International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-2, July 2020

Breast Cancer Prediction based on Deep Neural Network Model Implemented AWS Machine Learning Platform

L. D. P Cuong, Wang Dong, D. T. Hoang, L. M. N Uyen

Abstract*: Breast cancer in women is one of the most dangerous cancers leading to death in women by developing breast tissue. In this work, the application of the Deep Neural Network (DNN) model is implemented on AWS machine learning platform, besides, a comparison with other ML techniques includes XGBoost and Random Forest on a public dataset. Breast cancer prediction based on DNN model with Hyperparameter tuning has the best results of the plot of model accuracy for the training and validation sets and performance evaluation metrics to test the model.

***Keywords*: Breast cancer, Deep Neural Network, Deep Learning, AWS SageMaker, Docker containers.**

I. INTRODUCTION

Breast cancer is one of 10 cancers in women or a small proportion in men and it is also one of the worldwide diseases. The disease is on the rise in developing countries where the disease is diagnosed at an advanced stage, such as: an estimated 45,000 women in Vietnam are living with breast cancer or according to a survey in India, out of 100 patients, 28% of women of age group 40 to 50 and the rest for other age groups [1]. Therefore, a recommendation from leading experts in breast cancer diagnosis of women is to screen for breast cancer between the age of 40 and 90% of the disease if detected early.

In addition, current treatments for breast cancer have made great strides such as surgery, radiation, chemicals and combination with hormone therapy, biological therapy (targeted treatment). The quality of treatment has thus improved significantly. However, the key to effective treatment is early detection of cancer. Accurate diagnosis of breast cancer in images and information of histopathology is a challenge due to the heterogeneity of cancer cell growth as well as a series of benign breast tissue proliferation lesions. Many studies have taken step by step in predicting and preventing cancer, the exact part is still a challenge.

Additional tools have been added to help physicians facilitate accurate diagnosis. These tools focus on eliminating possible diagnostic errors and provide the easiest way to analyze large amounts of data. In this paper, a deep learning model of breast cancer detection to diagnose breast cancer is discussed using the Deep Neural Network (DNN).

The DNN is neural networks with multiple layers between the input and output layers. It finds the correct mathematical operations to turn inputs into outputs, whether it is linear or non-linear. The network moves through layers that calculate the probability of each output.

Source: https://www.cancer.org/

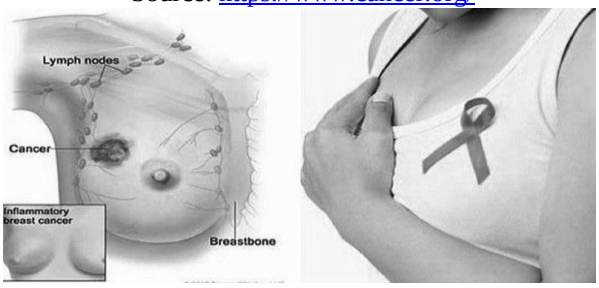

Figure 1: Breast Cancer Disease

Also, DNN is an artificial neural network (ANN) is an effective technology to implement a computational system with a large number of interconnected processing units to compute information. It applied neural network for regression of continuous target attributes. It has three layers as: Input layer, Hidden layer and Output layers [2,3].

Based on our research, many neural networks designed, the transitional neural network was used for the purpose of predicting cancer disease [4].

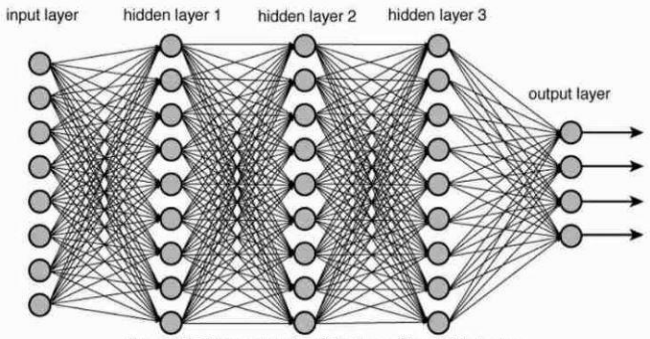

Figure 2: Deep Neural Network architecture with multiple layers

The above figure shows the architecture of a deep neural network combining layers.

