**TELKOMNIKA**, Vol.15, No.3, September 2017, pp. 1215~1222
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
**DOI:** 10.12928/TELKOMNIKA.v15i3.3638


# Decision Support System for Bat Identification using Random Forest and C5.0

**Deden Sumirat Hidayat[1], Imas Sukaesih Sitanggang*[2], Gono Semiadi[3]**
[1,2]Computer Science Department, Faculty of Natural Science and Mathematics,
Bogor Agricultural University, Indonesia
[1,3]Research Center for Biology-Indonesian Institute of Sciences (LIPI), Indonesia
Corresponding author, e-mail: d2n.scriptproject@gmail.com[1], imas.sitanggang@apps.ipb.ac.id*[2],
semiadi@gmail.com[3]

***Abstract***

*Morphometric and morphological bat identification are a conventional method of identification and requires precision, significant experience, and encyclopedic knowledge. Morphological features of a species may sometimes similar to that of another species and this causes several problems for the beginners working with bat taxonomy. The purpose of the study was to implement and conduct the random forest and C5.0 algorithm analysis in order to decide characteristics and carry out identification of bat species. It also aims at developing supporting decision-making system based on the model to find out the characteristics and identification of the bat species. The study showed that C5.0 algorithm prevailed and was selected with the mean score of accuracy of 98.98%, while the mean score of accuracy for the random forest was 97.26%. As many 50 rules were implemented in the DSS to identify common and rare bat species with morphometric and morphological attributes.*

***Keywords****: bat identification, classification, C5.0, decision support system, random forest*



## 1. Introduction

In biology, scientist relies heavily upon accurate identification of species as the basic unit of the natural hierarchy. In identifying species, there are numbers of arrangement of morphological and morphometric characters as the baseline data. Their identification works require significant amount of experience, encyclopedic knowledge, adequate specimen references and relevant literatures. The process to be an expert in identifying taxonomical group takes a lot of time and effort with the number of experts in taxonomy are limited to cover different groups and zoogeographical areas [1]. The level of taxonomy researchers is categorized into two, taxonomist, being the expert one and para-taxonomist refers to beginner level or non-professional taxonomist. Para-taxonomist frequently encounters difficulties in making quick identification in the field, due to a limitation in expertise and references in particular the species identification key and most of the times relies on the taxonomists help. Therefore a simple yet accurate enough methodology is needed in order to overcome this situation.

All species has distinctive morphological variance as the result of genetics and/or environment as phenotype plasticity [2], that functions as supporting mechanism towards change in the environment. The process of identifying a species can sometimes be difficult to do in conventional method. With the vast and diverse features, a species can sometime has one set of distinctive attribute with original pattern that cannot be identified using a simple statistical approach.

Bats have such important role for human, as pest control, pollinator, plant seed dispersal, and fertilizer from guano [3]. Therefore, efforts need to be taken to conserve the bat population. One of the pivotal field on identification of bat is morphometric [3]. Currently, identification of bat through morphometric feature has been the standard procedure with morphological observation as the supporting characters. This conventional method takes so much effort and time when deal with numbers of individual to identify.

Studies in applied computation, especially data mining, to identify animals and microorganism have been frequently conducted with visual, audio imaging and based on their

*Received May 27, 2017; Revised July 28, 2017; Accepted August 13, 2017*

morphological features. The most current computation studies have been conducted to help the identification or categorizing species of animals, plants and microorganism. Neural network approach has been developed for animal morphological identification [1], [4-6], others using nearest neighbor and Gaussian Process Learning [7, 8], random forest algorithm [9-16], and C5.0 algorithm [17] to identify animals and microbial. Compared to other fabricated neuro system and other statistical approaches, the decision tree, when applied to morphometric category variable, seems can provide as the most effective solution for species identification. The ability of the decision tree to study the pattern in multivariate data makes the approach as the most suitable one for bat identification. From those methods, the random forest and C5.0 have more than 80% accuracy.

