Journal of Physics: Conference Series

RapidMiner and Machine Learning Techniques for Classifying Aircraft Data

RapidMiner and Machine Learning Techniques for Classifying Aircraft Data

Syahaneim Marzukhi[1], Norfatimah Awang[2], Syed Nasir Alsagoff[2] and Hassan Mohamed[1]

[1]Pusat Keselamatan Siber, Universiti Pertahanan Nasional Malaysia (UPNM), Kem Sg. Besi, 5700 Kuala Lumpur
[2]Jabatan Sains Koputer, Fakulti Sains dan Teknologi Pertahanan, Universiti Pertahanan Nasional Malaysia (UPNM), Kem Sg. Besi, 5700 Kuala Lumpur

syahaneim@upnm.edu.my

Abstract. Machine learning is an important technique that helps companies, organizations and individuals to improve the quality of decision making. In today scenario, especially with the emerged of data science, it can see how machine learning techniques are used for data analytics. There are various machine learning techniques for data science tasks that can be categorized as follows: classification, prediction, regression, association analysis, clustering, time series forecasting, and many others. As there are many different free tools available for machine learning, the selection of the appropriate analysis technique is crucial to solve problem in hand. This study compares the performance of machine learning algorithms especially Naïve Bayes, Decision Tree, Random Forest and ID3 for classification task (i.e. classifying aircraft to certain category and into country of origin) using RapidMiner tool. Those algorithms are compared based on their accuracy rate, error rates, precision and recall for classifying aircraft. The results reveal that that Random Forest and ID3 algorithms given good classification accuracy due to the nature of the algorithms that is progressively improved apart from Decision Tree.

1. Introduction

Technology changes the way on how people make decisions and build workflows. For instance, the use of data science for business application by adapting machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics are widely used [1]. Nevertheless, data science is useful for various aspects of airline operation management. Air traffic control (ATC) is a service provided by ground-based air traffic controllers to direct aircraft on the ground and also control the airspace. Mainly, the function of ATC are: 1) preventing collisions, 2) managing the traffic flow, and 3) supporting information related to aircraft. In most countries, ATC monitors the location of aircraft in it allotted airspace and provides services to private, military, and commercial aircraft. While, in some countries, ATC plays a security or defensive role to monitor aircraft in their airspace.

With continuous growth of the aviation industry, the traffic management poses a huge challenge. Statistical analysis alone is not sufficient to handle huge amount of air traffic activities and aircraft data, as the aircraft data is collected at every single second and archived in the database. Here, data science has capability of handling large volumes of data with multiple attributes through several processes start with data preparation, mining data, statistics, data visualisation and experimentation. Further, machine learning techniques is implemented to search for knowledge and patterns from the data. This research compares the performance of machine learning algorithms especially Naïve Bayes, Decision Tree,

Random Forest and ID3 for analysing the aircraft data using RapidMiner tool. Thus, the data can be represented in meaningful way for traffic management (i.e. classifying aircraft to certain category and into country of origin). In this paper, we also highlight that Random Forest and ID3 algorithms offer better performance in terms of accuracy compared to other algorithms that is available in RapidMiner. The rest of this paper is organized as follows. Section 2 provides research background related to the research. Section 3 presents the methodology for the research, Section 4 discusses findings and results, and Section 5 contains concluding remarks.

2. Research Background

2.1 Machine Learning for Classification

In general, machine learning algorithms can be classified based on its learning methods. Thus, machine learning algorithms can be categorised into: supervised, unsupervised, and semi-supervised learning. In supervised learning, the algorithms develop a predictive model based on the given input and output. The algorithms receive a training dataset that consists of a set of labelled input with its correct output [2]. Then, after the learning process using the training dataset, the algorithms will predict the value of the output of the unseen data or new data. Typical supervised learning techniques are: Decision Tree, Naive Bayes, Support Vector Machines, Artificial Neural Networks, K-Nearest Neighbours (K-NN) and many others. In contrast, in unsupervised learning, the algorithms receive only the input that is unlabelled. The algorithms need to group and interpret the unseen data based on only the input (i.e. clustering, dimensionality reduction) [2]. Here, the algorithms will discover patterns in the data in order to define the relationships between the inputs. Some examples of unsupervised learning algorithms include: K-Means, Hierarchical Clustering, and many others.

