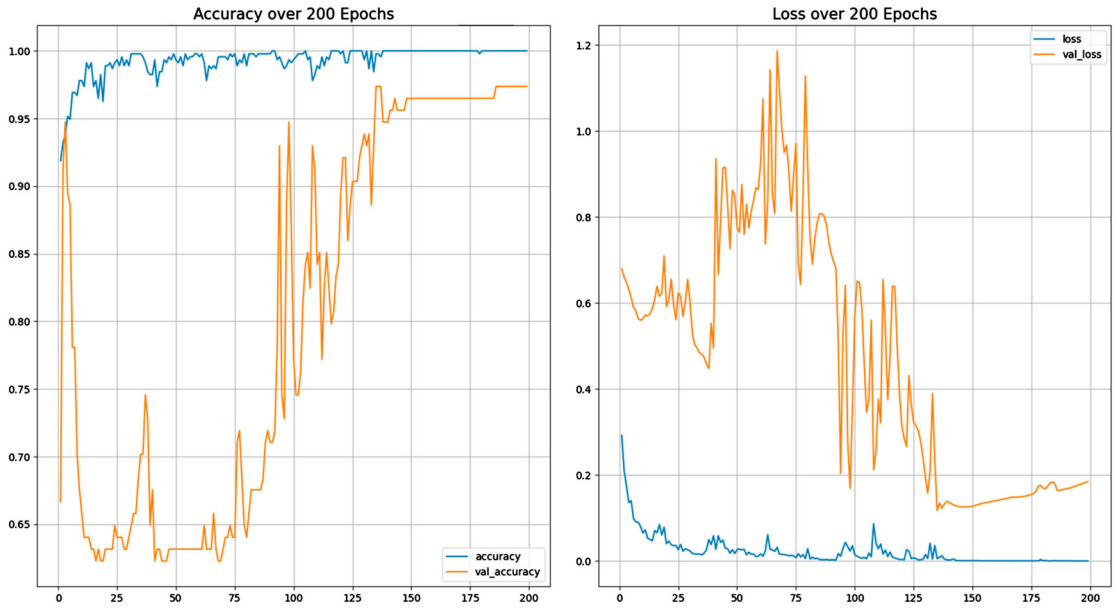

Figure 15. Average accuracy and loss per epoch for LR-CNN-LSTM classifier.

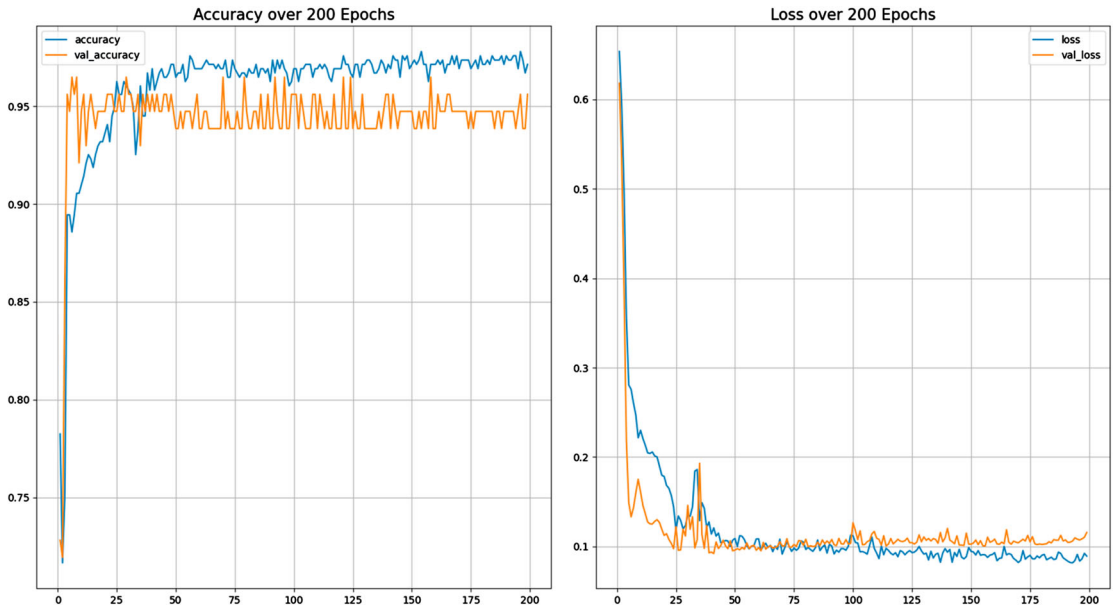

Figure 16. Average accuracy and loss per epoch for ANOVA-GRU classifier.

