Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NOVEL APPROACH FOR BREAST CANCER PREDICTION USING IMPROVED GATED RECURRENT UNIT

TINTU P B[1], DR. S VENI[2]

[1]Research Scholar,Department of Computer Science,Karpagam Academy of Higher Education Coimbatore,Tamil Nadu, India
[2]Professor,Department of Computer Science,Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
E-mail: [1]tintupadikkal@gmail.com,[2]venics@kahedu.edu.in

ABSTRACT

The cells that make up the breast tissue can transform into a tumorous malignancy known as breast cancer. Although males are not immune, it is more frequent in women and is among the most common malignancies globally. In most cases, the illness starts in the breast's ducts or lobules but, if left untreated, can metastasize to other organs. Predicting the likelihood of breast cancer is essential for both early detection and treatment planning. This study presents an advanced approach to breast cancer prediction using an Improved Gated Recurrent Unit (GRU) model. The methodology begins with preprocessing a CSV dataset of numerical records, where Recursive Feature Selection with Extra Tree Classifier (RFET) is employed to identify the most relevant features, enhancing the model's predictive accuracy. Following feature selection, an Improved GRU model is utilized for classification and predictive modelling. The improved GRU architecture incorporates optimizations to improve learning efficiency and accuracy, utilizing temporal dependencies within the data. High predictive performance was achieved by the suggested method, according to the findings, providing a useful tool for early identification and diagnosis of breast cancer.

Keywords: *Breast Cancer, Early Diagnosis, Extra Tree Classifier, Improved Gated Recurrent Unit, Recursive Feature Selection*

1.INTRODUCTION

Cancer can enter the body if the cell divides uncontrollably and subsequently spreads to other areas. When it comes to female-specific killers, breast cancer ranks second [1]. The World Health Organization (WHO) reports that there are more than 2.3 million cases of breast cancer each year, making it the most frequent kind of cancer globally [2]. Breast cancer is first or second among cancers affecting women worldwide in 95% of nations [3]. Ten million individuals lost their lives to cancer, and an estimated twenty million new cases were recorded worldwide [4]. Breast cancer can be defined as either malignant or benign, with aberrant cell proliferation being a defining feature of the illness [5]. Much research has looked at machine learning as a potential tool for breast cancer detection [6]. Improving prediction accuracy to allow correct diagnosis has continuously been the focus of the research, regardless of the dataset's features [7]. While ML algorithm modalities have shown promise on different breast cancer datasets, they still lack the consistency and accuracy needed for a reliable diagnosis [8-9] without the use of targeted data mining techniques.

There is an immediate need for efficient early detection and prediction modelling approaches since breast cancer is still among the most common cancers globally [10, 11]. Because medical data typically contains complex temporal patterns, traditional prediction approaches that depend on static variables and linear models can fail to do so [12–13]. The possibility of using sophisticated machine learning methods to raise the bar for prediction accuracy is investigated in this research [14]. It focuses on the use of an Improved Gated Recurrent Unit (GRU) model, which can detect patterns and relationships in sequential data over time [15]. The study's goal is to improve the GRU model's performance [16–17] by preprocessing and selecting important features from

a numerical dataset using Recursive Feature Selection with Extra Tree Classifier (RFET). This method has the potential to enhance diagnostic results and patient care by providing a more reliable tool for early breast cancer prediction [18].

The main contribution of the paper is:

- Feature Selection Using Recursive Feature Selection with Extra Tree Classifier
- Classification Using an Improved Gated Recurrent Unit

The rest of the paper is followed with Multiple authors address various approaches to breast cancer prediction in Section 2. The proposed model is discussed in Section 3. Section 4 provides a summary of the investigation's results. Section 5 concludes with an analysis of the results and suggestions for future scope.

1.1 Motivation of the paper

While useful, traditional techniques of diagnosis often encounter issues with efficiency and accuracy. This study seeks to address these challenges by utilizing advanced machine-learning techniques. Specifically, the use of an Improved Gated Recurrent Unit (GRU) model, combined with Recursive Feature Selection and an Extra Tree Classifier, aims to enhance predictive performance and learning efficiency. Early diagnosis and improved treatment planning are made possible by this method's emphasis on maximizing feature selection and model accuracy, which can lead to better patient outcomes and more progress in breast cancer research.