Decision tree is a famous method for classification tasks and it has been applied to a broad range of applications including identification problems. Some of decision tree algorithms are C5.0, ID3, C4.5 as a successor of ID3, and CART (Classification and Regression Tree). These algorithms are designed for nonspatial datasets like morphological and morphometric datasets [18]. Classification of categorical data is usually and very accurate using the c4.5 and c5.0 algorithms. The repetition of attributes in this algorithm can be simplified when changing the decision tree into a set of rules. The C5.0 algorithm has an accuracy of 0.9% greater than C4.5 algorithm. Yet, result from C5.0 has larger rules and larger trees than C4.5 [19]. For bat identification, others than morphological and morphometric features, the audio imaging has also been done, but this techniques requires high technology and expensive equipments and not practical for the para-taxonomist works. Therefore a simple developing decision support system (DSS) of morphometric bat identification is needed. The DSS is used to solve the problems in bat identification since it is able to provide best identification of morphological features and classification.

## 2. Research Method

### 2.1. Data

The source of data for the study is from the Indonesian Biodiversity Information System (IBIS), Research Centre for Biology, Indonesian Institute of Sciences (LIPI) data base of mammals with specific group of bats from Chiropteran. The data consist of several attributes namely taxonomy, age group, location coordinate, sample, morphology (body characters) and physical measurement (morphometric) consisted of body weight, length of tail, body (from the rear-end to the head), hind legs without claws, ears, radius-ulna and tibia-fibula. The method of classification used is data mining technique that is the random forest and C5.0 algorithm. The random forest develops rules of classification in the form of decision tree and later the rules are going to be used for the new dataset. C5.0 functions as contrasting algorithm of which purpose is to classify the data.

### 2.2. Framework

The theoretical framework in developing the Decision Support System is described in the flowchart on Figure 1. During the stage, the researchers identify some problems found in the observed objects in order to find alternative solutions to overcome the problems. Literature used is in the form of textbooks and research journals that discuss species identification, more particularly bat identification, that use the decision tree classification as the approach. In order to obtain relevant information, the researchers ask to access data in the form of results of the observations of bats in their natural habitat whose authors are bat experts in the Indonesian Biodiversity Information System (IBIS). As an addition, the researchers also ask to get an access to the database of the Indonesian Biodiversity Information System (IBIS), Research Center for Biology, Indonesian Institute of Sciences. Besides the data, acquiring knowledge from the experts is conducted in order to find out some important attributes for the study.

