Comparative analysis of classification algorithms for critical land prediction in agricultural cultivation areas

Deden Istiawan[*)]

Department of Statistics, Akademi Statistika Muhammadiyah Semarang
Jl. Prof. Dr. Hamka Km 01 Tambakaji Ngaliyan, Semarang, Indonesia 50185

Abstract – The identification of critical land that has been physically, chemically, and biologically damaged usually uses a geographic information system. However, it requires a high cost to get the high resolution of satellite images. In this study, a comparison framework is proposed to determine the classification algorithms' performance, namely C.45, ID3, Random Forest, k-Nearest Neighbor, and Naïve Bayes. This research aims to find the best algorithm for the classification of critical land in agricultural cultivation areas. The results show that the Random Forest algorithm has the highest accuracy of 93.10 % in predicting critical land. The naïve Bayes has the lowest performance, with 89.32 % of accuracy in predicting critical land.

Keywords – critical land; watersheds; classification algorithm; algorithm comparison

I. INTRODUCTION

Watershed is a complex system built on physical systems, biological systems, and human systems interconnected and interact with each other [1]. The condition of watersheds in Indonesia is getting worse, indicated by the increasing number of priority watersheds from year to year [2]. The damage of watersheds is caused by the use of natural resources that exceed the carrying capacity. Population growth is a consequence, and natural resource utilization policy is not principled for sustainable development [3]. The upstream part of the watershed should function as a water catchment area. However, forest and land damage and deforestation in the upstream will cause a long dry season and flooding during the rainy season and cause critical land [4].

Critical land is a land that has been damaged physically, chemically, and biologically. These damages cause the land to experience a decline in fertility [5]. Land that should be a place of production and a place for water catchment cannot properly function because it is damaged. Mapping and identification of critical land are crucial to be carried out for the planning and determination of watersheds, which are the priority of rehabilitation [6].

The indicators that cause critical land include land management, land use, rainfall, slope, and land erosion [5]. However, forest and land damage information and data damage often do not refer to database formats and structures that can be accounted for. One of the important aspects in determining the success of land mapping is determining the classification of land. However, the lack of spatial data and information affects the evaluation of the validity of critical land data [7]. The availability of accurate and informative information about the amount and distribution of degraded land has a significant meaning. Updating the degraded land data will continue to be carried out concerning the criteria and standard standards for determining and processing degraded land data. Data processing of critical land is essential to obtain the results of an inventory of critical land with high validity.

The previous method used for land classification is Geographic Information System (GIS) [8]-[10]. GIS method captures the condition of land from satellite images. The land condition score is then calculated according to the Decree of the Director-General of RRL No: 041/Kpts/V/1998 on determining critical lands [7]. The advantage of GIS is that it can capture the condition of land or area spatially by using satellite images [9]. However, it requires high costs to take the high-resolution satellite images and to know the rainfall levels, at least three different monitoring stations (Landsat-5 TM 1985, Landsat-7 ETM+ 2000 and Landsat-8 OLI–TIRS 2015) with 30 m Spatial resolution were used for the analysis of the studied watershed [11], [12]. For critical land mapping, images with high spatial resolution are needed to obtain information about the earth's surface [13].

Presently, only attributes data regarding critical land are available at the Indonesia Ministry of Forestry, such that spatial distribution is difficult to know. So, the synchronization of forest and land rehabilitation programs that are multi-sectoral is difficult because the spatial analysis is one of the main tools [7]. Unavailability of spatial data and information affects the assessment of the validity of critical land data.

[*)] Correspondence author (Deden Istiawan)
Email: dedenistiawan@gmail.com

However, to get reliable information from satellite data on the right target, the right classification technique is needed. Several classification methods have been optimized for the past few years [14], one of which is the data mining approach [15], [16]. Data Mining in the agriculture field is a relatively novel research field to predict critical land.

