DEVELOPMENT OF NOVEL DATA MINING ALGORITHM FOR THE PREDICTION OF RECURRENCE AND SURVIVABILITY OF BREAST CANCER PATIENTS.

NURUDEEN, A. A.[1], UMAR U.[2], ASARE B. K.[3], AND ABDULKARIM B.[4]

1 DEPARTMENT OF MATHEMATICS AND STATISTICS COLLEGE OF SCIENCE AND TECHNOLOGY, KADUNA POLYTECHNIC

2,3 DEPARTMENT OF STATISTICS USMANU DANFODIYO UNIVERSITY SOKOTO, SOKOTO.

4 DEPARTMENT OF COMPUTER SCIENCE USMANU DANFODIYO UNIVERSITY SOKOTO, SOKOTO.

ABSTRACT

Globally, breast cancer is currently the most common cancer, accounting for one-eighth of all new annual cancer cases, and it is one of the leading causes of cancer-related death in women, second only to lung cancer. The prediction of the recurrence and the survivability of breast cancer patients is important as it will assist patients in knowing about the recurrence and survivability pattern, and thereby encourage them to visit doctors promptly, so more lives can be saved. This study developed an ensemble learning model, ANN-SVM, that can predict breast cancer patients' recurrence and survivability. A total of 2,469 patients with breast cancer dataset were obtained from Barau-Dikko Teaching Hospital (BDTH), Kaduna, Cancer Registry Department. The results showed that the conventional Machine learning (ML) models- Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), and the proposed model- ANN-SVM could predict the recurrence of breast cancer respectively with 82.29%, 94.84%, 90.49%, and 95.65% accuracy, also they could predict survivability of breast cancer patients respectively with 63.29%, 90.46%, 81.93%, and 91.47% accuracy in the tested dataset. The ANN-SVM model outperformed the conventional ML models regarding recurrence and survival prediction of breast cancer patients. In this study, family history and chemotherapy, respectively, turned out to be the most important features for recurrence and survivability of breast cancer patients. The outstanding performance of the proposed model in terms of precision, recall and F1 score highlights the model's effectiveness in accurately predicting both “yes” and “no” for recurrence prediction and both “alive” and “dead” for survivability prediction. Both conventional ML models and the proposed ensemble learning model predict the recurrence of breast cancer and the survivability of breast cancer patients with high accuracy.

Keywords: Machine learning, Ensemble learning, Accuracy, Recurrence, Survivability.

1.0 INTRODUCTION

The most common types of cancer among women are breast, colorectal, lung, and cervical (Marjan *et al.,* 2016). Globally, the cancer burden is estimated to have risen to 18.1 million new cases and 9.6 million deaths in the year 2018 (Musa and Aliyu, 2020). One in 5 men and one in 6 women worldwide develop cancer during their lifetime, and one in 8 men and one in 11 women die from the disease. However, worldwide, the total number of people who are alive within 5 years of a cancer diagnosis is estimated to be 43.8 million (Musa and Aliyu, 2020). Cancer occurs due to mutations and abnormal changes in the genes responsible for regulating the growth of cells and keeping them healthy (Global Cancer Observatory, 2020). Breast cancer is a malignant disease that originates in the breast cells. Patients with a family history of breast or ovarian cancer have the possibility of developing breast cancer (Musa and Aliyu, 2020). Some of the risk factors for breast cancer are gender (more in females), heredity, genetic mutation, smoking, alcohol consumption, obesity (As in a sedentary lifestyle), canned foods, chemical carcinogens used as preservatives and in cosmetics (Musa and Aliyu, 2020). The genes are in each cell's nucleus, which acts as the "control room" of each cell. Normally, the cells in our body replace themselves through an orderly process of cell growth: healthy new cells take over as old

ISSN NUMBER: 1116-249X

ones die out. But over time, mutations can "turn on" certain genes and "turn off "others in a cell. The changed cell gains the ability to keep dividing without control or order, producing more cells just like it and forming a tumour. These cancers are abnormal cells that divide uncontrollably and can invade other tissues. A breast tumour is an abnormal growth of tissues in the breast, and it may be felt as a lump or discharge or a change of skin texture around the nipple. Breast cancer remains the world's leading type of cancer. (Adebamawo and Ajayi, 2000).

There are over 200 types of cancer but breast cancer is one of the most dangerous reproductive cancers that affects mostly women. The first breast cancer case was recorded in Egypt in 3000 BC (Jabbar, 2021). Breast cancer is currently the most common cancer, accounting for 12.5% of all new annual cancer cases, and it is one of the leading causes of cancer-related death in women, second only to lung cancer (Breast Cancer Organisation, 2023).

1.2 Statement of the Problem

The integration of ML models into predicting the recurrence and survival of patients with breast cancer has emerged as a promising approach to enhance the performance of metrics. Numerous studies have been conducted to predict the recurrence and survival of patients with breast cancer; the majority of these studies were carried out using statistical methods such as parametric, semi-parametric models, or machine learning techniques but a few of them employed ensemble learning. Some authors like (Gupta 2022; Izci *et al*., 2023; Pechprasarn *et al.,* 2023; Moyasebi *et al.,* 2020; Mostafa Atlam *et al.,* 2021; Ahmad *et al.,* 2013; Kalafi *et al.,* 2019; Khaoula et al., 2023; Ganggayah *et al.,* 2019; Gupta 2022; Li *et al.,* 2021; Maria *et al.,* 2019; Hugo *et al.,* 2019; Moyasebi *et al.,* 2020; Dawngliani *et al*, 2019; 2020; 2021, Shikha and Jitendra 2015; Haque *et al.,* 2022 and Carlos *et al.,* 2023) had captured the performance of some ML techniques like Artificial Neural Network, Decision tree, K-Nearest Neighborhood, Bayesian Networks, Support Vector Machine, ensemble learning and so on in predicting recurrence or survival of breast cancer patients. However, it is evident that the ensemble learning algorithm outperforms a single ML model and also enhances the performance of multiple weak classifiers to a strong classifier (Anwar *et al.,* 2014; Shahzad and Levesson 2013; Prusa *et al.,* 2015; Dawngliani *et al*, 2019; 2020; 2021).

It was observed that none of the authors reviewed have combined the prediction of recurrence and survivability of breast cancer patients using ML models and an ensemble learning model in their studies. Hence, developing an ensemble model that can predict the recurrence and survivability of breast cancer patients is important as it will assist patients in knowing about the recurrence and survivability pattern and thereby encourage them to visit doctors promptly, so more lives can be saved. In this study, we developed an ensemble learning model, evaluated and compared the performance metrics of conventional ML models with our developed ensemble (hybrid) model in predicting the recurrence and survival of patients with breast cancer.