For this research, the supervised learning algorithms are explored for solving classification task. Classification is a task of predicting the unseen data or new data to a predefined class or label. The goal of classification is to accurately predict the target class or label for each case in the collection of data. Here, classification tasks are to predict the output that can be either categorical or polynomial. Classification task can be solved using various techniques of machine learning algorithms for instance Decision Tree, Artificial Neural Network, K-Nearest Neighbours (K-NN), and even some Regression algorithms.

Therefore, machine learning algorithm is applied in this research for solving the classification problem. The machine learning algorithm learns to recognize the hidden patterns during training and predict the new data to the correct class. However, only the following machine learning algorithms will be used for solving the classification problem: Naïve Bayes, Decision Tree, Random Forest and ID3. Each of the algorithm is discussed as follows.

Naïve-Bayes technique is based on Bayes' Theorem with an assumption of independence among predictors [3]. Naïve-Bayes classifier assumes that the presence of a particular predictors or features in a class is unrelated to the presence of any other predictors or features. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naïve-Bayes algorithm is built based on the probabilistic theorem. Naïve-Bayes algorithm will predict the target class or label by calculating the probability of the predicted class for each member of the test instance. The class with the highest posterior probability is the outcome of prediction.

Decision tree is a tree-structured classifier that organized a series of test questions and conditions in a tree structure. Decision tree comprises of leaf nodes, internal nodes and links. A tree node represents a class label or output, an internal node represents the predictor or feature, and the link from a parent node to a child node represents a rule or decision [4]. Decision tree is an algorithm for learning, where the decision tree classifiers organized a series of test questions and conditions in a tree structure. Decision tree comprises of leaf nodes, internal nodes and links. In general, the algorithm for the decision tree is implemented in 2 phases as follows [5]: (1) Tree building is the first phase, where the tree is divided until all the data have its class in a top-down fashion and (2) Tree pruning is the second phase, where predictions and accuracy are improved in a bottom-up fashion. Thus, the main goal of decision tree is to build an optimal decision tree for solving any classification problem.

Whereas, ID3 algorithm (Iterative Dichotomiser 3), is an algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy (H) [4]. While, Random Forest, consists of a large number of individual decision trees that operate as an ensemble, where each individual tree in the random forest will have a class prediction and the class with the most votes becomes the selected model's prediction.

2.2 Related Works

Based on current research [1], Cross-Industry Data Mining (CRISP-DM) is the most widely used model because of its advantages for solving existing problems in directing businesses and organizations towards making incredible profits. For example, experts and practitioners as in [10-12] using CRISP-DM to solve different types of (e.g. prediction and classification) in various domains and fields including science, business, healthcare, engineering and many others. CRISP-DM is chosen compared to other framework because it uses a non-rigid sequential framework that consists of a six-step phases.

With the increasing demand to analyse huge volume of data, thus, a tool to analyse those data is needed; especially data analytics tools that can produce accurate results with less effort. Some features that is provided by the data analytic tools are: data retrieval, data manipulation, data analysis and data visualization. Whereas, other features include generate reports and dashboard with better visualization. There are various data analytics tools available including RapidMiner, Weka, KNIME, R tool, Orange, and many others. Each of them is examined and assessed in [6-9], where the papers are aimed to evaluate the most popular open source and free data analytics tools among user, developer, and researcher. Furthermore, researchers as in [13-17], comparing various data analytic algorithms particularly machine learning algorithms namely; Naïve Bayes (NB), Decision Tree (DT), K Nearest Neighbour (KNN), Random Forest and many others for solving various classification problems. Most of the papers either used dataset that is available for public (i.e. UCI-repository) or different datasets from real-problem data in various fields in order to evaluate the performance.

In this paper, CRISP-DM Framework is used as a based-line in conducting the research. Next, the framework is implemented using the open source data analytics tool Rapidminer. RapidMiner is choosen as it is one from the most popular open source and free data analytic tools that is easy to use and has different graphical capabilities to present the results. Further, machine learning algorithms (i.e. Naïve Bayes, Decision Tree, Random Forest, ID3) are applied in RapidMiner for performing data analytics task specially for addressing classification problems (i.e. classifying type of aircraft and country of origin).