2. LITERATURE REVIEW

Derangula, S et al. [2] Using the machine learning algorithms LGBM, CATBBOOST, and XGB, the author has determined the optimal parameters for the Wisconsin Breast Cancer Diagnostic dataset in this research. By providing all characteristics and optimizing them, the author evaluated how well the Naïve Bayes classifier performed.

Ghosh, P. et al. [6] Based on the findings shown above, it seems that a combination of Deep Learning Models might provide reliable estimates for Breast Cancer Detection. When looking at the big picture, it was clear that LSTM and GRU were crucial in producing valuable outcomes. Perhaps the reason these two algorithms outperform the others in terms of accuracy is that they have already discovered characteristics that significantly affect training success.

Kayikci, S., & Khoshgoftaar, T. M. [7] Thanks to improvements in AI technology, especially CNNs and deep learning approaches, disease detection has come a long way. These methods can accomplish classification results that surpass those of human specialists without necessitating a feature domain definition. The outcomes of the deep learning method for evaluating the risk of breast cancer using mammograms were encouraging.

Khamparia, S. et al. [8] An MVGG network pre-trained on ImageNet achieves 94.3% accuracy and 93.3% AUC, making it the top-performing design. Therefore, the author used a clinical classification criterion that was much lower than the mathematical one. The number of mammograms that come back negative will drop significantly with the use of this algorithm. Additionally, this will improve the odds of surviving for five years.

S, Rajdarsan et al. [12] The purpose of this paper was to discuss significant improvements in the methods used to deal with missing data during pre-processing. For some reason, instead of normalizing the 'Bare-Nuclei' values to fill in the gaps caused by missing data, the author ended up with models in which the relative importance of various attributes was skewed, with the 'Bare-Nuclei' attribute's weight moving closer to the normalized value. This leads to inaccurate accuracy calculations.

Saoud, H. et al. [14] these authors' research work aimed to use feature selection strategies to increase breast cancer classification accuracy. In this study, the author drew on the original (WBC) and diagnostic (WBCD) Wisconsin breast cancer datasets. The author sees the inverse for some classifiers, such as SVM, while for others, like Bayes net, the feature selection method enhanced accuracy in WBC and WBCD. Because of the feature selection approach, classification accuracy has been diminished. While Support Vector Machines without feature selection were the beast for WBCD, Bayes Network with feature selection was the best for WBC breast cancer classification.

Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Survey of Methodologies and Techniques in Breast Cancer Prediction and Detection

Author	Year	Methodology	Advantage	Limitation
Chaurasia & Pal	2021	Stacking-based ensemble framework with feature selection	Improved accuracy in breast cancer detection through ensemble learning	Can involve complex computation and longer processing time
Dutta et al.	2020	Stacked GRU-LSTM-BRNN for breast cancer prediction	Effective in capturing temporal dependencies, leading to better prediction performance	High computational cost and requires large datasets for training
Eroltu	2023	Genetic algorithm for feature selection	Efficient in reducing dimensionality, leading to faster and more accurate predictions	Performance can degrade if the genetic algorithm is not properly tuned
Sharma et al.	2022	Extra Tree Classifier	Enhanced prediction accuracy through feature ensemble and model combination	Complex model that might be challenging to implement and interpret
Taghizadeh et al.	2022	ML methods	High accuracy in prediction by utilizing genetic data	Requires high-quality and large-scale datasets for effective performance

2.1 Motivation of the paper

The pressing need for more precise early diagnosis and prognosis of breast cancer, the most common cancer killer of women, is the driving force behind this research. Although they work, traditional diagnostic approaches have their limitations when it comes to speed and precision. The goal is to improve learning efficiency and prediction performance using a combination of models, including an Improved Gated Recurrent Unit (GRU), Recursive Feature Selection, and an Extra Tree Classifier. Early diagnosis and improved treatment planning are made possible by this method's emphasis on maximizing feature selection and model accuracy, which can lead to better patient outcomes and more progress in breast cancer research.

.

3.MATERIALS AND METHODS

Here, we go into the specifics of the breast cancer prediction approaches that have been suggested. To get the data ready for analysis, we start

3.1 Dataset collection

by preprocessing a CSV dataset that contains numerical information. The next step in improving the model's performance is to use RFET, which stands for Recursive Feature Selection with Extra Tree Classifier, to find and choose the most important features. Following feature selection, an Improved Gated Recurrent Unit (GRU) model is utilized for classification and predictive modelling.