2 RELATED WORKS

According to Gupta (2022), in a study titled Predicting the time of breast cancer tumour recurrence using machine learning. The author pointed out that predicting the time of recurrence of breast cancer tumours using ML is very important as it can assist patients to consult doctors timely. The study analysed data from 198 patients and reported the performance of SVM, DT, and RF in terms of accuracy, with SVM having the highest accuracy of 78.7%.

Mazo *et al.* (2022) conducted a systematic review on the application of artificial Intelligence techniques to predict the risk of recurrence of breast cancer. The study identified the challenges associated with the prediction of the recurrence of breast cancer that the use of artificial

ISSN NUMBER: 1116-249X

Intelligence can address but due to the non-publicly available and insufficiently large datasets, the use of AI still remains difficult. However, the authors conclude that despite the potential of AI, predicting breast cancer recurrence accurately remains an unresolved challenge, necessitating further research and development in this critical area.

Moyasebi *et al.* (2020) reported a study on modelling and comparing data mining algorithms for the prediction of the recurrence of breast cancer. The objective of the study was to compare the performance of some data mining algorithms for predicting breast cancer recurrence. Data was collected from the period of June 2018 to June 2019, including 5,471 independent records of breast cancer patients for a minimum of 5 years' follow-up. The important features for prediction included LN involvement rate, Her2 value, tumour size, and tumour margin status. The authors emphasise the importance of selecting appropriate data mining tools for predicting disease recurrence, which can help physicians in making better-informed decisions. It was reported that KPCA-SVM (ensemble learning) recorded the highest accuracy of 78.5% in predicting the recurrence of breast cancer.

According to Kalafi *et al.* (2019), survival prediction of breast cancer can have a great effect on the selection of the best treatment approaches. The study employed ML and deep learning methods to predict breast cancer survival using 4,902 patient records. The authors reported that artificial neural network (ANN) could predict the survival of breast cancer with an accuracy of 88.2%, and was highest among the ML, such as RF, DT, and SVM and also tumour size turned out to be the most important feature for breast cancer survivability prediction.

Dawngliani *et al.* (2020) published an article on the breast cancer recurrence prediction model using a voting technique. The study employed a data mining classification technique called voting. The authors utilised different combinations of four base classifiers: DT, MLP, Naïve Bayes, and SMO and the performance of the classifiers was evaluated and compared. The research demonstrates that the voting classifiers achieve a high and consistent performance accuracy.

Illiyan *et al.* (2019) reported a study titled The application of machine learning models for survival prognosis in breast cancer studies. The study discusses the application of machine learning models to predict survival time in breast cancer based on clinical datasets. Various machine learning methods were compared with linear support vector regression, lasso regression, Kernel ridge regression, K-nearest neighbourhood regression, and decision tree regression, showing the most accurate results for survival prognosis.

3.0 METHODOLOGY

In the preparation of this manuscript, the researchers had carefully undertaken the following steps:

Data collection, Data preprocessing, Feature selection, Data splitting (Training and Testing), data mining models, evaluation of models and the method of data analysis employed in this study.

3.1 Method of Data Collection

The data used for this study were extracted from the records of the hospital's cancer registry department. The breast cancer data include variables like identification number, Age, Marital status, Menopausal status, Family history, Classification of breast cancer, Laterality, breast cancer stage classification, Estrogen receptor status, Progesterone receptor status, c-er-b2 status, Primary treatment type, Surgery type, Status, Tumour size (cm), Total axillary nodes removed, Number of positive lymph nodes and date of diagnosis, (date of clinical diagnosis).

3.2 Data preprocessing: The datasets in today's real world are highly susceptible to noise, missing values, and inconsistency due to their typically huge size. As a result of this, the dataset

ISSN NUMBER: 1116-249X

used for this study underwent thorough preprocessing to improve its quality and consequently improve the data mining results.

3.3 Data Cleaning and Balancing: This involves routine work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies. This step is very important because dirty data can cause confusion in the mining procedure and, hence, result in unreliable output. In this study, all patients with missing values were removed.

3.4 Feature Selection: It involves reducing the number of attributes to improve the accuracy of the outcome. Here, the irrelevant and redundant features were removed. Random forest is an excellent classifier to determine the importance of variables in a classification problem. This study employed a random forest classifier to select relevant features. Features such as family history, age at diagnosis, method of diagnosis, time and laterality were selected for the recurrence case, also twenty-four features were selected for the survivability case. The importance of feature selection is to improve the performance of the classification techniques (Figures 2 and 3).

3.5 Data Splitting: The preprocessed dataset was divided into training and testing. However, 80% of the dataset was for training while the remaining 20% was for testing.

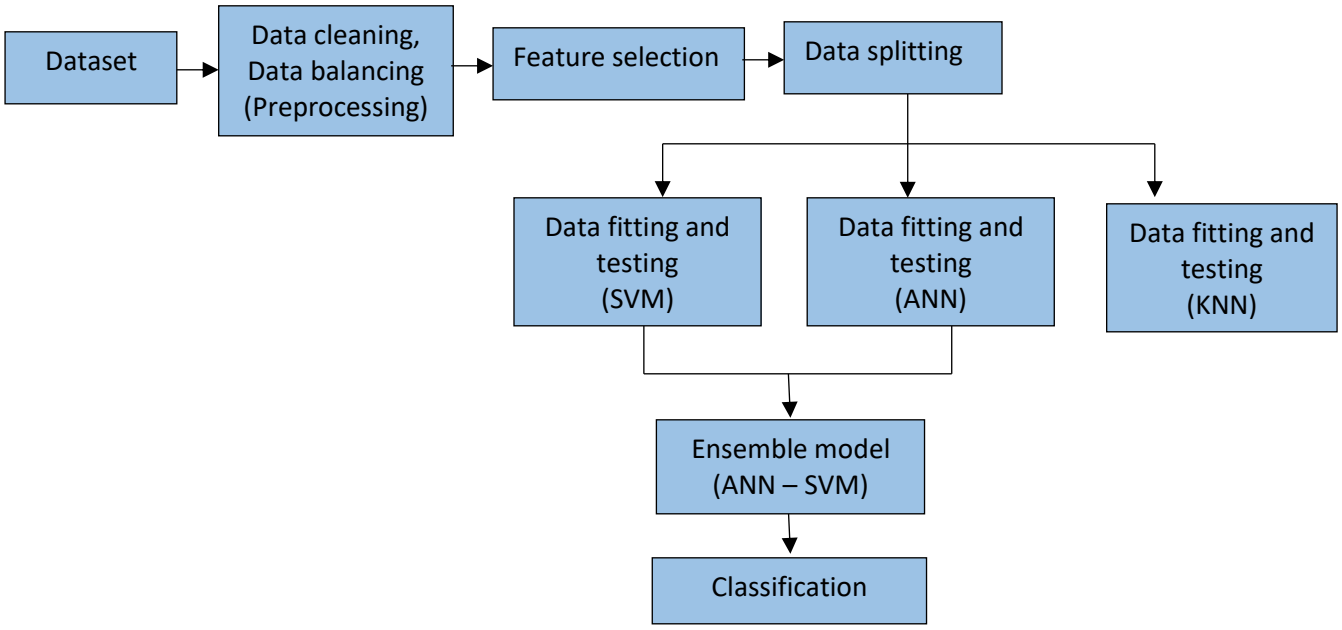

Figure 1: System Block Diagram

3.6 Data Mining Models Used in this Study

In this study, Artificial Neural Network (ANN), K-nearest Neighbors (KNN), Support Vector Machine (SVM) models and the proposed ANN- SVM model were employed as conventional machine learning models and an ensemble learning model (hybrid model) respectively to predict the recurrence and survivability of women with breast cancer. However, the selection of these data mining models met two criteria. Those that have shown the best performance in the related studies and the most frequently used in clinical datasets for classification problems. Let us provide a brief mathematical representation of each model:

ISSN NUMBER: 1116-249X

3.6.1 Artificial Neural Network (ANN) Model:

Artificial Neural Networks are computational models inspired by the structure and function of biological neural networks. They consist of interconnected nodes (neurons) organised in layers (input layer, hidden layers and output layer). The output of a neuron is typically calculated using an activation function, such as the sigmoid function (often used in binary classification problems). The forward propagation in an ANN can be represented mathematically as follows:

$$z^{(l+1)} = W^{(l)} a^{(l)} + b^{(l)} \quad (1)$$

$$a^{(l)} = g\left(z^{(l)}\right) \quad (2)$$

Where

$a^{(l)}$ is the activation of layer l,

$z^{(l+1)}$ is the weighted sum of activations of a layer,

$b^{(l)}$ are the weights and biases of layer l,

and g is the activation function.

3.6.2 K-Nearest Neighbours (KNN) Model:

K-Nearest Neighbours is a non-parametric classification algorithm that classifies an input by a majority vote of its neighbours, with the input being assigned to the class most common among its k-nearest neighbours (where k is a hyperparameter). Mathematically, the classification of a new data point x can be represented as:

$$c(x) = majority\ \text{vote}\left(C(x_1), C(x_2), ..., C(x_k)\right) \quad (3)$$

Where Is the class label of the i nearest neighbor, x, and $C(x)$ is the predicted class label of x.

3.6.3 Support Vector Machine (SVM) Model:

Support Vector Machine is a supervised learning algorithm that separates classes by finding the hyperplane that maximises the margin between classes. Mathematically, SVM aims to solve the optimisation problem:

$$\text{Minimize: } \frac{1}{2} \| w \|^2 \quad (4)$$

subject to:

$$y_i \left(w \,.\, x_i + b\right) \geq 1 \text{ for all i} \quad (5)$$

Where w is the weight vector, b is the bias term, x_i is the training sample, and y_i is its corresponding class label.

3.7 Ensemble Learning Model

The Ensemble Learning (EL) method creates multiple instances of conventional ML methods and combines them to evolve a single optimal solution to a problem. EL is based on the idea that a group of models can make better decisions than a single one, by leveraging the diversity and complementarity of their predictions. EL methods are also known as “committee of machines” or “committee of experts” with the latter following the assumption that each base learner is an “expert” and its output is an “expert opinion. This approach is capable of producing a better predictive model compared to the conventional approach. The top reasons to employ the EML method include situations of uncertainties in data representation, solution objectives, modelling techniques, or random initial seeds in a model. The instances of candidate methods are called

ISSN NUMBER: 1116-249X

base learners. Each base learner works independently as a conventional ML method, and the eventual results are combined to produce a single robust output. The combination could be done using any of the averaging (simple or weighted) methods and voting (majority or weighted) for regression and classification methods, respectively.

3.7.1 Proposed ANN-SVM (Hybrid Model)

In this hybrid approach, the ANN is used to learn high-level features from the input data, which are then fed into the SVM algorithm for classification. The ANN is trained to extract features from the input data, and the output of one of its layers can be used as the feature vector. The SVM algorithm then separates the classes based on these feature vectors using a hyperplane. Mathematically, the hybrid model can be represented as follows:

$$\left.\begin{aligned} ANN: h = f_{ANN}(x) \\ SVM: \hat{y} = SVM(h) \end{aligned}\right\} \quad (6)$$

Where x represents the input data, h represents the feature vector extracted by the ANN, f_{ANN} represents the function learned by the ANN, and $\hat{y}$ represents the predicted class label. These representations illustrate how the outputs of the ANN are used as input features for the SVM algorithm, resulting in a hybrid model that combines the strengths of both approaches. These mathematical representations provide a basic understanding of how the ensemble learning technique (ANN-SVM) works in predicting the recurrence and survivability of women with breast cancer. For a hybrid model combining Artificial Neural Network (ANN) with Support Vector Machine (SVM), the mathematical representation depends on the specific architecture and methodology used. The ensemble learning algorithm offers several advantages, such as improved accuracy and performance, especially for complex and noisy problems over a single model. In this study, we have done voting with the ANN-SVM model. Voting is the simplest ensemble algorithm and it is evident that it is very effective for classification problems (Dawngliani *et al*, 2019).

3.8 Evaluation of Models

The performance of the conventional ML models and the proposed model employed using the testing dataset in predicting the recurrence and survivability of breast cancer was evaluated using standard classification metrics such as precision, recall, F1 score and accuracy.

Accuracy: The ratio of correctly predicted instances to the total cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Precision: The ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{TP}{FP + TP} \quad (8)$$

This is a useful metric in situations where it is necessary to minimise the number of false positives.

Recall: The ratio of correctly predicted observations to all observations in the actual class.

$$\text{Recall} = \frac{TP}{FN + TP} \quad (9)$$

It is a useful metric in situations where it is important to minimise the number of false negatives.

F1Score: The harmonic mean of precision and recall.

ISSN NUMBER: 1116-249X

$$\text{F1Score} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (10)$$

The acronyms are as follows:
TP: True Positive,
TN: True Negative,
FP: False Positive, and
FN: False Negative
Confusion Matrix: A table used to describe the performance of a classification algorithm, displaying the true positives, true negatives, false positives and false negatives.

3.9 Method of Data Analysis

The study employed data mining techniques to analyse clinical data on women with breast cancer. All analyses were performed using Python 3.7. The following are the primary libraries employed in this study.
Jupyter Notebook: Interactive coding and documentation.
Pandas: For data preprocessing.
Sklearn: For implementing and evaluating models.
Numpy: For numerical calculations.
Matplotlib: For data visualisation and presentation of results graphically.

4.0 RESULTS AND DISCUSSION

4.1 Evaluation of Model Performance

The performance of the three ML models- ANN, KNN, SVM and ensemble model- ANN-SVM (proposed model) was evaluated using the testing dataset. The models were evaluated based on standard classification metrics stated above, which provided a deeper understanding of their capability to predict the recurrence and survivability of breast cancer.