3. Research Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM) Framework is among the most widely used framework for solving data science problems [10]. In this paper [10-12], the author discussed on how this CRISP-DM Framework translating business problems into data mining tasks through executing data mining projects independently from the application area and the used technology. Figure 1 shows 6 phases of the model: (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation and (6) deployment and is described as the following [3].

1) Business understanding. In this phase, the research's goals, objectives and requirements will be investigated thoroughly. Next, those information is translated into a data science problem for further action. Here, any suitable data science technique and method will be identified for solving classification problems.
2) Data understanding. In this phase, other related information regarding the research will be collected and gathered. This information will be analysed in order to understand the data better especially the hidden pattern or unseen data.
3) Data preparation. As there is too many information has been gathered from previous phase, this phase only extracts the relevant data. The data will be enhanced in terms of its quality in order to prepare it for the next phase. Some of the process include the pre-processing task and the data exploration task. Normally, this phase will be executed many times without any specific order until the correct dataset is formed.

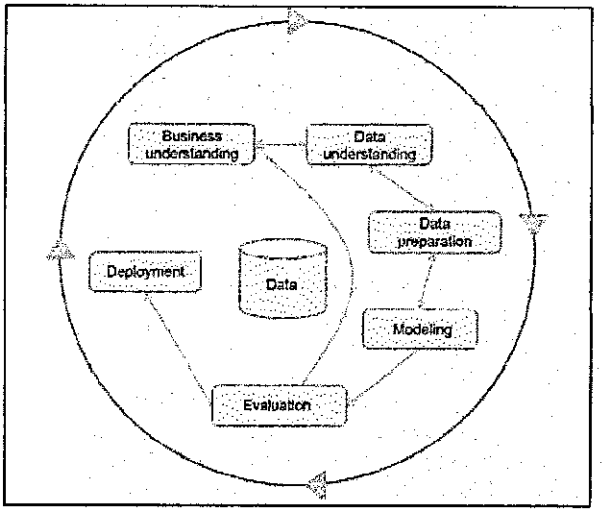

Figure 1. Cross Industry Standard Process for Data Mining (CRISP-DM) process [5].

4) Modelling. In this phase, various techniques from the fields of statistics, machine learning, operational research and many others are studied. Next, a suitable model is designed and applied to the dataset for solving classification problems. This phase is an iterative process as it involves selecting the variables for the model, executing the model, and diagnostic the model until the desired results is found.
5) Evaluation. In this phase, the model's performance is evaluated before it can be fully utilized. The model is also assessed against the research's goal and objectives to ensure that it served the purpose. The decision of selecting the model for solving the problem than is made.
6) Deployment. In this phase, the selected model is deployed. If the model does not fulfil the research's requirements in producing the desired results, thus going back to the data preparation phase and modelling phase is necessary.

4. Results and Discussion

4.1 Experimental Design

For this research, the dataset is downloaded from the Aircraft Monitoring System (AMS) that is located at the Cyber Security Center, National Defense University Malaysia. The system is used to monitor and record the aircraft activities within the radius of 450km from the center. The system collects information related to the aircraft and store in its database. Here, the dataset consists a real time aircraft data before the COVID-19 incident, within two hours' duration. The dataset contains 1,000 records with the following attributes: International Civil Aviation Organization (ICAO) number, flight date, flight time, aircraft type and aircraft registration country.

The experiments are performed on a computer with the following specification: Windows 10, Intel(R) Core i5-8250U processor and 8 GB memory. RapidMiner Studio Educational 9.8.01 is used for analysis. The tool can be downloaded freely from the website. RapidMiner is selected due to several reasons [6]: (1) an open-source tool for data analysis, (2) a powerful tool for data integration, Extract Transform Load (ETL), data analytic and reporting in a single package, (3) a complete visualization tool as it has Graphical User Interface (GUI) for designing the analytical processes and presenting the results, and (4) easy to use without needing to write the code and it is a complete and versatile package that consist of various operators for data integration, data exploration and machine learning algorithms.