Table 3. Performance comparison.

Methodology	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naïve Bayes [18]	92.94	93.85	90.89	92.63
Random Forest [27]	96.71	96.77	95.14	94.36
KNN [26]	95.03	95.24	94.18	95.52
SVM [28]	95.70	96.28	93.48	95.83
Decision Tree [29]	90.46	91.34	90.18	90.89
Logistic Regression [30]	95.74	96.83	95.32	95.10
General CNN [31]	85.83	87.34	83.13	86.38
LR-CNN-LSTM	98.24	99.14	98.30	98.14
ANOVA-GRU	96.49	97.04	96.67	96.67

algorithm is able to discriminate between both classes in an efficient way.

The confusion matrix depicted in Figure 18 provides a clear idea about the performance of proposed ANOVA-GRU classification algorithm in discriminating various classes in the validation data. The total number of patients considered for validation is 114. In the case of no-recurrence, 70 patients are correctly categorized and only 1 patient is wrongly categorized. In the case of recurrence, 40 patients are correctly categorized and 3 patients are wrongly categorized. This indicates that, the proposed algorithm is able to discriminate between both classes in an efficient way.

The Receiver Operating Characteristic (ROC) curve serves as a visual representation of the binary prediction algorithm's performance. This curve plots the true positive rate (TPR) along the y-axis against the false positive rate (FPR) along the x-axis, while systematically varying the discrimination threshold. The area under the curve (AUC) stands as a key metric that assesses the algorithm's overall performance. Higher values of AUC indicate improved discrimination between positive and negative classes. In Figure 19, we present the ROC curve and AUC for the proposed CNN-LSTM Classifier, offering insights into its classification capabilities.

The ROC curve is a vital tool for evaluating a model's capacity to separate classes, offering valuable insights as the decision threshold varies. This graphical representation showcases the trade-off between sensitivity and specificity. The resultant curve visually depicts how the model's performance fluctuates across various decision thresholds. In Figure 20, we present the ROC curve and AUC for the proposed ANOVA-GRU Classifier, shedding light on its classification performance characteristics.

The performance of the proposed breast cancer recurrence prediction scheme is compared with existing schemes. The mean values of accuracy, precision, recall and F1-score for the proposed LR-CNN-LSTM model were calculated as 98.24%, 99.14%, 98.30% and 98.14% respectively. The mean values