Manuscript received on May 25, 2020.
Revised Manuscript received on June 29, 2020.
Manuscript published on July 30, 2020.
***** Correspondence Author
L. D. P Cuong*, College of Electrical and Information Engineering, Hunan University, Changsha 410082, China. E-mail: ledinhphucuong.dalat@gmail.com
Wang Dong, College of Electrical and Information Engineering, Hunan University, Changsha 410082, China. E-mail: wangd@hnu.edu.cn
D. T. Hoang, College of Electrical and Information Engineering, Hunan University, Changsha 410082, China.
E-mail: duyenthehoang1995@gmail.com
L. M. N Uyen, College of Life science, Hunan Normal University, Changsha 410082, China. E-mail: uyenlmn@yersin.edu.vn.

II. LITERATURE REVIEW

Author Researchers are adopting various machine learning methods, including machine learning, deep learning and Cloud AutoML (Google or Amazon) technology [5,6,7], which they predict breast cancer disease with best results.

The survey showed some machine learning models to predict breast cancer from public data. This has used until now includes logistic regression, support vector machines, neural network and random forests. The performance of all the models tested is evaluated using several metrics, including Accuracy, Recall, F1 score, AUC score (Area under the Curve) [8]. These metrics were number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). This gives Accuracy (Eq.1), Recall (Eq.2), F1 (Eq.3) and AUC stands for "Area under the ROC curve" that is full dimension measurement is underneath the entire ROC curve from (0.0) to (1.1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (3)$$

In section II, our appreciation shows that the review of previous implementation in breast cancer prediction which has more details by different researchers with achieved their the results and methods below Table 1 [1,9,10,11,12,13].

Table 1: The reviewed results from different researcher

Reference Used	Applied Method	Achieved Prediction	Accuracy (%)	Limitations
Dhanalakshmi G. et al. [1]	Logistic Regression, k-Nearest Neighbor (k-NN) algorithm, Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF)	Breast cancer, SVM with accuracy 98%, k-NN algorithm with accuracy 87%, NN achieved training accuracy ranging from 93-97%. Value result of evaluation metrics result (Sensitivity and Specificity).	93-97	No value results of evaluation metrics (Recall, F1-Score, Precision, AUC). The results of evaluation metrics comparing the methods are not really clear.
Medisetty, H. K., Kunjam, N. R. [9]	SVM, RF, Gradient Boosting, Naïve Bayes (NB), Cart model, NN and Linear Regression.	Breast cancer, gave a performance comparison these algorithms showed that SVM was the best algorithm with accuracy 98.23%. This was the best performance in terms of precision and low error rate. Besides, features engineering for dataset. Value result of evaluation metrics result (Sensitivity and Specificity).	98.23	No value results of evaluation metrics (Recall, F1-Score, Precision, AUC)
Sivapriya, J. [10]	LR, SVM, NB and RF.	Breast cancer, gave a performance comparison these algorithms showed that Random Forest was the best algorithm with accuracy 99.76%. Especially, it gave time result to build each model.	99.76	It is not give some results of evaluation metrics and features engineering.

Yeulkar, K., Sheikh, R. [11]	Data mining (DM). C4.5 and NB.	Breast cancer. The performance of C4.5 (accuracy of 98.09%) is better than Naive Bayes (accuracy of 95.85%). Besides, features engineering for dataset. Especially, Pre-Classification.	98.09	No value results of evaluation metrics.
Atrey, K. et al. [12]	Statistical significance analysis (SSA) uses SPSS tool. NB, ANN, and SVM comparison, dataset pre-processing.	Breast cancer. Performed independent t-test to get the p-value by SPSS, mean and standard deviation. Besides, feature section. ANN is the best classifier than others in the range of 98.9% to 99.7% of the three feature sets	99.7	Classification for the models is clear but this is perceived to be a complex task
Dhahri, H. et al. [13]	Combining feature preprocessing methods (WEKA software to extract the features based on the EA and BF) and classifier algorithms (SVM, k-NN, DT, LR, AB, GNB and LDA).	Breast cancer. The result of the accuracy evaluation for the models by combining classifier accuracy and classifier log loss.	98.23	The accuracy evaluation by combining the two goals for the models is clear but this is considered a complex task.