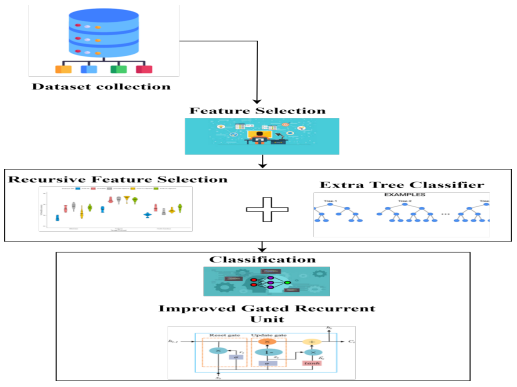

Figure 1: Proposed workflow architecture

The dataset used in this study was obtained from Kaggle, specifically from the "Breast Cancer Wisconsin (Diagnostic) Data Set" available

at https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. This dataset, originally sourced from the University of Wisconsin-Madison, provides a comprehensive collection of features related to breast cancer diagnoses. It includes numerical records of various characteristics extracted from breast cancer biopsies, such as mean values of cell nuclei features, standard deviations, and other statistical measures.

3.2 Feature Selection Using Recursive Feature Selection with Extra Tree Classifier

An efficient feature selection method that integrates Recursive Feature Elimination (RFE) with the Extra Tree Classifier (ETC) is known as Recursive Feature Selection with Extra Tree Classifier (RFET). Using the model's performance as a metric, RFE repeatedly removes characteristics that are not relevant before determining the most relevant subset. One such method is the Extra Tree Classifier, which ranks features according to their relevance and uses an ensemble of decision trees to handle complicated feature interactions while reducing overfitting. By combining these techniques, RFET improves prediction performance while decreasing computing complexity by honing down the most important elements of the model.

A ranked feature list is what SVM-RFE produces. Selecting a set of highly valued characteristics is the first step in feature selection. Similarities between the SVM model and the SVM-RFE ranking criteria are strong. SVM's great generalizability and excellent accuracy make it a popular approach for classification. Several e-nose applications have found success with it. Consequently, ranking criteria generated from this model will likely perform well.Finding a separating hyperplane with the biggest margin is the rationale behind support vector machines (SVM). A margin of two times the distance between the training sample nearest the separating hyperplane is used in linear separable scenarios.

$$f(x) = w.x + b \quad \text{---------- (1)}$$

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{ij=1}^{n} \alpha_i \quad \text{--------- (2)}$$

where the Lagrange multipliers are denoted by ¸i. By maximizing LD with the restrictions that ¸i > 0 and n i=1¸iyi = 0, we can find solutions of ¸i. Support vectors are the samples that correlate to nonzero ¸'s. It is thus possible to derive the weight vector w by

$$w = \sum_{i=1}^{n} a_i y_i x_i \quad \text{---------- (3)}$$

$$j(k) = w_k^2 \quad \text{---------- (4)}$$

The attribute that has the least impact on categorization is eliminated because it has the lowest ranking criteria. The next cycle uses the retained characteristics to train an SVM model. This procedure is carried out again and again until every characteristic has been eliminated. The traits are then organized based on the sequence in which they were removed. The significance of a feature should increase as its removal date approaches. It takes a lot of time to remove features one by one when the feature dimension is high.

3.3 Classification Using an Improved Gated Recurrent Unit

The GRU architecture's improved capabilities are used in classification employing an Improved Gated Recurrent Unit (GRU) to efficiently manage sequential and temporal data for predictive modelling. To capture dependencies in input sequences, GRUs—a kind of RNN—use gating mechanisms to keep track of previous information. To increase speed and learning efficiency, the upgraded GRU model expands upon the conventional GRU with improvements such as extra layers, better gating schemes, or sophisticated training approaches. Particularly well-suited for sequence-analysis tasks, this improved architecture is skilled at processing data including complicated patterns and temporal correlations. With these enhancements implemented, the model becomes more accurate and resilient in classification tasks, allowing for more accurate predictions and improved handling of different data sequences.

