Analysis of Breast Cancer Survivability Using Machine Learning Predictive Technique for Post-Surgical Patients

Tahsien Al-Quraishi
VIT Melbourne
Melbourne, Australia
tahsien.a@vit.edu.au

Lamyaa Al-Omairi
VIT Melbourne
Melbourne, Australia
lamyaa.omairi@vit.edu.au

Rahul Thakkar
VIT Melbourne
Melbourne, Australia
rahul.thakkar @vit.edu.au

Chetanpal Singh
VIT Melbourne
Melbourne, Australia
chetanpal.singh@vit.edu.au

Johnson Agbinya
MIT Melbourne
Melbourne, Australia
jagbinya@mit.edu.au

Osama A. Mahdi
MIT Melbourne
Melbourne, Australia
omahdi@mit.edu.au

Bhagwan Das
MIT Melbourne
Melbourne, Australia
bdas@mit.edu.au

***Abstract*— The primary objective of this study is to predict thelikelihood of long-term survival for breast cancer patients who have received surgical treatment for a duration of five years or more. The aim is to provide healthcare providers with accurate predictions that can guide treatment plans and medication decisions. Despite existing breast cancer survivability prediction techniques, their accuracy remains low, limiting their practical utility. Additionally, there is a lack of research specifically addressing the survivability prediction of breast cancer post- surgery. Therefore, this study proposes a deep learning-based approach to predict survivability in this context. The effectiveness of the proposed model in predicting survival rates is evaluated using Haberman's survival dataset, obtained from the University of Chicago's Billings. Different evaluation measures, including accuracy, sensitivity, and specificity, are employed to assess the model's performance. Experimental results demonstrate that the proposed approach outperforms other models, achieving an accuracy of 83.18%, sensitivity of 85.54%, and specificity of 97.19%. The high accuracy of the proposed approach makes it suitable for use by healthcare professionals in predicting breast cancer survivability outcomes. It enables physicians to adjust treatments based on individual patient predictions. Consequently, the suggested method is advisable for practical implementation in systems designed to predict the survival chances of breast cancer patients after undergoing treatment.**

Keywords— ***Breast Cancer; Breast cancer survivability prediction; Deep Neural Network Classifier***

I. INTRODUCTION

Presently, breast cancer stands as the leading cause of mortality among women in Australia[1] . It is also the second leading cause of death overall [2]. In 2021, the mortality rate for breast cancer in Australia was more than 3,000 instances of breast cancer have been reported, affecting 36 males and 3,102 females. Annually, approximately 20,000 Australians are diagnosed with breast cancer, and around 1 in 7 women receive a breast cancer diagnosis.at some point in their lives [3]. Several factors, including age, family history, and genetic risk, have been identified as increasing the likelihood of developing breast cancer. Treatment options for breast cancer include local treatments such as surgery and radiation therapy, as well as systematic treatments such as chemotherapy. A range of treatment approaches, such as hormone therapy, chemotherapy or a combination of both, have been utilized to minimize the spread of breast cancer and reduce distant metastases by a 33.33% reduction rate [4]. Nonetheless, accurately forecasting the survival grade of patients after undergoing specific treatment regimens continues to pose a notable difficulty in the field of breast cancer treatment. Predicting the chances of survival for breast cancer patients can significantly assist healthcare providers in making well- informed decisions regarding alternative treatment options andenhancing the patients' quality of life. Furthermore, it enables the development of personalized treatments to enhance treatment effectiveness [5]. Consequently, there has been significant research focused on predicting breast cancer survivability [6-12].

While numerous techniques for predicting breast cancer survivability have been proposed in academic literature, their accuracy has been deemed relatively low, rendering them unreliable for practical use. Reference [6] focused on predicting the survivability status of breast cancer patients in Taiwan to aid treatment decisions, but our study differs by specifically targeting survivability status following surgicaltreatment within a five-year timeframe. In [7], the objective was to enhance survivability classification for breast cancer patients with highly imbalanced data over five years, but the results were unsatisfactory. In a separate study [8], the prediction of five-year survival, both with and without imputation, was examined, yet their predictions proved less accurate compared to our own work. Another study [9] employed a combination of techniques to improve survival rate prediction for breast cancer patients in Thailand by generating high-quality datasets; however, the achieved accuracy in that study was disappointingly low.