Through a more detailed comparison of previous researchers with recent years of 2020, showed that as: both the algorithms and the models were applied to predict breast cancer with some parameters such as: radius, texture, perimeter, and concavity, so on, such as Logistic Regression, k-NN, RF, NB, GNB, SVM, NN and used WEKA and SPSS tools are the tools in data mining, so on. However, there are still limitations: only calculated sensitivity and specificity metrics, features engineering, classification for the models and the accuracy evaluation by combining the two goals for the models are clearly but these are considered to be a complex task.

III. METHODS AND MATERIALS

In this paper, we propose Amazon SageMaker technology [14,15] predicts the breast cancer in few years recently. So, we chose this method to make evaluation of DNN model. Besides, sequential model in Keras [16] is built by stacking layers sequentially which allows us to build a model layer by layer. Each layer has weights that expropriates to the layer. We use *add ()* function to add layers to our model, specifically, 'Dense' is layer type. Dense is a standard layer type that works for most cases. In this layer, all nodes in the previous layer connect to the nodes in the current layer [17].

In addition, the authors compare the exact evaluation results of DNN with other models as follows: XGboost and Random Forest. XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework which is used implementations of boosted decision trees in the word. Now, it is available in Amazon Sagemaker. It provides a very powerful tool is called "*Hyperparameter tuning jobs*" which helps to tuning DNN model faster and more effectively. Fig. 3 shows a Notebook instances in Amazon SageMaker, namely cuong-dalat.

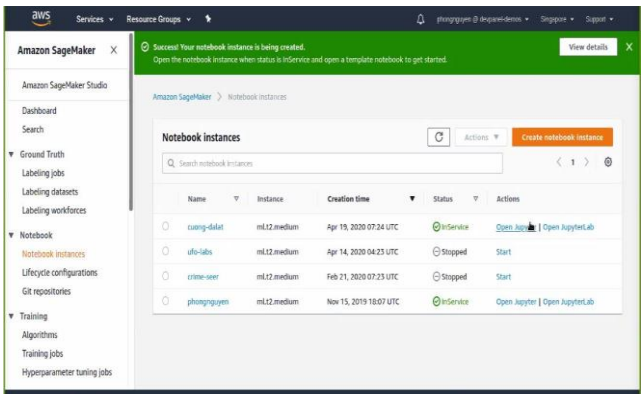

Figure 3: Amazon SageMaker

For many years, [18] reported XGBoost algorithm that is good at dealing with high dimensional data in this paper.

Based on cleaning for excessive number of variables and missing values of high dimensional data, therefore, we apply XGBoost algorithm which is using multi-observation data cleaning has advanced accuracy in prediction. It has also been considered a recent phenomenon of excellence in various cases in which the concept originated from the construction of additive models [19].

Algorithm 1 XGBoost algorithm

1: Data: Dataset and Hyperparameters
2: Initialize $f_0(x)$;
3: For i= 1,2, …, M do
Calculate $g_i = \frac{\partial L(y,f)}{\partial f}$;
Calculate $h_i = \frac{\partial^2 L(y,f)}{\partial f^2}$;
Determine the structure by choosing splits with maximized gain
$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} - \frac{G_L^2}{H}\right];$$
Determine the leaf weights $w^* = -\frac{G}{H}$;
Determine the base learner $\hat{b}(x) = \sum_{j=1}^{T} wI$;
Add trees $f_i(x) = f_{i-1}(x) + \hat{b}(x)$;
4: Result: $f_i(x) = \sum_{i=0}^{M} f_i(x)$;

Random forest also knew as random decision forests to build a large number of trees that achieve their output through ensemble learning methods. The random forest does not over-fit the data [20].

