THE USE OF DATA MINING TO IMPROVE BREAST CANCER DIAGNOSIS

Shorouq Eletter, Al Ain University
Tahira Yasmin, Al Ain University
Ghaleb Elrefae, Al Ain University
Abdullah Elrefae, Al Bashir Hospital Amman

ABSTRACT

Healthcare organizations currently employ predictive analytics to improve patient care. Using data mining models to predict breast cancer is significant for improved care outcomes and patient experience.Providing data-driven decision-making process. Furthermore, historical records help to create models that aid disease detection at an early stage, which might also positively impact care outcome and patient experience. Breast cancer screening is a common service offered by healthcare providers. The study aims to evaluate the performance of five data mining algorithms for the prediction of breast cancer. Five data mining namely deep learning, naïve base, generalized linear models, support vector machines and random forest are applied to predict breast cancer. Four metrics will be used to assess the performance and highlight the best classification model. Deep learning yielded higher performance results on all metrics and ranked the variables based on their importance in building the classifier. Deep learning algorithms can be used successfully to predict breast cancer. Glucose and Resistin were the most important in the classification process. Therefore, high levels of these variable need to be monitored.

Keywords: Data Mining, Deep Learning, Breast Cancer, Confusion Matrix, Classification, Naïve Base, Generalized Linear Models, Random Forest, Support Vector Machines

INTRODUCTION

Breast cancer is one of the fatal diseases globally (Araújo et al., 2017), and is the common widespread cancer in women that has attracted notable attention from doctors and experts (Li & Chen, 2018). Moreover, breast cancer screening is extremely pivotal to detection at an early stage and increases the likelihood of auspicious treatment outcomes (Silva et al., 2019). Early detection of the disease is crucial to improving the survival rate of patients (Rahman et al., 2020). In order to detect the disease, women worldwide are encouraged to undergo blood testing and other diagnostic procedures regularly based on their age. Traditionally, the approach to breast cancer research has been biological and clinical in nature. Research has showed that it is essential to identify the underlying mechanisms and characteristics, as well as the most significant genetic factors responsible for the disease through diagnosis of patients with cancer (Maniruzzaman et al., 2019). However, breast cancer diagnosis typically entails detection using mammographic examination and self-reported symptoms (Ryerson et al., 2015). Subsequently, a breast tissue biopsy is performed and examined in abnormal cases to check the possibility of malignant tissue growth. Despite the popularity of mammography, it is not regarded as an effective technique owing to its low sensitivity and low detection accuracy (Rahman et al., 2020). Therefore, more sensitive detection techniques have been developed based on other biomarkers measured using blood analysis to improve the accuracy of breast cancer diagnosis (Mathelin et al., 2006).

Citation Information: Shorouq, E., Tahira, Y., Ghaleb, E., & Abdullah, E. (2021). The use of data mining to improve breast cancer diagnosis. *Journal of Management Information and Decision Sciences, 24*(S1), 1-6.

Recently, physicians in the medical domain have initiated new diagnostic procedures that rely on data mining tools to refine the traditional approach and improve the probability of early detection of the disease. This will better inform the physician's decision, improve efficiency of the process, and reduce treatment costs. The application of such efficient models is vital to obtaining accurate and timely predictions and reducing mortality rate. Artificial intelligence, data mining, machine learning techniques, and other tools are utilized to create efficient and robust prediction models (Livieris et al., 2020). Argue that applying predictive models to routine consultation and blood analysis data will significantly advance the screening process of breast cancer (Patrício et al., 2018).

This research aimed to explore the classification accuracy of five data mining algorithms namely deep learning, naïve base, generalized linear models, support vector machines and random forest for the breast cancer dataset. Four performance metrics will be used to assess the performance and highlight the best classification model.

LITERATURE REVIEW

The information industry is currently utilizing data mining methods and tools to generate actionable intelligence from large masses of data produced in numerous specialties, including the medical field. Practitioners have become more interested in using data mining algorithms in view of increased demand to transform masses of data into useful information and knowledge (Li & Chen, 2018). Experts have also started using data mining models to advance the diagnostics process for better accuracy in the physician's decision. For instance, (Patrício et al., 2018) used machine learning algorithms (logistic regression, random forests, and support vector machines) to predict breast cancer employing different variables. The resulting models were then evaluated using the Monte Carlo Cross-Validation method to determine 95% confidence intervals for the sensitivity, specificity, and AUC of the models. Furthermore, (Hung et al., 2018) applied PySpark and its machine learning frameworks to a dataset collected during routine blood analysis in the prediction of breast cancer. The results revealed approximately 72% and 83% accuracy rates for detection and classification, respectively. (Livieris et al., 2018) also used ensemble methodologies (an improved semi-supervised and self-labeled algorithm) to predict cancer. Additionally, (Li & Chen, 2018) employed five classification models, including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Neural Network (NN), in their breast cancer prediction. The comparative analysis revealed superiority of the RF model in achieving the optimum performance and adaptation compared with other classification models. A study by (Silva et al., 2019) using a hybrid model structure that integrates artificial neural networks, fuzzy systems, and pruning methods also confirmed high prediction accuracy in breast cancer and a high degree of interpretability in the detection of the disease among people.

