VOL. 12, NO. 2, JANUARY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ENSEMBLE BASED MAJORITY VOTING FOR POINT-TO-POINTMEASUREMENTS OF *GYRODACTYLUS* SPECIES IDENTIFICATION

Rozniza Ali[1], Amir Hussain[2] and Andrew Abel[2]
[1]School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Malaysia
[2]Institute of Computing Science and Mathematics, University of Stirling, United Kingdom
E-Mail: rozniza@umt.edu.my

ABSTRACT

In the 21st Century, a key challenge in both wild and cultured fish populations for control and management of disease is to securely and consistently perform pathogen identification. To provide automated accurate classification for the challenging *Gyrodactylus* species, we introduce an ensemble based majority voting approach for their classification. In this system, an ensemble classification approach is created that utilises a combination of multiple feature sets and classifiers for *Gyrodactylus* species identification. The classifier base makes use of K-Nearest Neighbor (K-NN) and Linear Discriminant Analysis (LDA) approaches; with three different feature sets used for successful multi-species classification, considering 25 point-to-point data measurements, as well as smaller feature sets chosen using different feature selection techniques. The results show that our proposed ensemble based approach is accurate and robust, with ensemble based majority voting of classifiers and feature sets together found to be more effective than only combining feature sets.

Keywords: gyrodactylus, classification, feature selection, ensemble, majority voting.

INTRODUCTION

The continued worldwide expansion of Aquaculture has been accompanied by increased disease problems. Of particular interest are the disease issues related to the spread of ectoparasitic monogenean worms.

In this paper, the focus is on Monogenea of genus *Gyrodactylus.* These are widespread and can inhabit marine, brackish, and fresh water environments. Specifically, *G. salaris* primarily lives and reproduces in fresh water environments, However, it can tolerate brackish water for different time periods, depending on the salinity levels present. In several countries, *Gyrodactylus* salaris (*G. salaris*) has the reputation and nickname of 'the salmon killer' due to its significant impact on the mortality rate of salmon. Bakke *et al.* (Bakke, *et al.* 2007) stated that previously there has been a limited focus on this topic due to the complexity of *G. salaris* taxonomy. However, due to the effect on Atlantic salmon of *G. salaris* research interest has increased to such a level that gyrodactylids are now the most heavily researched type of monogeans.

Gyrodactylus are generally very small (<1mm), ectoparasitic monogenetic fish flukes (Harris, *et al.* 2004), with over 440 known species. The majority of *Gyrodactylus* species are non-pathogenic, which means that their presence causes very little harm or damage to the fish that hosts them, however this is not always the case.

There is incomplete knowledge of the majority of *Gyrodactylus* species, with descriptions often limited to an incomplete morphological description of their attachment hooks. In recent years, molecular techniques have contributed significantly to species discrimination (Cunningham, *et al.* 1995), however *Gyrodactylus* definitions often still rely on morphological characteristics (i.e. the morphology of attachment hooks, focusing on the sickle shape of the 16 small peripheral marginal hooks, due to these being considered the most important taxonomic feature) (Malmberg, 1970).

The expansion of managed fish culture into new environments due to continued decline in wild stocks because of anthropogenic activities, over-fishing, and other environmental changes, with the subsequent increase in ectoparasitic worms, has outstripped our ability to recognise a number of individual parasite pathogen species within a manageable timescale. There is also the issue of a wide variance in the pathogenicity seen between very closely related species and therefore, a key challenge in both cultured and wild fish populations is having the capability to securely and consistently identify pathogens.

As discussed previously, when it comes to species definitions, the key feature is often the morphological characteristics (i.e. the morphology of attachment hooks). These may be have a correlation with pathogenicity but research has identified that it is not always the case that there is a relationship between these characteristics and other recognised discriminatory molecular markers (Cunningham, *et al.* 1995). It has been found that some pathogenic species, including many monogeneans (whose discrimination from congeners can be particularly difficult due to the relatively small number of discrete morphological characteristics) can only be classified separately from other related non-pathogenic species by the process of morphological characterisation. This has the result of species classification being extremely time consuming and challenging.

