Original Article

Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data

(breast cancer / survival prediction / deep learning / machine learning)

E. Y. KALAFI[1], N. A. M. NOR[1], N. A. TAIB[2], M. D. GANGGAYAH[1], C. TOWN[3], S. K. DHILLON[1]

[1]Data Science and Bioinformatics Laboratory, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
[2]Department of Surgery, University Malaya Medical Centre, Kuala Lumpur, Malaysia
[3]Computer Laboratory, University of Cambridge, Cambridge, United Kingdom

Abstract. Breast cancer survival prediction can have an extreme effect on selection of best treatment protocols. Many approaches such as statistical or machine learning models have been employed to predict the survival prospects of patients, but newer algorithms such as deep learning can be tested with the aim of improving the models and prediction accuracy. In this study, we used machine learning and deep learning approaches to predict breast cancer survival in 4,902 patient records from the University of Malaya Medical Centre Breast Cancer Registry. The results indicated that the multilayer perceptron (MLP), random forest (RF) and decision tree (DT) classifiers could predict survivorship, respectively, with 88.2 %, 83.3 % and 82.5 % accuracy in the tested samples. Support vector machine (SVM) came out to be lower with 80.5 %. In this study, tumour size turned out to be the most important feature for breast cancer survivability prediction. Both deep learning and machine learning methods produce desirable prediction accuracy, but other factors such as parameter configurations and data transformations affect the accuracy of the predictive model.

Introduction

Breast cancer causes an important cancer-related mortality among women. Increasing prevalence and prominence of breast cancer in most of the Asian countries has been reported over the last decade (Sim et al., 2006; Hirabayashi and Zhang, 2009; Chaturvedi et al., 2015). However, breast cancer survival prediction can significantly affect selection of best treatment protocols. Population-based survival rates of Malaysian women for breast cancers indicated the overall 5-year survival rate for the cohort of 2000 to 2005 to be 49 % with median survival time of 68.1 months (Abdullah et al., 2013). In another study by Nordin et al. (2018), the median survival time for patients at stage three was 50.8 months and at stage four, 6.9 months. These studies were performed using traditional statistical methods (Ganggayah et al., 2019).

Censoring data is an important step in survival analysis; the longer is the follow-up time, the more meaningful is the information. The 5-year threshold is important to standardize reporting and to identify survivability. Labelling a patient record as 'survived' or 'not survived' takes at least five years (Kim and Shin, 2013); therefore, some previous studies used a 5-year threshold (Delen et al., 2005; Park et al., 2013) to identify the cohort's survivability. In the study by Boughorbel et al. (2016), thresholds of 2, 5, 8 and 11 years were used to conduct survival prediction in four separate analyses. Shukla et al. (2018) claimed that when the cut-off year for surviv-

Received June 4, 2019. Accepted August 28, 2019.

This project was supported by the University of Malaya Research Grant (PRGS) Program Based Grant (PRGS 2017-1) to the fourth author.

Corresponding authors: Sarinder Kaur Dhillon, Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. Phone: (+603) 79676741; Fax: (+603) 79674178; e-mail: sarinder@um.edu.my

Nur Aishah Taib, UM Cancer Research Institute and Department of Surgery, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia. Phones: (+603) 79492070, 79493642, (+601)92405856; e-mails: nuraish@gmail,com, naisha@um.edu.my, nur@ummc.edu.my.

Abbreviations: Adam – adaptive moment estimation algorithm, DT – decision tree, ER – oestrogen receptor, FP – false positive, FN – false negative, MLP – multilayer perceptron, PR – progesterone receptor, RBF – radial basis function, ReLU – rectified linear units, RF – random forest, RMSProp – root mean squared prop, SEER – surveillance, epidemiology and end results, SGD – stochastic gradient descent, SVM – support vector machine, TP – true positive, TN – true negative, UMMC – University Malaya Medical Centre, UMMCBR – University Malaya Medical Centre Breast Cancer Registry, ability period was changed to 3, 5 and 7 years, the prediction model performance considerably improved.

Many cancer survival analysis studies (Rathore et al., 2014; Lotfnezhad Ashar et al., 2015; Shukla et al., 2018) used the Surveillance, Epidemiology and End Results (SEER) cancer incidence dataset to identify patterns that were associated with the survivability of breast cancer patients. In each study, different prognostic variables were chosen, such as age, race, site, marital status, primary site, laterality, behaviour code, histology, grade, tumour size, lymph node, extension, TNM stage, radiation and surgery, but dealing with unknown or missing values was another issue that has been differently addressed by different studies. Acuña and Rodriguez (2004) used mean values instead of missing values in their data pre-processing. Some studies (Delen et al., 2005; Park et al., 2013; Boughorbel et al., 2016) removed the missing values of related subjects. In the study by Lotfnezhad Afshar et al. (2015), the authors replaced missing values with the multiple imputation method by using the average of each value in a complete dataset as a single datum. Many studies have focused on selecting appropriate learning algorithms. However, improving medical data quality is also important in building favourable prediction models. Quality problems of medical data, including missing, outlier, or skewed data, are common because they are collected without any specific research purpose. Many algorithms can manage missing data, but handling outlier and skewed data is challenging and affects the performance of the prediction model (Fielding et al., 2008).