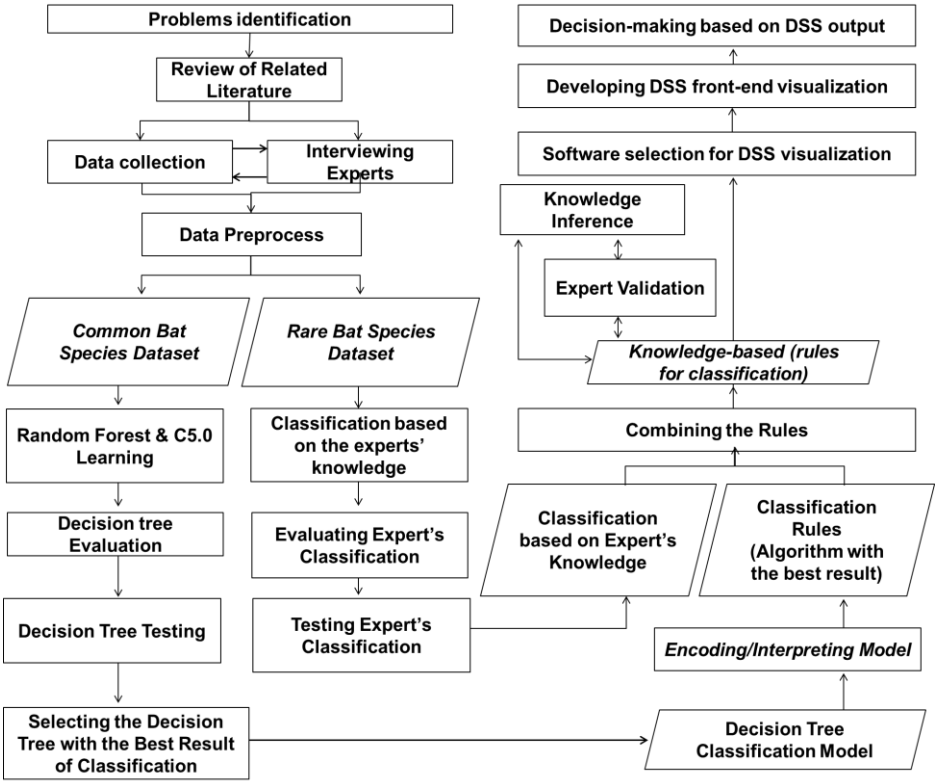


Figure 1. Theoretical Framework for Decision Support System Development

### 2.3. Data Preprocess and Classification

There are some processes conducted in the stage namely taking care of the missing value as well as the imbalance class, data integration, data dimension reduction, data consistency, selecting the features to use, data cleaning and data transformation. The outcome of the stage is two types of dataset to be used in the following stage; they are the common bat species dataset (complete one) and rare bat species dataset (limited or incomplete one). The complete dataset is, then, divided with the 5-fold cross validation method. In this stage, the researchers conduct learning process using the Random Forest and C5.0 Decision Tree Algorithm to the complete common species bat training dataset that will result in a decision tree. On the other hand, considering that the dataset for rare bat species is still limited or incomplete, classification is carried out based on recommendation from the experts.

### 2.4. Analysis and Evaluation of Decision Tree and Rules

The purpose of the stage is evaluation towards the result of decision tree classification based on the result of algorithm and rules of classification. The result is obtained based on the result of expert's classification using the confusion matrix as representation of correctly classified data and incorrectly classified ones as well as accuracy rate for each of the algorithm.

### 2.5. Decision Tree Testing and Expert Knowledge Rules

The goals of the stage are to test the result of decision tree classification based on the result of the algorithm and rules based on the result of the expert's classification using the tested data. For the algorithm, the following step is to select one with the best result, while the expert's classification will result in some rules that have been tested.

### 2.8. Selecting the Decision Tree with the Best Result

In the stage, specifically for the complete common species dataset, the researchers are selecting decision tree (Random Forest or C5.0) that gives the best classification result. The model of classification with the best accuracy is going to be selected as the chosen model.

### 2.9. Model Interpretation and Merging the Rules

The goal of the stage is to conduct interpretation of the best model in order to get the expected rules. In the stage, merging the rules that is one becomes the result of the algorithm with the best result and the rules of classification obtained based on the experts. The outcome is knowledge in the form of rules that have accommodated both common and rare species of bats.

### 2.10. Justification, Inference and Expert Validation

Justification process, drawing conclusion and validation are taking place in the stage. Having had the established knowledge in the previous stage, justification that is the process to formulate reasons and explanations of the existed process is conducted. The next stage is drawing some conclusion, and final stage is validation by the experts in taxonomy/taxonomists.

### 2.11. Developing Decision-Making Support System based on DSS Output

In the stage, the researchers select software to visualize the knowledge that has been validated in the previous stage. Besides that, they also select the language for programming and DBMS. The next step is to develop visualization or Decision Support System front end that applies the knowledge which have been validated by the experts. After application for Decision Support System has been developed, the application is going to be tested to determine system output based on the input being given.

## 3. Results and Analysis

### 3.1. Data Preprocessing

There were several processes the researchers did in the stage. The first step was separating the dataset of each attribute that is still written in one column or cell. The second step was data cleansing. The third step was data consistency where the researchers use the same form for decimal figures, for example changing the comma into period. Taking care of the missing value in some attributes is the following step. The fifth step is selecting features based on the expert's recommendation; the researchers reduce the number of attributes from 58 into 21 and 21 attributes are the minimum standardized documentation for adult bat. The outcome of the stage is two types of datasets readily used for the next stage; the first is common species bat dataset which consists of 2500 records and 25 classes where each class has 100 records and the second is rare species bat dataset with 11 classes and each class has one record. The complete dataset is going to be divided using the 5-fold cross validation method where the datasets are randomly divided into five parts, four of them as training data and the other one as testing data.