Some of the key challenges include the often small size of specimens, there only being very small differences in morphology in some of the most important taxonomic features, a lack of distinctive colouration and patterning, and also the fragile nature of some specimens, which means they have to be living or immediately preserved in order to be classified. Kay *et al.* (Kay, *et al.* 1999) provide a detailed discussion of the challenges associated with parasite classification by shape because of the often very slight differences between each species.

This paper presents a new method to alleviate this problem by creating an automated classification approach that uses Ensemble Classification and Majority Voting.

In the remainder of this paper, the background section provides further background information into the research problem. The Morphometric Dataset section presents a full description of the data collection process using the point-to-point measurement approach. After this, Ensemble Classification and Majority Voting are introduced in the following section. The proposed method, Ensemble Classification for *Gyrodactylus* Species Identification is the presented. The Results are presented and discussed in the following section and finally, the paper is concluded.

BACKGROUND

There are a range of significant multi-disciplinary issues with regard to classification of the *Gyrodactylus* species group. Firstly, manual classification is highly time consuming and labour intensive, and related to this, there is the issue of access to a level of expert knowledge that is capable of successfully distinguishing superficially similar species such as *G. arcuatus*, and *G. salaris* from each other. Thirdly, given the small size of the specimens, the error free acquisition of the required point-to-point measurements can be extremely challenging. If the measurements are incorrect, this can lead to inaccurate classification. Finally, when the final classification decision has to be made, only specialist domain experts can determine this, with the aid of their own expertise, experience, and vision.

To assist with accurate classification of Gyrodactylus species, machine learning techniques can be used to recognise and classify an image of a specimen, which will provide a big contribution with regard to parasite recognition and classification. To date, many different algorithms that can be used for classification and detection have been proposed, including parasite detection in fish (Choudhury & Bublitz, 1994), identification of mammalian species (Moyo, *et al.* 2006), leaf species recognition (Du, *et al.* 2007), and IHHN virus detection in shrimp tissue with the use of digital colour correlation (Alvarez-Borrego & Chavev-Sanchez, 2001).

To improve *G. salaris* identification, which as discussed earlier, is time consuming and extremely challenging for even domain experts to do manually, a number of morphometric techniques have been proposed that aim to distinguish this particular species from others that are closely related and can also be found on salmonids. These include molecular techniques (Cunningham, *et al.* 1995), (Cunningham, *et al.* 1995), (Meinila, *et al.* 2002), (Hansen & Bachmann, 2002), and also approaches based on statistical classification (Kay, *et al.* 1999), (McHugh, *et al.* 2000), (Shinn, *et al.* 2000).

MORPHOMETRIC DATASET

In this paper, we utilise a morphometric dataset based on point-to-point feature extraction. This was collected and prepared by the Parasitology Laboratory at the Institute of Aquaculture, University of Stirling. The specimens were collected from a wide geographic range. In this paper, nine species of *Gyrodactylus* in were used for investigation. These are *G. derjavinoides, G. arcuatus, G. salaris, G. kherulensis, , G. sommervillae, G. thymalli, G. truttae*, *G. cichilidarum, and G. gasterostei*. In total, 557 specimens from nine different *Gyrodactylus* ectoparasite groups were sampled, using light microscopy.

As discussed in Shinn *et al.* (Shinn, *et al.* 2000), the main method of attaching to the host species is with the opisthaptor of the attachment hook. This is also considered to be the most significant feature for identification of distinct *Gyrodactylus* Malmberg species. In this paper, feature information was extracted from the tree part of the opisthaptor attachment hooks; hamuli, ventral bar and marginal hook.