Classification is one of the primary roles of data mining that can be used to classify data with a label or class, which is called supervised learning since it requires a label or class in the process [17]. The classification predicts the label or class of a dataset. The dataset is divided into two parts in the classification process, namely training data and testing data [18]. In the training data, some data that are known are applied to create a classification model. Whereas the testing data is used to test the classification model to determine how well the accuracy of the classification model is formed. The classification model with good accuracy can be used to predict data of labels or class that has not yet been known [19].

In recent years, there have been several studies on algorithm comparison to determine classification algorithms' performance. Hall et al. [20] show that the naïve Bayes (NB) algorithm has the highest accuracy among Logistic Regression (LR), and Neural Network (NN). Kim et al. [21] show that the k-Nearest Neighbor (kNN) algorithm has better performance than the Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA). Narayanan et al. [22] reported that Decision Tree (DT), NB and kNN algorithms are the most commonly used classification algorithms. Five classification algorithms considerably have the best performance, namely C.45, ID3, Random Forest (RF), NB, and kNN. C.45, ID3, and RF algorithms are induction of DT.

With the motivation to seek an efficient approach for critical land prediction, we propose a new approach to classify critical land with data mining. In this study, we propose a comparison framework to compare the C.45, ID3, RF, kNN, and NB algorithms' performance to predict critical land in agricultural cultivation areas. This research is expected to provide benefits in the form of a framework for further research, especially for critical land prediction.

II. RESEARCH METHODS

The proposed algorithm comparison framework can be seen in Figure 1. The proposed framework consists of critical land datasets in the agricultural cultivation area, classification algorithm, model validation, model evaluation, and model comparison. This study aims to find out the best algorithms for critical land classification.

A. Datasets

The data used for the experiment in this study is a dataset from BPDAS Pemali Jratun. This data is in the form of critical land parameter data in agricultural

Table 1. Dataset structure

No	Attribute	Description
1	Land Productivity	1: Very Poor
		2: Poor
		3: Moderately
		4: Good
		5: Very Gdod
2	Slope Level	1: Very Steep
		2: Steep
		3: Rather Steep
		4: Sloping
		5: Flat
3	Erosion Level	2: Very Heavy
		3: Heavy
		4: Moderately
		5: Light
4	Land Management	5: Good
		3: Moderately
		2: Poor
5	Critical Land	5: Very Critical
		4: Critical
		3: Rather Critical
		2: Potential to Critical
		1: Not Critical

cultivation, with four attributes. The attributes used include land productivity, slope level, erosion hazard level, and land management, while the class used is a land criticality with a total data of 111,003 [23]. The dataset structure is shown in Table 1.

B. Classification algorithms

In this study, five classification algorithms are compared to get the best model for the classification of critical land in agricultural cultivation, namely algorithm C.45, ID3, Random Forest (RF), K-Nearest Neighbor (K-NN) and Naïve Bayes (NB). Those algorithms were chosen because those algorithms are most commonly used, as stated in [22]. The selection aims to achieve a balance between established classification algorithms used in critical land prediction.

Figure 1 shows a proposed framework for the comparison of classification algorithms. The initial stage is a collection of datasets, then a comparison of classification algorithms. The distribution of training data and testing data uses stratified 10-fold cross-validation to evaluate classification algorithms' performance using accuracy, recall, and precision. In the final stage, a statistical test is performed to determine the significant differences between the classification models.

C. Model evaluation

Evaluation of the experiments' results assesses or measures how well the proposed method against other methods and whether the proposed method has a significant difference in other models' results. In this study, the evaluation models used are accuracy, recall, and precision.

The accuracy value (AC) is calculated by taking the correct prediction percentage from the whole data (Eq. 1). Sensitivity or recall (R) in the field of information search measures the proportion of original positives that are correctly predicted as positive (Eq. 2). Sensitivity is related to the ability of testing to identify positive results from a number of actually positive data. Precision (P) or positive predictive value is a matrix to measure system performance in getting relevant data (Eq. 3). Precision is the amount of true positive data divided by the amount of data that is recognized as positive. *TN* denotes true negative, *TP* as true positive, *FP* as false positive, and *FN* as a false negative.

$$AC = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

$$P = \frac{TP}{TP+FP} \tag{3}$$

D. Model comparison

Model comparison is used to compare the performance of classification algorithms. The classification model is validated using 10-fold cross-validation, which means that the algorithm's performance is calculated and compared directly. This approach is difficult to know whether the difference in the average value of accuracy is significantly different or not. In this study, a statistical significance test was used to determine differences in the classification model's performance. Statistical test results are test statistics and p-values, both of which can be interpreted and used in the presentation of results to measure the level of confidence or significance in differences between models. It allows stronger claims to be made as part of the model selection rather than not using a statistical hypothesis test.