MATERIALS AND METHODS

The dataset generated by (Patrício et al., 2018), which was used by the research, involved the collection of blood samples for 116 women, including healthy (control) women. The data that were gathered for each woman indicated age, height, weight and menopausal status, where the latter showed if the participant was at least one year post-menopausal, or signified a bilateral oophorectomy. In order to obtain the BMI (in kg/m^2) the weight (in kg) was divided by the height (in m^2). The dataset was randomly separated into two subsets, 70 percent of which was training data utilized for constructing breast cancer classifiers, whereas 30 percent represented testing data applied in paradigm evaluation. Subsequently, both data mining models were applied and the best paradigm was chosen by using several performance metrics.

Citation Information: Shorouq, E., Tahira, Y., Ghaleb, E., & Abdullah, E. (2021). The use of data mining to improve breast cancer diagnosis. *Journal of Management Information and Decision Sciences, 24*(S1), 1-6.

Researchers have been motivated by the evolution in data mining and data science to capitalize on vast quantities of data (known as "big data"), and also to use the available data to generate value. It is only since big data recently appeared that data mining has become prominent, despite that fact that it is not a new phenomenon (Moro et al., 2019). Progression in data science and the availability of massive volumes of data have been advantageous to the healthcare sector in the same way as other disciplines have. Furthermore, in healthcare operations, big data occupy an essential function which involves various illnesses as well as the availability of personalized treatment of patients leading to enhanced results (Alexander & Wang, 2017). The effectiveness and safety of such treatment is helped by the development of predictive paradigms (Bibault, Giraud & Burgun, 2016).

Deep learning (DL) may be defined as is a non-linear and multiple-layered algorithm (Nosratabadi et al., 2020) which imitates the human brain and learns tasks by applying artificial networks (Zhang et al., 2019). In order to create "deep networks", the layers, each of which is composed of basic units known as neutrons, are connected and sequentially stacked (Serre, 2019). For the purpose of resolving tasks, such connections are able to obtain information from raw data because they are trained on data. Training data are used in order to prepare the weighted connections by modifying the connection weights, and consequently used to establish the link between input and output (Zhang, 2019).

Generalized Linear Models (GLM) can be described as statistical processes grounded on advanced regression analysis which extend the linear paradigm to enable the dependent variable to be linearly associated with the covariates and factors by means of a specific connection function. Furthermore, GLM can be successful in categorizing data with a discrete dependent variable (Crisci, Ghattas & Perera, 2012).

Naïve Bayes (NB), which is grounded on Bayes' theorem, can be defined as a probabilistic classification algorithm. It belongs to a group of algorithms which assume class conditional independence, whereas every feature makes an important contribution to the result. In displaying a high performance, NB functions better than more sophisticated classification algorithms, particularly with categorical data, and it computes the posterior probability of each term representing a class:

$$P(c|\,X) = \frac{P(X\,|c) * P(c)}{P(X)} \qquad (1)$$

P(c) indicates the prior probability of the class within the dataset; P (X | c) represents the prior probability of a feature given a class; and *p(X)* indicates the prior probability of a feature which has already happened (Vembandasamy, Sasipriya & Deepa, 2015).

The Random Forest (RF) algorithm which merges multiple decision trees utilizes an independent subset in the course of the training phase. It also examines every decision tree's output, and its eventual result will be according to a majority vote and, subsequent to training, is used to categorize new cases. It is also able to process high data multicollinearity and dimensionality successfully (Guidotti et al., 2018; Belgiu & Drăguţ, 2016; Kononenko, 2001).

As a non-probabilistic algorithm, the Support Vector Machine (SVM) performs binary classification assignments by taking a training data subset as input and constructing a hyperplane which serves as a decision boundary. It also seeks an optimal hyperplane with the ability to retain the maximum distance between the various classes (Guidotti, 2018). Table 1 depicts a Confusion Matrix (CM) which indicates the real labels as against the predicted ones for two dataset classes. In order to investigate the DL paradigms' prediction performance, four metrics were employed. These were sufficiently accurate to evaluate a classifier's preductive power as well as the precision, sensitivity and F1-score which indicates the harmonic mean of precision and recall (Rahman, 2020; Akben, 2019; Livieris, 2018). The value of the F1-score was between 0 and 1, with 1 being the best

Citation Information: Shorouq, E., Tahira, Y., Ghaleb, E., & Abdullah, E. (2021). The use of data mining to improve breast cancer diagnosis. *Journal of Management Information and Decision Sciences, 24*(S1), 1-6.

and 0 the worst, and an increase in this score signified that the paradigms were more effective (Li & Chen 2018).