A dataset of Morphometric features containing 25 point-to-point measurements was measured from glass slide mounted specimens. The number of each species is presented in Table-1. This table shows the distribution of the multiple species of *Gyrodactylus* in the dataset. Some species have more than 50 specimens (e.g.: *G. salaris*), while others have much less (e.g.: *G. arcuatus*)

Table-1. Detailed breakdown of the *Gyrodactylus* species and their number of specimens.

Gyrodactylus species name	Specimen numbers
G. arcuatus (a)	24
G. derjavinoides (d)	137
G. gasterostei (g)	30
G. kherulensis (k)	30
G. salaris (s)	71
G. sommervillae (sm)	30
G. thymalli (t)	85
G. truttae (tr)	80
G. chichilidarum (c)	70
Total	557

In the dataset, all measured point-to-point distances are categorised as scale type data and measured in micrometres (µm). 6 were taken from the ventral bar which spans the two hamuli (anchors), 11 from one of the paired central hamuli, and the remaining 8 were measured from one of the 16 peripheral marginal hooks. The total list of point-to-point measurements is given in Table-2.

Table-2. Detailed list of all point-to-point measurements.

Measured Object	Measurement Name
ventral bar	total length (VBTL) total width (VBTW) process-to-mid length (VBPML) median length (VBML) process length (VBPL) membrane length (VBMBL)
hamulus	total length (HTL) point length (HPL) shaft length (HSL) root length (HRL) aperture distance (HAD) proximal shaft width (HPSW) inner angle (HIA) distal width (HDSW) inner curve length (HICL) aperture angle (HAA) point curve angle (HHPCA)
Marginal hook	total length (MHTL) shaft length (MHSHL) sickle length (MHSL) sickle proximal width (MHSPW) sickle distal width (MHSDW) sickle toe length (MHSTL) sickle aperture (MSHAD) instep height (MHIH).

ENSEMBLE CLASSIFICATION

Single classifier based approaches discussed in previous research (Ali, *et al.* 2011), investigated the use of a range of single classifiers and different single feature sets. It was found that none of these approaches (i.e., those utilising one feature set combined with only one classifier) produced significantly better Gyrodactylus classification results. The key limitation of using a single classifier together with a single feature set is that this may only learn to completely correctly classify certain species within the subset of possible species, whereas a different classifier may completely classify other species, and generate different errors. It is therefore expected that the combined use of different feature and classifier sets for Gyrodactylus species identification will allow for robust classification performance and overcome weaknesses with individual classifiers.

As discussed by Yu and Xu (Yu & Xu, 2011), the use of an ensemble classification approach has merits due to difficulties with producing fully accurate classification with single classifier approaches, for a number of reasons, with one important scenario being when a large dataset, or one with a large number of features or data points, is used. In order to be able to learn interpretable multi-target models capable of classifying several classes simultaneously, the FIRE (Fitted Rule Ensembles) method was proposed (Aho, *et al.* 2009), which can learn multi-target regression rule ensembles. Published results confirm that in general, there is a trend of larger models being more accurate.

With regard to multi-disciplinary research, one previous example is field line proteomic mass spectra classification (Geurts, *et al.* 2005), which proposed a decision tree ensemble based systematic approach to determine proteomic biomarker and predictive models, with promising results and more efficient processing times.

MAJORITY VOTING

In a range of relevant examples of existing ensemble models, majority voting has been applied. Based on their research, Kainulainen (Kainulainen, 2010) recommend the use of majority voting if the output consists of class labels, and therefore, for the research presented in this paper, this approach was followed for *Gyrodactylus* species classification. We also use simple majority voting, which is a decision rule that selects one of a number of alternative choices, by choosing the prediction that has the most "votes" (Kim, *et al.* 2011). To demonstrate this, given a hypothetical ensemble consisting of three different classifiers, which we call $h1$, $h2$ and $h3$, and a sample x to be classified, classification is first performed with all three. In one scenario, $h2(x)$ and $h3(x)$ may both classify correctly but $h1(x)$ produces an incorrect result. In this scenario majority voting will correctly classify x by counting the votes of each classifier and selecting the majority opinion. In theory, if the errors that the classifiers make are independent (i.e. have misclassifications with different labels), the majority vote should outperform the best single classifier.