Table 1: METRICS OF THE CONFUSION MATRIX ON TESTING DATASET

Algorithm	Recurrence	Correct predictions	Incorrect predictions
SVM	0: Yes 1: No	229 282	22 88
ANN	0: Yes 1: No	317 272	32 0
KNN	0: Yes 1: No	291 271	33 26
Proposed ANN-SVM	0: Yes 1: No	317 277	27 0
Algorithm	**Survivability**	**Correct guesses**	**False predictions**
SVM	0: Alive 1: Dead	209 229	116 138
ANN	0: Alive 1: Dead	316 310	35 31
KNN	0: Alive 1: Dead	243 324	21 104
Proposed ANN-SVM	0: Alive 1: Dead	303 330	15 44

Source: Author's computation, February 21, 2025, Python 3.7

ISSN NUMBER: 1116-249X

Table 2: PERFORMANCE METRICS OF MODELS ON TESTING DATASET

	Algorithm	Precision (%)	Recall (%)	F1score (%)	Accuracy (%)
RECURRENCE	SVM	72.24	91.24	80.63	82.29
	ANN	100	90.83	95.19	94.84
	KNN	91.80	89.81	90.80	90.49
	Proposed ANN-SVM	**100**	**92.15**	**95.92**	**95.65**
	Algorithm	**Precision**	**Recall**	**F1score**	**Accuracy**
SURVIVABILITY	SVM	60.23	64.31	62.39	63.29
	ANN	91.07	90.03	90.54	90.46
	KNN	70.03	90.05	79.54	81.93
	Proposed ANN-SVM	**87.32**	**95.28**	**91.13**	**91.47**

Source: Author's computation, February 21, 2025, Python 3.7

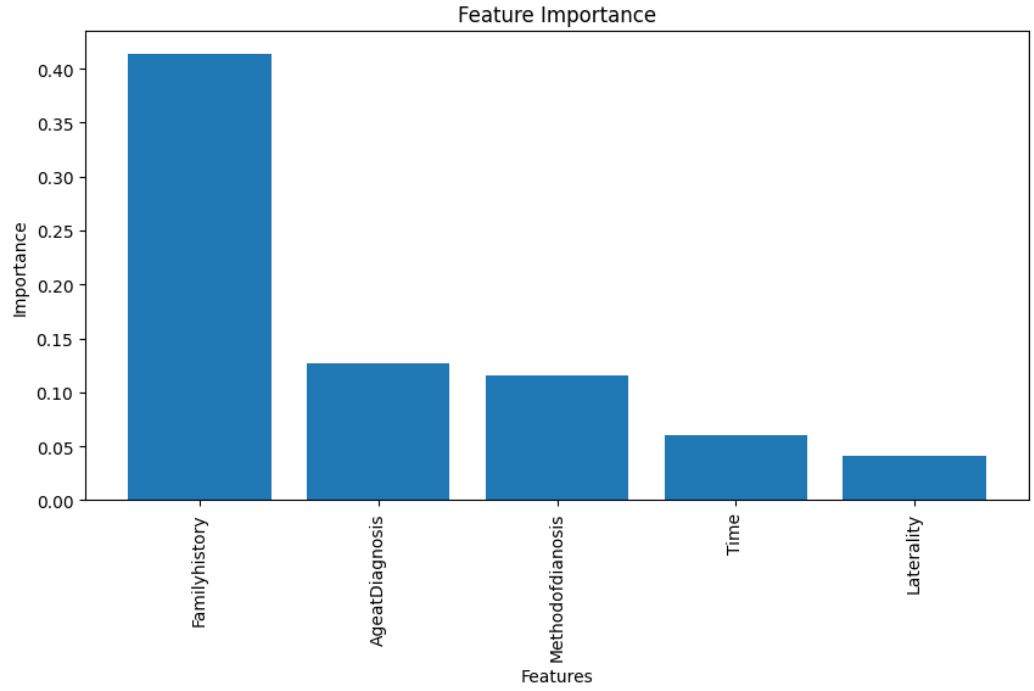

Figure 2: Feature importance in predicting recurrence of breast cancer.

ISSN NUMBER: 1116-249X

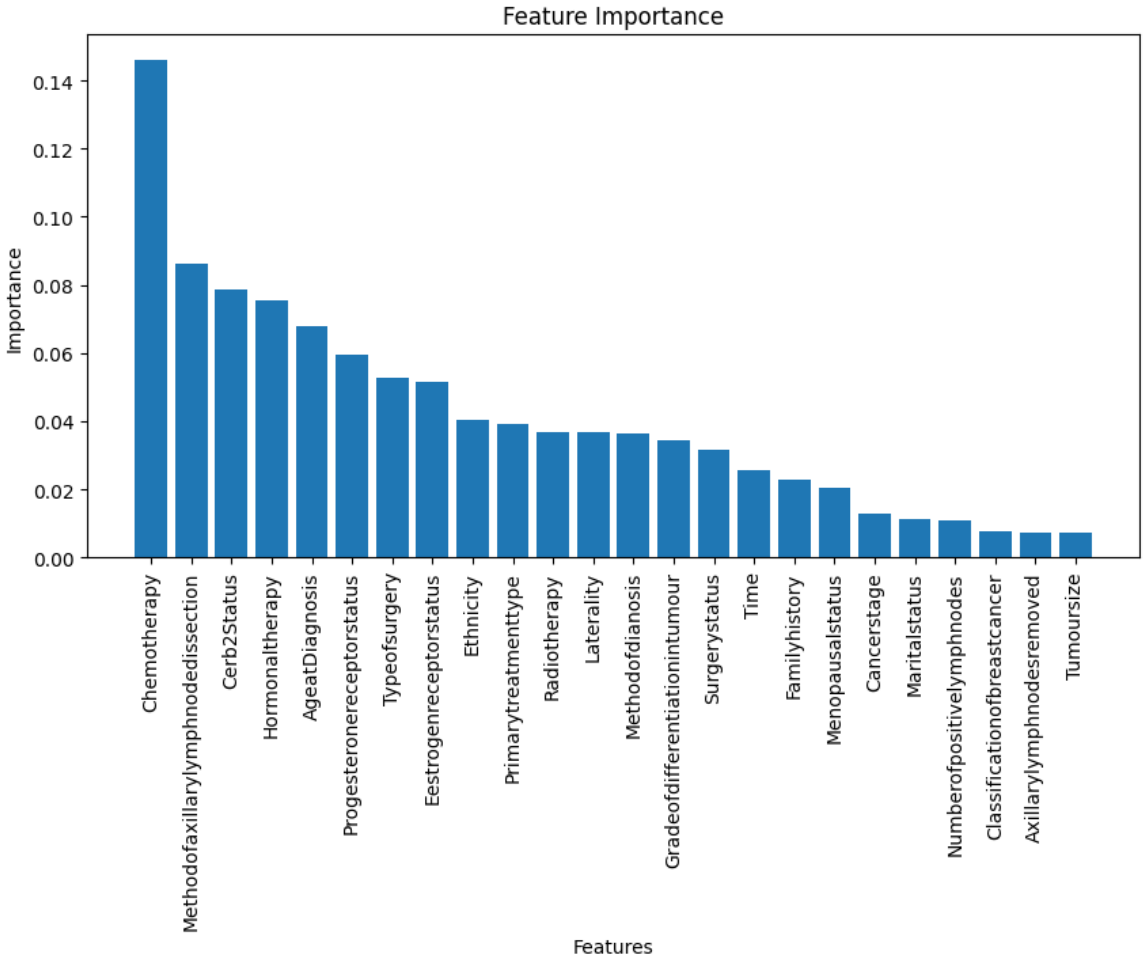

Figure 3: Feature importance in predicting the Survivability of breast cancer patients.