The machine learning algorithms that are used for the classification task as follows: Naïve-Bayes, Decision Tree, Random Forest and ID3. In order to solve the problem using machine learning algorithm, the dataset is divide into two subsets (i.e. training set and testing set). The training set is used to train a

model, whereas the test set is used to test the trained model (i.e. predict the value of class based on its experience in the training set). Here, Split Validation and Cross Validation methods are performed to divide the dataset into training and testing set. Finally, the machine learning algorithms (i.e. Decision Tree, Random Forest, Naïve-Bayes and ID3) are assessed based on their classification accuracy rate, error rates, precision and recall for classifying the data. Using these results, the algorithms are compared to find the best algorithm for classifying the data (i.e. correctly classified aircraft to its type or country of origin).

4.2 Results and Analysis

In this section, results of the experiment are discussed and presented. The classification accuracy rate, error rate, precision and recall, for each algorithm is used as the measurement parameters. For example, if the algorithm has a higher accuracy rate and lower error rate in classifying the data, then it is highly recommended. First, the dataset is divided into training set and testing set. The training set is used to construct the model, and the testing set is used to validate the model. A standard rule of thumb is two-thirds of the data as training set and one-third as testing set. Next, two set of experiments are performed using different techniques to split the dataset into training set and testing set. The first set of experiment is performed to investigate the capability of machine learning algorithms (i.e. Naïve Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft data into three types of aircraft (i.e. general, military, unknown) and for classifying aircraft data based on registration country (i.e. Malaysia, Indonesia, etc) using Split Validation technique. Whereas, the second set of experiment is performed on the same dataset using Cross Validation technique. Finally, the performance of each model is recorded in terms of classification accuracy rate, error rate, precision and recall.

Table 1 and 2 show the results of machine learning algorithms (i.e. Naïve-Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft into three types of aircraft (i.e. general, military, unknown) and for classifying aircraft based on registration country (i.e. Malaysia, Indonesia, etc) using Split Validation technique. Considering accuracy rate and error rate as the performance measure, it shows that Random Forest and ID3 algorithms are the best algorithm for solving this problem. Even though the result of Decision Tree algorithm achieved 100% accuracy for classifying aircraft into three types of aircraft in Table 1, but the algorithm does not achieve it best performance in the other dataset (i.e. for classifying aircraft based on registration country in Table 2).

Table 1. Performance of machine learning algorithms (i.e. Naïve-Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft into three types of aircraft using Split Validation.

	Algorithms			
	Naïve-Bayes (%)	Decision Tree (%)	Random Forest (%)	ID3 (%)
Accuracy	**99.65**	**100**	**100**	**100**
Classification Error	0.35	0	0	0
Weighted Mean Precision	83.33	100	100	100
Weighted Mean Recall	99.84	100	100	100

Table 2. Performance of machine learning algorithms (i.e. Naïve-Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft into country of origin using Split Validation.

	Algorithms			
	Naïve-Bayes (%)	Decision Tree (%)	Random Forest (%)	ID3 (%)
Accuracy	**64.81**	**67.60**	**92.84**	**100**
Classification Error	35.19	32.40	7.16	0
Weighted Mean Precision	13.83	53.37	80.70	81.82
Weighted Mean Recall	36.36	30.91	77.11	81.82

Therefore, details analysis of the results is performed (see Table 3). Based on the class precision rate (i.e. the number of instance that is correctly labelled belong to the positive class), Decision Tree algorithm achieved 100% only for three countries from eleven (i.e. India, Brunei and Qatar in Table 3). Also, given the recall rate (i.e. the number of items that are not labelled as belonging to the positive class, but it should have been) in Table 3, Decision Tree algorithm achieved 100% only for Brunei.

Table 3. Details performance of machine learning algorithms (i.e. Naïve-Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft into country of origin using Split Validation.