A RNN variation, GRU is similar to LSTM in that it is primarily designed to address issues related to gradients in backpropagation and long-term memory. But with one less gate, GRU's construction is simpler than L STM's. It is the responsibility of the reset gate to regulate the storage of past data. The less data from the past is disregarded, as the reset gate's value increases.

$$g_r = \sigma(W_r.[s_{t-1}, x_t] + b_r) \quad \text{---------- (5)}$$

$$g_z = \sigma(W_z.[S_{t-1}, x_t] + b_z) \quad \text{---------- (6)}$$

$$s_t = \tanh(W_h.[g_r * s_{t-1}, x_t] + b_z) \quad \text{--------- (7)}$$

the reset gate, the update gate, the candidate cell state, and the current new cell state are represented by gr, gz, sˆt, and st, respectively. Here the input is

denoted as tx = [Vt, It, Tt], the weight is denoted as W, the bias is denoted as b, the sigmoid function is presented in Equation (5) as σ(•), and the tanh function is shown in Equation (6) as tanh(•).

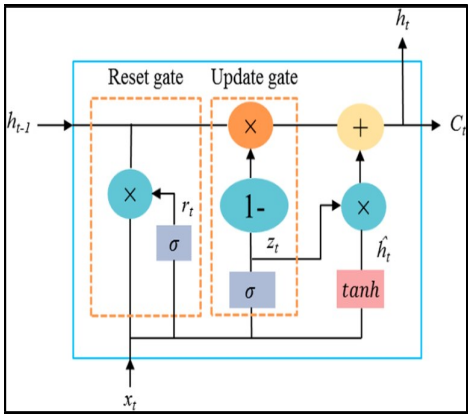

Figure 2: Improved GRU architecture

$$\sigma(x) = \frac{1}{1+e^{-x}} \text{ ---------- (8)}$$

The GRU design can exchange parameters at multiple time steps and does a good job of processing large time series data. Time series data with just one dimension is what the battery data—voltage, current, and temperature—are. It is logical to estimate SOC.

using GRU. The dataset $\{(x1,y1),\{(x2,y2),\bullet\bullet\bullet,\{(xt,yt)\}$ was used to train the GRU network in this article. The GRU's input vector was $xt = [Vt, It, Tt]$ and its output value was $yt = SOC_t$

Algorithm 1: Improved Gated Recurrent Unit
Input: Breast cancer dataset containing features such as mean radius, mean texture, mean perimeter, and others. Steps Weights and Biases: Weight Matrices: W_r, W_z, W_h Bias Vectors: b_r, b_z, b_h Sigmoid Function Process: 2. $g_r = \sigma(W_r.[s_{t-1}, x_t] + b_r)$

Reset Gate Calculation:

Controls how much of the previous hidden state s_{t-1}, x_t should be forgotten.

Update Gate Calculation:

Determines how much of the previous state S_{t-1}, x_t will be carried forward to the next state.

$$g_z = \sigma(W_z.[S_{t-1}, x_t] + b_z)$$

Generates the potential new state for the current time step.

$$s_t = \tanh(W_h.[g_r * s_{t-1}, x_t] + b_z)$$

Combines the previous state and candidate state to form the new hidden state.

Output:

The predicted classification output y_t at time step t, representing the breast cancer status based on the input features.

4. RESULTS AND DISCUSSION

In this section, we present the results obtained from the performance evaluation of the proposed methods implemented by using python programming language. The analysis focuses on the key metrics of accuracy, precision, recall, and F-measure across the different models—RNN, LSTM, GRU, and IGRU.

Figure 3 shows the relationship between perimeter_mean and area_mean the x-axis shows the perimeter means and the y-axis shows the area mean

Journal of Theoretical and Applied Information Technology
15th May 2025. Vol.103. No.9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