Reference [10] examined an approach based on artificial neural networks (ANN) and C4.5 with logistic regression to predict survivability status in breast cancer patients. However, the accuracy of the method was still constrained.

Similarly, in [11], a model was proposed that incorporated various algorithms such as C4.5 decision tree, back-propagated neural network, and Naive Bayes for predicting survivability in breast cancer patients, but the achieved accuracy remained low. Additionally, [12] this study investigated the application of support vector machines (SVM) for estimating the post- surgery treatment's impact on survivability status, but the accuracy was relatively low. Overall, there is a lack of research specifically focused on breast cancer survivability prediction following surgery treatment, and the analysis emphasizes the immediate requirement for a highly precise model that can predict the breast cancer patients' survivability.The objective of this paper is to develop a predictive model capable of determining the survival status of patients who have undergone surgical treatment for breast cancer. The approach proposed in this study is based on a deep learning neural network (DNN) classifier. According to Zahng et al. [13] to demonstrated that the DNNs are a promising approach for cancer survival prediction, and they can be used to improve the early detection and diagnosis of cancer. The evaluation of the approach was conducted using Haberman's survival dataset. The contributions of this study can be outlined as follows:

- The implementation of a deep learning-based method for forecasting the survivability of breast cancer patients after treatment.
- A comparison between the proposed DNN-based model, a random forest (RF)-based model, and the model presented in reference [12]. The results showed that the proposed DNN model achieved higher accuracy compared to the existing models.

The rest of the paper is structured as follows: initially, weexamine the existing research on predicting survival rates in breast cancer datasets. Next, we outline the approach employed in this study, describing the procedures and algorithm utilized to predict the survival outcomes of breast cancer patients following surgical treatment. Subsequently, wepresent the framework employed for predicting survival in breast cancer patients and provide an analysis of the performance of the survival evaluation method. Lastly, the conclusion section summarizes the outcomes obtained in this study.

To summarize, this paper introduces a new method utilizing deep learning to predict the survival outcomes of breast cancer patients after undergoing surgery. The proposed model outperforms existing models, as demonstrated through the evaluation and comparison with a random forest-based model and the work presented in reference [11]. The paper provides insights into the methodology, algorithm, framework,and performance analysis, ultimately leading to the conclusionand findings of the study.

II. RELATED WORKS

Over the recent years, a multitude of predictive models has emerged in the healthcare field, specifically aimed at assisting physicians in forecasting the survival rates of cancer patients. In this section of the paper, we present an overview of machine learning techniques employed in the prediction of breast cancer survival rates. An exemplary study conducted by Chao et al. [6] introduced a predictive model to classify the survivability of breast cancer. The research encompassed 1,340 breast cancer patients diagnosed in Taiwan, and theobjective was to offer a point of reference for making treatment decisions based on their survivability. The researchers utilized classification methods, such as Support Vector Machines (SVM), C5.0 decision tree, and logistic regression, to identify and predict the survival rates of breast cancer patients. The models were evaluated using a 10-fold cross-validation method, and their performances were assessed. based on accuracy. The results indicated that among the three classification techniques, SVM exhibited the highest accuracy rate of 95.22% in predicting breast cancer survival. This finding suggests that SVM is effective in classifying breast cancer survivability and can be a valuable tool for physicians in making treatment decisions. It is important to consider that the study focused specifically on breast cancer patients inTaiwan, and the findings may not be directly generalizable to other populations. Additionally, the evaluation metric usedwas accurate, which provides a measure of overall correctness but may not capture the full picture of model performance. Nonetheless, the work of Chao et al. [5] contributes to the body of research exploring machine learning techniques for breast cancer survivability prediction and highlights the potential of SVM as a reliable classifier in this domain.

In their study, Wang et al. [7] proposed a combined algorithm that integrates Synthetic Minority Oversampling Technique (SMOTE) and Particle Swarm Optimization (PSO) to enhance the accuracy of 5-year survivability classification for breast cancer patients, particularly when dealing with imbalanced data. They employed logistic regression, 1-Nearest Neighbor search, and decision tree (C5) methods for theiranalysis. The study utilized the SEER breast cancer dataset, which included 973,125 patients and 118 variables spanning from 1973 to 2007. The dataset was divided into training and test sets, and a 10-fold cross-validation was conducted to assess the performance of the classifier. The combined classifier of SMOTE + PSO + C5 achieved the highest accuracy compared to other combined algorithms.