An example of this was proposed by Bouziane (Bouziane, *et al.* 2011) , who applied majority voting for predicting the secondary structure of globular proteins. They combined Artificial Neural Networks (ANNs) and K-Nearest Neighbor (K-NN) with Multi-class Support Vector Machines and used these classifiers to compare three different strategies for voting; Influence Majority Voting (IMV), Weighted Majority Voting (WMV), and Simple Majority Voting (SMV).

ENSEMBLE CLASSIFICATION FOR *GYRODACTYLUS* SPECIES IDENTIFICATION

For *Gyrodactylus* species identification, a number of combinations of feature sets and classification methods have been considered for the creation of an ensemble based approach. Three different feature sets have been utilised, and these are used two different classifiers, K-NN and Linear Discriminant Analysis (LDA), which together comprise the classifier base. These were chosen based on prior research (Ali, *et al.* 2011), which identified that when comparing LDA and K-NN approaches to Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) methods, the resulting misclassifications were not noticeably different, meaning that integrating all four classification methods into one ensemble held little additional value as the individual classifiers did not all make independent errors. The success of an ensemble system is dependent on being able to correct the errors made by individual classifiers within it (Kainulainen, 2010). If all classifiers provide the same output (i.e. are not independent of each other), it is impossible to correct individual mistakes.

In addition, as MLP and SVM are non-linear approaches, it can be argued that the chosen (linear) classifiers are less computationally complex than implementing additional non-linear approaches.

The chosen features are the 25 full feature point-to-point measurement set, 21 selected features using Sequential Forward Selection (SFS) (Karagiannopoulos, *et al.* 2007), (Gheyas & Smith, 2010) and 20 selected features from Sequential Backward Selection (SBS) (Karagiannopoulos, *et al.* 2007), (Kolodyazhniy, *et al.* 2011). Majority voting (Kim, *et al.* 2011) is implemented to combining the classifiers and feature sets.

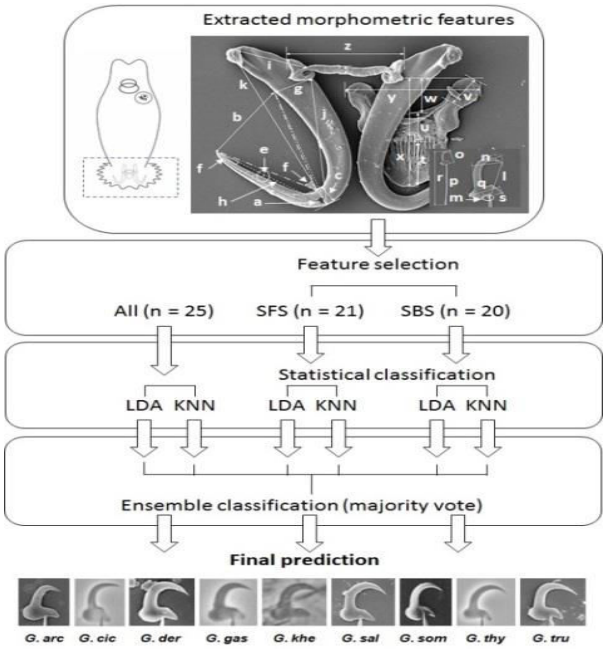

Figure-1. Proposed ensemble based feature selection and majority voting approach for *Gyrodactylus* species classification.

