ICIC Express Letters
Part B: Applications
Volume **12**, Number **8**, August **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 757–764

BIG DATA ANALYSIS USING RAPIDMINER STUDIO TO PREDICT SUICIDE RATE IN SEVERAL COUNTRIES

Evaristus Didik Madyatmadja[1], Samuel Imanuel Jordan[2]
and Johanes Fernandes Andry[2]

[1]Information Systems Department
School of Information Systems
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
emadyatmadja@binus.edu

[2]Information Systems Department
Faculty of Technology and Design
Bunda Mulia University
Jl. Lodan Raya No. 2, Ancol, Jakarta Utara 14430, Indonesia
samueljordan322@gmail.com; jandry@bundamulia.ac.id

Received November 2020; accepted February 2021

Abstract. *Suicide may be a growing public health concern with a worldwide prevalence of roughly 800,000 deaths per annum. The present process of evaluating suicide risk is very subjective, which may limit the efficacy and accuracy of prediction efforts. During this research, authors are going to take advantage of data mining in optimizing suicide risk prediction in each country occurring, and each generation which has committed suicide in each country. Prediction analytics in big data helps to spot individuals in crisis to intervene with emotional support, crisis and psych educational resources, and alerts for emergency help. Data mining and machine learning have additionally been used to support the clinical management of suicide across medicine and analysis, medication management, and the activity of medical care delivery. Data mining involves exploring and analyzing large amounts of knowledge to seek out patterns for giant data. RapidMiner is one of the well-known data mining tools and is used for data mining. With RapidMiner, authors can compare datasets containing socio-economic information at a level that is appropriate to the year and country. For analyzing the suicide data, the methods that can be used such as deep learning, decision tree for modeling the data, and KNN to classify the data attributes into groups so it is easier to analyze. The authors use accuracy and Kappa result produced by methods to show how good the data is for prediction and the prediction number will be produced for further analysis.*

Keywords: Predict, Big Data Analytics, Data mining, RapidMiner, Suicide

1. **Introduction.** Big data is driving radical changes in traditional and conventional data analysis platforms. Current technological developments such as big data technology have become an important role in helping society, especially suicides which are increasingly occurring [1]. Performing any evaluation and assessment on such voluminous and complicated data, scaling up the hardware platforms, becomes imminent and choosing the proper hardware or software program platforms becomes an important decision if the user's requirements are to be satisfied in a reasonable amount of time [2]. There are numerous big data platforms available with different characteristics and selecting the proper platform requires in-depth knowledge about the capabilities of these platforms [3]. The RapidMiner is one of big data tools that has many advantages, a number of which are that it is provided freely as an open source software, but it is provided also under a commercial license suitable for closed-source commercial applications development, experienced free

DOI: 10.24507/icicelb.12.08.757

community but also a highly professional paid support is also available out there, mature and growing user and developer community, the value of development with RapidMiner is comparatively low in comparison with other data mining solutions [4].

In response to the issues of analyzing a very large scale of data, quite a few efficient methods [5], as well as sampling, data condensation, density-based approaches, grid-based approaches, divide and conquer, progressive learning, and distributed computing, have been presented. Of course, these methods are constantly used to improve the performance of the operators of data analytics process [6]. Big Data Analytics poses other particular and unique challenges for machine learning and data analysis, including format variation of the data, the highly distributed input sources, the very fast moving streaming data, the trustworthiness of the data analysis, the noisy and poor quality data, high dimensionality, scalability of algorithms, unbalanced input data, unsupervised and un-categorized data, where the supervised/labelled data is limited, etc. [7]. Nowadays, there are scientific research and commercial literature [8]. In the area of healthcare industry it traditionally has generated big quantities of data, driven by record keeping, compliance and regulatory requirements, and patient care [9]. Suicide is the 17th leading cause of death in the world and the second leading cause of death among 15-29 years old people. The suicide rate increases from year to year and a psychological autopsy studies find that up to 90% of people who died by suicide in Western countries met criteria for a mental disorder [10]. The use of traditional statistical methods of analyzing suicides rates, which are limited in analyzing complex data, has failed to predict suicide behaviours above chance levels. Given these current limitations, the statistically advanced machine learning (ML) approaches, as a sub-field of artificial intelligence (AI), are being developed with increasing frequency to improve suicide care [11].

In this paper, authors are going to examine the suicide data from various countries in the world. The objective is to analyze the suicide rate and the results of this study are expected to make a positive contribution to the mechanism for preventing suicide and reducing mortality in a country. Also, the results can be used as a reference for further research in the field of Big Data Analytics for predicting studies.

