A ROBUST REGRESSION MODEL BASED ON OPTIMAL FEATURE SET FOR SIMPLE DECISION MAKING IN INDOOR FARMS

Sam, Nguyen-Xuan[*], **Nguyen Ngoc Giang**[+]
[*] Posts and Telecoms Institute of Technology at HCM city campus
[+] Banking University at HCM city

Abstract: This paper proposed a robust regression model for simple decision making in smart indoor farms. In our proposal, there are several steps to ensure the time-series data set which collected from sensor nodes in smart indoor farms are expanded to its features into new data set. The step tries to maximize features, then high corelated features with outcome in new data set will be filtered with strong threshold value. Moreover, we use statistical tests to remove the features in original regression model for finding out the final model. The approach not only interprets curve fitting but also produces small features for equation in the final equation. Simulation results shown that R-square value of the final model is close to R-squared value of original model while outcome in the final equation just depends on small features. The results shown that our proposal can make optimized decisions making in practical applications of agricultural systems.

Keywords: Multiple Regression (MR), Smart Indoor Farms (SIF), Optimal Feature Set (OFS), Simple Decision Making (SDM).

I. INTRODUCTION

Recently, it is very essential to integrate new technologies such as artificial intelligence (AI), internet of things (IoT) for monitoring and controlling agriculture systems because climate change and complex environmental problems impact and change rapidly. Based on the technologies, collected data from IoT devices can be transformed to information at end-devices. The agriculture systems not only help monitor environmental problems but also deliver information to enhance farmers' decisions [1]. In the context, a decision-making for the smart systems may prefer quick and simple reactions to outcome. To solve the problem, a robust multiple regression modeling with specifies variables is necessary. In general, the multiple regression models determine the simple relationships of variables in which outcome is a dependent variable and the other ones are independent variables [2]. A new concept of smart indoor farms

technology is introduced [3, 4] by using IoT devices such as solar radiation, temperature, relative humidity, and wind speed, etc. The raw data of the variables are useful for analyzing the relationship between the independent variables and outcome. Moreover, the more independent variables, the best performance of the model are generated, then various decision making at outcome if we can expand more features from the raw data.

On the other hand, the correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. [5]. It means that there are several independent variables or features can contribute for optimal outcome in multiple regression. Thus, a practical method for controlling the outcome in the smart indoor systems should consider correlation coefficient between the independent variables and outcome. Therefore, a maximizing outcome in the model need to find out the strong positive correlation features from the data set and a minimizing outcome requires strong negative correlation features.

Fig.1 presents our proposed concept [6, 7] of smart indoor farms for smart agriculture, where module 1 is farm side, providing actuator and sensor devices, module 2 contains processes data, stores data at firebase cloud, and module 3 is client side, providing data visualization. In module 1, our prototype sensors are deployed across farming area to collect various data relating to temperature, relative humidity, precipitation, solar radiation, wind speed, and actuators. The raw data is forward to firebase cloud, where the raw data is pre-processed before feeding to the learning algorithm. The module 3 present various types of information such as real time measurement, location, prediction of temperature in short term and long term.

Tác giả liên lạc: Nguyễn Xuân Sâm,
Email: samnx@ptithcm.edu.vn
Đến tòa soạn: 10/2020 chỉnh sửa: 11/2020, chấp nhận đăng: 12/2020
Nghiên cứu này được tài trợ bởi PTIT có mã số 08-HV-2020-RD_TH2.

However, in this work, we focus on how to maximize temperature outcome while keeping small independent variables. Therefore, we proposed a robust final equation where strong correlation features in data set can present the relationship between outcome and independent variables clearly and simply.