In thousands of individual measurements that have been collected to predict cancer survival rate, a spectrum of machine learning methods helps to identify good models for predictions (Beam and Kohane, 2018). Support vector machine (SVM), random forests (RF) and decision trees (DT) are machine learning algorithms that are becoming increasingly popular with the growth of data mining in the field of information systems. SVM is able to perform pattern recognition and regression according to the theory of statistical learning and the principle of structural risk minimization (Idicula-Thomas et al., 2006).

According to many studies, it could be seen that analysing and predicting survivability of breast cancer is a challenging task. Although several studies have been conducted using machine learning techniques, still the demand for getting better results forces researchers to explore enhanced prediction techniques. Deep learning methods have made prominent contribution to cancer analysis by using multi-dimensional data for cancer prognosis prediction (Sun et al., 2018).

Since deep learning models are able to learn the task with little human instruction or prior assumptions, they rank at the top of machine learning methods. Generally, in traditional machine learning, input features must be hand-crafted from raw data according to practitioner expertise and domain knowledge to determine explicit patterns of prior interest. The machine learning "black art" (Domingos, 2012) includes building, analysing, selecting, and evaluating proper features, which can be difficult and time consuming and requires trial-and-error, and most of the time luck. On the other hand, deep learning techniques learn and select optimal features from the data itself, without any human interference, empowering automatic discovery of relationships between data that might be unknown or hidden (Shickel et al., 2017). In short, deep learning approaches can compute cohorts on the basis of all attributes of cancer data (Shukla et al., 2018). Previously, deep learning has been used in survival prediction of many cancer studies, but only for analysis of images (Li et al., 2017) and genomic data (Angermueller et al., 2016). In this study, we used a deep learning method, multilayer perception neural network, for analysing clinical data to predict breast cancer survivability at the University Malaya Medical Centre (UMMC). We also compared our deep learning prediction outcome with machine learning techniques such as SVM, RF and DT on the same dataset and parameters.

Material and Methods

The University Malaya Medical Centre Breast Cancer Registry (UMMCBCR) consists of 8,066 patients' records for the years 1993–2017. This pathologically confirmed breast cancer dataset consists of female patients in the age range of 21–95 years, who had been followed up until March 2017. In total, 37 demographic and clinical characteristic attributes were collected from UMMCBCR, but not all were used as the feature set in this study. Preprocessing was performed to normalize and reduce the dimensions of the dataset and prepare data for building the best possible predictive model. The prediction models were designed and implemented by machine learning (SVM, RF and DT) and deep learning MLP techniques. Fig. 1 illustrates the workflow of this study.

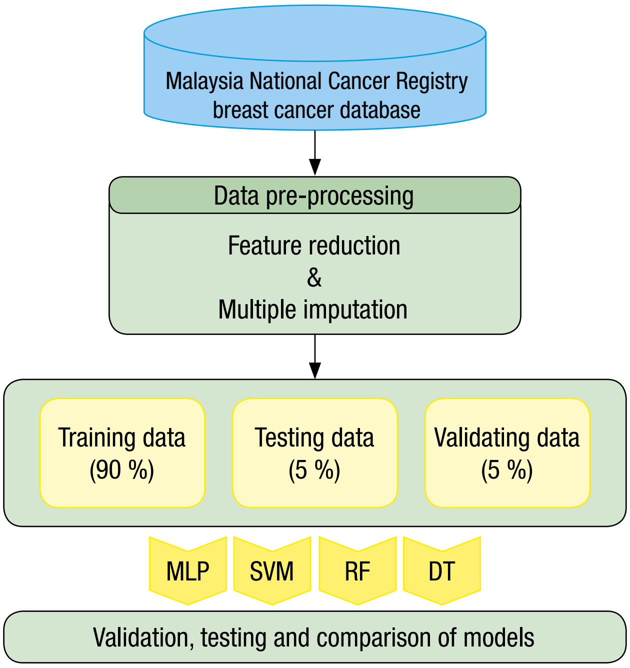

Fig. 1. Overview of methodological approach

Data pre-processing

One of biggest challenges in data mining is data preprocessing (Han and Kamber, 2000), which includes data reduction, data cleaning, data transformation, and data integration (Zhang et al., 2003). Determining and dealing with inaccuracy, inconsistency and missing values of datasets is another challenge that affects the performance of the prediction model and reduces the statistical accuracy of the study. For handling inconsistent and missing values, all patients with missing values were removed. Out of 37 variables, only 23 concerning breast cancer were selected in the analysis. In the dataset of 8,066 patients, 69.6 % were alive, whereas 30.4 % were dead; thus, we normalized the data by selecting all dead patients (2,451) and randomly selecting 2,451 cases from the alive patients' dataset. Hence, in this study, a total of 4,902 patients' records were used in the prediction model. The feature set containing demographic and clinical characteristics of patients is shown in Table 1.