4.2 Accuracy of the Models

4.2.1 ANN-SVM (Proposed Model): Of all the data mining models employed, the proposed model had the highest accuracy of 95.65% and 91.47%, respectively, for recurrence and survivability of breast cancer. This implies that the proposed model correctly predicted the recurrence of breast cancer in nearly 96% of cases. Similarly, it correctly predicted slightly above 91% of cases of the survivability of women with breast cancer. The confusion matrix shown in Figure 4 reveals that there are 594 correct predictions and 27 false predictions. However, this model predicted 317 data points as 0 and 277 data points as 1, which represents its correct prediction. The model also predicted 27 data points as 0 and no data as 1, which is its wrong prediction. In the case of survivability, the confusion matrix shown in Figure 8 depicts that there are 633 correct guesses and 59 false predictions. However, this model predicted 303 data as 0 and 330 data as 1, which represents its correct prediction. The model also predicted 15 data points as 0 and 44 data points as 1, which is an absolutely wrong prediction.

4.2.2 Support Vector Machine Model: This model achieved 82.29% accuracy for recurrence prediction and 63.29% accuracy for survivability prediction of breast cancer patients. This means that the model correctly predicted 82.3% and 63.3% of cases of the recurrence and survivability of breast cancer patients, respectively. Here, for the recurrence of breast cancer. The confusion

ISSN NUMBER: 1116-249X

matrix shown in Figure 6 shows that there are 511 correct predictions and 100 erroneous predictions. The model predicted 229 data points as 0 and 282 data points as 1. So, this is its correct prediction. This same model also predicted 22 and 88 data points to be 0 and 1, respectively. So, this is an absolutely wrong prediction. In the case of survivability, the confusion matrix shown in Figure 10 reveals that there are 438 correct guesses and 254 incorrect predictions. However, this model predicted 209 data as 0 and 229 data as 1, so this represents its correct prediction. The model also predicted 116 data points as 0 and 138 data points as 1, this is an absolutely wrong prediction. Here, the number of wrong predictions by SVM is greater than the number of wrong predictions by the proposed model. For this reason, its accuracy is less than that of ANN-SVM for both recurrence and survivability cases.

4.2.3 K-Nearest Neighbourhood model: This model correctly predicted 90.49% of breast cancer recurrence. Hence, its accuracy is nearly 90.5% of the cases, which is better than that of SVM but lower than that of ANN and the proposed models. Similarly, the KNN model has achieved 81.93% accuracy, which indicates that it has correctly predicted the survivability of breast cancer in almost 82% of the cases, which is better than SVM but lower than ANN and the proposed models. In the case of recurrence, there are 562 correct predictions and 59 incorrect predictions, as shown by the confusion matrix in Figure 5. This model predicted 291 data as 0 and 271 data as 1, this is its correct prediction. However, it also predicted 33 and 26 data as 0 and 1, respectively, which is a wrong prediction. In the case of survivability, the confusion matrix shown in Figure 9 reveals that there are 567 correct guesses and 125 false predictions. However, this model predicted 243 data points as 0 and 324 data points as 1, representing its correct prediction. The model also predicted 21 data points as 0 and 104 data points as 1, which is a wrong prediction. It can be seen that the number of wrong predictions is higher than that of the proposed and ANN models but lower than the SVM model, and this is the reason why its accuracy is less than the proposed and ANN models, but greater than the SVM model.

4.2.4 Artificial Neural Network model: The model correctly predicted 94.84% of the recurrence of breast cancer. Hence, its accuracy is approximately 95%, which is less than the proposed model and better than the SVM and KNN models. Similarly, the ANN model correctly predicted 90.46% of cases of survivability of breast cancer patients, so its accuracy is close to 90.5%. In the case of breast cancer recurrence prediction. The confusion matrix shown in Figure 7 depicts that, the number of correct and false predictions in this case is 589 and 32, respectively. This model predicted 317 as data 0 and 272 as data 1. This is a correct prediction. However, the model also predicted 32 data as 0 and no data as 1. So, this is a wrong prediction. However, in the case of survivability of breast cancer patients. The confusion matrix shown in Figure 11 indicates that there are 626 correct predictions and 66 erroneous predictions. Here, the model predicted 316 data points as 0 and 310 data points as 1, and this is its correct prediction. However, the model also predicted 35 and 31 data as 0 and 1, respectively. This is a wrong prediction. Here, it can be seen that the number of wrong predictions for both recurrence and survivability of breast cancer is more than that of the proposed model but less than the SVM and KNN models, and this is the reason why its accuracy is less than that of the proposed models but greater than SVM and KNN models.

4.3 Precision and Recall of the Models

A deeper understanding of models' performance in predicting the recurrence and survivability of breast cancer patients was provided by precision and recall metrics. The proposed model achieved a precision of 100% and a recall of 92.15%, highlighting the model's effectiveness in accurately predicting both “yes” and “no” on recurrence prediction. However, the same model

ISSN NUMBER: 1116-249X

achieved a precision of 87.32% and a recall of 95.28%, expressing the model's effectiveness in accurately predicting both “alive” and “dead” on survivability prediction. However, the ANN model displayed a precision of 100% and a recall of 90.83%, the SVM model achieved a precision of 72.24% and a recall of 91.24%, and the KNN model demonstrated a precision of 91.80% and a recall of 89.81% for recurrence prediction. Similarly, the ANN model displayed a precision of 91.07% and a recall of 90.03%, the SVM model achieved a precision of 60.23% and a recall of 64.31%, and the KNN model demonstrated a precision of 70.03% and a recall of 90.05% for survivability prediction (Table 2).

4.4 F1 Score of the Model

This metric balances precision and recall. The proposed model had the highest F1 score of 95.92% and 91.13% for recurrence and survivability, respectively. However, this confirmed the superiority of the proposed model in the recurrence and survivability prediction of breast cancer patients over the conventional ML models (Table 2).

4.5 Confusion Matrix

The matrices and Table 1 display the number of correct predictions and the number of false predictions obtained by the models employed. The proposed model had the lowest number of incorrect predictions when compared with conventional ML models and this validates its accuracy.