III. RESULTS AND DISCUSSION

The experiments in this study used a computing platform based on Intel Core i5 2.3 GHz CPU, 4 GB RAM and Microsoft Windows 10 64-bit operating system, and RapidMiner Studio 9.3. RapidMiner was used to measure Accuracy, Recall, and Precision. This experiment used a critical land dataset consisting of five classification algorithms.

The 10-fold cross-validation test results of the five classification algorithms are expressed as confusion matrix in Table 2 for the kNN algorithm, Table 3 for C4.5, Table 4 for ID3, Table 5 for RF, and Table 6 for NB. The classification algorithms for predicting critical land in agricultural cultivation areas are summarized under the selected statistical criteria, as shown in Table 7. The accuracy of the prediction of critical lands indicates that the RF algorithm has the highest accuracy, while the lowest is the NB algorithm. It confirms [24], [25] that RF is an
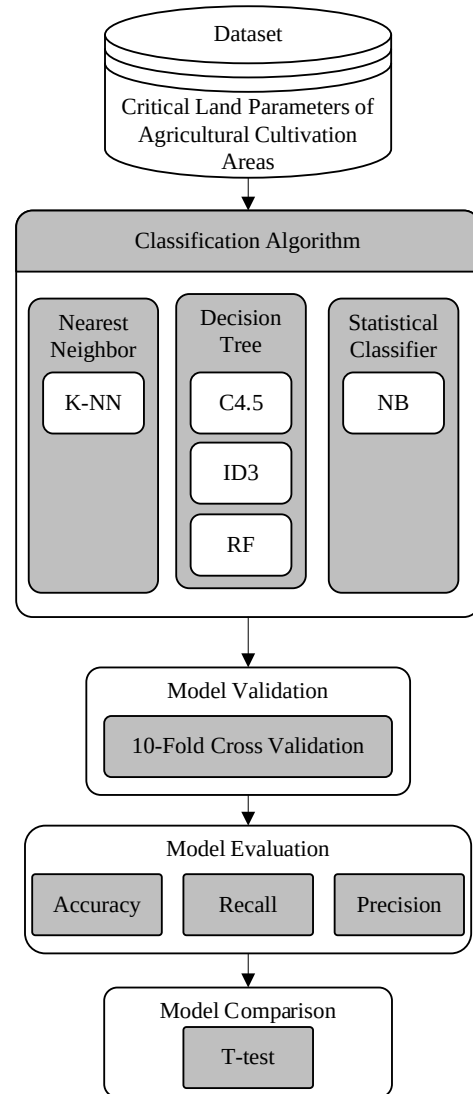

Figure 1. Proposed comparison framework

ensemble learning method and considered a reference due to its excellent performance.

Table 7 also shows that the RF algorithm has the highest averaged recall and precision, with 92,74 % and 94.94 %, then followed by the kNN algorithm with averaged recall and precision values of 92,02 % and 93.51 %, respectively. It is consistent with [21], stating that the kNN has better performance than QDA and LDA algorithms. In contrast, the lowest performance is the NB algorithm with an accuracy value of 89.32 %, averaged recall and precision of 65.45 % and 77.68 %, respectively.

In this study, the NB has the worst classification performance due to obtaining the classification results by calculating the attribute probability values independently, which means the attribute values are not interdependent. This can disrupt the classification performance of NB in [20], [26]-[29].