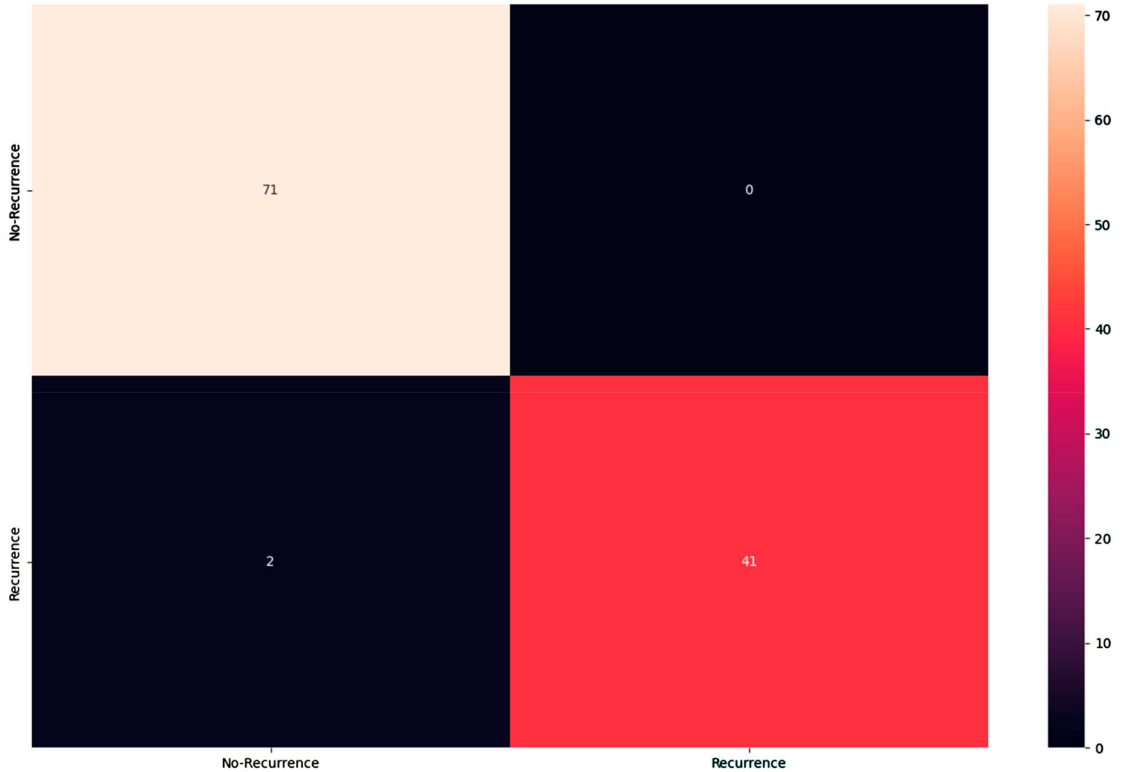

Figure 17. Confusion matrix for LR-CNN-LSTM classifier.

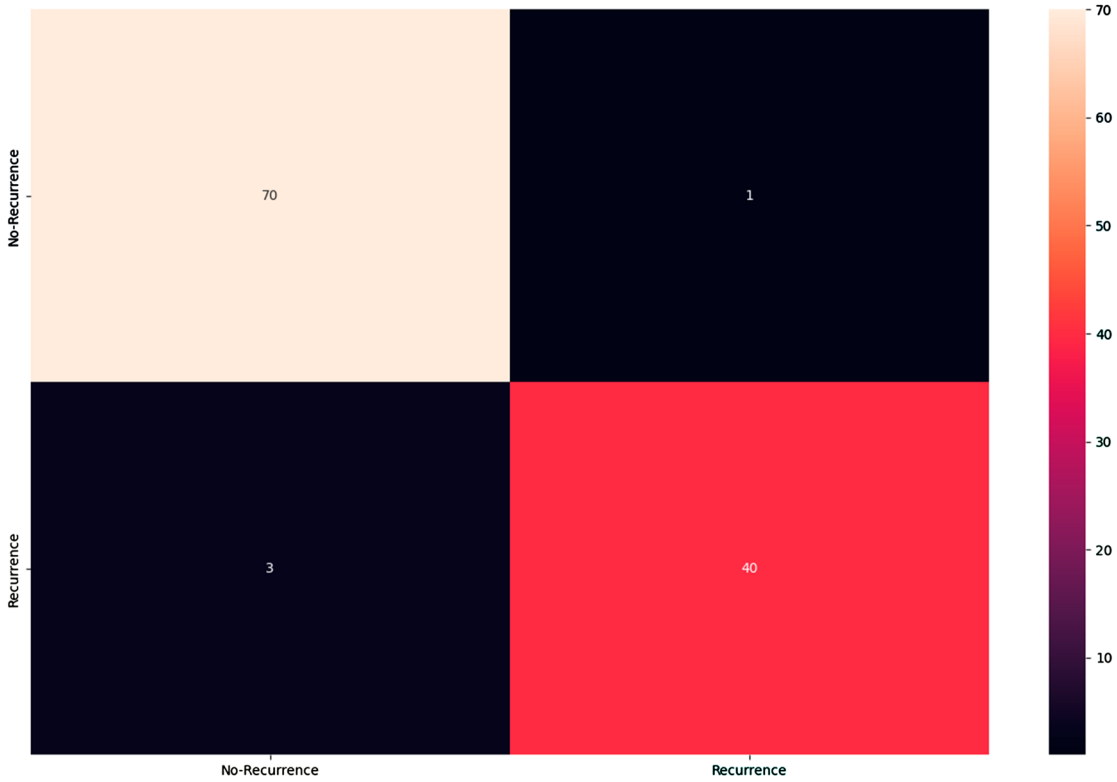

Figure 18. Confusion matrix for ANOVA-GRU classifier.