Dataset is used from Wisconsin breast cancer data on public dataset. It consists of 32 features, with the ID number column which is removed in data pre-processing, the next column being the diagnosis result (*benign or malignant*). We seem that this is a binary classification problem then we change the label to 0 and 1 for features.

Out[8]:

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419
1	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812
2	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069
3	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597
4	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809

5 rows × 31 columns

Figure 4: Preprocessed Breast Cancer Dataset

IV. RESULT

To experiment the proposed model implemented AWS Machine Learning Platform with the public dataset has taken and a comparison with the other models. The dataset contains many breast cancer attributes like radius, texture, perimeter, and concavity, etc. They are the given information to the neural network and trained using DNN model. Since, our results have achieved the following highlights as:

Firstly, the comparison of other models is XGBoost and Random Forest in Fig.5 and Fig. 6.

Secondly, the DNN model layers in the training process shows in Fig.7.

Thirdly, plot accuracy and loss values of DNN model gives in Fig.8 and Fig. 9.

Finally, the evaluation metrics highlights indicated that the DNN model has improved results compared to the majority of previous models with evaluation metrics (Accuracy, Recall, F1 and AUC score) Fig.10.

```
Out[16]: XGBClassifier(base_score=0.5, booster=None, colsample_bylevel=1,
                       colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
                       importance_type='gain', interaction_constraints=None,
                       learning_rate=0.300000012, max_delta_step=0, max_depth=6,
                       min_child_weight=1, missing=nan, monotone_constraints=None,
                       n_estimators=100, n_jobs=0, num_parallel_tree=1,
                       objective='binary:logistic', random_state=0,
                       reg_lambda=1, scale_pos_weight=1, subsample=1,
                       validate_parameters=False)

In [17]: y_pred = model.predict(X_tes
         predictions = [round(value) for value in y_pred]

In [18]: from sklearn.metrics import accuracy_score

In [19]: accuracy = accuracy_score(y_test, predictions)
         print("Accuracy: %.2f%%" % (accuracy * 100))

         Accuracy: 94.05%
```

Figure 5: XGBoost model result

```
In [17]: pred = tree.predict(X_test)

In [18]: from sklearn.metrics import classification_report,confusion_matrix

In [19]: print(classification_report(y_test,pred))

              accuracy  precision  recall  f1-score

                  0.91       0.95    0.81      0.87
```

Figure 6: Random Forest model result

```
In [19]: model.summary()

Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 60)                1860
_________________________________________________________________
dense_5 (Dense)              (None, 30)                1830
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 31
=================================================================
Total params: 3,721
Trainable params: 3,721
Non-trainable params: 0
```