In another study by Garcia-Laencina et al. [8], various frameworks were evaluated for predicting breast cancer survivability in the presence of incomplete or missing data. The research included 399 instances of breast cancer, with 16 variables obtained from the Institute Portuguese of Oncology of Porto (IPO). Logistic Regression, Classification Trees, SVM, and KNN were employed to develop predictive models.The evaluation employed nested 10-fold cross-validation, revealing that the KNN technique achieved the highest accuracy rate of 81.73% compared to the other methods.

In their study, Thongkam et al. [9] employed a hybrid approach to improve the prediction of survival rates for breast cancer patients in Thailand by generating high-quality datasets. They utilized C-Support Vector Classification (CSVC) to eliminate outlier cases and performed over- sampling with replacement to augment the number of cases in the minority class. Evaluation metrics such as accuracy, sensitivity, specificity, ROC, and F-measure were utilized to assess the performance, with SVM demonstrating the highest accuracy performance. Delen et

al. [10] introduced an ANN technique called MLP plus back-propagation and combined logistic regression with C4.5 decision tree for predicting survivability in breast cancer cases. They utilized the SEER dataset from 1973 to 2000, comprising 433,272 patients and 72 variables. Through a 10-fold cross-validation, the decision tree (C4.5) achieved the highest accuracy rate of 93.6%, while logistic regression had the lowest accuracy rate of 89.2%.

Bellaachia and Guven [11] utilized C4.5 decision tree algorithms, back-propagated neural network, and Naive Bayes algorithm to determine the most predictive survivability model Abbreviations and Acronyms for breast cancer patients. They considered a dataset of 482,052 cases from the SEER dataset spanning from 1973 to 2002, including Survival Time Recode (STR), Vital Status Records (VSR), and Cause of Death (COD) fields. In their experiment, the C4.5 decision tree algorithms achieved the highest accuracy rate of 68.7%. Aljawad et al. [12] applied Support Vector Machines (SVM) and Bayesian network (BN) models to predict the survivability status of breast cancer patients who underwent surgery treatment. They compared the performance of SVM and BN using Haberman's survival dataset. The SMOTE technique was employed to address the issue of imbalanced data. The Weka software package was used for these applications, and a greedy approach was applied to optimize the classifier parameters and improve accuracy.

The performance evaluation included metrics such as accuracy, recall, and precision. SVM exhibited a higher accuracy rate of 74.44%, while Bayesian Network presented a lower accuracy rate of 67.56%. It is important to note that none of the papers [6-11] discussed the prediction of survivability specifically for breast cancer patients after surgical treatment. Our study contributes to decision-making in the post-surgery treatment phase and demonstrates better performance with higher accuracy compared to the study conducted by Aljawad et al. [12]. therapy and targeted therapies are often used to block the growth of hormone receptor-positive breast cancer cells. Chemotherapy utilizes medication to eradicate cancer cells, while radiotherapy is employed selectively to decrease the likelihood of tumor metastasis or recurrence. Surgery is a frequently utilized treatment choice that encompasses lumpectomy) or mastectomy as viable options. Our study aims to predict the survivability outcomes of breast cancer patients who have undergone surgical treatment for a period of five years or longer. By developing a prediction model, we hope to provide valuable insights into the likelihood of survival for these patients. This information can aid healthcare professionals in making informed decisions about the patient's treatment plan and overall care. Figure 1 depicts a patient with breast cancer who received surgical treatment. In our analysis, we utilized the publicly accessible Haberman's survival dataset for our research [14]. The dataset employed in our study contains information pertaining to breast cancer patients who underwent surgery at the University of Chicago's Billings Hospital. The Haberman's survival dataset exhibits imbalanced data, with samples categorized into two classes: majority and minority. In Table II, we present an analysis of