of accuracy, precision, recall and F1-score for the proposed ANOVA-GRU model were calculated as 96.49%, 97.04%, 96.67% and 96.67% respectively. The reason for obtaining good performance is due to the incorporation of recurrence and additional layers. The comparison of performance among proposed schemes with existing schemes is presented in Table 3 and depicted in Figure 21.

The average performance of WBC dataset is compared with seven existing techniques. From the above

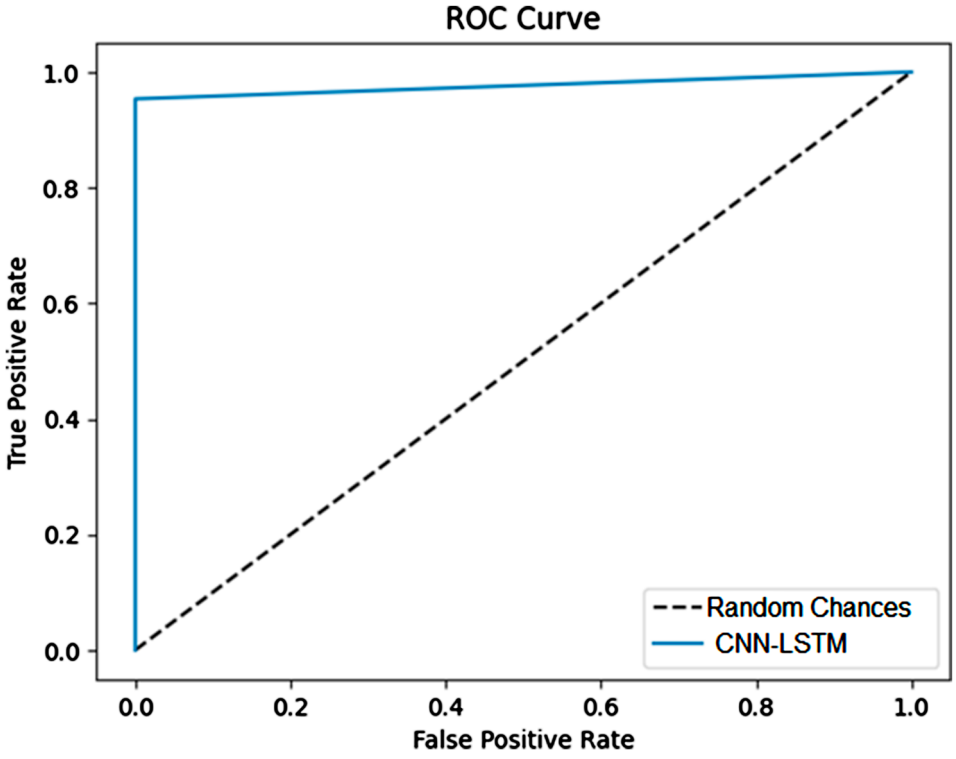

Figure 19. ROC curve of CNN-LSTM classifier algorithm.