Performance Features		Algorithms			
		Naïve-Bayes (%)	Decision Tree (%)	Random Forest (%)	ID3 (%)
Accuracy		**64.81**	**67.60**	**92.84**	**100**
Class Precision	Malaysia	90.91	65.84	87.72	100
	India	0	100	100	100
	Indonesia	0	63.64	100	100
	USA	0	66.67	100	100
	Singapore	0	90.91	100	100
	Brunei	7.41	100	100	100
	Thailand	0	0	100	100
	Qatar	41.67	100	100	100
	Netherland	12.12	0	100	100
	Philippine	0	0	0	0
	China	0	0	0	0
Class Recall	Malaysia	100	94.12	100	100
	India	0	10	100	100
	Indonesia	0	28	83.33	100
	USA	0	22.22	91.11	100
	Singapore	0	35.7	88.06	100
	Brunei	100	100	100	100
	Thailand	0	0	85.71	100
	Qatar	100	50	100	100
	Netherland	100	0	100	100
	Philippine	0	0	0	0
	China	0	0	0	0

Therefore, based on the previous results, second experiment is performed to investigate the capability of machine learning algorithms (i.e. Naïve-Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft data based on registration country (i.e. Malaysia, Indonesia, etc) using Cross Validation technique (see Table 4). Cross-validation technique is usually preferred as it provides the algorithm a chance to train on multiple train-test dataset. Using this technique, the algorithm performance on classifying unseen data can be tested in order to see how well the algorithm performed. Again, it shows that Random Forest and ID3 algorithms are the best classifier for classifying the data. In contrary, Decision Tree algorithm shows no significant change for classifying aircraft data based on registration country (see Table 4).

Table 4. Performance of machine learning algorithms (i.e. Naïve Bayes, Decision Tree, Random Forest, and ID3) for classifying aircraft into country of origin using Split Validation

Performance Features	Algorithms			
	Naïve-Bayes (%)	Decision Tree (%)	Random Forest (%)	ID3 (%)
Accuracy	**76.70**	**63.51**	**100.00**	**100**
Classification Error	23.30	36.49	0	0
Weighted Mean Precision	29.76	33.55	92.73	92.73
Weighted Mean Recall	54.55	30.67	92.73	92.73

It is expected that Decision Tree algorithm should achieve good result compared to Naïve-Bayes for the tested dataset. Given a fact that Decision Tree is a powerful and popular tools for classification and prediction. Decision tree is a tree-structured algorithm that consists of node and it classify instances by starting at the root of the tree and moving through it until a leaf node [18].

The algorithm divides the instances in dataset recursively using depth-first greedy approach or breadth-first approach, until all the instances is classified into a particular label or class [19]. Yet, ID3 and Random Forest achieved better performance and outperformed Decision Tree.

This is due to, that ID3 (Iterative Dichotomiser 3) creates simple and efficient tree with the smallest depth follows a greedy approach. ID3 basically built on the Concept Learning System (CLS) algorithm; the basic algorithm for building a decision tree by selecting a best attribute using information gain and entropy. Whereas, Random Forest is an ensemble of unpruned decision trees that grows multiple trees that creates a forest [19]. Each individual tree in the Random Forest will have a class prediction and the class with the most votes becomes the selected model's prediction. Therefore, based on the behaviour of ID3 and Random Forest that is far advanced from Decision Tree, it can be concluded that these two algorithms are the most accurate among other algorithms that being investigated in this research for classifying the dataset.

5. Conclusion

Classification is the task of predicting the class (i.e. target, label or category) of given dataset. Classification task includes the process of learning past examples in order to make predictions in the future or unseen data. This research compares the performance of machine learning algorithms especially Naïve-Bayes, Decision Tree, Random Forest and ID3 for classification task that is available in RapidMiner (i.e. classifying aircraft to certain category and into country of origin). Those algorithms are compared based on their accuracy rate, error rate, precision and recall for correctly classifying the data. Findings show that Random Forest and ID3 algorithms outperformed other two algorithms (i.e. Naïve-Bayes and Decision Tree) either using Split Validation or Cross Validation technique for classifying the data. For this research, Random Forest and ID3 algorithms have good classification accuracy due to the nature of the algorithms that is progressively improved from Decision Tree. But, one thing that need to consider is that Decision Tree, ID3, and Naïve-Bayes are sensitive to unbalance dataset. So, that is why the result of Decision Tree and Naïve-Bayes achieve 100% for certain experiments, meanwhile, 0% for the others experiment, especially when considering that Naive Bayes can quickly learn to use high dimensional features with limited training data. Therefore, Decision Tree, Random Forest and ID3 algorithms will be further explored in future for solving other classification problems in order to determine the best controlling conditions in the results.