4.5.1 CONFUSION MATRIX OF DATA MINING MODELS FOR RECURRENCE OF BREAST CANCER.

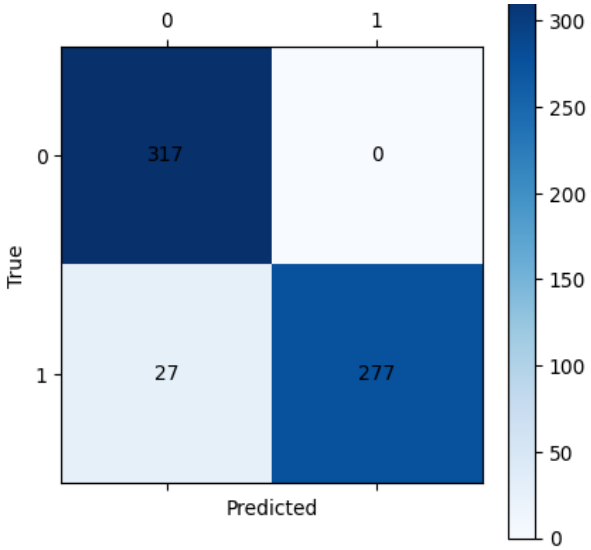

Figure 4: Confusion matrix of ANN-SVM model.

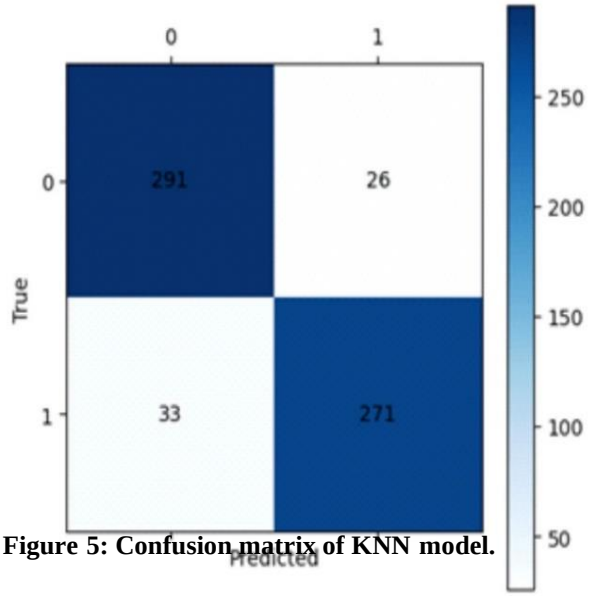

Figure 5: Confusion matrix of KNN model.

ISSN NUMBER: 1116-249X

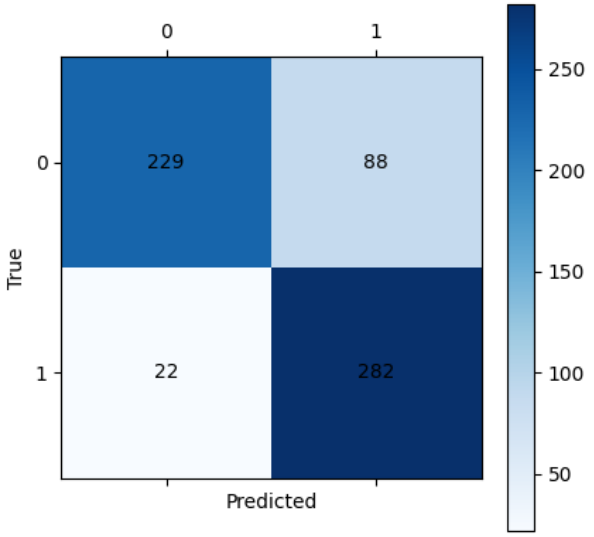

Figure 6: Confusion matrix of SVM model.

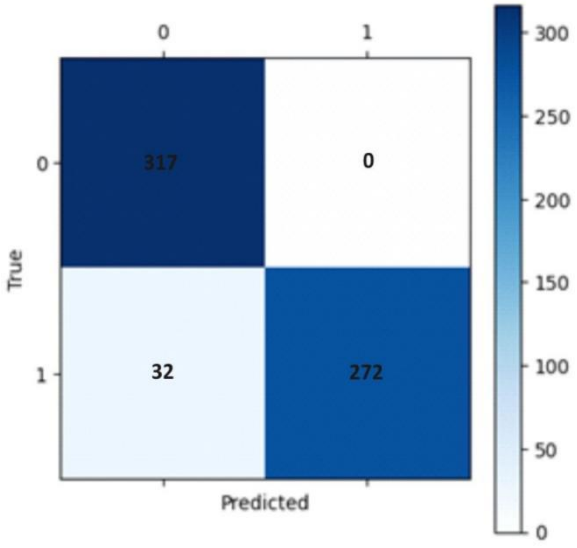

Figure 7: Confusion matrix of ANN model.

4.5.2 CONFUSION MATRIX OF DATA MINING MODELS FOR SURVIVABILITY OF BREAST CANCER PATIENTS.

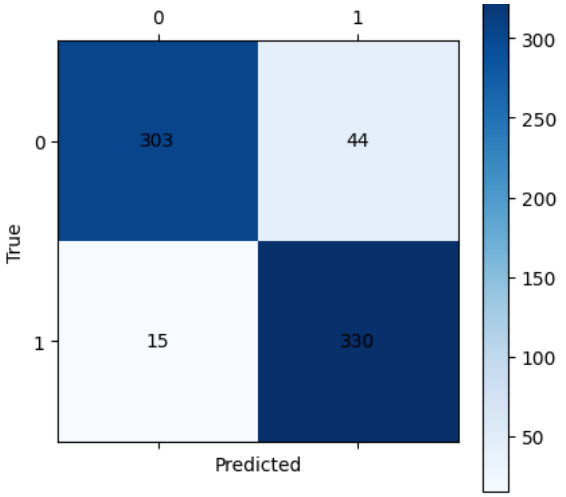

Figure 8: Confusion matrix of ANN-SVM model.

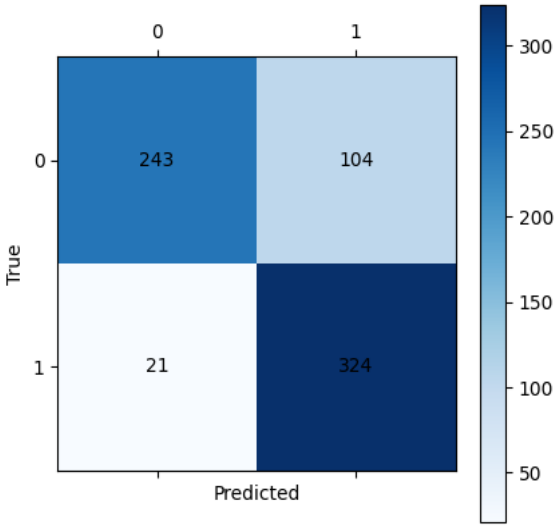

Figure 9: Confusion matrix of KNN model.