### 3.2. Learning with Random Forest & C5.0 and Classification with the Expert's Knowledge

Random Forest and C5.0 learning of the complete common bat species training dataset using the R Studio software result in decision tree. The Random Forest and C5.0 learning consists of 5-fold classification. Having been run, it results in 5-fold .csv Random Forest files and 5-fold.csv C5.0 files. The result of classification with the Random Forest and C5.0 of the common bat species is presented in confusion matrix table. Since the dataset for the rare species of bats is limited in number, classification is carried out based on recommended identification key from the experts in order to establish the rules.

### 3.3. Analysis and Evaluation of Decision Tree and Rules

After the 5-fold common bat species training datasets for both the Random Forest and C5.0 have been run and formulated into confusion matrix table, the following steps are calculating the accuracy of each fold of confusion matrix table, calculating mean accuracy rate of 5-fold confusion matrix table, and comparing mean accuracy rate between the mean of the Random Forest and that of C5.0. Based on the steps, accurate results of classification for the common species with algorithm can be presented in Table 1 and 2.

Table 1. Accuracy of Classification Result for the Common Species with C5.0 Algorithm

| Dataset | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy (%) | 98.7 | 98.6 | 99.1 | 99.4 | 99.1 |

Table 2. Accuracy of Classification Result for the Common Species with Random Forest Algorithm

| Dataset | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
|---|---|---|---|---|---|
| Accuracy (%) | 97.49248 | 97.19073 | 97.62266 | 97.47275 | 96.56912 |

The next stage is selecting fold with the highest accuracy from the C5.0 algorithm. Fold 3 and 5 are selected because their accuracy rates are the closest to the mean. Therefore, the researchers make comparison of the rules in fold 3 and fold 5. Fold 3 has fewer rules than fold 5, and as the result fold 3 is selected. Fold 3 consists of 38 rules; the examples of rule 2 and 3 are as follow:

```
Rule 2: (307/226, lift 6.5)
 X4 <= 1
          -> class Cynopterus brachyotis [0.265]
```

```
Rule 30: (73, lift 26.8)
 X3 <= 1
 X13 <= 1
 -> class Megaderma spasma [0.987]
```

Within 38 rules to identify the common bat species, there are some explanations as given in the following example. Rule 30: (73, lift 26.8) is the example of the rule that can identify the bat species accurately. It means out of 73 records, all of them can be identified correctly. On the other hand, Rule 2 (307/226, lift 6.5) is the example of the rule with poor ability to identify the bat species. Out of 307 records, rule 2 is unable to identify 226 records accurately. The higher the lift score is, the better ability a rule has to identify the common bat species. Within the established rules in the study, rule 4 (*Cynopterus titthaecheilus*) and 30 (*Megaderma spasma* ) have the highest lift score of 26.8. These are evidence that C5.0 algorithm is the most suitable algorithm to be applied in the dataset of the both species. Another reason is that there is good documentation for the *Cynopterus titthaecheilus* and *Megaderma spasma* bat species.

### 3.4. Decision Tree and Expert's Rules Testing

Based on the testing result, it is revealed that the rules in fold 3 of the C5.0 algorithm are the best ones to identify and predict the testing data. The identification test for 514 records is successful and accurate

### 3.5. Selecting Decision Tree and Formulating Rules with the Best Result of Classification

Based on the explanation in the previous section, the selected decision tree is the decision tree in fold 3 from the C5.0 algorithm. The decision tree (in the form of rules) is going to be used for the implementation of web-based DSS.

### 3.6. Combining the Rules

The chosen rules for the common species identification are combined with the expert's recommendations for the rare species identification. The detailed information in the combination of the two rules of identification is as follow: One default class rule is added to the 38 chosen rules for identifying the common species of bats so that there are 39 rules to identify the common bat species. These rules are then combined with 11 rules from the expert's recommendation. There are 50 rules to be implemented in DSS to identify the common bat species (with morphological and morphometric attributes) and rare bat species (with morphological attributes). Besides the rules for the common species generated from the C5.0 algorithm, which have been described in the analysis, the example of the expert's recommendation of one rare species is as follow:

```
//spesies 26
        if ($model->ukuranbadan == 2 && $model-
>mata == 2 && $model->hidung == 2 && $model-
>garistelinga == 3 && $model->tonjolantelinga == 3 &&
$model->moncong == 2 && $model->leher == 3 &&
$model->sayap == 3 && $model->cakar == 1 &&
$model->rambutpunggung == 1 && $model-
>garispunggung == 2&& $model->corakpunggung == 2 &&
$model->ekor == 4 && $model->anus == 2
        ) {
            $this->redirect(array('/spesies/view', 'id' =>
$model2->id = 26));
        }
```

On the recommendation of the expert rules, can only accommodate morphological attributes to identify endangered species.