2. **Methodology.** The methodology is used by the authors to analyze, work on, and solve the problems at hand. The theoretical framework or scientific framework is the scientific methods that will be applied in conducting research. The research process in this paper is studying the literature, collecting data, processing data, analyzing data, and making reports.

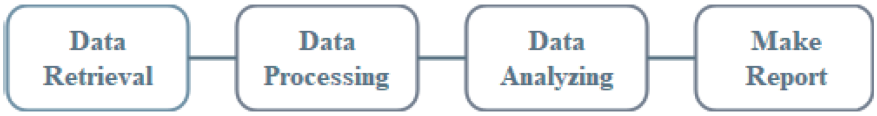


Figure 1. Research process

2.1. **Research process.** Based on Figure 1, it is known that this research starts from:

1) Data retrieval, data collection on suicide rates by year and country to be processed and analyzed using secondary data from Kaggle website;
2) Data processing, the data that has been taken is then processed by entering the dataset from Kaggle into the RapidMiner Studio application to take it to the next stage, namely the data analysis stage;
3) Data analyzing, at this stage the data is analyzed in the RapidMiner Studio application. First, the data is analyzed using the advanced chart feature to view data in graphical form based on predefined labels. Second, data analysis is carried out by making a decision tree to display or modeling the data. The third is analyzing the data by

making a decision tree classification model and using the applying model operator to display the prediction and confidence of the data. In the fourth stage, the data is analyzed by testing the model using the applying model operator. Then the last stage of data analysis is to filter the data not missing value, cross validation, and its use;

4) Make a report, this stage is carried out after the data has been analyzed. The results of the analysis are made into a research journal containing abstracts, introduction, research methods, results and discussions, conclusions, and bibliography based on data that has been taken, processed, analyzed, discussed, and obtained from the research.

2.2. **Method of collecting data.** The method of data collection used in this research is literature study. Literature study is an activity to retrieve information in accordance with the topic or problem that is the object of research. This piece of information can be obtained from scientific works, encyclopaedias, the Internet, and other sources that can support the results of this research. The data is obtained from open datasets, namely from the Kaggle Internet site which provides data that can be retrieved and processed to be the result of the analysis of this research as, can be seen in Table 1.

Table 1. Data overview

No.	Attribute	Description	No.	Attribute	Description
1	Country	The country where the suicide occurs	7	Suicides/100k pop	Suicide rate per 100,000 population in every country
2	Year	The year the suicide occurs	8	Country-year	Combined column between the country column and the year column
3	Sex	The gender of the suicide victim	9	HDI for year	Comparison of life expectancy per year
4	Age	Age range for suicide victims	10	GDP for year	National income per year in a country
5	Suicide no.	Number of people who commit suicide in each country	11	GDP per capita	National income per capita
6	Population	Population in each country	12	Generation	Generation of residents in a country

2.3. **Data analysis technique.** RapidMiner Studio uses predictive data analysis techniques and descriptive data analysis techniques in providing knowledge to each user so that they can make excellent decisions. Statistical techniques are also used in RapidMiner. The processing of the data uses descriptive statistics or inferential statistics. Descriptive statistics are statistics that are used to analyze data by making a description of the data which has been combined into one as is without the intention of making general conclusions.

Prediction analysis is the process in which outcome will be predicted based on current data. Predictive analytics indicates a focus on making predictions. The main alternative to predictive analytics can be called descriptive analytics where this area is often also called “knowledge discovery in data” or KDD. Decision tree is a one of prediction models that uses a tree structure or hierarchical structure. Decision tree combines data search and modelling, so it is a good model as the first step in the modelling process and when used as the final model of various other techniques. Besides decision tree, authors also used KNN as classification method in this paper because classification method usually produces high quality models [12].

3. **Analysis and Discussion.** As previously discussed in the Research Methods chapter, in this chapter authors will talk about the implementation of retrieving, processing, and

analyzing the data that has been taken from Kaggle about suicide rates overview from 1985 to 2016 with several data mining techniques found in the RapidMiner Studio application.

3.1. **Data retrieval.** First of all, before carrying out the data processing, authors need to retrieve the data authors need. The data which authors use is from Kaggle, where the following table contains the attributes and description of the data sets that will be used in this study.