Feature importance

Forests of trees is an excellent model to determine the importance of variables in classification. Feature importance can give a sense of which variables have the strongest effect in classification models. The information

Table 1. Demographic and cancer-specific information of patients

Variables	Value	Numbers	Proportion (%)
Marital status	Married	0	81.6
	Not married	1	18.4
Menopausal status	Natural menopause	0	42.8
	Pre-menopause	1	50.6
	Surgical menopause	2	6.6
Presence of family history	Yes	0	81.2
	No	1	18.8
Race	Chinese	0	68.4
	Malay	1	19.7
	Indian	2	11.9
Method of diagnosis	Excision	0	20.8
	FNAC (Fine needle aspiration cytology)	1	24.5
	Imaging only	2	0.5
	Trucut	3	54.2
Classification of breast cancer	Invasive	0	95.3
	In-situ	1	4.7
Laterality	Left	0	45.5
	Right	1	49.5
	Bilateral	2	1.3
	Unilateral	3	3.7
Cancer stage classification	Pre-cancer (Stage 0)	0	4.6
	Curable cancer (Stage 1, 2, 3)	1	84.2
	Metastatic cancer (Stage 4)	2	11.2
Grade of differentiation in tumour	Good	0	32.9
	Moderate	1	37.1
	Poor	2	30
Oestrogen receptor (ER) status	Positive	0	58.9
	Negative	1	41.1
Progesterone receptor (PR) status	Positive	0	46
	Negative	1	54
c-er-b2 status	Positive	0	24.1
	Negative	1	65.4
	Equivocal	2	10.5
Primary treatment type	Chemotherapy	0	12.6
	Hormone therapy	1	3.4
	Surgery	2	77.8
	None	3	6.2
Surgery status	Surgery done	0	85.5
	No surgery	1	14.5
Type of surgery	Breast-conserving surgery	0	24.3
	Mastectomy	1	61.1
	No surgery	2	14.6
Method of axillary lymph node dissection	Yes	0	70.6
	SLNB (Sentinel lymph node biopsy)	1	6.7
	SLNB to AC axillary	2	0.4
	None	3	22.3
Radiotherapy	Yes	0	49.4
	No	1	50.6
Chemotherapy	Yes	0	54.3
	No	1	45.7
Hormonal therapy	Yes	0	54.9
	No	1	45.1
Status	Alive	1	69.6
	Dead	0	30.4
Age	Age at diagnosis	21–96	
Axillary lymph node	Total axillary lymph nodes removed	1–14	
Positive lymph nodes	Number of positive lymph nodes	1–19	
Tumour size	Size of the tumour (cm)	0.1 to 30	

could be used to engineer new features, drop out features that look like noise, or just to continue building models. In this study, the Python package of scikit-learn (scikit-learn: machine learning in Python – scikit-learn 0.21.3 documentation," n.d.,, https://scikit-learn.org/stable) were used to determine the feature importance. Sklearn instances of the "Forests of trees" model have a .feature_importances_ attribute, which returns an array of each feature's importance in determining which split of variables will most effectively help distinguish the classes of survivability. We illustrated the most important features that affect survival in breast cancer, using all 23 features from the dataset, to perform classification using deep learning and machine learning.

Classification: deep learning

Multilayer perceptron (MLP)

The multilayer perceptron (MLP) model consists of one input layer with 23 neurons, two hidden layers with 100 and 32 neurons, and one output layer with two neurons. The output is the prediction of survivability of patients with the status of dead or alive. The relationship between input and output is identified by calculating the weights in the neural network. Rectified linear units (ReLU) were used as an activation function in the MLP model, with the Softmax function as their classification function. Regularization techniques were applied to help in improving the accuracy of MLP model and prevent model overfitting. Table 2 illustrates the summary of MLP architecture, while the structure of the MLP neural network in this study is shown in Fig. 2.

The choices of loss function and optimization algorithms can play an important role in the MLP model performance. Cross entropy loss measures the performance of the classification model where the output is either 0 or 1, and it indicates the distance between the model prediction and reality. Optimization algorithms update weights and biases in a deep learning model and minimize the error (cost) criterion at each time stage. Adaptive moment estimation algorithm (Adam) is an optimization algorithm that is computationally efficient, robust, and makes fast progress in lowering the cost. In this study, the training process of multilayer neural network was tested by different optimization algorithms such as Adam, AdaGrad, Adaδ, Root mean squared prop (RMSProp), and stochastic gradient descent (SGD).

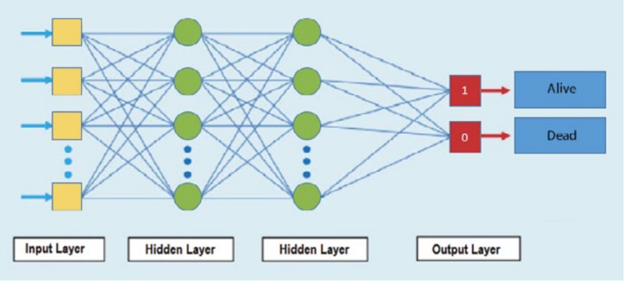

Fig. 2. Multilayer perceptron neural network structure. The model consists of one input layer with 23 neurons, two hidden layers with 100 and 32 neurons, and one output layer with two neurons.