### 3.7. Justification, Inference and Expert Validation

The experts give a suggestion to give maximum and minimum scores for every input of the attributes on the initial dataset. Limiting the scores functions when users input data in the following format, minimum score>x> maximum score; the DSS is going to send notification that the input cannot be identified.

### 3.8. Visualization/DSS Front End Software and Development

The Yii web framework is selected as the software for the implementation of DSS visualization. The software used in the study are the 5.4.0 PHP language programming, Yii version 2, DBMS MYSQL, Windows and Linux. The purpose of the stage is to design and develop database. DSS database consists of 19 tables, 14 tables for morphological attribute details, one table which consists of all morphological attributes of the field, 1 table which consists of two common and rare species categories, one table of classification which contains morphometric and morphological field, one table of species which consists of field detail of the species, and one user table which gives information about level of DSS users. Because the dataset of the bats consists of the one of adult bats only, the database is limited to the data of the adult bats. The implementation of the combined rules of DSS is at the "KlasifikasiController.php" file, while the implementation of maximum and minimum scores of each attribute is located at the "Klasifikasi.php" file.

### 3.9. Developing Public User Interface and Administrator Interface

The DSS has two types of users; the first type is public user who can only access three main menus namely "Home" that consists of the main menu of bat identification, "Bat Species" that consists of lists of bat species DSS identifies, "About" that consists of brief description about DSS and the second type is administrator who have full access of all DSS features as long as he/she has logged into the application. There are two menus that only administrator has access to; "Species Data Management" that is able to update the data of the bat species and "Data Attribute Management" that can update data attribute of the bats. Figure 2 gives some general description on the system being developed.

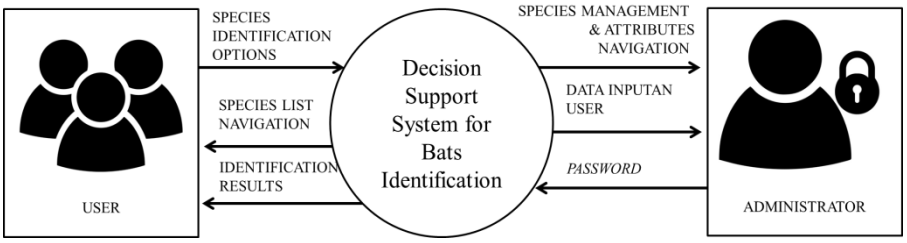


Figure 2. DSS for Bat Identification Contextual Diagram

DSS outlook is developed using responsive outlook so that it can be accessed using various mobile devices since responsive outlook will adjust to size of mobile device screen that will help users to use the application. Figure 3 is example of DSS outlook.