Figure 7: DNN model layers

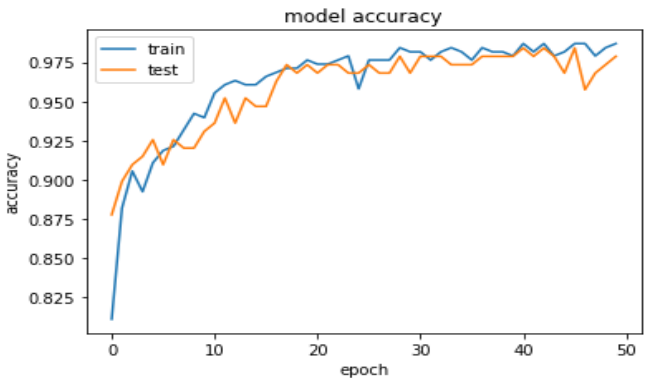

Figure 8: Plot accuracy values of DNN model

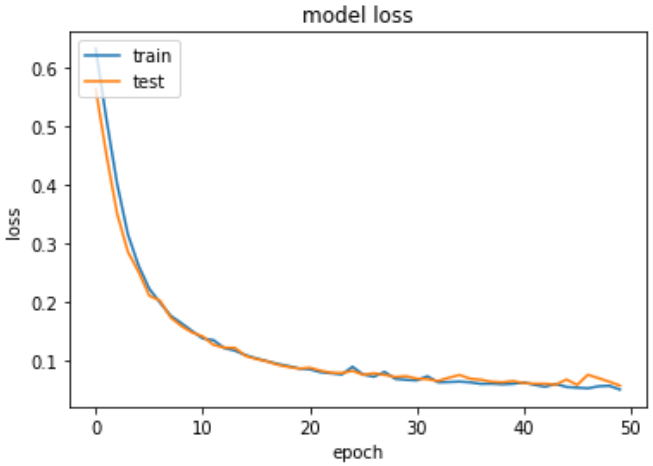

Figure 9: Plot loss values of DNN model

```
# Đánh giá kết quả trên tập Train
f1 = f1_score(np.around(model.predict(X_train)), y_train)
recall = recall_score(np.around(model.predict(X_train)), y_train)
precision = precision_score(np.around(model.predict(X_train)), y_train)
accuracy = accuracy_score(np.around(model.predict(X_train)), y_train)

print("Train F1 score: {}".format(f1))
print("Train Recall: {}".format(recall))
print("Train Precision: {}".format(precision))
print("Train Accuracy: {}".format(accuracy))
```

```
Train F1 score: 0.9826989619377162
Train Recall: 0.9861111111111112
Train Precision: 0.9793103448275862
Train Accuracy: 0.9868766404199475
```

```
# Đánh giá kết quả trên tập Test
f1 = f1_score(np.around(model.predict(X_test)), y_test)
recall = recall_score(np.around(model.predict(X_test)), y_test)
precision = precision_score(np.around(model.predict(X_test)), y_test)
accuracy = accuracy_score(np.around(model.predict(X_test)), y_test)

print("Test F1 score: {}".format(f1))
print("Test Recall: {}".format(recall))
print("Test Precision: {}".format(precision))
print("Test Accuracy: {}".format(accuracy))
```

```
Test F1 score: 0.9777777777777777
Test Recall: 0.9705882352941176
Test Precision: 0.9850746268656716
Test Accuracy: 0.9840425531914894
```

```
model.save('model_2.h5')
```

Figure 10: The result of evaluation metrics of DNN model

V. CONCLUSION

The SageMaker is Amazon's machine learning platform for building and deploying machine models. The platform includes innate support many common machine learning methods and algorithms including XGBoost, Random Forest, and Deep Learning, etc. The highlight of SageMaker platform is using Docker containers [21] technology to train and deploy models more effectively.