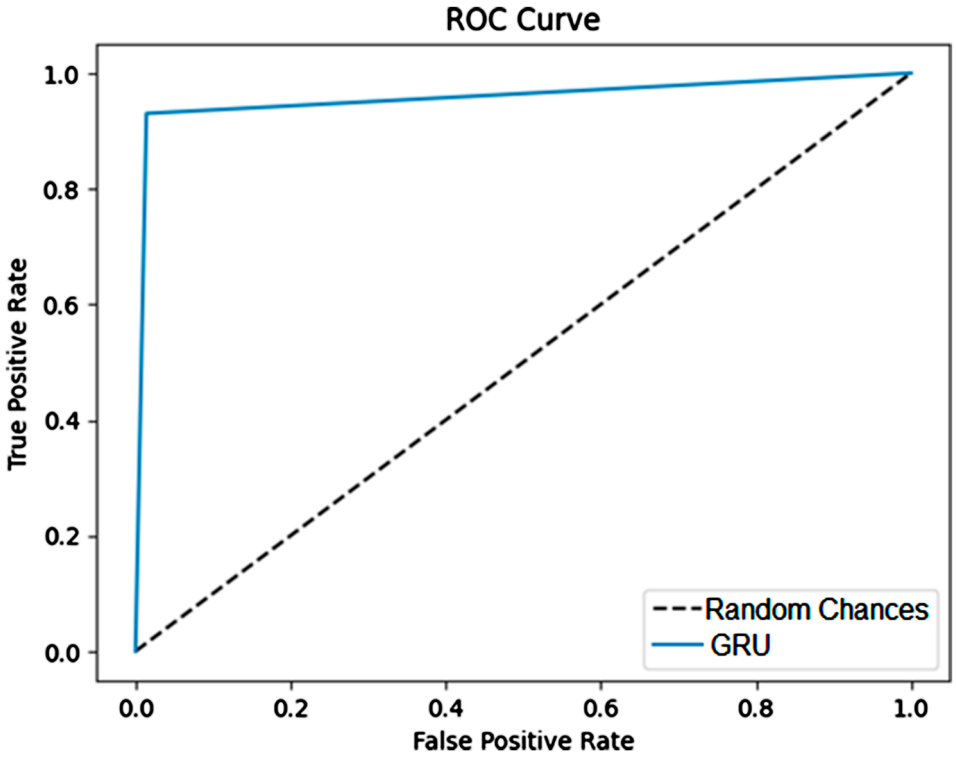

Figure 20. ROC curve of ANOVA-GRU classifier algorithm.

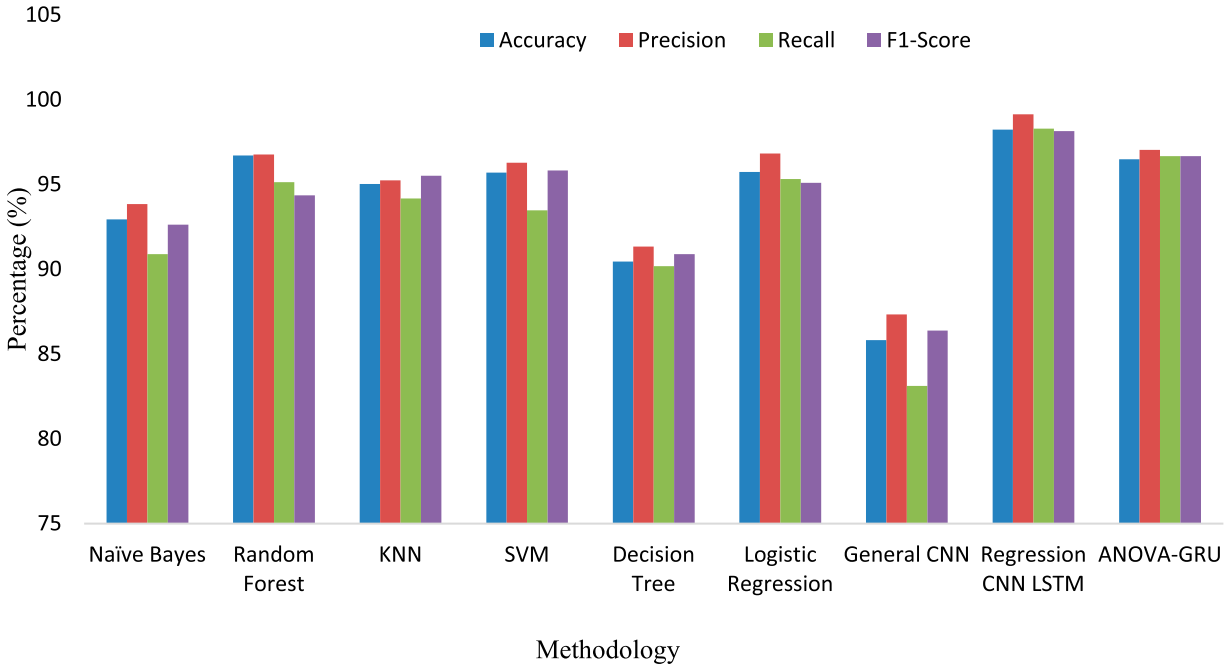

Figure 21. Comparison of performance.