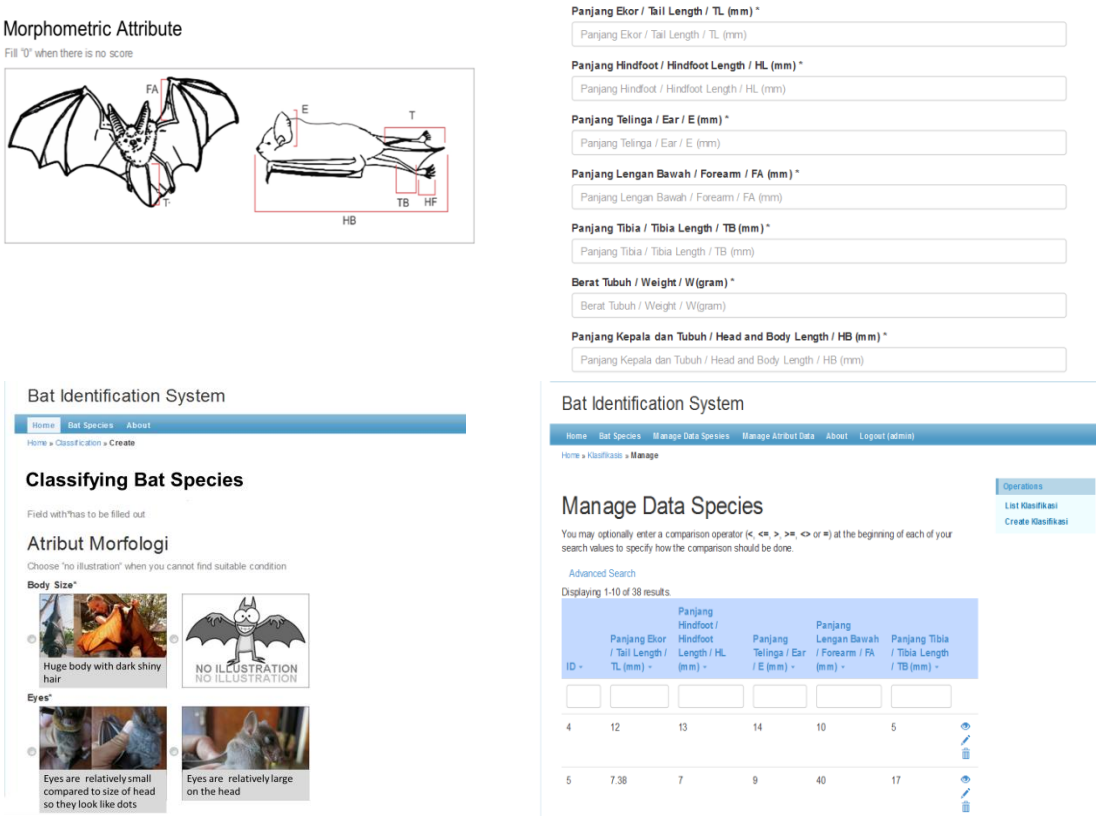


Figure 3. DSS Interface for Bats Identification

### 3.9. Decision Making based on DSS Output

Testing is carried out four times. The first testing is to input data but limited to morphometric attributes. The results are unidentified (notification for morphological field is not complete) because, in reference to the rules, there is no species that can be identified using morphometric attributes. The second is to input data but limited to morphological attributes only. Morphological attributes of the bats can successfully identify the rare species of bats because based on the rules, rare bat species can only identified based on their morphological attributes. The third test is to input both morphometric and morphological attributes data. The result of identification is both common and rare species. The final test is to input both morphometric and morphological attributes data using (minimum score>x> maximum score) format. The result is unidentified (notification to fill out morphometric field with the ”minimum score>x> maximum score” format) since, in reference to the rules, there is no species that can be identified based on the morphometric attribute only (Table 3).

Table 3. Results of the testing on Decision Making

| Testing | Morphological Data Input | Morphometric Data Input | Result of Identification |
|---|---|---|---|
| Testing 1 | No | Yes | Notification/ unidentified |
| Testing 2 | Yes | No; x=0 | Rare Species |
| Testing 3 | Yes | Yes | Common/Rare Species |
| Testing 4 | Yes | Yes (minimum score>x> maximum score) | Notification/ unidentified |

## 4. Conclusion

The Random Forest and C5.0 algorithms are applicable and able to find out characteristics and identify common species of bats. The C5.0 algorithm is effective and chosen for the study because it has 98.98% accuracy rate, while the Random Forest algorithm has accuracy rate of 97.26 %. The Decision Support System (DSS) is developed using the

combination of the findings of the C5.0 algorithm and expert's recommendation. The DSS has been tested using morphological and morphometric data inputs and has been able to identify common and rare species of bats successfully. To combine the algorithm and the expert's recommendation, one default class rule is added to the 38 selected rules of the species so that there are 39 rules to identify the common species of bats to be implemented in the DSS. Next, the 39 rules are combined to 11 rules from the expert's recommendation so the total rules to be implemented in DSS to identify common species (with morphological and morphometric attributes) and rare species (with morphological attributes) of the bats are 50 rules. The DSS testing has proven that DSS is able to identify both common and rare bat species when users input data about morphological attribute, morphometric attribute, as well as both morphological and morphometric attributes of the bats.