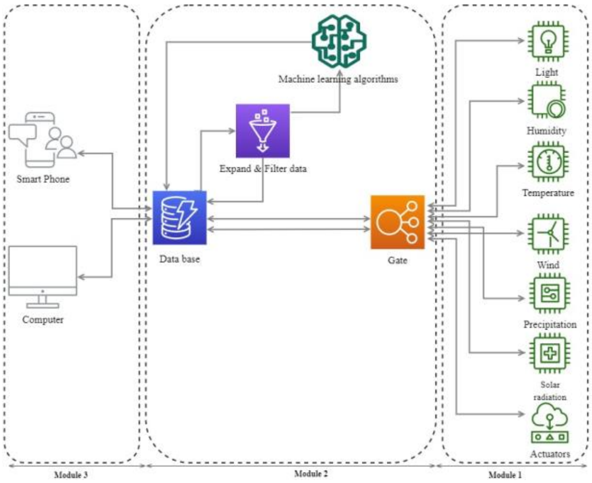

Fig 1. Smart farming at UCLAB [6]

In the work, Our proposal differs to previous work [3] by three-fold 1) the variables from raw data are expanded by feature representation [8] in time series, 2) threshold-based feature selection algorithm to find the optimal subset for learning algorithm, and 3) statistical test to remove the features in original regression model for finding out the final model. The first and second steps aim to select the “best” features that are described in the regression model have strongest correlation relationship between independent variable as an outcome, while the last one keeps the model is simple.

The rest of this paper is organized as follows: Section 2 is related works, section 3 is proposed model, then section 4 is our simulation results on different scenarios, and the last section shows conclusions and future works.

II. RELATED WORKS

Recently, projects relating with internet of thing (IoT) based smart indoor farming are proposed [3, 9, 10]. The projects aim to design and develop a smart control system using sensor devices and actuators with suitable flatforms for monitoring, controlling, and managing independent and dependent variables anytime and anywhere. With correct solution and method, it is possible to save and allow a better efficiency in the process of outcome. In the projects, light, relative humidity, temperature, wind speed, solar radiation, precipitation, etc. sensor devices have produced very huge raw data. Moreover, the relationship of variables with the outcome are determined via coefficients in the equations of multivariate regression [11].

Basically, related humidity and temperature are crucial conditions which not only reflect for growing plants but also influence on the other variables. Raw data, including temperature, relative humidity collected inside a farm uses the humidity and temperature sensors [12] with platforms such as Arduino, nodeMCU [13], etc. The devices are not only low cost but also easy to use. Basically, the accuracy of DHT22 sensors is ± 0.5 ºC for temperature and ± 5 % for relative humidity and the sensor devices deploy different positions. Relative humidity has both negative and positive correlations with temperature depending on seasons, time, period of day.

Precipitation is a major component of the water cycle and is responsible for depositing the fresh water on the planet. Precipitation has both negative and positive correlations with temperature [14]. On the other hand, wind speed and temperature have strong relationship in term of outdoor condition but in the smart indoor farm, increasing in wind speed from 1 to 3 m/s, temperature decreases to 0.78 ºC [15]. According to research [16] solar radiation is positive correlation with temperature range on the daily.

A new concept of smart indoor farms is used the sensors and actuators devices, cloud flatform, and visualization technology to provide forecasting and predicting accuracy. Some models of farm temperature requirement have been formulated, which based on the collecting data. Introducing frameworks which employ a context aware into IoT is expected to be a critical solution. These contextual data along with the incoming rules are provided in report [17], the rules are based on the context data such as temperature, humidity, wind, and so on. Thus, the service rules can be easily described with control actions.

A simple system are introduced [18] by controlling on-off outcome via smart phone, tablet and desktop. Thus, a new way to manage and control outcome based on on/off decisions depending on correlation values of outcome and independent variables. For example, we decide speed up air temperature inside smart farm by turning on the light (as first option) if we find out the correlation between light and temperature are very strong. It is worth to noting that the model can help you find out the best solutions for interrupt, speedups, timing delays, etc. [19, 20].