comparison, the accuracy for proposed LR-CNN-LSTM method is 12.41% higher than general CNN model and 1.53% higher than Random Forest. The precision measure of proposed LR-CNN-LSTM method is 11.8% higher than CNN and 2.31% higher than LR. The recall measure for proposed LR-CNN-LSTM method is 15.17% higher than CNN and 2.98% higher than Logistic Regression. The F1-score for LR-CNN-LSTM scheme is 11.76% higher than CNN and 2.31% higher than SVM. The ANOVA-GRU method outperforms the general CNN model, showcasing a significant 10.66% increase in accuracy. Moreover, in terms of precision, the proposed ANOVA-GRU method excels by 9.7% compared to CNN. When considering recall, the ANOVA-GRU method boasts an impressive 13.54% improvement over CNN and a 1.35% edge over LR. Additionally, the F1-score for the ANOVA-GRU approach surpasses CNN by 10.29% and SVM by

0.84%. Optimized layer selection effectively mitigates overfitting concerns while enhancing network performance.

6. Conclusion

This study introduces two innovative systems for the prediction and classification of breast cancer recurrence, leveraging CNN, LSTM, and GRU models. Fuzzy Computation. Many system performance criteria, parameters, and decision factors are not always essential, nor is it always practicable, to quantify them precisely. Variables are considered uncertain or fuzzy when their values are not well-defined.

The evaluation is conducted on the WBC dataset, encompassing both training and validation phases. In this framework, 30 features extracted from the WBC dataset are employed to predict breast cancer recurrence likelihood. Feature optimization is achieved through LR and ANOVA techniques. The training phase involves minimizing validation loss by optimizing the number of epochs. Various combinations of training and validation datasets are explored in the experimentation process.

The mean performance metrics for the proposed LR-CNN-LSTM model are as follows: accuracy (98.24%), precision (99.14%), recall (98.30%), and F1-score (98.14%). Similarly, the mean metrics for the proposed ANOVA-GRU model are as follows: accuracy (96.49%), precision (97.04%), recall (96.67%), and F1-score (96.67%). Empirical findings affirm that the LR-CNN-LSTM model-based breast cancer recurrence prediction system outperforms alternative algorithms. To further enhance the algorithm's performance, fine-tuning of network parameters is a potential avenue. Future research directions involve refining the CNN model by incorporating additional layers and conducting fine-tuning for continued performance enhancement.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin. 2018;68(6):394–424. doi:10.3322/caac.21492

[2] Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. doi:10.1016/j.media.2017.07.005

[3] Wolberg WH, Street WN, Heisey DM, et al. Computer-derived nuclear features distinguish breast cytology. Hum Pathol. 1992;23(4):422–429.

[4] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–1182.

[5] Zhang J, Wu X, Zhang M. Genomic sequencing capacity, data retention, and personal access to Raw data in europe. Front Genet. 2020;11:303), doi:10.3389/fgene.2020.00303

[6] Wang H, Li Y, Khan SA, et al. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. Artif Intell Med. 2020;110:101977), doi:10.1016/j.artmed.2020.101977

[7] Macías-García L, Martínez-Ballesteros M, Luna-Romera JM, et al. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. Artif Intell Med. 2020;110:101976), doi:10.1016/j.artmed.2020.101976

[8] Yang J, Ju J, Guo L, et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. Comput Struct Biotechnol J. 2022;20:333–342. doi:10.1016/j.csbj.2021.12.028

[9] Huang SH, Loh JK, Tsai JT, et al. Hydrochloric acid-enhanced radiofrequency ablation for treating a large hepatocellular carcinoma with spontaneous rapture: a case report. Chin J Cancer. 2017;36(1):1–9. doi:10.1186/s40880-016-0161-8

[10] Chatterjee CC, Krishna G. (2019). A novel method for IDC prediction in breast cancer histopathology images using deep residual neural networks. In 2019 2nd international conference on intelligent communication and computational techniques (ICCT) (pp. 95–100). IEEE.