REFERENCES

1. Dhanalakshmi, G., Keerthana, P., Rohini, M. and Karunamoorthy, Y. (2019). Decision Support System for Breast Cancer Predicttion. IJRASET, 7(3), March 2019, pp 816–821. Retrieved from http://doi.org/10.22214/ijraset.2019.3142
2. Subashini, A., Thamarai, S. M. and Meyyappan, T. (2019). Advanced Weather Forecasting Prediction using Deep Learning. IJRASET, 7(8), Aug 2019, pp 939-945. Retrieved from http://doi.org/10.22214/ijraset.2019.8139
3. Specht, D. F. A general regression neural network, IEEE transactions on neural networks, 2 (6), 1991
4. Niaei, A., Towfighi, J., Khataee, A. R. and Rostamizadeh, K. (2007). The use of ANN and the mathematical model for prediction of the main product yields in the thermal cracking of naphtha. Petroleum science and technology, 25 (8), Aug 2007, pp 967-982. Retrieved from http://doi.org/10.1080/10916460500423304
5. Cloud AutoML https://cloud.google.com/automl/
6. L. Faes et al (2019). Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study, The Lancet Digital Health, 1(5), 2019
7. Bisong, E. (2019). Google AutoML: Cloud Vision, Building Machine Learning and Deep Learning Models on Google Cloud Platform. 2019
8. Hoang, D. T., Yang, P. L., Cuong, L. D. P., Trung, P. D., Tu, N. H., Truong, L. V., ... and Nha, V. T. (2020). Weather prediction based on LSTM model implemented AWS Machine Learning Platform. IJRASET, 8(5), May 2020, pp 283–290. Retrieved from http://doi.org/10.22214/ijraset.2020.5046
9. Medisetty, H. K., Kunjam, N. R. (2018). Prediction of Breast Cancer using Machine Learning techniques. IJMTE, 8(12), December 2018, pp 5261-5269. Retrieved from http://doi.org/16.10089.IJMTE.2018.V8I12.17.2594
10. Sivapriya, J., Aravind, K. V., Siddarth, S. S. and Sriram, S. (2019). Breast Cancer Prediction using Machine Learning. IJRTE, 8(4), November 2019, pp 4879-4881. Retrieved from http://doi.org/10.35940/ijrte.D8292.118419
11. Yeulkar, K., Sheikh, R. (2017). Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R. IJRASET, 5(3), March 2017, pp 406-410. Retrieved from http://doi.org/10.22214/ijraset.2017.3074
12. Atrey, K., Sharma, Y., Bodhey, N. K., & Singh, B. K. (2019). Breast cancer prediction using dominance-based feature filtering approach: A comparative investigation in machine learning archetype. Brazilian Archives of Biology and Technology, 2019, 62. Retrieved from http://dx.doi.org/10.1590/1678-4324-2019180486
13. Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. Journal of Healthcare Engineering, 2019. Retrieved from https://doi.org/10.1155/2019/4253641
14. AWS https://aws.amazon.com/vi/
15. Amazon Sagemaker https://aws.amazon.com/vi/sagemaker/
16. Moolayil, J., Moolayil, J., and John, S. (2019). Learn Keras for Deep Neural Networks. Apress.
17. Ketkar, N. (2017). Introduction to keras. In Deep learning with Python Apress, Berkeley, CA, 2017, pp 97-111.
18. Ma, X., et al. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, Electronic Commerce Research and Applications, 2018, 31, 24-39.
19. Choi, D. K (2019). Data-Driven Materials Modeling with XGBoost Algorithm and Statistical Inference Analysis for Prediction of Fatigue Strength of Steels, International Journal of Precision Engineering and Manufacturing, 2019, 20(1), pp 129-138.
20. Jean, S. et al. (2020), Breast Cancer Classification and Prediction using Machine Learning, IJERT, February 2020, 9(2). Retrieved from http://dx.doi.org/10.17577/IJERTV9IS020280
21. Cito, J., Ferme, V. and Gall, H. C. (2016). Using Docker containers to improve reproducibility in software and web engineering research, International Conference on Web Engineering, 2016.

AUTHORS PROFILE

L. D. P Cuong is currently a PH.D candidate in Computer Science from College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. He is also a lecturer in Department of Information Technology, Yersin University, Da Lat, Vietnam. He received his Master degree in Information Technology, Paris VI University. His research interests include Artificial Intelligence, Cloud Computing, Wireless Network and Data analytics.
Email: ledinhphucuong.dalat@gmail.com

Wang Dong received the B.S. and Ph.D. degrees in computer science from Hunan University, in 1986 and 2006, respectively. From 2004 to 2005, he was a Visiting Scholar with the University of Technology Sydney, Australia. Since 1986, he has been with Hunan University, China, where he is currently a Professor. His main research interests include network test and performance evaluation, wireless communications, and mobile computing.
Email: wangd@hnu.edu.cn

D. T. Hoang is currently a Master candidate in College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. His research interests Artificial Intelligence.
Email: duyenthehoang1995@gmail.com

L. M. N Uyen is currently a PH.D candidate in Biochemistry and Molecule Biology from College of Life Science, Hunan Normal University, Changsha 410082, China. She is also a lecturer in Department of Nursing, Yersin University, Da Lat, Vietnam. She received her Master degree in Biology, Da LatUniversity. Her research interests include Life Sciences and Biotechnology.
Email: uyenlmn@yersin.edu.vn