Table 2. Architecture of MLP network

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 100)	2,400
dropout_1 (Dropout)	(None, 100)	0
activation_1 (Activation)	(None, 100)	0
batch_normalization_1	(Batch (None, 100)	400
dense_2 (Dense)	(None, 32)	3,232
dropout_2 (Dropout)	(None, 32)	0
activation_2 (Activation)	(None, 32)	0
batch_normalization_2	(Batch (None, 32)	128
dense_3 (Dense)	(None, 2)	66
Total params: 6,226 Trainable params: 5,962 Non-trainable params: 264		

Classification: machine learning

Support vector machine

Support vector machine (SVM) (Vapnik, 1995) is a supervised machine learning algorithm based on statistical learning theory using the concept of structural risk minimization. The SVM solves binary classification problems by fitting a maximum margin discriminator to the dataset in a kernel-induced feature space as shown in Fig. 3. It has been applied to many medical diagnosis and disease classifications (Blumenthal et al., 2017; Selvaraj et al., 2007). The implementation of SVM in this study was based on the libsvm library in the sklean Python package.

Decision tree

In decision tree (DT), instances (data points) are classified by sorting them based on feature values. Each node

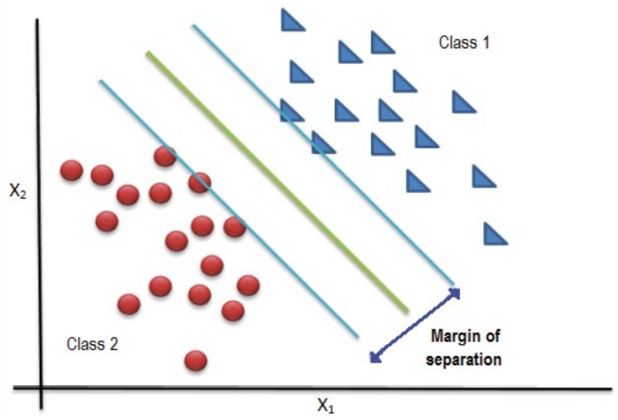

Fig. 3. SVM in linear classification. X_1 and X_2 are two parameters for binary classification and the green line is the hyperplane as a decision boundary.

in a decision tree represents a feature of an instance to be classified, and each branch represents the value that the node can assume. Classification of instances starts at the root node and the data are sifted according to their feature values. The root node of the tree would be the feature that divides the training data in the best way. In this study, the Gini index algorithm was used to identify the corresponding threshold to split the input data to sub-branches. As a result of repeating this step, we find the threshold that has maximized the homogeneity of subgroups of samples.

Random forests

Random forest (RF) is an ensemble type classification method, which tends to perform better than traditional decision tree classification methods (Ganggayah et al., 2019). Decision trees are the fundamental classifiers in RF that vote for each of the predictions, and the survivability prediction is based on the majority voting method in each tree (Breiman, 2001). The accuracy of each individual tree and independency of the trees from each other lead to robustness of classification. We used 100 trees in predicting two target classes, survival or not survival of breast cancer patients

Performance evaluation

Classification performance of models in this study was measured by sensitivity, specificity, accuracy, precision, F1 score, and Matthews correlation coefficient (Powers, 2011), which were obtained from confusion matrix entries. In a confusion matrix, the relation between classification outcomes and predicted classes are illustrated. The level of classification performance is calculated by the number of correctly and incorrectly classified samples in each class. Accuracy is computed based on the total number of correct predictions, defined as:

$$\frac{TP + TN}{\mathrm{TP + FN + TN + FP}} \quad (1)$$

Sensitivity is the proportion of true positive predictions that have been identified correctly, defined as:

$$\frac{TP}{\mathrm{TP + FN}} \quad (2)$$

Specificity is the proportion of true negative cases that have been predicted correctly, defined as:

$$\frac{TN}{\mathrm{TN + FP}} \quad (3)$$

Precision, or positive predictive value, is the ratio of correctly predicted positive observations to total predicted positive observations, defined as:

$$\frac{TP}{\mathrm{TP + FP}} \quad (4)$$

F1 score is the weighted average of precision, which is calculated as:

$$\frac{2TP}{\mathrm{2TP + FP + FN}} \quad (5)$$

Matthews correlation coefficient (MCC) is the correlation coefficient between the observed and predicted classifications, defined as:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (6)$$

where true positive (TP) and true negative (TN) stand for the number of correct predictions, and false positive (FP) and false negative (FN) stand for incorrect predictions. In this study, the true positive class indicated the group of patients who survived and the true negative class signifies the group of patients who did not survive.

Results

In this study, feature importance was calculated according to forests of trees, using ensembles of decision trees, which computed the relative importance of each variable based on the strongest relationship of variables and survival time. The score of importance of each variable used in this study is illustrated in Fig. 6, which demonstrated that the tumour size, stage, age at diagnosis, total axillary lymph node removed, and number of positive lymph nodes, respectively, are the most relevant variables to explain the survival of breast cancer. The prediction models of MLP, SVM, DT, and RF were evaluated by different measurements such as sensitivity, specificity, precision, negative predictive value, false positive rate, false discovery rate, false negative rate, accuracy, F1 score, and Matthews correlation coefficient. The results illustrated in Table 3 and boxplots in Fig. 4 show the outperformance of MLP, using 10-fold cross-validation for each model. The calibration plot is demonstrated in Fig. 5. In classifiers that were well calibrated, the output of the predicted probability could be directly interpreted as a confidence level, and the best calibrated classifier was MLP among all.