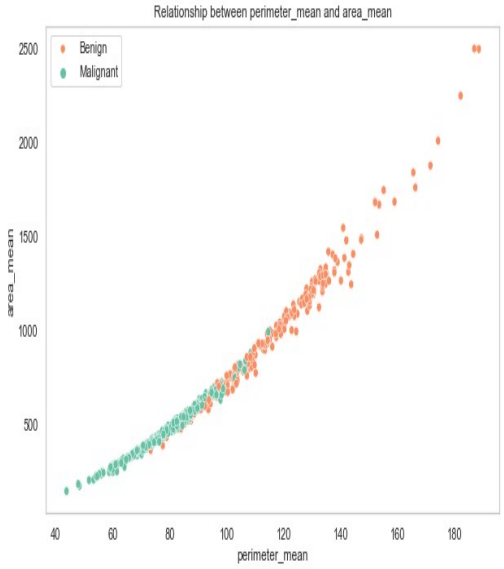

Figure 3: Relationship between perimeter_mean and area_mean

Figure 4: Relationship between radius mean and texture mean

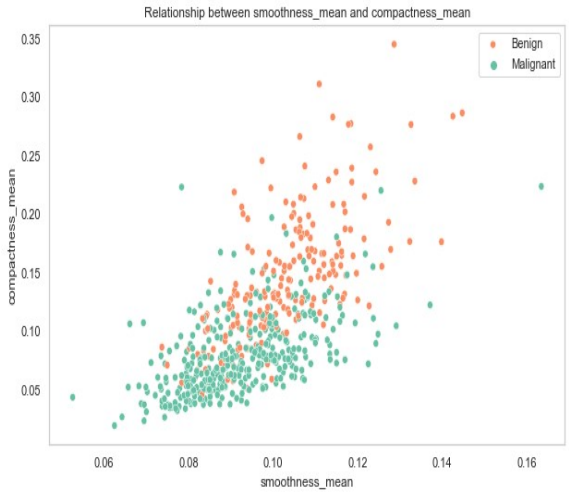

Figure 5: Relationship between smoothness mean and compactness mean

In Figure 5, we can see the connection between the mean of smoothness and the mean of compactness. Mean compactness is shown on the y-axis while mean smoothness is shown on the x-axis.

The correlation between the average radius and the average texture is seen in Figure 4. The x-axis displays the average radius, while the y-axis displays the average texture.

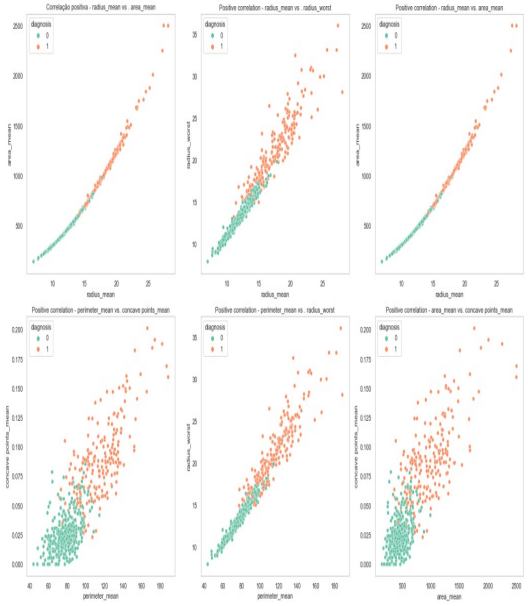

Figure 6: Plot graph

Figure 6 consists of six scatter plots that visually represent the correlations between various features used in breast cancer classification. Each plot compares two different

features, with data points coloured by the "diagnosis" category, indicating whether the sample is benign or maligna

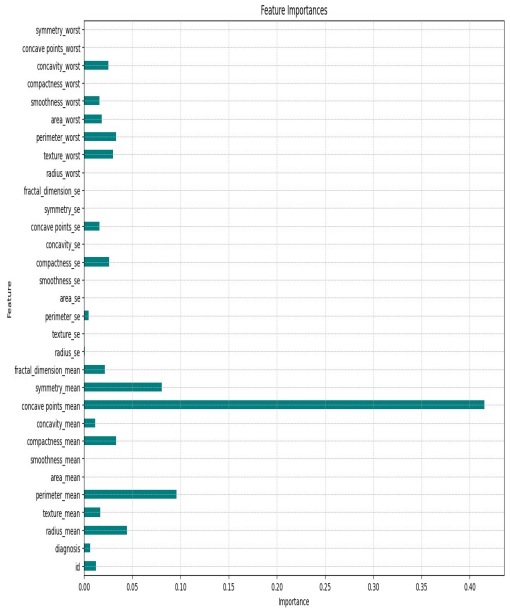

Figure 7: Feature importance chart

As seen in Figure 7, In a feature importance chart, features are shown along the y-axis and significance is shown along the x-axis.