TABLE I. COMPARATIVE ANALYSIS OF RELATED WORKS

Author(s)	Dataset	Best Method	Prediction Rate
Chao et al. (2014)	Specific Hospital inCentral Taiwan	SVM	90%
Wang, et al. (2014)	SEER	SMOTE +PSO +C5	94.1%
Garcia-Laencinaet al.(2015)	Institute Portuguese ofOncology of Porto	KNN	81.3%
Thongkam et al. (2009)	Srinagarind Hospital in Thailand	SVC+SVM	93.34%
Delen et al. (2005)	SEER	C5	93.6%
Bellaachia andGuven (2006)	SEER	C5	68.7%
Aljawad, et al.(2017)	Haberman ’s Survival Dataset	SVM	74.44%
Our Proposed Method	Haberman ’s Survival Dataset	DNN	83%

this dataset, which comprises 306 samples. Each sample is characterized by four numerical variables, including the patient's age at the time of the operation (ranging from 30 to 83 years) and the year of the operation (ranging from 0 to 52, using the year-1900 notation). Among the 306 samples, 225 belong to class 1, representing patients who survived for five years or longer after the surgery. Additionally, there are 81 samples in class 2, indicating patients who passed away within five years after the surgery. It is important to highlight that class 1 consists of a considerably larger number of samples in comparison to class 2, indicating the imbalanced distribution of the dataset. The presence of imbalanced data introduces an intriguing and demanding classification problem for researchers, often requiring the utilization of data mining tools specifically designed to address data imbalance. However, in our study, we did not employ any of these data balancing techniques and worked with the original imbalanced dataset [15]. Figure 2 illustrates the correlation analysis conducted among the variables in the Haberman's survival dataset, focusing on their association with the class variable.

III. BREAST CANCR SURVIVABILITY FRAMEWORK

Once an individual receives a breast cancer diagnosis, the treatment approach depends on the stage of the disease. For stages 1-3, the main goal of treatment is to eliminate cancer cells and reduce the risk of cancer reoccurrence. On the other hand, for stage 4 breast cancer, the focus is on alleviating pain,improving the patient's quality of life, and prolonging their life. There are various therapies available for breast cancer treatment, including hormone therapy, targeted therapies, chemotherapy, radiotherapy, and surgery. Hormone therapy and targeted therapies are often used to

block the growth of hormone receptor-positive breast cancer cells. Chemotherapy utilizes medication to eradicate cancer cells, while radiotherapy is employed selectively to decrease the likelihoodof tumor metastasis or recurrence. Surgery is a frequently utilized treatment choice that encompasses lumpectomy) or mastectomy as viable options. Our study aims to predict the survivability outcomes of breast cancer patients who have undergone surgical treatment for a period of five years or longer. By developing a prediction model, we hope to provide valuable insights into the likelihood of survival for these patients. This information can aid healthcare professionals in making informed decisions about the patient's treatment plan and overall care. Figure 1 depicts a patient with breast cancer who received surgical treatment. In our analysis, we utilized the publicly accessible Haberman's survival dataset for our research [14]. The dataset employed in our study contains information pertaining to breast cancer patients who underwent surgery at the University of Chicago's Billings Hospital. The Haberman's survival dataset exhibits imbalanced data, with samples categorized into two classes: majority and minority. In Table II, we present an analysis of this dataset, which comprises 306 samples. Each sample is characterized by four numerical variables, including the patient's age at the time of the operation (ranging from 30 to 83 years) and the year of the operation (ranging from 0 to 52, using the year-1900 notation). Among the 306 samples, 225 belong to class 1, representing patients who survived for five years or longer after the surgery. Additionally, there are 81 samples in class 2, indicating patients who passed away within five years after the surgery. It is important to highlight that class 1 consists of a considerably larger number of samples in comparison to class 2, indicating the imbalanced distribution of the dataset. The presence of imbalanced data introduces an intriguing and demanding classification problem for researchers, often requiring the utilization of data mining tools specifically designed to address data imbalance. However, in our study,we did not employ any of these data balancing techniques and worked with the original imbalanced dataset [15]. Figure 2 illustrates the correlation analysis conducted among the variables in the Haberman's survival dataset, focusing on their association with the class variable.