Fig. 7 shows the level of prediction and results of significance testing for survival prediction in MLP, SVM, RF, and DT. The confusion matrix illustrates that the accuracy of MLP, RF, DT, and SVM is 88.2 %, 83.3 %, 82.5 %, and 80.5 %, respectively. SVM prediction was the lowest among all classifiers, and MPL achieved the highest accuracy for prediction of survivability in breast cancer using the dataset from UMMC BCR. The best performance was exhibited by MLP, which showed superiority in all evaluation measurements reported in Table 3, Fig. 4 and Fig. 5.

Discussion

In this study, the survival prediction among UMMC Malaysian patients was assessed by using machine and deep learning techniques such as SVM, DT, RF, and

Table 3. Evaluation measurements for MLP and SVM

Measure	**MLP**	**DT**	**RF**	**SVM**	**Derivations**
Sensitivity	**0.960**	1.000	0.937	0.977	TPR = TP / (TP + FN)
Specificity	**0.830**	0.739	0.768	0.709	SPC = TN / (FP + TN)
Precision	**0.792**	0.653	0.718	0.652	PPV = TP / (TP + FP)
Negative Predictive Value	**0.968**	1.000	0.951	0.983	NPV = TN / (TN + FN)
False Positive Rate	**0.170**	0.261	0.232	0.291	FPR = FP / (FP + TN)
False Discovery Rate	**0.208**	0.347	0.282	0.349	FDR = FP / (FP + TP)
False Negative Rate	**0.040**	0.000	0.063	0.023	FNR = FN / (FN + TP)
Accuracy	**0.882**	0.825	0.833	0.805	ACC = (TP + TN) / (P + N)
F1 Score	**0.868**	0.790	0.813	0.782	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	**0.775**	0.695	0.687	0.660	TP × TN – FP × FN / sqrt((TP+FP) × (TP+FN) × (TN+FP) × (TN+FN))

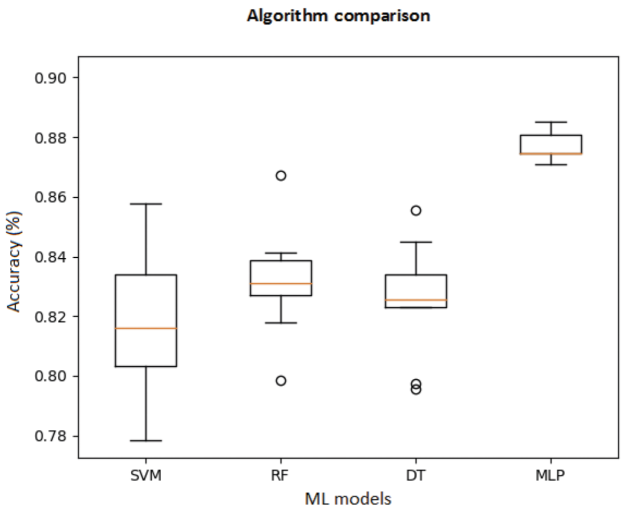

Fig. 4. Comparison of multilayer perceptron, support vector machine, random forests, and decision trees performance based on the accuracy with 10-fold cross validation

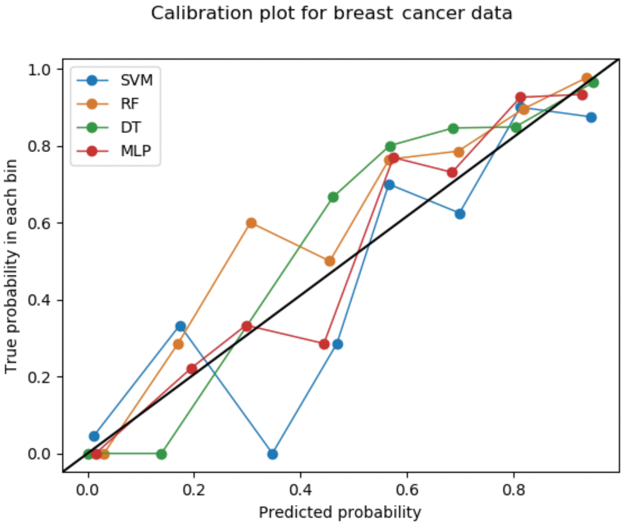

Fig. 5. Calibration curve of multilayer perceptron, support vector machine, random forests, and decision trees

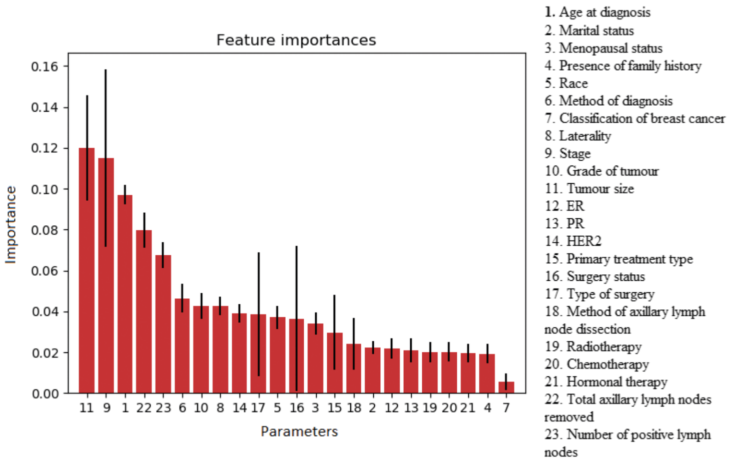

Fig. 6. The importance score of predictor variables in predicting breast cancer survival. The variables are explained in detail in Table 1.