The proposed ensemble voting framework is presented in Figure-1. This consists of:

1. **Feature selection**. 25 features were extracted from SEM images using manual point-to-point measurement techniques, as discussed previously. From these, two feature selection techniques are then applied to produce two further feature-sets.
2. **Classification**. 2 classifiers are used in this paper, LDA, and K-NN.
3. **Voting system**. Prior research found that the different classifiers produced different misclassification results with the use of different feature sets. These results are used as an input into the voting system, and are not individually weighted to adjust the contributions of the classifier to the voting system, but all combinations of classifier and feature set are weighted equally.

After the statistical step, to maximise specimen classification accuracy, a majority voting based ensemble is applied which combines the individual classifier and feature set combination results. The complete methodology is presented algorithmically as follows:

1. Given $D = X_1Y_1, \ldots, X_nY_n$ where$Y_j \in (1,2,\ldots,9)$.
2. Carry out feature selection, retaining the original feature set of 25 measurements.
   a. Sequential forward selection (SFS), choosing 21 of the 25 original measurements. Beginning with an empty feature set, this method sequentially adds the feature (A^+) that maximises $E(B_j + A^+)$ when combined with the features B_j that have previously been added to the feature set.
   i. Begin with empty set $B_0 = (\emptyset)$.
   ii. Identify the best feature $A^+ = arg\ max_{A \notin B_j} E(B_j + A)$ that has not alreadt been chosen.
   iii. Update $B_j + A^+; j = j + 1$.
   iv. Return to step ii, and repeat this process until the optimum set of features has been identified.
   b. Sequential backward selection (SBS), choosing 20 measurements to use as features. Here, beginning with a the full feature set, remove the feature A^- that has the least negative effect on the objective function$F(B - A\text{-})$.
   i. Begin with set $B_0 = A$.
   ii. Eliminate the least relevant feature $A^- = \arg max_{A \in B_j} F(B_j - A)$.
   iii. Update $B_{j+1} = B_j - A^-; j = j + 1$.
   iv. Return to step ii, and repeat this process until the optimum set of features has been determined.
3. Ensemble classification is then performed with K-NN and LDA, using the three different sets of features, SFS, SBS, and the original full set of 25 measurements. Each set is then used as an input into the two classifiers:
   a. LDA
   $$\widehat{N}_i = W_0 + W_{i1}D_1 + W_{i2}D_2 + \cdots + W_{in}D_n \quad (1)$$
   b. K-Nearest Neighbor (KNN)
   $$(X_t, X_u) = \sqrt{\sum_{mm=1}^{m} (X_t^{mm} - X_u^{mm})^2} \quad (2)$$
4. These individual results are then combined in the majority voting ensemble system, combining the output from both the K-NN and LDA classifiers, as given with:
   $$Q = \arg max_y \sum_{t=1}^{T} I(\hat{H}_t(D) = y), y \in Y \quad (3)$$
5. In the scenario where there is an equal split in the voting calculation, meaning that a simple majority is not identified, the result of SFS-LDA is given the casting vote, due to this technique being shown to outperform all others in an earlier study by Bakke *et al*. (Bakke, *et al.* 2007)).
6. The identity of the specimen Q is then determined as being one of the species listed in Table-1

RESULTS AND DISCUSSION

To demonstrate our proposed ensemble based majority voting algorithm, the nine class morphometric point-to-point measurement dataset discussed earlier is used for all experiments, and the three point-to-point

measurement feature sets discussed in the previous section are selected.

Rather than split the dataset into test, validation, and training subsets, we utilise 10-fold cross validation. This is because of the unbalanced number of specimens present per species (see Table-1), and consequently, other research has recommended the use of cross validation (Refaielzadeh, *et al.* 2008). The complete dataset was therefore randomly split into *b (10)* subsets, with *B-1* subsets allocated for training, and the final one being the testing subset. This was repeated 10 times (with all subsets being recalculated accordingly), and from this the overall average performance was calculated. For the purposes of statistical classification, the representation of 10-fold cross validation is given by *10-fold = accuracy/b*. Here, *accuracy* is the number of correct classifications in *b* experiments. First, the individual classifier results are given in Table-3.