ISSN NUMBER: 1116-249X

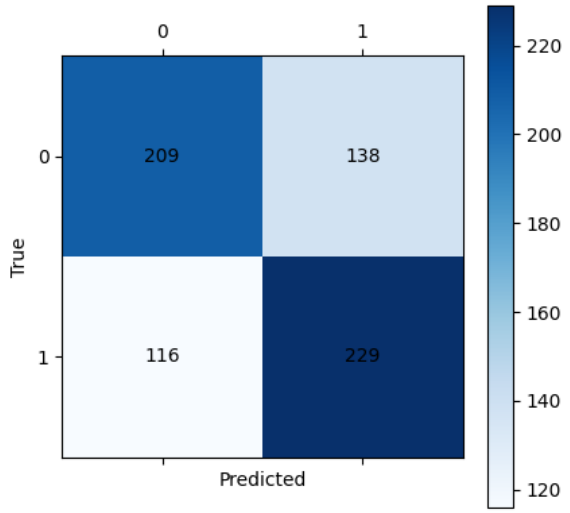

Figure 10: Confusion matrix of SVM model.

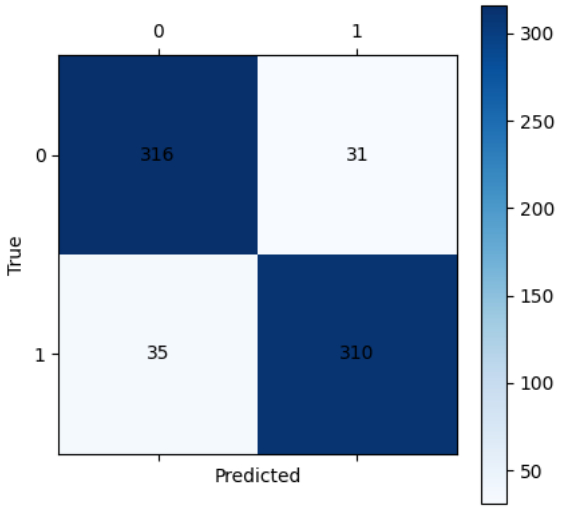

Figure 11: Confusion matrix of ANN model.

Table 3: Model Comparison (Recurrence)

Current paper (Model name)	Accuracy (%)	Referenced paper	(Model name)	Accuracy (%)
SVM	82.29	Gupta, (2022)	SVM	78.70
ANN	94.84	Ahmad *et al.*, (2013)	ANN	94.70
KNN	90.49	Wang *et al.*, (2020)	KNN	88.88
ANN-SVM (PROPOSED)	95.65	Dawngliani *et al.*, (2021)	Ensemble	81.75

Table 4: Model Comparison (Survivability)

Current paper (Model name)	Accuracy (%)	Referenced paper	(Model name)	Accuracy (%)
SVM	63.29	Haque *et al.*, (2022)	SVM	85.00
ANN	90.46	Maabreh *et al.*, (2021)	ANN	91.60
KNN	81.93	Pilaftis and Rubio	KNN	83.90
ANN-SVM (PROPOSED)	91.47	Jabbar (2021)	Ensemble	97.42

4.6 Model Comparison.

The models in Tables 3 and 4 are a comparison between the current study and existing studies on the prediction of the recurrence of breast cancer or the prediction of the survivability of breast cancer patients. However, it can be seen that all four models we employed have a good accuracy level. The proposed model demonstrated outstanding performance over the conventional ML models and is consistent with findings in other domains where the ensemble models have been shown to improve predictive accuracy. The findings in this current study align with existing studies such as Anwar *et al.* (2014), Shahzad and Levesson (2013); Prusa *et al.*, (2015); Dawngliani *et al*, (2019; 2020; 2021) on the prediction of the recurrence of breast cancer or predicting the survivability of breast cancer patients.

5.0 CONCLUSION

This study has used only four data mining models, comprising three conventional ML models: ANN, SVM, KNN and a proposed model (ANN-SVM) to predict the recurrence of breast cancer and survivability of breast cancer patients. The results obtained revealed that these models could

ISSN NUMBER: 1116-249X

predict the recurrence of breast cancer and the survivability of breast cancer patients with varying degrees of accuracy. This study demonstrated that the performance of the ensemble model (ANN-SVM) was the best among the models in terms of accuracy, precision, recall and F1 score in predicting recurrence and survivability of breast cancer patients, followed by the ANN model. It was observed that an ensemble model can enhance the performance of weak models like SVM and KNN. However, this prediction can encourage patients to consult doctors promptly, thereby saving their lives. The key contribution of this study is that a precise literature review of the related works was carried out. Development of the hybrid ML (ensemble learning model) approach, employing feature selection, voting with the proposed model and classification techniques for predicting the recurrence and survivability of breast cancer patients. Lastly, comparing the performance metrics of the conventional ML models with the ensemble (hybrid) model indicates the novelty and significance of the study. Further extensive evaluation of other machine learning models can be carried out using some combinations, such as decision tree and random forest, to predict the recurrence and survivability of breast cancer patients. Finally, it would be of interest to test the proposed model in a different real-world dataset.

REFERENCES

Adebamawo C.A., Ajayi O.O. (2000). Breast Cancer in Nigeria. *West Africa Journal of Medicine*.
Jul-Sep;19(3):179-91. PMID: 11126081.

Ahmed L.G., Eshiaghy A.T., Poorebrahim A., Ebrahimi M. and Razavi A.R. (2013). Using three Machine learning techniques for predicting breast cancer recurrence, *Journal of Health and Medical Informatics*, 4: 124. Doi:10.4172/2157 7420.100124.

Anwar Hina, Qamar Usman, Muzaffar Qureshi, Abdul Wahab (2014). "Global Optimization Ensemble Model for Classification Methods", *The Scientific World Journal*, vol. 2014, Article ID 313164, 9 pages https://doi.org/10.1155/2014/313164.

Breast Cancer.org (2023). Retrieved online on May 15, 2023.

Carlos D. C. L., Deevyankar A., Isabel D T., and Lourdes M. R. (2023). Automatic detection of Breast Cancer by using Ensemble Learning, DOI: https://doi.org/10.21203/rs.3.rs-2934498/v1.

Dawngliani M.S, Chandrasekaran, N. & Samuel Lalmuanawma (2019). A Comparative Study between Data Mining Classification and Ensemble Techniques for Predicting Survivability of Breast Cancer Patients. *International Journal of Computer Science and Mobile Computing,* Vol.. 8 Issue.9 September 2019, pg. 01-10.

Dawngliani, M.S, Chandrasekaran, N., Lalmawipuii, R. & Thangkhanhau, H.. (2021). Breast Cancer Recurrence Prediction Model Using Voting Technique. 10.1007/978 3-030-49795-8_2.