RESULTS

This section presents the results of the proposed classifiers, namely DL, GLM, NB, RF and SVM. Presented in Table 1 and Figure 1 a summary of the performance metrics for the classification models.

Table 1
CLASSIFICATION PERFORMANCE OF THE DATA MINING MODELS

Classification model	DL	GLM	NB	RF	SVM
Accuracy	85.71 %	77.14%	74.29%	74.29%	74.29%
Sensitivity	85.83%	78.33%	70.00%	73.58%	75.83%
Precision	85.36%	77.96%	84.48%	74.17%	75.83%
F1-score	85.59%	78.14%	76.56%	74.00%	75.83%

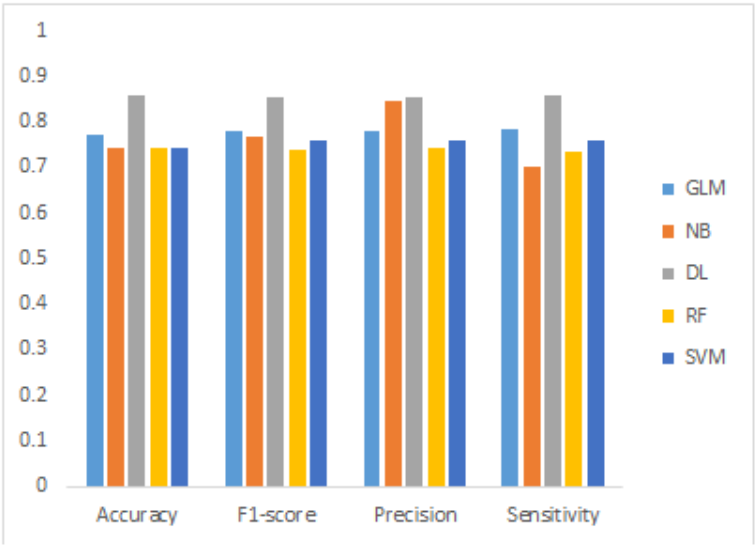

FIGURE 1
CLASSIFIERS PERFORMANCE METRICS

The bar charts in Figure 2 display the importance of the variables used in developing the DL ranked in descending order. It appears that glucose, Resistin, BMI, and Insulin had a higher effect on the classification of cases by the DL which highlights these variables as the most important indicators for diagnosis of breast cancer. High levels of these variables may require further careful attention.

Citation Information: Shorouq, E., Tahira, Y., Ghaleb, E., & Abdullah, E. (2021). The use of data mining to improve breast cancer diagnosis. *Journal of Management Information and Decision Sciences, 24*(S1), 1-6.

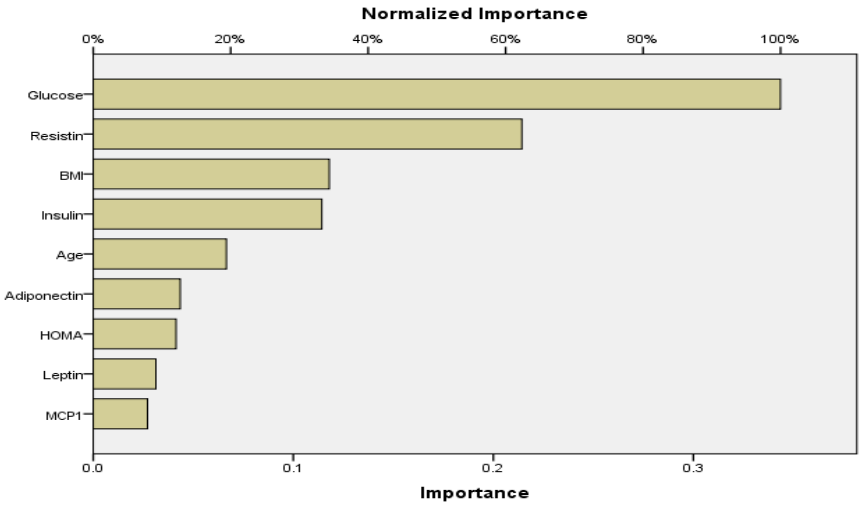

FIGURE 2
VARIABLE IMPORTANCE OF DL MODEL

DISCUSSION

The DL algorithm generated the highest performance with 85.71% accuracy, GLM yielded 77.14%. Conversely, NB, RF and SVM achieved 74.29% classification accuracy. The sensitivity of the SVM outperformed the other classifiers. Sensitivity of DL 85.83%. NB achieved the lowest sensitivity of 70.00%. Additionally, DL and NB precisions were 85.36% and 84.48% respectively which outperformed the other models. Furthermore, DL yielded the highest F1-score of 85.59% which outperformed the other classifiers. Overall, the results of the four metrics suggest that DL algorithm reveals better performance than other classifiers. This is in contrast to the study by (Li & Chen, 2018) which achieved F1-score of and 78% but the accuracy was 74.3% for the RF model which is close to the results of current study. Yet, the SVM model in their study yielded 71.4% and 76.2% accuracy and F1-score, respectively. The study by (Silva, 2019) used an artificial neural network model yielded 73.8% for accuracy, however the DL model used here achieved 85.71%.