MLP. As compared with SVM, DT, and RF, the MLP approach was superior as it yielded better performance. The highest accuracy of 88.2 % was reported by the MLP model. SVM, DT and RF performed well in breast cancer survival prediction in previous studies (Bai and Latecki, 2008; Huang et al., 2008; Ganggayah et al., 2019; Hosseini and Kesler, 2014). It is also worth noting that SVM as a supervised learning method with radial basis function (RBF) kernel has the least favourable outcome in this study. Fig. 8 demonstrates the MLP model performance on the training and validation set. The best accuracy and lowest loss was accomplished at epoch 430, and the model obtained high generalization ability. In Fig. 9, it is notable that Adam and RMSProp algorithms were the best optimizers in minimizing the loss function in 500 epochs. Although the performance of MLP in prediction of survivorship was better, it was more costly in terms of time consumption for designing the best architecture, while implementation of RF, DT and SVM was faster with less hyperparameter tuning.

While comparing the results in this study with previous studies, it is important to note that there are other factors that affect the performance of machine learning or deep learning models such as validation of the dataset, the method of handling missing values, and correlation of variables that have to be taken into consideration. Hence, these factors will be considered in future work in comparing the results of breast cancer survival prediction.

The prediction of survivorship among Malaysian breast cancer patients recorded for the years 1993–2017 was assessed. The survivability prediction was assessed according to 23 demographic and clinical variables such as breast cancer class (invasive or in-situ), family his-

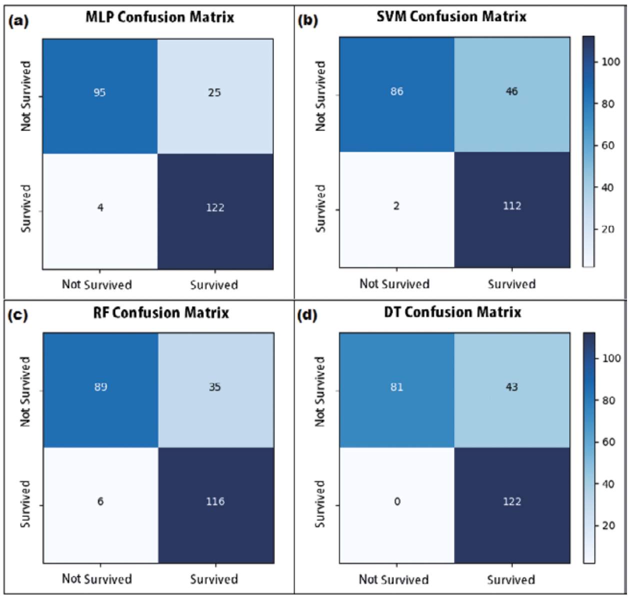

Fig. 7. Confusion matrix of (**a**) multilayer perceptron, (**b**) support vector machine, (**c**) random forests, and (**d**) decision trees classification results

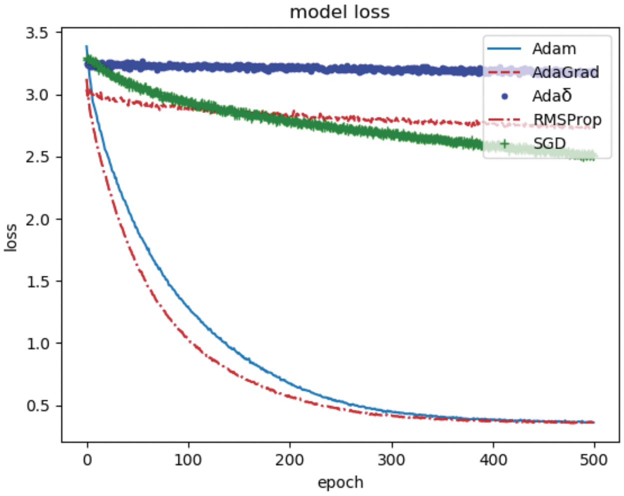

Fig. 9. Behaviour of different algorithms for optimizing the gradient descent in training MLP