The dataset is categorized into two classes: majority and minority. The majority class contains a significantly larger number of samples compared to the minority class, indicating the presence of imbalanced data. This aspect presents an intriguing and demanding classification scenario for researchers, as the typical data mining techniques used for balancing data are not applied to tackle the problem of imbalanced data [15]. When employing these tools, they often achieve higher accuracy in the majority class but exhibit lower accuracy in the minority class, as stated in reference [16]. Encouragingly, advancements in cancer diagnosis and treatment have resulted in a decline in the mortality rate and an extension of survival duration. Consequently, a greater proportion of patients are now experiencing survival compared to those who do not. In the latter part of our suggested framework, the classification task consists of several stages, commencing with the evaluation of the model.

To validate the DNN and RF classifiers separately, we divided them into a training set (80%) and a testing set (20%). The training set was used to approximate the function

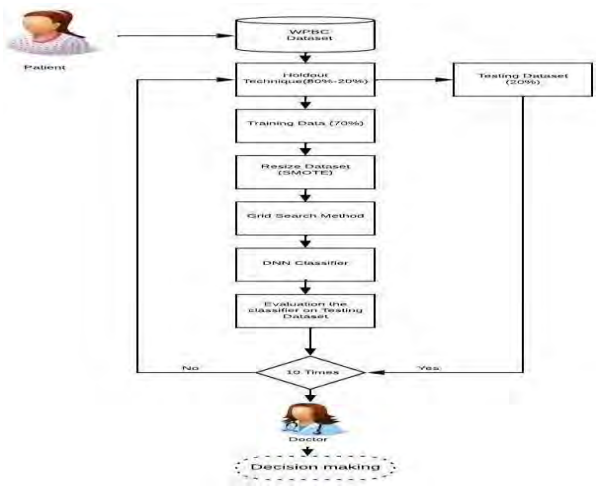

Fig. 1. Breast Cancer Survivability Framework

estimator, which was then used to predict survivability outcomes for the testing data. The model was evaluated by assessing the accumulated errors using the test set. This approach was chosen because it requires less computational time. The evaluation criteria depended on how the data was split between the training and test sets, and researchers have focused on addressing this aspect. Addressing the challenge of imbalanced data classification can be accomplished through two steps: resampling the data at the data level and employing a sophisticated classification method at the algorithm level. One resampling technique is researchers [17]. SMOTE is a technique used in data pre- processing to address the issue of over-fitting by augmenting the Synthetic

TABLE II. INVESTIGATION OF HABERMAN'S DATASET

Statistical Measure	Age of Patient	Year of Operation	Number of Positive Nodes Detected
Mean	52.458	62.853	4.0262
Median	52	63	1
Standard deviation	10.803	3.249	7.119
Maximum	83	69	52
Minimum	30	58	0

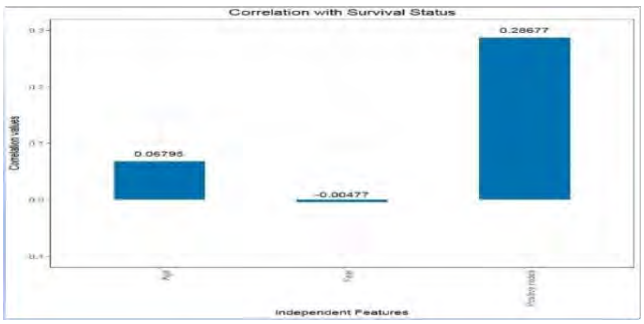

Fig. 2. The correlation analysis between variables in Haberman's survival dataset

Minority Over-sampling Technique (SMOTE), which was introduced by the number of samples in the minority class through the generation of synthetic instances within the original dataset. Unlike basic replication, SMOTE conducts over-sampling in the feature space, leading to wider decision regions for the minority class. This approach helps alleviate the constrained decision region that would arise from simple over-sampling with replacement.

To generate new synthetic samples, certain parameters such as the number of nearest neighbors (k) and the over-sampling rate (percentage) are defined. In their work, Chawlaet al. [18] outlined the steps involved in generating these synthetic samples for continuous data:

- Calculate the distance between a feature vector and one of its k nearest neighbors in the minority class.
- Multiply the distance obtained in the previous step by a randomly selected number from the range of 0 to 1.
- Add the resulting value from the previous step to the feature value of the initial feature vector. This process creates a new set of features by:

$$sn = so + \delta.\ (soi - so) \quad (1)$$

In equation (1), sn represents the new synthetic data sample, while so denotes a feature vector of a sample in the minority class.