Table-3. Average result of species identification between individual classifications.

Feature Set	Individual Classifier			
	LDA (%)	K-NN (%)	MLP (%)	SVM (%)
Original feature (25F)	96.38±1.95	95.32±2.71	97.67±2.33	96.41±2.34
SFS (21F)	96.74±1.69	95.34±2.55	96.59±2.37	96.59±2.16
SBS (20F)	96.55±1.95	93.89±2.11	97.13±2.17	96.95±2.12

Table-3 summarises individual classifier performance (including the classifiers not chosen to be part of the ensemble) with the different feature sets, the results are compared to identify the best performance.

To evaluate the misclassification error in more detail, a confusion matrix is used. A confusion matrix of size $n \times n$ for a single classifier provides the classification results, with n being the number of different possible classifications present in the original dataset. (Freitas, *et al.* 2007), (Sofia Visa, *et al.* 2011). Two examples of individual classifier performance are given in Table-4 and Table-5.

Table-4. Confusionmatrix for the original set of 25 measurements, combined with an MLP.

	a	*c*	*d*	*G*	*K*	*s*	*sm*	*t*	*tr*
a	**24**	0	0	0	0	0	0	0	0
c	0	**70**	0	0	0	0	0	0	0
d	0	0	**131**	0	0	2	1	0	3
g	0	0	0	**30**	0	0	0	0	0
k	0	0	0	0	**30**	0	0	0	0
s	0	0	0	0	0	**71**	0	0	0
sm	0	0	0	0	0	0	**29**	0	1
t	0	0	2	0	0	0	0	**80**	1
tr	0	0	1	0	0	0	0	0	**79**

Table-5. Confusion matrix for the SFS feature set, combined with an LDA classifier.

	a	*c*	*d*	*G*	*K*	*s*	*sm*	*t*	*tr*
a	**24**	0	0	0	0	0	0	0	0
c	0	**70**	0	0	0	0	0	0	0
d	0	0	**129**	0	0	1	0	1	6
g	0	0	0	**30**	0	0	0	0	0
k	0	0	0	0	**30**	0	0	0	0
s	0	0	0	0	0	**67**	0	3	1
sm	0	0	0	0	0	0	**30**	0	0
t	0	0	1	0	0	0	0	**82**	2
tr	0	0	3	0	0	0	0	0	**77**

It is not easy to accurately predict FS performance of multiple species. As discussed previously, some classifiers manage to accurately classify different species completely correctly with different feature sets, and with different errors identified in a number of these. For reasons of space, we show some examples in Table-4 and Table-5 of different feature set results. An investigation identified that certain features could cause the boundaries distinguishing species from each other to be obscured, and were therefore not always suitable for use for classification, but this was not always consistent. This therefore justified the use of feature selection, and also the decision to make use of several different feature sets.

Due to there being remaining species that are notfully classified by a single classifier, an ensemble method is therefore justified.

Using the combination of classifiers and features sets as part of the ensemble discussed previously in the paper, the results of all individual evaluations are input into the ensemble voting method to determine the final classifier output.It was then calculated that the overall accuracy of the ensemble based approach is 97.29% ±1:98. The overallconfusion matrix of the ensemble resultsareshown in Table-6. The results show that there remain a small number of misclassification, with 15 individuals from 7 different species remaining misclassified.

Table-6. Confusion matrix of proposed ensemble model.

	a	*c*	*d*	*G*	*k*	*s*	*sm*	*t*	*tr*
a	**24**	0	0	0	0	0	0	0	0
c	0	**70**	0	0	0	0	0	0	0
d	0	0	**130**	0	0	2	0	0	5
g	0	0	0	**30**	0	0	0	0	0
k	0	0	0	0	**30**	0	0	0	0
s	0	0	0	0	0	**68**	0	2	1
sm	0	0	0	0	0	0	**30**	0	0
t	0	0	1	0	0	0	0	**83**	1
tr	0	0	3	0	0	0	0	0	**77**

An analysis of the results presented in Table 6 shows that when ensemble based majority voting is applied, of the 70 *G. salaries* examples in the dataset, 68 are identified correctly, but on two occasions, *G. salaries*

was identified as *G. tnymalli*. Regarding the *G. thymalli* classification results, there were only two misclassification results (83 out of 85 correct), with *G. thymalli*being identified once as *G. derjavinoides* and once as *G. truttae*. In addition, misclassification errors remain for *G. derjavinoides* and *G. truttae*.