III. PROPOSED MODEL

A. Mathematical Model

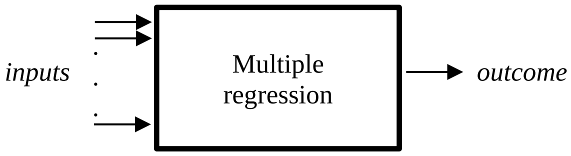

Fig 2. Multiple regression model

To describe the relationships between a dependent output as an outcome and independent inputs. The general components of proposed model are presented in Fig. 2. Our general learning model for Fig.2 is shown in equation (1) as following:

$$meanT = f(\,minP, maxP, meanP, minRH, maxRH, meanRH, minWS, maxWS, meanWS, minSR, maxSR, meanSR) \quad (1)$$

Where *meanT* is a outcome, as dependent variable (°C), *P* is denoted as precipitation variable (mm), *RH* is denoted as humidity variable (%), *P* is denoted as precipitation variable (mm), and *WS* is denoted as wind speed variable (m/s), and *SR* is denoted as solar radiation variable (W/m^2). Many decisions can be formulated for outcome as temperature depending on the independent variables and a decision can be make based on the feature set. It ranges from a strongest correlation set to a weakest correlation set. For example, we can either maximize outcome by controlling the actuators related to strongest positive correlation of the independent variables or minimize outcome by control the actuators related to strongest negative correlation of the independent variables. To simple investigating, we summarize collected from sensors in time series (daily) that are *mean*, *max*, and *min*.

To find best fit and high correlation among the variables, we proposed two steps to optimal feature set, namely feature expansion and selected feature steps. In the first step, new data points in time series are generated from an existing data points [8], Intuitively, the first step not only add more independent variables or features buts also generate time series inputs that will be used to make predictions for future time steps. From this point, we proposed first step to shift off data set of independent variables three days (within confident interval) to generate new features for all original variables. For example, time series of *meanT#-1*, *meanT#-2*, and *meanT#-3* are generated from *meanT*, etc. By this way, 45 features, including *meanT#-1*, *meanT#-2*, *meanT#-3*, *minT#-1*, *minT#-2*, *minT#-3*, *maxT#-1*, *maxT#-2*, and *maxnT#-3*, are available for learning model in fig.2.

In the second step, an optimal feature selection using statistical technique to evaluate the relationship between features which are collected from the first step and outcome. Thus, the step remove redundant features using correlation method [5]. In general, correlation coefficient, denoted as r_i, has the range between -1 and +1. If a feature has strong positive correlation when its correlation value is larger than 0.7. The correlation coefficient is determined in equation (2) as following:

$$r_i = \frac{\sum_{i=1}^{n}(f_i - \bar{f})}{\sqrt{\sum_{i=1}^{n}(f_i - \bar{f})^2 - \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where r_i is correlation coefficient of the outcome and i^{th} feature, n is a sample size, and f_i *(i=1, 2,…,n)* is the values of the features, $\bar{f}$ is the mean value of the feature, y_i *(i=1,2,…,n)* the values of outcome, $\bar{y}$ is the mean value of the outcome.

Our proposed algorithm for expanding and selecting features steps with threshold value (0.7) is shown in fig.3. As a result, the strong positive correlation values of the features can be selected in table 1.

Table 1. The correlation values of selected features

	meanT
maxT#-3	0.856301
maxT#-2	0.869892
minT#-3	0.889736
minT#-2	0.902798
maxT#-1	0.907211
meanT#-3	0.918951
minT#-1	0.928184
meanT#-2	0.931690
meanT#-1	0.961724
meanT	1.000000

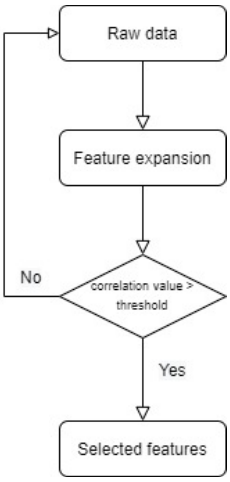

Fig 3. Proposed algorithm for expanding and selecting features

The equation (1) is then rewritten into general form as following:

$$\hat{y} = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + \cdots + \beta_n f_n \tag{3}$$

where β_0 is regression constant, and $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients to be determined from the selected variables as inputs $f_1, f_2, \ldots, f_n$.