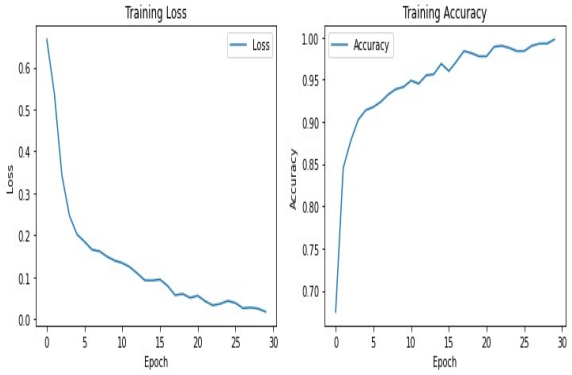

Figure 8: Training loss and accuracy comparison chart

If you look at picture 8, you can see a graph that compares training loss with accuracy. The y-axis displays training loss and accuracy value, while the x-axis displays epochs.

Table 2: Performance metrics comparison table

Methods	Accuracy	Precision	Recall	F-measure
RNN	**95.12**	**95.16**	**95.28**	**95.31**
LSTM	**96.15**	**96.27**	**96.11**	**96.10**
GRU	**97.31**	**97.02**	**97.15**	**97.14**
IGRU	**98.10**	**93.02**	**95.31**	**94.36**

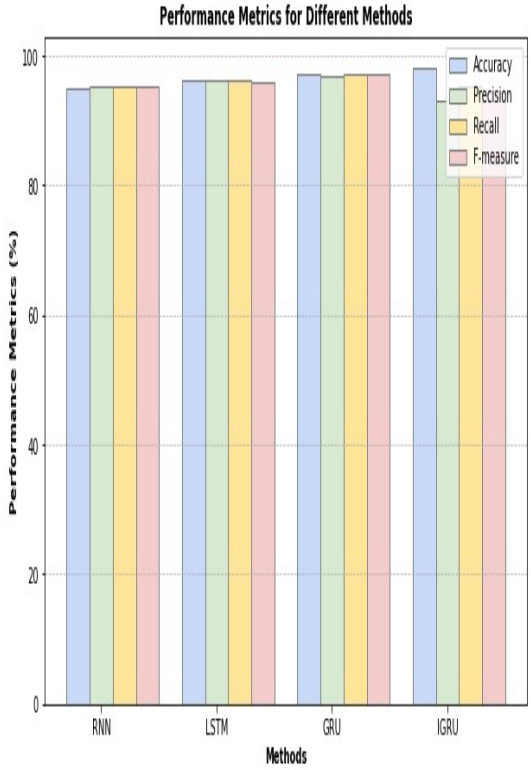

Figure 9: Performance metrics comparison chart

Table 2 and Figure 9 show the performance evaluation of the four methods—RNN, LSTM, GRU, and IGRU—revealing distinct variations across key metrics: accuracy, precision, recall, and F-measure. The IGRU model outperformed the others in terms of accuracy, achieving a high of 98.10%, indicating its superior capability in correctly classifying instances overall. However, it displayed a lower precision of 93.02%, suggesting some trade-offs in terms of the model's ability to correctly identify positive instances. The recall and F-measure for IGRU were 95.31% and 94.36%, respectively, showing balanced but slightly reduced performance compared to its accuracy. In contrast, the GRU model demonstrated consistent performance across all metrics, with 97.31% accuracy, 97.02% precision, 97.15% recall, and 97.14% F-measure, highlighting its robustness. Meanwhile, LSTM showed strong overall performance with an accuracy of 96.15% and precision of 96.27%, but slightly lower recall and F-measure values of 96.11% and 96.10%, respectively. RNN, while performing well, lagged with an accuracy of 95.12%, precision of 95.16%, recall of 95.28%, and F-measure of 95.31%.

5. CONCLUSION

In this study, the integration of Recursive Feature Selection with Extra Tree Classifier (RFET) and an Improved Gated Recurrent Unit (GRU) model has shown significant promising results for revolutionizing breast cancer prediction systems.By efficiently selecting the most relevant features and utilizing the optimized GRU architecture to extract temporal patterns, The RFET method effectively highlighted the most relevant features from the numerical dataset, which were then utilized by the improved GRU model to improve predictive performance. The optimized GRU architecture successfully captured temporal patterns in the data, leadingtosignificant gains in accuracy and efficiency. These results highlight the promise of state-of-the-art machine learning methods for improving breast cancer detection systems that identify the disease at an earlier stage. By improving prediction accuracy, this approach not only contributes to more reliable diagnostic tools but also holds the promise of enhancing patient outcomes through timely and precise intervention.