The nearest neighbor of so in the *ith* iteration is denoted as soi, and δ represents a random number between zero and one. This sequence of three steps is repeated nine times, resulting in the creation of a new synthetic instance on each iteration. During each repetition, one of the five nearest neighbors is randomly chosen.

In the field of machine learning, algorithms often have hyperparameters that need to be optimized for achieving the best prediction performance. One common approach for this optimization is grid search, where all possible combinations ofparameter values are tried to determine the optimal values.

In our proposed framework, we employ grid search to identifythe optimal number of epochs and nodes in the hidden layersof the deep neural network.

A deep neural network is a specific type of artificial neural network that consists of multiple layers, including one or more hidden layers between the input and output neurons [19]. It utilizes forward propagation and backpropagation techniques. The classification process involves both training and testing phases. In the training phase, the deep learning classifier is used to train the variables. Neural networks are known for their adaptability and reliance on data. Deep neuralnetworks can learn more complex models compared to shallow networks, thanks to their deeper architectures. The logistic (or sigmoid) derivative functions are applied to activate the hidden layers *(h)* and propagate information from the *zk* layer.

$$x_h = logistic(y_h) = \frac{1}{1 + f^{-yh}}, y = j_h + \sum_k z_k w_{kh} \quad (2)$$

Equation (2) represents the components involved in the equation. The term *jh* refers to the bias associated with unit h, whereas k represents an index that iterates through the units in the layer below. The weight assigned to the connection from unit k to unit h in the layer is represented by *wkh*. Additionally, the index (i) represents an iteration over all the classes.

$$CF = -\sum i\ ph\ log\ ch \quad (3)$$

The cost function, denoted as *CF*, calculates the cross-entropy (p) between the target probabilities and the output (c) of the softmax function.

$$\Delta wkh(q) = \Delta wkh\ (q-1) - \epsilon\ \partial CF/\partial wkh(q) \quad (4)$$

To assess the effectiveness of the proposed method on the testing data, we employ a 10-fold cross-validation technique to classify and enhance the model's performance. This technique, also referred to as rotation estimation, was reported by Kohavi [20]. The process includes randomly splitting the dataset, referred to as D, into K equally sized folds, labeled as $D1$, Dt, Dk. The classifier is trained and tested K times, where $D1$, Dt is trained during each iteration t $\in \{1, 2, \ldots, k\}$.

Furthermore, it is evaluated on Dt. The accuracy, which indicates the total number of correct classifications, is estimated through cross-validation. Let $D(i)$ represent the test set samples $xi\,(vi, yi)$. In this case, the accuracy of cross-validation is estimated using the following formula:

$$acccv = 1/n \sum (v, y) \in D\ \delta(I(D(D(i), vi), yi) \quad (5)$$

The classification task is repeated for an average of 10 iterations. Once the survivability prediction results are obtained, it is up to the doctors to make decisions based on these outcomes.

IV. PERFORMANCE EVALUATION

In this section, we will discuss key metrics utilized to assess the model's performance and outline the steps involved in establishing the experimental procedure.

A. Evaluation measures

To evaluate the predictive performance of the model, weutilised three metrics. These metrics include accuracy, recall, and precision, each with its own definition:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$
$$Recall = TP / (TP + FN) \quad (7)$$
$$Precision = TP / (TP + FP) \quad (8)$$

The equations mentioned above involve various terms: TP represents the count of true positives, TN corresponds to the count of true negatives, FP denotes the count of false positives, and FN represents the count of false negatives.

B. Experimental setup

In this section, we present the validation experiments conducted on two fronts. Firstly, we validate the previous research conducted by Aljawad et al. [12], where they compared the performance of Support Vector Machine (SVM) and Bayesian Network (BN) classifiers using Haberman's survival dataset. They addressed the challenge of imbalanced data by applying the SMOTE technique prior to classification. Their model was evaluated using 10-fold cross-validation, and they optimized the classifier parameters using a greedy technique.