B. Modelling Analysis

In general, linear regression finds the smallest residuals that is possible for the dataset and the most common method to measure closeness is to minimize the residual sum of squares (*rss*). Generally, the difference between the true and the predicted value are presented j^{th} residual, $\epsilon_j = y_j - \hat{y_j}$. We define the residual sum of squares as:

$$rss = \sum_{j=1}^{n} \epsilon_j^2 \tag{4}$$

where $\epsilon_j \ (j = 1,2,\ldots,n)$ a vector of residual terms. The equation (4) is equivalent as:

$$rss = \sum_{j=1}^{n} (y - X\beta)^T (y - X\beta) \tag{5}$$

where X is data matrix with an extra column of ones on the left to account for the intercept, y = $(y_1, \ldots, y_n)^T$, and β = $(\beta_0, \ldots, \beta_n)^T$. The parameters are shown in equations

(5).

$$y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23}. \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{pmatrix}$$

IV. EVALUATION RESULTS

A. Data simulation

We use our proposed concept of smart indoor farms for agriculture to collect data set for above 6 variables. The sensor nodes collect the 500 samples in which each sample is delivered in every ten minutes from 2PM to 3PM from April 2019 to July 2020. The raw data then is forwarded directly to firebase database via IEEE 802.11n/g wireless channel integrated in nodeMCU [13]. According to the raw data, the proposed the algorithm for expanding and selecting features extracts to get new data set including 9 features in table 1. The specific features are used as inputs for multiple regression model.

Because we try to find out the maximum numbers of features that have strong positive relationship to outcome, thus correlation value of i^{th} feature is larger than 0.7 [5]. The images illustrate what the relationships might look like at different degrees of strength are shown in the fig. 4, outcome describes very good positive linear relationships with selected features such as *minT#-2, maxT#-1,* and *maxT#-3.*

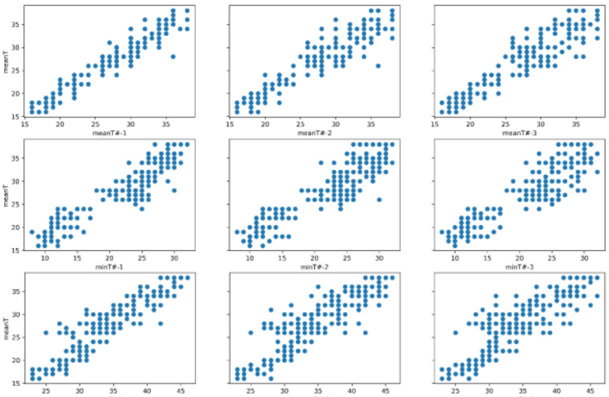

Fig. 4 Correlation between selected features and outcome

B. Evaluation and Discussions

In order to evaluate our proposal, we use statistical tests to evaluate the significance of the features [21]. In the work, we choose significant level (α = 0.05) for statistical tests to remove the features in new data set in the final equation. The regression summary consists of two tables. The first one is table 2, it presents the R-squared values for 9 selected features as the original model and 3 selected features as the final model in tables 3.

Table 2. Model summary of OLS Regression

3 selected features				**9 selected features**			
Dep. Variable:	T	R-squared:	0.934	Dep. Variable:	T	R-squared:	0.939
Model:	OLS	Adj.R-squared:	0.932	Model:	OLS	Adj.R-squared:	0.935

The 3 features in table 2 is selected because their *P* value (P>|t|) is smaller than significant level (α = 0.05). Because R-squared in 3 selected features (0.934) is very close to R-squared in 9 selected features (0.939) while their features are quite different. From this point, we can select the 3 selected features instead of 9 selected features in the final equation of model. By this way, the final model can support simple decision making because it deals with smaller features.

Table 3 presents the coefficients of the intercept and the constant for multiple regression. In addition, the other coefficients such as standard error (std err), t statistic, *P* value, confident interval are shown. Standard error refers to standard deviation and tell us how accurate the mean of any given sample from population, *t* statistic is given by the ratio of the coefficient of the predictor variable of interest, and its corresponding standard error. The confidence interval is the range of values that we would expect to find the features of interest. Thus, smaller confidence interval, the higher chance of accuracy.