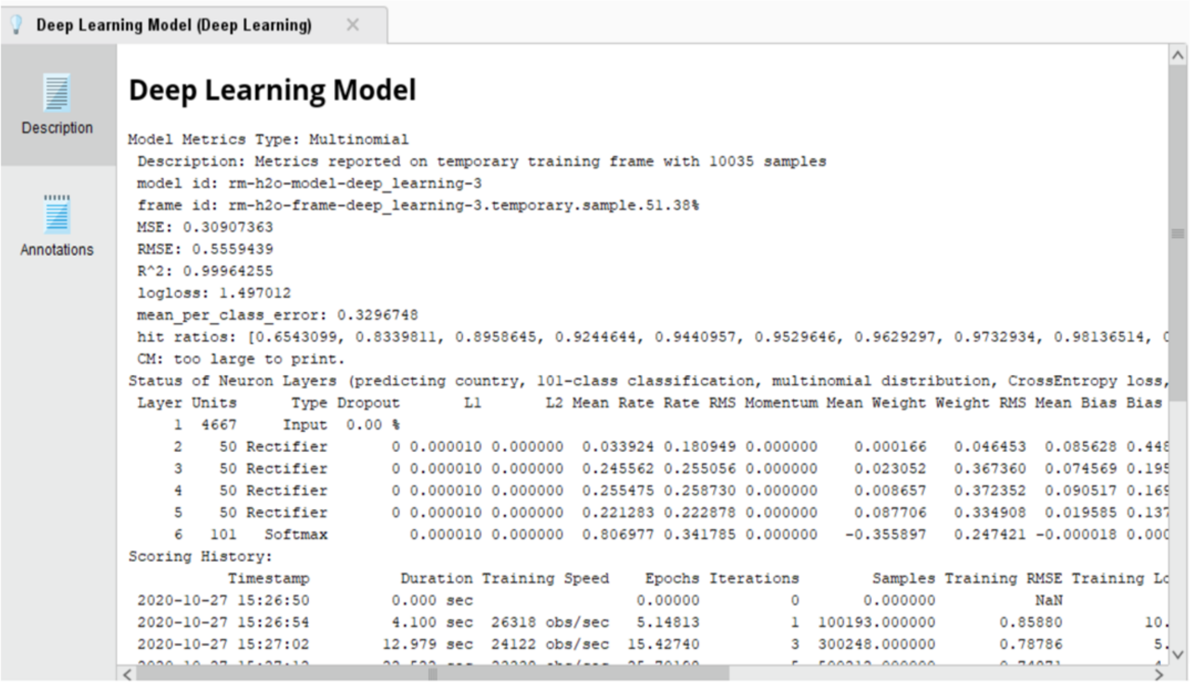
3.2. **Data pre-processing.** Before authors analyze the data, authors need to pre-process the data. The data pre-processing technique is a data mining technique which is used to transform and rearrange the raw data in our dataset into a useful and efficient format. Data pre-processing technique in the data mining process involves the remodelling and turns the raw data into a visible, useful, and efficient format. Real-world data is usually incomplete, inconsistent, lacking inbound behaviours or trends, and is probably going to contain several errors. Data pre-processing is an established methodology for breakdown of such issues.

There is a missing value in attribute HDI for year with a total of 19456 missing. Missing data, or missing values, occur when no data value is held or stored on the variable in an observation. Missing data are a typical prevalence and might have a major impact on the conclusions that may be drawn from the data. In order to get rid of the missing value, authors need to filter out those missing value data. First, the authors filter out the missing value by using the Replace Missing Value operator where this operator replaces missing values in dataset of selected attributes by a specified replacement which authors replaced those missing values with the average value of the attribute. The Reorder Attributes operator is used to reorder regular attributes of a dataset. Reordering can be done alphabetically, by user specification or with a reference dataset. Select Attribute operator is used to select a subset of attributes of a dataset. Last is Normalize operator to normalize the value of the selected attributes. Authors replace the HDI for year attribute missing value with the average value of the attribute. After there is no more any missing value in the dataset, authors can analyze the data and make a more accurate conclusion from the dataset.

3.3. **Data analyzing.** Data analyzing is outlined as a method of cleaning, transforming, and modelling data to get helpful data for business decision-making.