It can be concluded that combining a number of feature sets together with a number of classifiers in an ensemble, as presented in this paper, is arguably more effective than only combining features sets with individual classifiers. However, the misclassification errors have not been fully eliminated. Although the misclassification errors do not show a significant improvement when compared to individual optimised feature sets and classifiers, the number of errors has been minimised using our ensemble approach.

For better prediction and classification, other techniques in addition to ensemble based majority voting could be applied in future research. This would investigate further improvements and reduce the remaining misclassifications. There still remains the issue of certain classifiers being very good at classifying certain species, while performing poorly with other species, and a lack of complete independence in the results (which results in poor ensemble performance that majority voting does not always solve well). One potential approach to this would be to implement multi-target regression with rule ensembles (Aho, *et al.* 2009). This approach combines decision trees into a single large collection of rules, which can then be optimised to identify the best rule subset, and a suitable weighting to be used.

CONCLUSIONS

This paper presented a majority voting ensemble system that successfully integrates a range of previously evaluated classifiers and feature selection techniques to improve on the classification robustness identified using a single feature selection approach. We first presented a detailed background and review to the problem, as well as the dataset used, before presenting a full description of our new proposed approach. The results were presented in the Results and Discussion section, and show that our ensemble based approach was found to be accurate and robust.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the collaborative project between Universiti Malaysia Terengganu, Malaysia and Stirling University, United Kingdom. To the team of parasitology, Stirling University, thank you for allowing the use of *Gyrodactylus* dataset in this study. Andrew Abel is funded by EPSRC project EP/M026981/1.

REFERENCES

[1] Aho, T., Zenko, B. & Dzeroski, S., 2009. Rule ensemble for multi-target regression. Data Mining, pp. 21-30.

[2] Alvarez-Borrego, J. & Chavev-Sanchez, M. C., 2001. Detection of IHHN virus in shrimp tissue by digital color correlation. Aquaculture, Volume 64, pp. 161-376.

[3] Bakke, T. A., Cable, J. & Harris, P. D., 2007. The biology of gyrodactylid monogeneans: the "Russian-doll killers". Advances in Parasitology, Volume 64, pp. 161-376.

[4] Bouziane, H., Messabih, B. & Chouarfia, A., 2011. Profiles and majority voting-based ensemble method for protein secondary structure prediction. Evolutionary Bioinformatics, Volume 7, pp. 171-189.

[5] Choudhury, G. S. & Bublitz, C. G., 1994. Electromagnetic method for detection of parasite in fish. Aquaculture Food Producr Technology, 185(2), pp. 883-893.

[6] Cunningham, C., McGillivray, D., MacKenzie, K. & Melvin, W., 1995. Identification of Gyrodactylus (monogenea) species paratizing salmonid fish using DNA probes. Fish Disease, Volume 18, pp. 539-544.

[7] Cunningham, C. O., McGillivray, D. M., MacKenzie, K. & Melvin, W. T., 1995. Discrimiantion between Gyrodactylus salaris, G. derjavini dan G. truttae (Platyhelminthes: Monogenea using restriction fragment length polymorphisms and an logonuctide probe within the small subunit ribosomal RNA genes. Parasitology, Volume 111, pp. 87-94.

[8] Du, J. X., Wang, X. F. & Zhang, G. J., 2007. Leaf shape based plant species recognition. Applied Mathematics and Computation, 185(2), pp. 883-893.