REFERENCES

[1]. Chaurasia, V., & Pal, S. (2021). Stacking-based ensemble framework and feature selection technique for the detection of breast cancer. SN Computer Science, 2(2), 67.

[2]. Derangula, S.Edara & P. K. Karri. (2020) . Feature Selection Of Breast Cancer Data Using Gradient Boosting Techniques Of Machine Learning, Clinical Medicine. Available from: https://www.academia.edu/66487804.

[3]. Dutta, S., Mandal, J. K., Kim, T. H., & Bandyopadhyay, S. K. (2020). Breast cancer prediction using stacked GRU-LSTM-BRNN. Applied Computer Systems, 25(2), 163-171.

[4]. Eroltu, K. (2023). Using genetic algorithm for breast cancer feature selection. International Research Journal of Oncology, 6(2), 203-226.

[5]. F. A. Muhammet. (2020).A Comparative Analysis Of Breast Cancer Detection And Diagnosis Using Data

Visualization And Machine Learning Applications, Healthcare. Available from: https://doi.org/ 10.3390/healthcare8020111.

[6]. Ghosh, P., Azam, S., Hasib, K. M., Karim, A., Jonkman, M., & Anwar, A. (2021, July). A performance-based study on deep learning algorithms in the effective prediction of breast cancer. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[7]. Kayikci, S., & Khoshgoftaar, T. M. (2023). Breast cancer prediction using gated attentive multimodal deep learning. Journal of Big Data, 10(1), 62.

[8]. Khamparia, S. Bharati, P. Podder, D. Gupta, A. Khanna, T. K. Phung & D. N. H. Thanhl. (2021). Diagnosis Of Breast Cancer Based On Modern Mammography Using Hybrid Transfer Learning, Multidimensional Systems And Signal Processing,32,747. Available from:https://doi.org/ 10.1007/s11045-020-00756-7.

[9]. LooNL, ChiewYS, TanCP, MatNorMB, RalibAM. (2021). A Machine Learning Approach To Assess Magnitude Of Asynchrony Breathing.Biomedicalsignalprocessing And Control.Available from: https://doi.org/10.1016/j.bspc.2021.102505.

[10]. M. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan & M.N. Kabir. (2020). Breast cancer prediction: A comparative study using machine learning techniques, SN Computer Science.Available from: https://doi.org/10.1007/s42979-020-00305-w.

[11]. N. F. Idris & M. A. Ismail. (2021). Breast Cancer Disease Classification Using Fuzzy-ID3 Algorithm With FUZZYDBD Method: Automatic Fuzzy Database Definition, PeerJ Computer Science. Available from: https://doi.org/10.7717/peerj-cs.427.

[12]. S, Rajdarsan & S, Shreyas & Nikhil, U & Chinnappa Naidu, Rani & P, Bharath & Muthu, Rajesh. (2023). Hybrid Methods For Classification Of Breast Cancer Using Machine Learning Techniques.1-5.Available from: https://doi.org/10.1109/ViTECoN58111.2023.10157808.

[13]. S.Raj, S.Singh, A.Kumar, S.Sarkar & C.Pradhan. (2021).Feature Selection And Random Forest Classification For Breast Cancer Disease, Data Analytics In Bioinformatics; Available from: https://doi.org/10.1002/9781119785620.ch8.

[14]. Saoud, H., Ghadi, A., Ghailani, M., & Abdelhakim, B. A. (2019). Using feature selection techniques to improve the accuracy of breast cancer classification. In Innovations in Smart Cities Applications Edition 2: The Proceedings of the Third International Conference on Smart City Applications (pp. 307-315). Springer International Publishing.

[15]. Sharma, D., Kumar, R., & Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. Measurement: Sensors, 24, 100560.

[16]. Taghizadeh, E., Heydarheydari, S., Saberi, A., JafarpoorNesheli, S., & Rezaeijo, S. M. (2022). Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. BMC bioinformatics, 23(1), 410.

[17]. WangH, JiangW, DengX, GengJ. (2021). A New Method For Fault Detection Of Aeroengine Based On Isolationforest. Measurement. 185:110064. Available from:https://doi.org/10.1016/j.measurement.2021.110064.

[18]. Zhou S, Hu C, Wei S, YanX.(2024) Breast CancerPrediction Based on multiple MachineLearningAlgorithms. Technology in Cancer Research & Treatment.Available from:https://doi.org/10.1177/15330338241234471.