One of the methods authors are using in predicting the outcome is using deep learning method. Deep learning relies on a multi-layer feed-forward artificial neural network that is trained with random gradient descent using back-propagation. From Figure 2, the deep learning model that has been created by RapidMiner shows that the model metrics type used was multinomial distribution where the multinomial distribution is a type of distribution probability to calculate the outcomes of experiments involving two or a lot of variables. With the deep learning method, authors test and train the data authors used and get the result in accuracy and Kappa. Therefore, authors know how accurate our data is by knowing the accuracy percentage. As for the Kappa performance it is statistically used to calculate and measure the inter-rater reliability for qualitative (categorical) items and the statistical analysis is based on the suitability of interpretation or degree of agreement.

The authors used the deep learning operators in RapidMiner and set the epochs to 200, which means we train the deep learning model 200 times. Also, we set 4 hidden layers with each layer size as 50. Therefore, we can get the accuracy performance of the data processed by deep learning which is 64.69%. It means that the dataset that has been processed by the deep learning is neither bad nor good and on average. The Kappa performance for the deep learning model has the result 0.643. The Kappa performance here is quite good and shows that authors could get on a more accurate performance



Deep Learning Model

```
Model Metrics Type: Multinomial
 Description: Metrics reported on temporary training frame with 10035 samples
 model id: rm-h2o-model-deep_learning-3
 frame id: rm-h2o-frame-deep_learning-3.temporary.sample.51.38%
 MSE: 0.30907363
 RMSE: 0.5559439
 R^2: 0.99964255
 logloss: 1.497012
 mean_per_class_error: 0.3296748
 hit ratios: [0.6543099, 0.8339811, 0.8958645, 0.9244644, 0.9440957, 0.9529646, 0.9629297, 0.9732934, 0.98136514, 0
 CM: too large to print.
Status of Neuron Layers (predicting country, 101-class classification, multinomial distribution, CrossEntropy loss,
 Layer Units      Type Dropout       L1       L2 Mean Rate Rate RMS Momentum Mean Weight Weight RMS Mean Bias Bias
     1  4667     Input  0.00 %
     2    50 Rectifier       0 0.000010 0.000000  0.033924 0.180949 0.000000    0.000166   0.046453  0.085628 0.448
     3    50 Rectifier       0 0.000010 0.000000  0.245562 0.255056 0.000000    0.023052   0.367360  0.074569 0.195
     4    50 Rectifier       0 0.000010 0.000000  0.255475 0.258730 0.000000    0.008657   0.372352  0.090517 0.169
     5    50 Rectifier       0 0.000010 0.000000  0.221283 0.222878 0.000000    0.087706   0.334908  0.019585 0.137
     6   101   Softmax         0.000010 0.000000  0.806977 0.341785 0.000000   -0.355897   0.247421 -0.000018 0.000
Scoring History:
           Timestamp          Duration Training Speed    Epochs Iterations        Samples Training RMSE Training Lo
 2020-10-27 15:26:50         0.000 sec                  0.00000          0       0.000000           NaN
 2020-10-27 15:26:54         4.100 sec  26318 obs/sec   5.14813          1  100193.000000       0.85880          10.
 2020-10-27 15:27:02        12.979 sec  24122 obs/sec  15.42740          3  300248.000000       0.78786           5.
```