[9] Freitas, C. O. A., de Carvalho, J. M. & Oliveira, J. J., 2007. Confusion matric disagrement for multiple classifiers. Progress in Pattern Recognition, Image Analysia and Applicationd.

[10] Geurts, P. *et al.* 2005. Proteomic mass spectra classification using decision three based ensemble methods. Bioinformatics, 21(15), pp. 3138-3145.

[11] Gheyas, I. A. & Smith, L. S., 2010. Feature Subset Selection in large Dimensionality domains. Pattern Recognition, Volume 437, pp. 5-13.

[12] Hansen, H. & Bachmann, L. A., 2002. Mitochondrial DNA variation of Gyrodactylus spp. (Monogenea, Gyrodactylidae) populations infecting Atlantic salmon, grayling and rainbow trout in Norway and Sweden. Parasitology, Volume 33, pp. 1471-1478.

[13] Harris, P. D., Shinn, A. P., Cable, J. & Bakke, T. A., 2004. Nominal species of the genus Gyrodactylus v. Nordmann 1832 (Monogenea: Gyrodactylidae), with

a list of principal host species. Systematic Parasitology, Volume 59, pp. 1-27.

[14] Kainulainen, L., 2010. Ensemble of locally linear models: Application to bankcruptcy prediction. Data Mining, pp. 280-286.

[15] Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B. & Pintelas, P. E., 2007. Feature selection for regression problems. Hellenic European Research on Computer Mathematics & Its Applications, pp. 20-22.

[16] Kay, J. W., Shinn, A. P. & Sommerville, C., 1999. Towards an automated system for the indentification of notifiable pathogens using Gyrodactylus salaris as an example. Parasitology Today, 15(5), pp. 210-203.

[17] Kim, H., Kim, H., Moon, H. & Ahn, H., 2011. A weight-adjusted voting algorithm for ensemble of classifiers. Korean Statistical Society, 40(4), pp. 437-449.

[18] Kolodyazhniy, K., John, G. H., Gross, J. J. & Roth, W. T., 2011. An effective computing approach to physiological emotion specificity: Toward subject-independent and stimulus independent classification of film-induced emotions. Psychophysiology, pp. 1-15.

[19] Malmberg, G., 1970. The excretory systems and marginal hooks asa basic for the systematics of Gyrodactylus (Trematoda, Monogenea). Arkiv for Zoologi Series 2-Band, 2(23), pp. 1-235.

[20] McHugh, S. E., Shinn, A. P. & Kay, J. W., 2000. Discrimination of G. salaris and G. thymalli using statistical classifiers applied to morphometric data. Parasitology, Volume 121, pp. 315-323.

[21] Meinila, M., Kuusela, J. & Zietara, M., 2002. Brief report primers for amplifyig 820 bp of highly polymorphic mitochondrial COI gene of Gyrodactylus salaris. Hereditas, Volume 137, pp. 72-74.

[22] Moyo, T., Bangay, S. & Foster, G., 2006. The identification of mamalian species through the classification of hair patterns uisng image pattern recognition. Computer Graphic, Virtual Reality, Visualisation and Interaction, pp. 177-181.

[23] Refaielzadeh, P., Tang, L. & Liu, H., 2008. Cross-validation. Arizona State University.

[24] Shinn, A. P., Kay, J. W. & Sommerville, C., 2000. The use of statistical classifier for the discrimination of species of the genus Gyrodactylus (monogenea) paratizing salmonids. Parasitology, Volume 120, pp. 261-269.

[25] Sofia Visa, A. R., Ramsay, B. & van der Knaap, E., 2011. Confusion matrix-based feature selection. CEUR Workshop Proceeding, Volume 710.

[26] Yu, H. & Xu, S., 2011. Simple rule-based ensemble classification for cance dna microarray data classification. Computer Science and Service System, pp. 2555-2558.