[11] Aishwarja AI, Eva NJ, Mushtary S, et al. (2020). Exploring the machine learning algorithms to find the best features for predicting the breast cancer and its recurrence. In International conference on intelligent computing & optimization (pp. 546–558). Springer, Cham.

[12] Chang CC, Chen SH. Developing a novel machine learning-based classification scheme for predicting SPCs in breast cancer survivors. Front Genet. 2019;10:848), doi:10.3389/fgene.2019.00848

[13] Boeri C, Chiappa C, Galli F, et al. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med. 2020;9(9):3234–3243. doi:10.1002/cam4.2811

[14] Bakre MM, Ramkumar C, Attuluri AK, et al. Clinical validation of an immunohistochemistry-based can assist-breast test for distant recurrence prediction in hormone receptor-positive breast cancer patients. Cancer Med. 2019;8(4):1755–1764. doi:10.1002/cam4.2049

[15] Cespedes Feliciano EM, Kwan ML, Kushi LH, et al. Body mass index, PAM50 subtype, recurrence, and survival among patients with nonmetastatic breast cancer. Cancer. 2017;123(13):2535–2542. doi:10.1002/cncr.30637

[16] Kim W, Kim KS, Park RW. Nomogram of naive Bayesian model for recurrence prediction of breast cancer. Healthc Inform Res. 2016;22(2):89–94. doi:10.4258/hir.2016.22.2.89

[17] Huang SC, Pareek A, Seyyedi S, et al. Blockchain vehicles for efficient Medical Record management. NPJ Dig Med. 2020;3(1):1–9. doi:10.1038/s41746-019-0211-0

[18] Smith TP, Ezhilarasi TP, Balamurugan K. Bio-inspired weighed quantum particle swarm optimization and smooth support vector machine ensembles for identification of abnormalities in medical data. SN Appl Sci. 2019;1(10):1–10.

[19] Li JP, Haq AU, Din SU, et al. Heart disease identification method using machine learning classification in e-healthcare. IEEE Access. 2020;8:107562–107582. doi:10.1109/ACCESS.2020.3001149

[20] Garcia F, El-Sappagh S, Islam SR, et al. Real evaluation for designing sensor fusion in UAV platforms. Informat Fusion. 2020;63:208–222. doi:10.1016/j.inffus.2020.06.003

[21] Maniruzzaman M, Rahman M, Ahammed B, et al. Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 2020;8(1):1–14. doi:10.1007/s13755-019-0095-z

[22] Stephen O, Sain M, Maduh UJ, et al. An efficient deep learning approach to pneumonia classification in healthcare. J Healthc Eng. 2019;2019.

[23] Breast Cancer Wisconsin Dataset, Kaggle. Accessed Feb 13, 2020. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

[24] Khan S, Islam N, Jan Z, et al. A novel deep learning-based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognit Lett. 2019;125:1–6. doi:10.1016/j.patrec.2019.03.022

[25] Battineni G, Sagaro GG, Chinatalapudi N, et al. Applications of machine learning predictive models in the chronic disease diagnosis. J Pers Med. 2020;10(2):21), doi:10.3390/jpm10020021

[26] Magboo VPC, Magboo MSA. Machine learning classifiers on breast cancer recurrences. Procedia Comput Sci. 2021;192:2742–2752. doi:10.1016/j.procs.2021.09.044

[27] Quist J, Taylor L, Staaf J, et al. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. Cancers (Basel). 2021;13(5):991), doi:10.3390/cancers13050991

[28] Ahmad LG, Eshlaghy AT, Poorebrahimi A, et al. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform. 2013;4(124):3.

[29] Guo J, Fung BC, Iqbal F, et al. Revealing determinant factors for early breast cancer recurrence by decision tree. Informat Syst Front. 2017;19:1233–1241. doi:10.1007/s10796-017-9764-0

[30] Witteveen A, Nane GF, Vliegen IM, et al. Comparison of logistic regression and Bayesian networks for risk prediction of breast cancer recurrence. Med Dec Mak. 2018;38(7):822–833. doi:10.1177/0272989X18790963

[31] Desai M, Shah M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clin eHealth. 2021;4:1–11. doi:10.1016/j.ceh.2020.11.002