Table 3. The coefficients of OLS Regression

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.6373	0.714	0.893	0.373	-0.769	2.044
meanT#-1	-0.1200	0.262	-0.458	0.647	-0.636	0.396
meanT#-2	0.4497	0.264	1.706	0.089	-0.069	0.969
meanT#-3	0.1298	0.265	0.490	0.625	-0.393	0.652
minT#-1	0.5075	0.143	3.556	0.000	0.226	0.789
minT#-2	-0.2670	0.149	-1.789	0.075	-0.561	0.027
minT#-3	0.0302	0.145	0.208	0.835	-0.256	0.316
maxT#-1	0.5654	0.143	3.963	0.000	0.284	0.846
maxT#-2	-0.3967	0.150	-2.643	0.009	-0.692	-0.101
maxT#-3	0.0798	0.146	0.546	0.586	-0.208	0.368

C. Decision making equation

By removing unnecessary features if P value (P>|t|) of the features is larger than 0.05. Then, minT#-1, maxT#1, and maxT#-2 are chosen for the final equation of decision model and the other features are removed. Therefore, the relationship between outcome and features now can be modelled in equation (5) as follows:

$$T = 0.6373 + 0.5075*(minT\#\text{-}1) + 0.5654*(maxT\#1) - 0.3967*(maxT\#\text{-}2) \quad (5)$$

From the equation (5), if the output *T* will increase one unit, then the dependent inputs is expected to increase/decrease a unit corresponding to their coefficients. On the other hand, we can estimate *T* if we know the values of above collected independent variables. Because we have selected 3 features, the final decisions just only depend on the features. By this way, the model not only make final decision simply and efficiently but also remain good fit.

V. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we proposed a robust regression model for simple decision making based on optimal feature sets for simple decision making in smart indoor farms. As result outcome in our proposed model performs wells with decision making and easy of computation because the model is straightforward to interpret small but strong correlation with outcome.

The future work will implement scalability and online setting for making predictions and evaluate our model with a variety of metrics will be investigated and analyzed. Moreover, we try to find out the ways to optimal our final decisions that not only select strong positive correlation but also gather strong negative correlation among features. By this way, we can provide making decision solutions for both positive and negative relationships.

REFERENCES

[1] B. ÖhlméYr, K. Olson, and B. J. A. e. Brehmer, "Understanding farmers' decision making processes and improving managerial assistance," vol. 18, no. 3, pp. 273-290, 1998.

[2] C. Akinbile, G. Akinlade, A. J. J. o. W. Abolude, and C. Change, "Trend analysis in climatic variables and impacts on rice yield in Nigeria," vol. 6, no. 3, pp. 534-543, 2015.

[3] T. Popović *et al.*, "Architecting an IoT-enabled platform for precision agriculture and ecological monitoring: A case study," vol. 140, pp. 255-265, 2017.

[4] J. Gubbi, R. Buyya, S. Marusic, and M. J. F. g. c. s. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," vol. 29, no. 7, pp. 1645-1660, 2013.

[5] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013.

[6] *Smart farming at UCLAB*. Available: https://prediction-sys.firebaseapp.com/

[7] S. Nguyen-Xuan and N. L. Nhat, "A dynamic model for temperature prediction in glass greenhouse," in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, 2019, pp. 274-278: IEEE.

[8] A. A. J. I. J. o. K.-b. Jalal and I. E. Systems, "Big data and intelligent software systems," vol. 22, no. 3, pp. 177-193, 2018.

[9] A. Glória, C. Dionísio, G. Simões, J. Cardoso, and P. J. S. Sebastião, "Water Management for Sustainable Irrigation Systems Using Internet-of-Things," vol. 20, no. 5, p. 1402, 2020.

[10] B. King and K. J. A. w. m. Shellie, "Evaluation of neural network modeling to predict non-water-stressed leaf temperature in wine grape for calculation of crop water stress index," vol. 167, pp. 38-52, 2016.

[11] J. Muangprathub *et al.*, "IoT and agriculture data analysis for smart farm," vol. 156, pp. 467-474, 2019.