This study also used a t-test to see whether there are significant differences or not between the classification models and to see whether there is a significant difference seen from the p-value. If the p-value is less than α=0.05,

Table 2. The confusion matrix for kNN classification

Class Prediction	True Condition					Class Precision (%)
	Not Critical	Potential to Critical	Rather Critical	Critical	Very Critical	
Pred. Not Critical	71342	3783	25	41	8	**94.87**
Pred. Potential to Critical	3559	18919	403	3	2	**82.67**
Pred. Rather Critical	53	519	10353	39	0	**94.43**
Pred. Critical	5	0	26	1680	8	**97.73**
Pred. Very Critical	3	0	0	0	230	**97.87**
Class Recall (%)	**95.17**	**81.47**	**95.80**	**95.18**	**92.47**	

Table 3. The confusion matrix for C4.5 classification

Class Prediction	True Condition					Class Precision (%)
	Not Critical	Potential to Critical	Rather Critical	Critical	Very Critical	
Pred. Not Critical	72717	5066	32	33	8	**93.40**
Pred. Potential to Critical	2202	17844	499	1	2	**86.84**
Pred. Rather Critical	33	311	10237	77	0	**96.05**
Pred. Critical	7	0	39	1641	22	**96.02**
Pred. Very Critical	3	0	0	13	216	**93.10**
Class Recall (%)	**97.01**	**76.84**	**94.73**	**94.73**	**87.10**	

Table 4. The confusion matrix for ID3 classification

Class Prediction	True Condition					Class Precision (%)
	Not Critical	Potential to Critical	Rather Critical	Critical	Very Critical	
Pred. Not Critical	72727	5113	32	33	8	**93.34**
Pred. Potential to Critical	2197	17845	524	1	2	**86.76**
Pred. Rather Critical	28	263	10212	67	0	**96.61**
Pred. Critical	7	0	39	1643	18	**96.25**
Pred. Very Critical	3	0	0	21	220	**90.16**
Class Recall (%)	**97.02**	**76.85**	**94.49**	**93.09**	**88.71**	

Table 5. The confusion matrix for Random Forest classification

Class Prediction	True Condition					Class Precision (%)
	Not Critical	Potential to Critical	Rather Critical	Critical	Very Critical	
Pred. Not Critical	72717	4582	32	33	8	**93.98**
Pred. Potential to Critical	2204	18347	383	0	2	**87.63**
Pred. Rather Critical	31	292	10367	33	0	**96.68**
Pred. Critical	7	0	25	1698	1	**98.09**
Pred. Very Critical	3	0	0	1	237	**98.34**
Class Recall (%)	**97.01**	**79.01**	**95.93**	**96.20**	**95.56**	

Table 6. The confusion matrix for naïve Bayes classification

Class Prediction	True Condition					Class Precision (%)
	Not Critical	Potential to Critical	Rather Critical	Critical	Very Critical	
Pred. Not Critical	72695	5528	41	30	8	**92.84**
Pred. Potential to Critical	2211	17291	1963	1	2	**80.54**
Pred. Rather Critical	21	394	8735	1265	0	**83.87**
Pred. Critical	35	8	66	405	109	**65.01**
Pred. Very Critical	0	0	2	64	129	**66.15**
Class Recall (%)	**96.98**	**74.46**	**80.83**	**22.95**	**52.02**	

then H0 is rejected, meaning significant differences between the two classification models. Table 8 shows no difference in accuracy between the C4.5 algorithm with ID3 and the kNN algorithm since there is no difference with the ID3 algorithm. Based on the t-test, RF algorithm is the best algorithm for the classification of critical land in agricultural cultivation areas.

IV. CONCLUSION

RF algorithm has the best performance for classifying critical land in agricultural cultivation areas. C4.5, ID3, and kNN algorithms have good performance, but there is no significant difference, while the NB algorithm has the lowest performance for critical land prediction. Results of the prediction of the classification of critical land can be used for decisions related to land rehabilitation.

For further research, it is suggested to make attribute selection before applying the classification model to eliminate irrelevant information so that it can improve the performance of classification algorithms and reduce computing time.