TABLE III. COMPARISON OF THE EXISTING WORK AND THE PROPOSED WORK [12]

Technique	Accuracy	Precision	Recall
BN	67.7%	67.7%	67.8%
SVM	64.4%	64.6%	67.8%
RF	76.85%	90.17%	88.11%
DNN	83.18%	83.32%	98.71%

To assess the effectiveness of their model, various performance metrics such as accuracy. recall, and precision were utilized. In our study, we utilized the Weka platform to predict the survival status of breast cancer patients who underwent surgery for a duration of five years or more.

Secondly, we conducted our own experiments using theR programming language to evaluate the performance of our proposed framework. A comparison was made between our approach and the previous study [12]. We employed two classification evaluation methods: 10-fold cross-validation and a holdout technique with an 80% training set and a 20% test set.

V. EXPERIMENTAL RESULTS DISCUSSION

In this section, we examined the outcomes of the previous study [12] conducted on the imbalanced Haberman's survival dataset. To tackle the data imbalance, we employed the Synthetic Minority Oversampling Technique (SMOTE) filter to balance the dataset. The classifiers' performance was evaluated using accuracy, precision, and recall metrics. The results for the Bayesian network (BN) technique are presented in Table III, demonstrating the highest performance rates with an accuracy of 67.7%, precision of 67.7%, and recall of 67.8%. On the other hand, the Support Vector Machines (SVM) technique exhibited lower performance rates with an accuracy of 64.4%, precision of 64.6%, and recall of 67.8%.

When comparing our proposed method to the existing work [12], our approach achieved superior accuracy. The results are visualised in In Figure 3&4 respectively. In figure 3, the evaluation is achieved by utilising 80% Training and 20% test sets, it can be observed that the DNN classifier achieved the highest accuracy rate of 83.18%, while the RF classifier obtained the lowest accuracy rate of 79.97%. Additionally, the RF model demonstrated a precision rate of 89.34%, indicating its strong capability to correctly identify positive class instances within the positive samples. Despite having a precision rate of 85.54% and consequently more False Positives, the DNN classifier demonstrated the best recall rate of 97.19%. This indicates that DNN models are more reliable in accurately identifying instances of the positive class. On the other hand, a low recall rate suggests a higher number of False Negatives. In Figure 4, by using the 10-fold cross-validation technique, the DNN classifier achieved a higher accuracy rate of 83%, surpassing the RF model's rate of 76.58%. The RF approach exhibited a precision rate of 90.17%, while the DNN model had a slightly lower precision rate of 83.32%. Furthermore, the DNN classifier displayed a higher recall rate of 98.71%, while the RF model had the lowest recall rate of 88.11%.

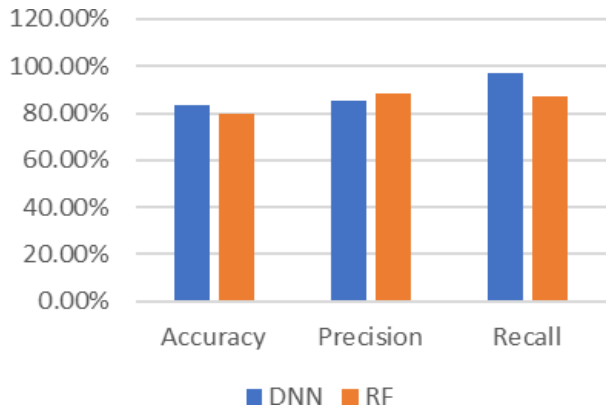

Fig. 3. Prediction summaries of DNN and RF using 80% Training/ 20% tests.

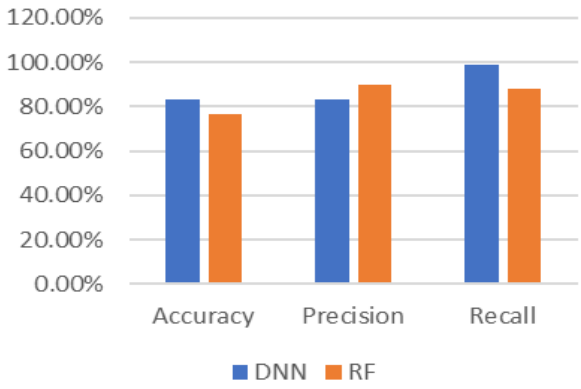

Fig. 4. Prediction summaries of DNN and RF using 10-fold cross-validation.