Dawngliani, M.S., Chandrasekaran, N, Lalmuanawma, Samuel & Thangkhanhau, H.. (2020). Prediction of Breast Cancer Recurrence Using Ensemble Machine Learning Classifiers. 10.1007/978-3-030-46828-6_20.

Ganggayah M.D., Taib N. A., Har Y. C., Lio P. and Dhillon S.K. (2019). Predicting factors for Survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision making*, **http://doi.org/10.1186/s12911-019-0801-4 19:48**.

Global Cancer Observatory (2020). Information on Cancer. Retrieved online on August 15, 2020.

Gupta S.R. (2022). Prediction time of breast cancer tumour recurrence using Machine Learning

ISSN NUMBER: 1116-249X

Learning. *Cancer Treatment and Research Communications*. 32. 100602. 10.1016/j.ctarc.2022.100602.

Haque MN, Tazin T, Khan MM, Faisal S, Ibraheem SM, Algethami H, Almalki FA. Predicting Characteristics Associated with Breast Cancer Survival Using Multiple Machine Learning Approaches. (2022). *Comput. Math. Methods Med.* doi: 10.1155/2022/1249692. PMID: 35509861; PMCID: PMC9060999.

Hugo Aerts J.W.L., Yiwen Xu, Ahmed Hosny, Roman Zeleznik, Chintan Parmar, Thibaud Coroller, Idalid Franco, Raymond H. Mak, (2019). Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *American Association for Cancer, Research Clin Cancer Res* 25:3266 75 doi: 10.1158/1078-0432.CCR-18-2495.

Iliyan Mihaylov, Maria Nisheva, and Dimitar Vassilev. (2019). "Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies" *Information* 10, no.3: 93. https://doi.org/10.3390/info10030093.

Izci, Hava & Macq, Gilles & Tambuyzer, Tim & De Schutter, Harlinde & Wildiers, Hans & Duhoux, Francois & Azambuja, Evandro & Taylor, Donatienne & Staelens, Gracienne & Orye, Guy & Hlavata, Zuzana & Hellemans, Helga & Rop, Carine & Neven, Patrick & Verdoodt, Freija. (2023). Machine Learning Algorithm to Estimate Distant Breast Cancer Recurrence at the Population Level with Administrative Data. *Clinical Epidemiology*. Volume 15. 559-568. 10.2147/CLEP.S400071.

Jabbar, Meerja, Akhil. (2021). Breast cancer data classification using ensemble machine learning, *Engineering and Applied Science Research* 2021;48(1):65-72.

Jason Brownlee (2017). *Master Machine Learning Algorithms 'Discover how they work and implement them from scratch'* edition, V1.12, Jason Brownlee publications, USA.

Jean Sunny, Nikita Rane, Rucha Kanade and Sulochana Devi (2020). Breast Cancer Classification and Prediction using Machine Learning, *International Journal of Engineering Research and Technology (IJERT)* http://www.ijert.org Vol. 9 Issue 02 ISSN: 2278-0181.

Kalafi EY, Nor NAM, Taib NA, Ganggayah MD, Town C, Dhillon SK. (2019). Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. Folia Biol (Praha). 2019;65(5-6):212-220. PMID: 32362304.

Khaoula Chtouki, Maryem Rhanouni, Mounia Mikram, Siham Yousfi and Kamelia Amazian (2023). Supervised Machine Learning for Breast Cancer Risk Factors Analysis and Survival Prediction. *Journal of arXiv:2304.07299v1* [cs.LG] 13 Apr 2023.

Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, Peng X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. PLoS One. 2021 Apr 16;16(4):e0250370. doi: 10.1371/*journal.pone.0250370. PMID: 33861809; PMCID:* PMC8051758.

Maria Nisheva, Iliyan Mihaylov, and Dimitar Vassilev (2019). Application of machine learning models for survival prognosis in breast cancer studies, *www.mdpi.com/journal/information*10, 93; doi:10.3390/info10030093.

Marjan M., Mohammad R. M., Hamid R. M., Miguel A. M., Fariborz M. (2016). A hybrid computer-aided diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning, *Journal of Computational and Structural Biotechnology* http://dx.doi.org/10.1016/jcsb2016.11.004.

Mazo C, Aura C, Rahman A, Gallagher WM, Mooney C. (2022). Application of Artificial

ISSN NUMBER: 1116-249X

Intelligence Techniques to Predict Risk of Recurrence of Breast Cancer: A Systematic Review. *J Pers Med.* 2022 Sep 13;12(9):1496. doi: 10.3390/jpm12091496. PMID: 36143281; PMCID: PMC9500690.

Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH. (2020). Modeling and comparing data mining algorithms for the prediction of recurrence of breast cancer. *PLoS One. Oct 15;15(10): e0237658. doi: 10.1371/journal.pone.*0237658. PMID: 33057328; PMCID: PMC7561198.

Mostafa Atlam, Hanaa Torkey, Nawal El-Fishwy and Hanaa Salem (2021). Coronavirus disease (COVID-19): Survival Analysis using deep learning and Cox regression model. Journal of pattern analysis and applications (2021) 24-993-1005. https://doi.org/10.1007/3/0044.021-00958-0.

Musa A. A., Aliyu U.M., (2020). Application of Machine Learning Techniques in Predicting Breast Cancer Metastases Using Decision Tree Algorithm, in Sokoto, Northwestern Nigeria, *Journal of Data Mining Genomics Proteomics.* 11:220. doi: 10.35248/2153-0602.20.11.220.

Pechprasarn, S. & Wattanapermpool, O. & Warunlawan, M. & Homsud, P. & Akarajarasroj, T. (2023). Identification of Important Factors in the Diagnosis of Breast Cancer Cells Using Machine Learning Models and Principal Component Analysis. Journal of Current Science and Technology. 13. 642-656. 10.59796/jcst.V13N3.2023.700.

Prusa, J., Khoshgoftaar, T. M., and Dittman, D. J. (2015). Using ensemble learners to improve classifier performance on tweet sentiment data. In *2015 IEEE International Conference on Information Reuse and Integration* (pp. 252-257). IEEE.

Rivera, D. L., Narvaez, R. A. (2023). Application of Machine Learning Approaches in Cancer Prediction. *Canadian Journal of Nursing Informatics,* 18(2). https://cjni.net/jo urn al/ ?p=11581.

Shahzad, R. K., and Lavesson, N. (2013). Comparative analysis of voting schemes for ensemble-based malware detection. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *4*(1), 98-117.

Shikha Agrawal and Jitendra Agrawal (2015). Neural Network Techniques for Cancer Prediction: A Survey, *Procedia Computer Science*, Volume 60, Pages 769-774, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.08.234.

WHO (2021). Bulletin on Breast Cancer. Retrieved on July 12, 2021.

ISSN NUMBER: 1116-249X