REFERENCES

[1] A. Susetyaningsih, "Pengaturan penggunaan lahan di daerah hulu DAS Cimanuk sebagai upaya optimalisasi pemanfaatan sumberdaya air," *Kontruksi*, vol. 10, no. 1, pp. 1-8, 2012.

[2] E. Junaidi and S. D. Tarigan, "Pengaruh hutan dalam pengaturan tata ruang air dan proses sedimentasi daerah aliran sungai: studi kasus di DAS Cisadane," *Jurnal Penelitian Hutan dan Konservasi Alam*, vol. 8, no. 2, pp. 155-176, 2011. doi: 10.20886/jphka.2011.8.2.155-176

[3] I. Mawardi, "Kerusakan daerah aliran sungai dan penurunan daya dukung sumberdaya air di pulau Jawa serta upaya penanganannya," *Jurnal Hidrosfir Indonesia*, vol. 5, no. 2, pp. 1-11, 2010.

[4] P. Pratiwi, I. S. D. Wayan, G. M. E. Hartoyo, and Y. Nugroho, "Kesesuaian tempat tumbuh jenis-jenis pohon di DAS Pemali Jratun Jawa Tengah," *Jurnal Penelitian Hutan dan Konservasi Alam*, vol. 9, no. 4, pp. 299-321, 2012. doi: 10.20886/jphka.2012.9.4.299-321

[5] H. Hendro, Z. Nadhi, S. Budiastuti, and D. Purnomo, "Pemetaan lahan kritis di kawasan Muria untuk meningkatkan daya dukung lingkungan yang berbasis pada sistem informasi geografis," *Jurnal Ilmu Pertanian*, vol. 17, no. 1, pp. 46-51, 2014.

[6] I. M. Parsa, U. S. Wiradisastra, and H. Pawitan, "Identifikasi dan pemetaan lahan kritis menggunakan teknik pengindraan jauh dan sistem informasi geografi," *Jurnal Manajemen Hutan Tropika*, vol. 9, no. 2, pp. 63-77, 2003.

Table 7. Performance comparison of all algorithms

Model	Model Evaluation (%)		
	Recall	Precision	Accuracy
kNN	92.02	93.51	92.36
C4.5	90.08	93.08	92.49
ID3	90.03	92.62	92.54
RF	**92.74**	**94.94**	**93.10**
NB	65.45	77.68	89.32

Table 8. P-Value of T-test

	C4.5	ID3	RF	NB	K-NN
C4.5		0.694	**0.000**	**0.000**	0.277
ID3			**0.000**	**0.000**	0.180
RF				**0.000**	**0.000**
NB					**0.000**
kNN					

[7] D. S. Candra, "Analysis of critical land in the musi watershed," *International Journal of Remote Sensing and Earth Sciences*, vol. 8, pp. 13-18, 2011.

[8] A. R. Oktaviani, A. L. Nugraha, and H. S. Firdaus, "Analis penentuan lahan kritis dengan metode fuzzy logic berbasis pengindraan jauh dan sistem informasi geografi," *Jurnal Geodesi Undip*, vol. 6, no. 4, pp. 332-341, 2017.

[9] L. H. Prabandaru, A. Lailanugraha, and A. Sukmono, "Pemetaan tingkat lahan kritis kabupaten wonosobo dengan penginderaan jauh dan sistem informasi geografis," *Jurnal Geodesi Undip*, vol. 5, vol. 4, pp. 65-72, 2016.

[10] L. A. Ramayanti, B. D. Yuwono, and M. Awaluddin, "Pemetaan tingkat lahan kritis dengan menggunakan pengindraan jauh dan sistem informasi geografis," *Jurnal Geodesi Undip*, vol. 4, no. 2, pp. 200-207, 2015.