## References


[1] P Fedor, I Malenovsky, J Vanhara, W Sierka, J Havel. Thrips (Thysanoptera) identification using artificial neural networks. *Bull. Entomol. Res.* 2008; 98(05): 437-447.

[2] TN Ananthakrishnan. Perspectives and dimensions of phenotypic plasticity in insects. *Insect Phenotypic Plast. Divers. Responses. Sci. Publ. Enfield, NH*. 2005: 1-23.

[3] TA Ransaleleh, RRA Maheswari, P Sugita, W Manalu. Identifikasi Kelelawar Pemakan Buah Asal Sulawesi Berdasarkan Morfometri (The Morphometric Identification of Celebes Fruit Bats. *J. Vet.* 2013; 14(4).

[4] J Vaňhara, N Muráriková, I Malenovsk\`y, J Havel. Artificial neural networks for fly identification: A case study from the genera Tachina and Ectophasia (Diptera, Tachinidae). *Biologia (Bratisl).* 2007; 62(4): 462-469.

[5] YA Kartika. Pengenalan Jenis Katak dan Kodok Berdasarkan Ciri Bentuk dan Penklasifikasi dengan Pendekatan Statistik, Fuzzy dan Jaringan Syaraf Tiruan. University of Indonesia. 2011.

[6] AH Serna, LFJ Sequra. Automatic identification of species with neural networks. *Peer J*. 2014; 2: 563.

[7] T Lucas. Bat Identification with Gaussian Process Learning. 2011.

[8] A Prawesti. Sistem Pakar Identifikasi Varietas Ikan Mas (Cyprinus carpio) Berdasarkan Karakteristik Morfologi dan Tingkah Laku. 2013.

[9] N Larios, B Soran, LG Shapiro, G Martínez-Muñoz, J Lin, TG Dietterich. *Haar Random Forest Features and SVM Spatial Matching Kernel for Stonefly Species Identification*. In ICPR. 2010; 1(2): 7.

[10] M Immitzer, C Atzberger, T Koukal. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data. *Remote Sens.* 2012; 4(9): 2661-2693.

[11] K De Bruyne, B Slabbinck, W Waegeman, P Vauterin, B De Baets, P Vandamme. Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. *Syst. Appl. Microbiol.* 2011; 34(1): 20-29.

[12] HR Huber, JC Jorgensen, VL Butler, G Baker, R Stevens. Can salmonids (Oncorhynchus spp.) be identified to species using vertebral morphometrics?. *J. Archaeol. Sci.* 2011; 38(1): 136-146.

[13] C Guisande, A Manjarrés-Hernández, P Pelayo-Villamil, C Granado-Lorencio, I Riveiro, A Acuña, E Prieto-Piraquive, E Janeiro, JM Matías, C Patti, IPez: an expert system for the taxonomic identification of fishes based on machine learning techniques. *Fish. Res.* 2010; 102(3): 240-247.

[14] ME Barkworth, DR Cutler, JS Rollo, SWL Jacobs, A Rashid. Morphological identification of genomic genera in the Triticeae. *Breed. Sci.* 2009; 59(5): 561-570.

[15] DR Cutler, TC Edwards Jr, KH Beard, A Cutler, KT Hess, J Gibson, JJ Lawler. Random forests for classification in ecology. *Ecology*. 2007; 88(11): 2783-2792.

[16] DW Armitage, HK Ober. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecol. Inform.* 2010; 5(6): 465-473.

[17] K Zhang, B Hu. Individual urban tree species classification using very high spatial resolution airborne multi-spectral imagery using longitudinal profiles. *Remote Sens.* 2012; 4(6): 1741-1757.

[18] IS Sitanggang, R Yaakob, N Mustapha, AN Ainuddin. A decision tree based on spatial relationships for predicting hotspots in peatlands. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2014; 12(2): 511-518.

[19] P Thariqa, IS Sitanggang, L Syaufina. Comparative Analysis of Spatial Decision Tree Algorithms for Burned Area of Peatland in Rokan Hilir Riau. *Indonesian Journal of Electrical Engineering.* 2016; 14(2): 684-691.