[12] Technical Specification of DHT22 [Online]. Available: https://www.sparkfun.com/datasheets/Sensors/Temperature/DHT22.pdf

[13] NodeMCU [Online]. Available: https://www.nodemcu.com/index_en.html

[14] M. Gocić *et al.*, "Soft computing approaches for forecasting reference evapotranspiration," vol. 113, pp. 164-173, 2015.

[15] A. Ganguly, S. J. E. Ghosh, and Buildings, "Model development and experimental validation of a floriculture greenhouse under natural ventilation," vol. 41, no. 5, pp. 521-527, 2009.

[16] B. T. Nguyen and T. L. J. R. E. Pryor, "The relationship between global solar radiation and sunshine duration in Vietnam," vol. 11, no. 1, pp. 47-60, 1997.

[17] E. Symeonaki, K. Arvanitis, and D. J. A. S. Piromalis, "A Context-Aware Middleware Cloud Approach for Integrating Precision Farming Facilities into the IoT toward Agriculture 4.0," vol. 10, no. 3, p. 813, 2020.

[18] N. Kaewmard and S. Saiyod, "Sensor data collection and irrigation control on vegetable crop using smart phone and wireless sensor networks for smart farm," in *2014 IEEE Conference on Wireless Sensors (ICWiSE)*, 2014, pp. 106-112: IEEE.

[19] H. Navarro-Hellín, J. Martínez-del-Rincon, R. Domingo-Miguel, F. Soto-Valles, R. J. C. Torres-Sánchez, and E. i. Agriculture, "A decision support system for managing irrigation in agriculture," vol. 124, pp. 121-131, 2016.

[20] M. Robert, A. Thomas, and J.-E. J. A. f. s. d. Bergez, "Processes of adaptation in farm decision-making models. A review," vol. 36, no. 4, p. 64, 2016.

[21] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *CVPR 2011*, 2011, pp. 785-792: IEEE.

MÔ HÌNH HỒI QUI ĐA BIẾN TĂNG CƯỜNG DỰA TRÊN TẬP TỐI ƯU ĐẶC TRƯNG ỨNG DỤNG CHO VIỆC RA QUYẾT ĐỊNH HIỆU QUẢ TRONG TRANG TRẠI NÔNG NGHIỆP

Tóm tắt: Bài báo này đã đề xuất giảm số biến độc lập trong mô hình hồi quy đa biến để đơn giản việc ra quyết định trong các trang trại thông minh. Trong đề xuất của chúng tôi, có một số bước để đảm bảo tập dữ liệu chuỗi thời gian được thu thập từ các nút cảm biến trong các trang trại thông minh được mở rộng. Dựa trên tập dữ liệu mở rộng này, các biến có hệ số tương quan mạnh với đầu ra sẽ được dùng cho mô hình hồi quy đa biến. Sau đó, chúng tôi sử dụng phương pháp thống kê để rút gọn các biến trong phương trình cuối cùng. Kết quả mô phỏng cho thấy giá trị R-squared của mô hình cuối cùng gần giống với giá trị R-squared của mô hình gốc trong khi kết quả trong phương trình cuối cùng chỉ phụ thuộc vào các có số biến ít hơn. Kết quả cho thấy rằng đề xuất của chúng tôi có thể đưa ra các quyết định được đơn giản hóa trong ứng dụng thực tế trong nông nghiệp.

Keywords: hồi qui đa biến (MR), trang trại thông minh (SIF), tập tối ưu đặc trưng (OFS), ra quyết định hiệu quả (SDM).

NGUYEN XUAN SAM received the B.Eng degree in Communications Engineering from Posts and Telecoms Institute of Technology (PTIT), Hanoi, Vietnam in 2002, the M.Sc. degree in Information and Communications Engineering from the Andong National University, and the Doctor degree in Computer Engineering from Korea University (Seoul campus), Republic of Korea in 2009 and 2016, respectively. His research interests include the distributed computing, real-time embedded systems, artificial intelligence for Internet of Things.

NGUYEN NGOC GIANG received the Doctor degree in Math Education from The Vietnam Institute of Educational Science, Hanoi city, Vietnam in 2017, respectively. His research interests include machine learning and deep learning.