CONCLUSIONS

This study used five data mining models, namely DL, GLM, NB, RF and SVM to detect the presence of breast cancer using a real dataset which evaluated attributes including age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin,Resistin, and MCP-1. Accuracy, sensitivity, precision and F1-score were the criteria used to assess the predictive accuracy of the developed models. The results indicated that DL yielded higher performance results on all metrics and ranked the variables based on their importance in building the classifier. Additionally, Glucose and Resistin were the most important in the classification process. This indicates the high levels of these variable need to be monitored.

These results could help to reduce misdiagnosis and provide suitable treatment for breast cancer patients (Li & Chen, 2018). Additionally, screening for breast cancer using a blood sample permits detection of the disease at an early stage increases the opportunity for auspicious treatment outcomes, if discovered in a timely manner, and decreases the overall cost.

Citation Information: Shorouq, E., Tahira, Y., Ghaleb, E., & Abdullah, E. (2021). The use of data mining to improve breast cancer diagnosis. *Journal of Management Information and Decision Sciences, 24*(S1), 1-6.

REFERENCES

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C. ... & Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks. *PloS one*, *12*(6).

Li, Y., & Chen, Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. *Appl. Comput. Math*, *7*(4), 212-216.

Silva Araújo, V.J., Guimarães, A.J., de Campos Souza, P.V., Rezende, T.S., & Araújo, V.S. (2019). Using resistin, glucose, age, and bmi and pruning fuzzy neural network for the construction of expert systems in the prediction of breast cancer. *Machine Learning and Knowledge Extraction*, *1*(1), 466-482.

Rahman, M.M., Ghasemi, Y., Suley, E., Zhou, Y., Wang, S., & Rogers, J. (2020). Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric and clinical features. *IRBM*.

Maniruzzaman, M., Rahman, M.J., Ahammed, B., Abedin, M.M., Suri, H.S., Biswas, M., ... & Suri, J.S. (2019). Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Computer methods and programs in biomedicine*, *176*, 173-193.

Ryerson, A.B., Miller, J., & Eheman, C.R. (2015). Reported breast symptoms in the national breast and cervical cancer early detection program. *Cancer Causes & Control*, *26*(5), 733-740.

Mathelin, C., Cromer, A., Wendling, C., Tomasetto, C., & Rio, M.C. (2006). Serum biomarkers for detection of breast cancers: a prospective study. *Breast cancer research and treatment*, *96*(1), 83-90.

Livieris, I., Pintelas, E., Kanavos, A., & Pintelas, P. (2020). An improved self-labeled algorithm for cancer prediction. In *GeNeDis 2018, Springer, Cham*, 331-342.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC cancer*, *18*(1), 29.

Hung, P.D., Hanh, T.D., & Diep, V.T. (2018). Breast cancer prediction using spark MLlib and ML packages. In *Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications*, 52-59.

Livieris, I., Pintelas, E., Kanavos, A., & Pintelas, P. (2018). An improved self-labeled algorithm for cancer prediction. *Advances in Experimental Medicine and Biology*.

Moro, S., Esmerado, J., Ramos, P., & Alturas, B. (2019). Evaluating a guest satisfaction model through data mining. *International Journal of Contemporary Hospitality Management*.

Alexander, C.A., & Wang, L. (2017). Big data analytics in heart attack prediction. *J Nurs Care*, *6*(393), 2167-1168.

Bibault, J.E., Giraud, P., & Burgun, A. (2016). Big data and machine learning in radiation oncology: State of the art and future prospects. *Cancer letters*, *382*(1), 110-117.

Nosratabadi, S., Mosavi, A., Duan, P., & Ghamisi, P. (2020). Data science in economics. *arXiv preprint arXiv: 2003.13422*.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, *99*, 238–248.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*, 399-426.

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, *240*, 113-122.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109.

Akben, S.B. (2019). Determination of the blood, hormone, and obesity value ranges that indicate the breast cancer, using data mining based expert system. *IRBM*, *40*(6), 355-360.

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*(1).

Citation Information: Shorouq, E., Tahira, Y., Ghaleb, E., & Abdullah, E. (2021). The use of data mining to improve breast cancer diagnosis. *Journal of Management Information and Decision Sciences, 24*(S1), 1-6.