VI. CONCLUSION

Breast cancer patients who have undergone surgery facea significant challenge in estimating their likelihood of surviving for five years or longer. Accurate prediction of survival outcomes can greatly assist physicians in making treatment decisions. The survival statuses are categorized as either one or two. In this study, we initially replicated and evaluated a previous work [12] using the Weka platform. Following that, we validated our proposed framework for survival prediction, which involved several components. The methods employed in this study consisted of several steps. Firstly, we utilized the Haberman's survival dataset from the UCI Machine Learning Repository. Secondly, we balanced thedataset by applying the SMOTE technique. Thirdly, the classifier's parameters were optimized using the grid search method. Lastly, we incorporated the DNN classifier toimprove the accuracy of our framework. To assess the performance of the framework, we employed the holdout technique, where 80% of the data was used for training andthe remaining 20% for testing. This process was repeated 10 times to ensure the reliability of our results. The framework's performance was assessed using accuracy, precision, and recall metrics. Overall, the framework incorporating the DNN classifier outperformed the RF classifier in both techniques. It demonstrated superior performance compared to the previous work, indicating its effectiveness.

REFERENCES

[1] T. A. J. H. C. M. U. R. S. &. A. A. S. Al-Quraishi, "Breast cancer recurrence prediction using random forest model," in Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, 2018.

[2] T. A. J. H. A.-Q. N. A. A. &. A.-O. L. Al-Quraishi, "Predicting breast cancer risk using subset of genes," in 6th international conference on control, decision and information technologies (CoDIT), Paris, 2019M. King and B. Zhu, "Gaming strategies," in Path Planning to the West, vol. II, S. Tang and M. King, Eds. Xian: Jiaoda Press, 1998, pp. 158-176.

[3] A. G. C. Australia. "Breast cancer in Australia statstics." https://www.canceraustralia.gov.au/cancer-types/breast- cancer/statistics (Accessed 2022).

[4] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," Journal of the American Medical Informatics Association, vol. 20, no. 4, pp. 613-618, 2013.

[5] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," Bioinformatics, vol. 23, no. 1, pp. 30-37, 2006.

[6] C.-M. Chao, Y.-W. Yu, B.-W. Cheng, and Y.-L. Kuo, "Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree," *Journal of medical systems,* vol. 38, no. 10, p. 106, 2014.

[7] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," *Applied Soft Computing,* vol. 20, pp. 15-24, 2014.

[8] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," *Computers in biology and medicine,* vol. 59, pp. 125-133, 2015.

[9] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward breast cancer survivability prediction models through improving training space," *Expert Systems with Applications,* vol. 36, no. 10, pp. 12200-12209, 2009.

[10] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial intelligence in medicine,* vol. 34, no. 2, pp. 113-127, 2005.

[11] A. Bellaachia and E. Guven, "1 Predicting Breast Cancer Survivability Using Data Mining Techniques," 2006.

[12] D. A. Aljawad *et al.*, "Breast cancer surgery survivability prediction using bayesian network and support vector machines," in *Informatics, Health & Technology (ICIHT), International Conference on*, 2017: IEEE, pp. 1-6.

[13] Zhang, Z., Wang, W., Zhang, Y., Zhang, J., & Zhang, J. (2019). DNNs for cancer survival prediction: A review. IEEE Access, 7, 116634-116646. doi:10.1109/ACCESS.2019.2945976.

[14] A. Asuncion and D. Newman, "UCI machine learning repository," ed, 2007.

[15] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering,* vol. 21, no. 9, pp. 1263-1284, 2009.

[16] Y. Chen, "Learning classifiers from imbalanced, only positive and unlabeled data sets. Project Report for UC *https://www. cs. iastate. edu/~ yetianc/cs573/files/CS573_ProjectReport_YetianChen. pdf,* 2008.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over- sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[18] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *European Conference on Principles of Data Mining and Knowledge Discovery*, 2003: Springer, pp. 107-119.

[19] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, vol. 14, no. 2: Stanford, CA, pp. 1137-1145.