Figure 2. Deep learning model

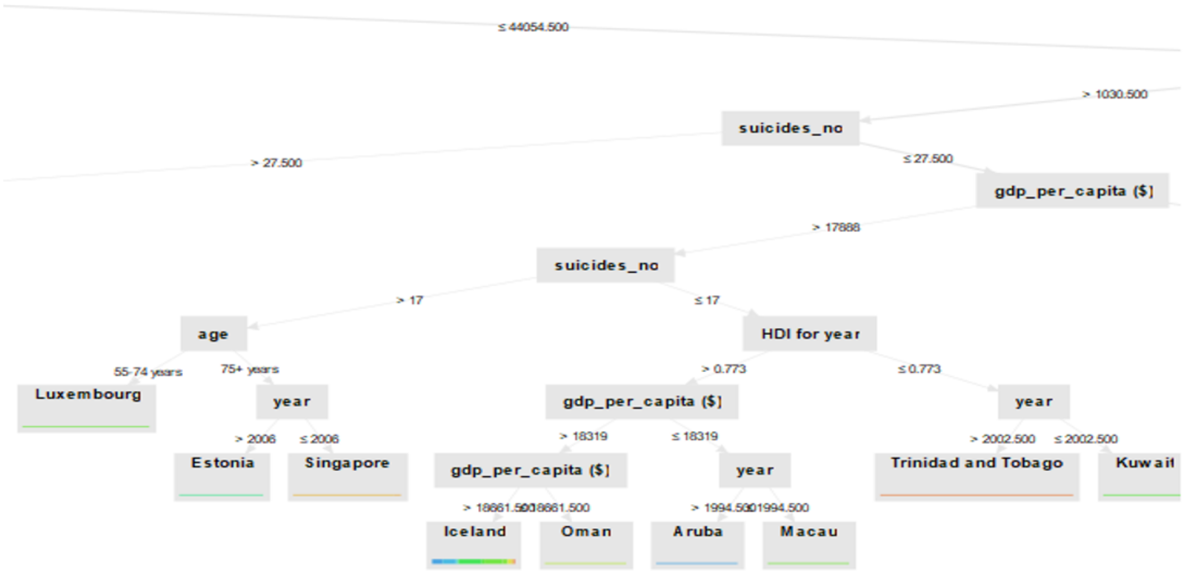


Figure 3. Decision tree

evaluation. Besides deep learning method, decision tree also been used to build the data predict model of the dataset. Decision tree is a structured branching flowchart. Each branch represents a test performed on an attribute. The branches represent the results of the tests and the leaves represent the class labels. There are four steps in the decision tree making process in the decision tree algorithm, namely

a) Select attributes as root/root;

b) Create a branch for each value;

c) Divide each case into a branch;

d) Iterate the process in each branch so that all cases in the branch have the same class.

Figure 3 shows a decision tree generated from the Kaggle data and has many branches and leaves which are divided into several classes so that decisions can be made. Based on the decision tree image, the root is owned by the suicide_no (suicide number), while the branches are owned by gdp_per_capita, age, and HDI for year then the leaves are owned by the country. The decision tree above is a support tool that produces a tree-like model of decisions and their potential consequences, where the model of decision and potential consequences together combine with occurrence outcomes, resource costs, and utility. Model of decision making is an interactive computer information system that can give alternate solution for the decision makers [13]. From the decision tree, authors can get the decision tree rules for prediction purposes. From the decision tree model, authors will know the decision tree rules produced by the model. The rules were made by the decision tree model of the dataset. The decision tree is linearized into decision rules, where the result is the contents of the leaf node, and therefore the conditions on the path form conjunction within the if clause. In general, the principles of the decision tree rules have the form of: if condition1 and condition2 and condition3 then the outcome. Decision tree rules are usually generated by constructing association rules with the target variable on the right where they will also denote temporal or causal relations.

KNN or K-Nearest Neighbours is one of the supervised learning algorithms that is widely used for classification problems in machine learning. In KNN method, authors classified the attribute of generation. Authors set the K to 5 for the nearest model of classification. We classified this attribute so we can get a more accurate predictive outcome. KNN works by finding the distances between a query and all of the examples within the data, choosing the desired number of examples (K) closest to the query, then votes for the foremost frequent label (in the case of classification) or averages the labels (in the case of regression).

In the case of classification authors are facing, authors have a tendency to try to select the correct K for our data which is completed by trying several Ks and choosing the one that works best where authors choose K equal to 5 and it works very fine and best with our dataset.

Table 2 shows that the accuracy of the KNN method is 100% with the micro average as 100% which means the accuracy performance of the method is incredibly good. The prediction of each generation reaches 100%, also the class precision of each generation reaches 100%; therefore, this method could predict effectively, accurately and very recommended using this method to get a very reliable output prediction. As for the Kappa performance, it has the same result as the accuracy performance, which is 1 meaning that the data is reliable and has a very good level of agreement. The generation attribute is very reliable to the country; therefore, the Kappa value is 1. By combining the attribute country and generation, authors can predict which country and generation are most likely to have suicide in their country in the future. Each generation column value represents the total data sample of each generation in the dataset which if it is combined together the total sample is 27820. With the result of the KNN method, it means the attribute

Table 2. KNN accuracy performance

	True Gen X	True Silent	True G.I. Gen	True Boomers	True Millennials	True Gen Z	Class precision
Pred. Gen X	6408	0	0	0	0	0	100%
Pred. Silent	0	6364	0	0	0	0	100%
Pred. G.I. Gen	0	0	2744	0	0	0	100%
Pred. Boomers	0	0	0	4990	0	0	100%
Pred. Millennials	0	0	0	0	5844	0	100%
Pred. Gen Z	0	0	0	0	0	1470	100%
Class recall	100%	100%	100%	100%	100%	100%	100%

of generation can be predicted accurately where the class precision reaching 100% means there are hardly any false or wrong predictions that will happen in the future.