6. References

[1] Vijay Kotu and Bala Deshpande 2019 Data Science: Concepts and Practice *Morgan Kaufman*
[2] Syahaneim Marzukhi 2014 Three-Cornered Coevolution Learning Classifier Systems for Classification *Thesis, Victoria University of Wellington, New Zealand*

[3] Vijay Kotu 2016 Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner *Morgan Kaufman*
[4] Hongbo Du 2010 Data Mining Techniques and Applications: An Introduction *Cengage Learning*
[5] Ghous and Kovács H 2020 Efficiency comparison of Python and RapidMiner *Multidisciplinary Sciences* **10** 3
[6] Shraddha Dwivedi, Paridhi Kasliwal and Suryakant Soni 2016 Comprehensive Study of Data Analytics Tools (RapidMiner, Weka, R tool, Knime) *Sym on Colossal Data Analysis and Networking*
[7] Ahmad Al-Khoder and Hazar Harmouch 2015 Evaluating four of the most popular Open Source and Free Data Mining Tools *International Journal of Academic Scientific Research* **3** (1) pp 13-23.
[8] Rangra K and Bansal K L 2014 Comparative Study of Data Mining Tools *International Journal of Advanced Research in Computer Science and Software Engineering* **4** 6.
[9] Jovic A Brkic K and Bogunovic N 2014 An Overview of Free Software Tools For General Data Mining *Information and Communication Technology, Electronics and Microelectronics (MIPRO)*
[10] Wiemer H Drowatzky and Ihlenfeldt L 2019 Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model *Application Science* **9.**
[11] Layth Almahadeen Murat Akkaya and Arif Sari 2017 Mining Student Data Using Crisp-DM Model *International Journal of Computer Science and Information Security* **15** 2.
[12] Syahaneim Marzukhi Nur Hidayah Mohammad Daud Zuraini Zainol and Omar Zakaria 2018 Framework of Knowledge-Based System for United Nations Peacekeeping Operations using Data Mining Technique *IEEE Explore.*
[13] Ginika Mahajan Bhavna Saini and Tai Almas 2019 Taxonomy on RapidMiner using Machine Learning *International Conference on Sustainable Computing in Science, Technology & Management.*
[14] Hemlata and Preeti Gulia 2019 Experimental Evaluation of Open Source Data Mining Tools: R, Rapid Miner and KNIME *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* **9** 1
[15] Sanusia and Juniana Husnab 2020 Utilization of Rapidminer using the K-Means Clustering Algorithm for Classification of Dengue Hemorrhagic Fever (DHF) Spread in Banda Aceh City *Jurnal Inovasi Teknologi dan Rekayasa* **5** (2) 146-152.
[16] Muhammad Rifqi Firdaus Abdul Latif2 Ipin Sugiyarto and Windu Gata 2020 Classification Of The Prospects For City Trees Life Expectancy Using Naive Bayes Method *Jurnal Ilmu Pengetahuan dan Teknologi Maklumat (JITK)* **6** 1.
[17] Nur Diyana Kamarudin Syarifah Bahiyah Rahayu Zuraini Zainol Mohd Shahrizal Rusli and Kamaruddin Abdul Ghani 2018 Performance Comparison of Machine Learning Classifiers on Aircraft Databases *STRIDE Technical Bulletin* **11** (2) 154-169.
[18] Prajwala T R 2015 A Comparative Study on Decision Tree and Random Forest Using R Tool *International Journal of Advanced Research in Computer and Communication Engineering* **4** 1.
[19] Khaled M Almunirawi and Ashraf Y A Maghari 2016 A Comparative Study on Serial Decision Tree Classification Algorithms in Text Mining *Journal of Intelligent Computing Research (IJICR)* **7** 4.