[11] T. Gashaw, T. Tulu, M. Argaw, and A. W. Worqlul, "Evaluation and prediction of land use/land cover changes in the Andassa watershed, Blue Nile Basin, Ethiopia," *Environmental Systems Research*, vol. 6, no. 1:17, pp. 1-15, 2017. doi: 10.1186/s40068-017-0094-5

[12] K. Amri, A. Halim, and M. F. Barchia, "Criticality analysis of recharge area and land in the catchment area of Musi hydropower Bengkulu Indonesia," *APCBEE Procedia*, vol. 10, pp. 235-240, 2014. doi: 10.1016/j.apcbee.2014.10.045

[13] H. Moayedi, A. Jamali, M. B. A. Gibril, L. Kok Foong, and M. Bahiraei, "Evaluation of tree-base data mining algorithms in land used/land cover mapping in a semi-arid environment through Landsat 8 OLI image; Shiraz, Iran," *Geomatics Natural Hazards and Risk*, vol. 11, no. 1, pp. 724-741, 2020. doi: 10.1080/19475705.2020.1745902

[14] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823-870, 2007. doi: 10.1080/01431160600746456

[15] J. R. Otukei and T. Blaschke, "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms," *International Journal of Applied Earth Observation and Geoinformation*, vol. 125, pp. 527-531, 2010. doi: 10.1016/j.jag.2009.11.002

[16] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sensing Environment*, vol. 86, no. 4, pp. 554-565, 2003. doi: 10.1016/S0034-4257(03)00132-9

[17] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1937-1946, 2014. doi: 10.1016/j.eswa.2013.08.089

[18] F. Gorunescu, *Data Mining Concepts, Model and Techniques, 12th ed*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[19] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann, 2012.

[20] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, "A systematic literature review on fault prediction performance in software engineering," IEEE Transactions on Software Engineering, vol. 38, no. 6, pp. 1276-1304, 2012. doi: 10.1109/TSE.2011.103

[21] K. S. Kim, H. H. Choi, C. S. Moon, and C. W. Mun, "Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions," *Current Applied Physics*, vol. 11, no. 3, pp. 740-745, 2011. doi: 10.1016/j.cap.2010.11.051

[22] U. Narayanan, A. Unnikrishnan, V. Paul, and S. Joseph, "A survey on various supervised classification algorithms," in *International Conference on Energy, Communication, Data Analytics and Soft Computing*, Chennai, India, Aug. 2017, pp. 2118-2124. doi: 10.1109/ICECDS.2017.8389824

[23] D. Istiawan and L. Khikmah, "Implementation of C4.5 algorithm for critical land prediction in agricultural cultivation areas in pemali jratun watershed," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, vol. 2, no. 2, pp. 67-73, 2019. doi: 10.24014/ijaidm.v2i2.7569

[24] M. Liu, M. Wang, J. Wang, and D. Li, "Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar," *Sensors and Actuators B: Chemical*, vol. 177, pp. 970–980, 2013. doi: 10.1016/j.snb.2012.11.071

[25] C. J. Mantas, J. G. Castellano, S. Moral-García, and J. Abellán, "A comparison of random forest based algorithms: random credal random forest versus oblique random forest," *Soft Computing*, vol. 23, no. 21, pp. 10739–10754, 2019. doi: 10.1007/s00500-018-3628-5

[26] J. Wu, S. Pan, X. Zhu, Z. Cai, P. Zhang, and C. Zhang, "Self-adaptive attribute weighting for naive Bayes classification," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1487-1502, 2015. doi: 10.1016/j.eswa.2014.09.019

[27] L. Jiang, Z. Cai, H. Zhang, and D. Wang, "Naive Bayes text classifiers: A locally weighted learning approach," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 25, no. 2, pp. 273-286, 2013. doi: 10.1080/0952813X.2012.721010

[28] L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," *Engineering Applications of Artificial Intelligence*, vol. 52, pp. 26-39, 2016. doi: 10.1016/j.engappai.2016.02.002

[29] J. Wu, S. Pan, Z. Cai, X. Zhu, and C. Zhang, "Dual instance and attribute weighting for Naive Bayes classification," in *International Joint Conference on Neural Networks*, Beijing, China, Jul. 2014, pp. 1675-1679. doi: 10.1109/IJCNN.2014.6889572