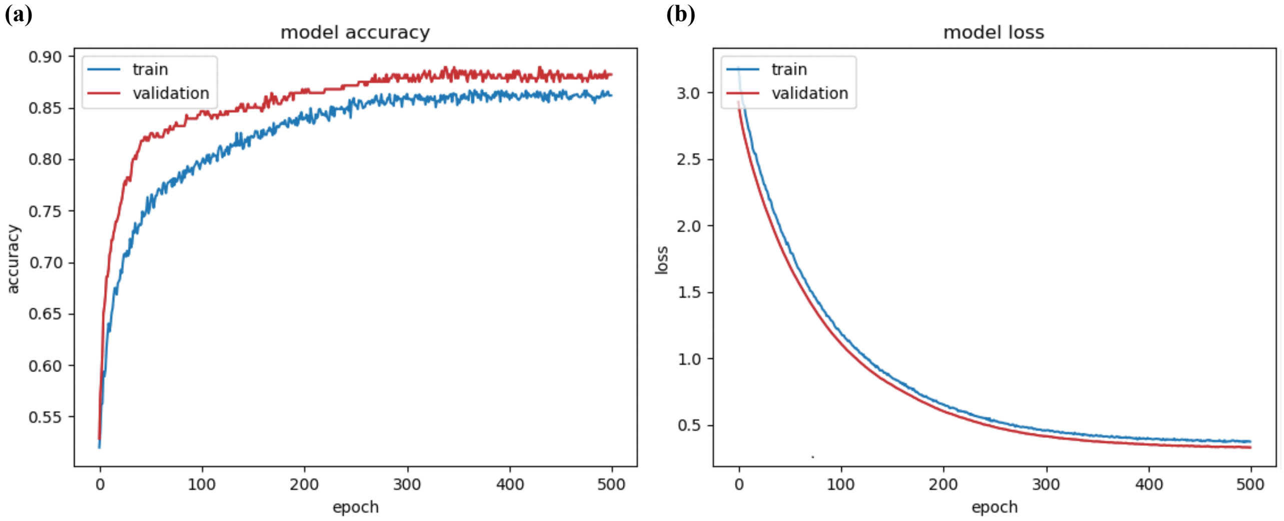

Fig. 8. MLP model performance. (**a**) Training and validation accuracy, (**b**) the model loss in training and validation

tory, hormonal therapy, chemotherapy, radiotherapy, oestrogen receptor (ER), progesterone receptor (PR), marital status, and method of axillary lymph node dissection. The five most important factors in survivability prediction are tumour size, stage, age, total axillary lymph node removed, and number of positive lymph nodes. In the study of survival and prognosis factors in breast cancer (Kong et al., 2017; Nordin et al., 2018), the authors claimed that cancer staging at diagnosis was an important factor in survival prediction in breast cancer, while the limitation of their study was that important factors such as tumour size and lymph node status were not considered in the survival analysis. In contrast, in the current study, the feature importance analysis showed that tumour size, stage, age, total axillary lymph node removed, and number of positive lymph nodes are, respectively, the most important factors in survival prediction in breast cancer. In another study (Ganggayah et al., 2019), the authors also found that the most important factors identified in their study were cancer stage classification, tumour size, total axillary lymph nodes removed, positive lymph nodes, primary treatment type, and method of diagnosis.

Analysing clinical records to predict cancer survivability of patients has been of significant interest among researchers in the last two decades. There are many challenges associated with the quantity and quality of data, prioritization of variables, missing data, and selection of survivability period, which render the analysis of survivability difficult to resolve. This study presents the deep learning technique as an alternative to machine learning for addressing breast cancer survivability prediction using selected variables. The survivability of breast cancer was compared in four cohorts of patients, since Taib et al. (2011) believed that improvement of oncology services in different time frames causes significant changes in the survivability rate of breast cancer patients. Therefore, diagnosis date in four years' time frames was set as a variable that facilitated improving the baseline accuracy of the MLP classifier.

Conclusion and future direction

In conclusion, this study presents a slight improvement in the accuracy of breast cancer survivability prediction using a deep learning technique. The feature importance analysis demonstrated that the absence of variables such as breast cancer class (invasive or in-situ), family history, hormonal therapy, chemotherapy, radiotherapy, ER, PR, marital status, and method of axillary lymph node dissection do not affect the survivability prediction significantly. The five most important factors in survivability prediction are tumour size, stage, age, total axillary lymph node removed, and number of positive lymph nodes.

Acknowledgment

E. Y. Kalafi and N. A. M. Nor are the first authors. The authors are thankful to Prof Emeritus Yip Cheng Har for providing the UMMC BC registry data.

Competing interests

The authors declare that they have no competing interest.

References

Abdullah, N. A., Wan Mahiyuddin, W. R., Muhammad, N. A., Ali, Z. M., Ibrahim, L., Ibrahim Tamim, N. S., Mustafa, A. N., Kamaluddin, M. A. (2013) Survival rate of breast cancer patients in Malaysia: a population-based study. *Asian Pac. J. Cancer Prev.* **14**, 4591-4594.

Acuña, E., Rodriguez, C. (2004) The treatment of missing values and its effect on classifier accuracy. In: *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation*, eds. Banks, D., House, L., McMorris, F. R., Arabie, P., Gaul, W. pp. 639-647. Springer Berlin, Heidelberg.

Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O. (2016) Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878.

Bai, X., Latecki, L. J. (2008) Path similarity skeleton graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1282-1292.

Beam, A. L., Kohane, I. S. (2018) Big data and machine learning in health care. *JAMA*, **319**, 1317-1318.