4. Conclusions.

Big Data Analytics can be used or potentially to gain insight and make informed decisions. Authors have a way of knowing how much accuracy in predicting suicide will be improved by applying analytical features and enhancements. Based on the analysis and discussion results in the previous section, there is a missing data in the dataset which is typical prevalence and may have a major impact on the conclusions that may be drawn from the data. Authors filter the missing values using the Replace Missing Value operator where this operator replaces the missing values with the mean of the attributes. The accuracy of the data obtained is 64.69%, which means that the dataset that has been processed by deep learning is not bad or good and on average. Meanwhile, the Kappa performance of deep learning method is 0.643 which is fairly good and shows that authors can get a more accurate performance evaluation. In addition to deep learning methods, decision trees are also used to build predictive models for data from these datasets. From the decision tree, authors know the decision rules that are usually created by creating association rules with the target variable on the right where the rule will also show a temporal or causal relationship. Finally, authors also use KNN or K-Nearest Neighbours which are supervised learning algorithms so that authors can get more accurate prediction results. Authors also get 100% accuracy of the KNN method with an average of 100% micro, which means that the accuracy of the method is incredibly good and has a perfect performance. The prediction of each generation reaches 100%, and the class precision of each generation reaches 100%; therefore, this method can predict effectively, accurately and it is highly recommended to use these methods to get a reliable prediction result.

REFERENCES

[1] E. D. Madyatmadja, J. F. Andry and A. Chandra, Blueprint enterprise architecture in distribution company using TOGAF, *J. Theor. Appl. Inf. Technol.*, vol.98, no.12, pp.2006-2016, 2020.

[2] D. Singh and C. K. Reddy, A survey on platforms for big data analytics, *J. Big Data*, vol.2, no.1, pp.1-20, DOI: 10.1186/s40537-014-0008-6, 2015.

[3] V. S. Agneeswaran, P. Tonpay and J. Tiwary, Paradigms for realizing machine learning algorithms, *Big Data*, vol.1, no.4, pp.207-214, DOI: 10.1089/big.2013.0006, 2013.

[4] R. Burget, J. Karásek, Z. Smékal, V. Uher and O. Dostál, RapidMiner image processing extension: A platform for collaborative research, *TSP 2010 – The 33rd Int. Conf. Telecommun. Signal Process.*, pp.114-118, 2010.

[5] D. Xu et al., Solution-based evolution and enhanced methanol oxidation activity of monodisperse platinum-copper nanocubes, *Angew. Chemie – Int. Ed.*, vol.48, no.23, pp.4217-4221, DOI: 10.1002/anie.200900293, 2009.

[6] C. W. Tsai, C. F. Lai, H. C. Chao and A. V. Vasilakos, Big data analytics: A survey, *J. Big Data*, vol.2, no.1, pp.1-32, DOI: 10.1186/s40537-015-0030-3, 2015.

[7] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, Deep learning applications and challenges in big data analytics, *J. Big Data*, vol.2, no.1, pp.1-21, DOI: 10.1186/s40537-014-0007-7, 2015.

[8] F. Soleimani-Roozbahani, A. R. Ghatari and R. Radfar, Knowledge discovery from a more than a decade studies on healthcare big data systems: A scientometrics study, *J. Big Data*, vol.6, no.1, DOI: 10.1186/s40537-018-0167-y, 2019.

[9] N. Savage, Digging for drug facts, *Commun. ACM*, vol.55, no.10, pp.11-13, DOI: 10.1145/2347736.2347741, 2012.

[10] R. C. Kessler et al., The role of big data analytics in predicting suicide, *Pers. Psychiatry Big Data Anal. Ment. Heal.*, pp.77-98, DOI: 10.1007/978-3-030-03553-2_5, 2019.

[11] T. M. Fonseka, V. Bhat and S. H. Kennedy, The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors, *Aust. N. Z. J. Psychiatry*, vol.53, no.10, pp.954-964, DOI: 10.1177/0004867419864428, 2019.

[12] Y. Omae, M. Mori, T. Akiduki and H. Takahashi, A novel deep learning optimization algorithm for human motions anomaly detection, *International Journal of Innovative Computing, Information and Control*, vol.15, no.1, pp.199-208, 2019.

[13] E. D. Madyatmadja, Decision support system model to assist management consultant in determining the physical infrastructure fund, *Journal of Theoretical and Applied Information Technology*, vol.62, no.1, pp.269-274, 2014.