Blumenthal, D. T., Artzi, M., Liberman, G., Bokstein, F., Aizenstein, O., Bashat, D. B. (2017) Classification of high-grade glioma into tumor and nontumor components using support vector machine. *Am. J. Neuroradiol.* **38**, 908-914.

Boughorbel, S., Al-Ali, R., Elkum, N. (2016) Model comparison for breast cancer prognosis based on clinical data. *PLoS One*, **11**, e0146413.

Breiman, L. (2001). Random forests. *Mach. Learn.* **45**, 5-32.

Chaturvedi, M., Vaitheeswaran, K., Satishkumar, K., Das, P., Stephen, S., Nandakumar, A. (2015) Time trends in breast cancer among Indian women population: an analysis of population based cancer registry data. *Indian J. Surg. Oncol.* **6**, 427-434.

Delen, D., Walker, G., Kadam, A. (2005) Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* **34**, 113-127.

Domingos, P. (2012) A few useful things to know about machine learning. *Commun. ACM*, **55**, 78-87.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., Dhillon, S. K. (2019) Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med. Inform. Decis. Mak.* **19**, 48.

Fielding, S., Maclennan, G., Cook, J. A., Ramsay, C. R. (2008) A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* **9**, 51.

Ghosh, R., Papapanagiotou, I., Boloor, K. (2014) A survey on research initiatives for healthcare clouds. In: *Cloud Computing Applications for Quality Health Care Delivery*, eds. Moumtzoglu, A., Kastania, A., IGI Global, pp. 1-18. Hershey, PA.

Han, J., Kamber, M. (2000) *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Hirabayashi, Y., Zhang, M. (2009) Comparison of time trends in breast cancer incidence (1973-2002) in Asia, from cancer incidence in five continents, Vols IV-IX. *Jap. J. Clin. Oncol.* **39**, 411-412.

Hosseini, S. M. H., Kesler, S. R. (2014) Multivariate pattern analysis of fMRI in breast cancer survivors and healthy women. *J. Int. Neuropsychol. Soc.* **20**, 391-401.

Huang, C.-L., Liao, H.-C., Chen, M.-C. (2008) Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Syst. Appl.* **34**, 578-587.

Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., Balaji, P. V. (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics* **22**, 278-284.

Kim, J., Shin, H. (2013) Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *J. Am. Med. Inform. Assoc.* ***20***, 613-618.

Kong, Y.-C., Bhoo-Pathy, N., Subramaniam, S., Bhoo-Pathy, N., Taib, N. A., Jamaris, S., Kaur, K., See, M. H., Ho, G. F., Yip, C. H. (2017) Advanced stage at presentation remains a major factor contributing to breast cancer survival disparity between public and private hospitals in a middle-income country. *Int. J. Environ. Res. Public Health* **14**, 326-453.

Li, H., Zhong, H., Boimel, P. J., Ben-Josef, E., Xiao, Y., Fan, Y. (2017) Deep convolutional neural networks for imaging based survival analysis of rectal cancer patients. *Int. J. Radiat. Oncol. Biol. Phys.* **99**, S183.

Lotfnezhad Afshar, H., Ahmadi, M., Roudbari, M., Sadoughi, F. (2015) Prediction of breast cancer survival through knowledge discovery in databases. *Glob. J. Health Sci.* **7**, 392-398.

Nordin, N., Yaacob, N. M., Abdullah, N. H., Hairon, S. M. (2018) Survival time and prognostic factors for breast cancer among women in north-east peninsular Malaysia. *Asian Pac. J. Cancer Prev.* **19**, 497-502.

Park, K., Ali, A., Kim, D., An, Y., Kim, M., Shin, H. (2013) Robust predictive model for evaluating breast cancer survivability. *Eng. Appl. Artif. Intell.* **26**, 2194-2205.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Shu C. C.,C., Iyengar, S. S. (2018) A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, **51**, 92.

Powers, D. M. (2011) Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Tech.* **2**, 37-63.

Selvaraj, H., Selvi, S. T., Selvathi, D., Gewali, L. (2007) Brain MRI slices classification using least squares support vector machine. *ICMED* **1**, 21–33.

Shickel, B., Tighe, P., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **1**, 1.

Shukla, N., Hagenbuchner, M., Win, K. T., Yang, J. (2018) Breast cancer data analysis for survivability studies and prediction. *Comput. Methods Programs Biomed.* **155,** 199-208.

Sim, X., Ali, R. A., Wedren, S., Goh, D. L.-M., Tan, C.-S., Reilly, M., Hall, P., Chia, K. S. (2006) Ethnic differences in the time trend of female breast cancer incidence: Singapore, 1968-2002. *BMC Cancer* **6**, 261.

Sun, D., Wang, M., Li, A. (2018) A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE Trans. Comput. Biol. Bioinform.* **1**, 10.

Taib, N. A., Akmal, M., Mohamed, I., Yip, C.-H. (2011) Improvement in survival of breast cancer patients - trends over two time periods in a single institution in an Asia Pacific country, Malaysia. *Asian Pac. J. Cancer Prev.* **12**, 345-349.

Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag.

Zhang, S., Zhang, C., Yang, Q. (2003) Data preparation for data mining. *Applied Artificial Intelligence*, **17**, 375-381.