Gender Identification of *Sitophilus oryzae* using Discriminant Analysis and Support Vector Machine: A comparison study

T M F Azis\(^{1,2,*}\) and K F Kasim\(^{2,3}\)

\(^1\)Faculty of Applied Sciences, UiTM Cawangan Perlis, 02600 Arau, Perlis, Malaysia
\(^2\)Faculty of Chemical Engineering Technology, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia
\(^3\)Centre of Excellence for Biomass Utilization, Universiti Malaysia Perlis, 02600 Perlis, Malaysia

\*'E-mail: tunmohdfirdausbinazis@gmail.com

Abstract. *Sitophilus oryzae* (rice weevil) known as severe pest to many stored products, including rice. Gender density of *S. oryzae* is a critical information in predicting the rate of stored grain lost. However, the techniques used in gender identification of *S. oryzae* is a destructive technique which involved dissection to identify its reproductive organ. It was a tedious work and very time consuming. Thus, this study focused on the use of non-destructive technique which only based on numerical information of *S. oryzae* morphological features to identify its gender. The numerical information was analysed and tested against two model i.e. Discriminant Analysis Model and Support Vector Machine Model. The result show that rostrum length and width were adequate to be used in the Discriminant Analysis Model for gender identification of *S. oryzae* with 91% correct classification, while Support Vector Machine Model perform poorly in classification with 62% correct classification. Gender identification of *S. oryzae* using numerical information features were more accurate and liable compared to normal identification which based on the internal reproductive organ.

1. Introduction

*Sitophilus oryzae* is a rice weevils widely known as major stored product pest to a variety of stored products [1]. It feeds directly on the commodity, especially the seed [2] that causes weight loss of commodities. From the larval stages to adults stages, *S. oryzae* continue to cause severe damage to stored grain [3]. The female adult of *S. oryzae* made a hole in the grain and lay egg inside the hole. After the egg hatch, the larvae that emerge out will start to feed internally inside the grain [4]. The development of larvae inside the single grain consumes up to 60% of grain weight [5]. Thus, the higher the population density of *S. oryzae*, the higher rate of weight loss of commodities infested. Existence of *S. oryzae* also cause other problem such as contamination when they cast skins, webbing, and eliminating faecal material that reducing the market value of the commodities [6]. This insect infestation also encourages mould growth, including those fungi that produce mycotoxins [7].

Fumigation of stored grain by using methyl bromide and phosphine is currently the standard way to manage this insect pest [8]. However, the extensive use of pesticides increases the insect resistance and
also cause environmental pollution [9]. The European has prohibited the use of methyl bromide since January 2005 because of it may cause depletion of stratospheric ozone layer [8].

In order to overcome pesticide usage, Insect Pest Management (IPM) recommends a systematic approach to control pests, with the goal of maximising the use of a variety of methods to handle a whole range of pests in a specific crop environment[9]. The IPM framework was focused on limiting pesticide usage by using action thresholds and alternative control options such as biological products or methods [9]. This action threshold is a point where pest populations indicate that pest control action is a must [10].

In pest population assessment, sex in the population is important information for understanding population growth. Female proportion of S. oryzae in the population could badly affect the grain compared to male because of their ability to lay about 150 eggs in their lifetime inside the grain [11]. The eggs will develop into larvae inside the grain and further damage the rice grain [4,12]. Thus, if the number of female offspring is larger than a previous female generation, it shows that the population is in increasing condition and vice versa. Understanding male density is also essential, especially during sampling and monitoring [13], because it affects the reproductive fitness of insect. A study conducted by Flay et al. [14] shows that female genital tract was more likely to be damaged when male density increased. This damage then will lead to disruption of the female reproduction system and affect the population of the next generation. Therefore, it is crucial to identify the males and females of S. oryzae to understand and forecast its population growth. From the identification of gender sex ratio and total fertility rate easily been measure. Both of the information very crucial the forecast population growth, whether the pest population is decreasing, increasing or in stable condition [15,16].

Currently, there a limited mathematical model can be used to distinguish between the gender of S. oryzae. Rees [11] notes the males and females of S. oryzae are similar in appearance, but on rostra, males are shorter, broader and irregular. This knowledge is not numerically definite and not statistically confirmed. This research, therefore, concentrated on developing a model to discern S. oryzae sex depending on its morphological properties.

2. Methodology

2.1. Culture of S. oryzae

The starter culture of S. oryzae was collected from infested stored rice. It has been reared and maintained in a plastic container (13 x 10 x 20 cm) contained 2 kg of rice with the temperature between 15 to 34°C and relative humidity between 58 to 89%. The container was closed, and new rice was introduced periodically for proper development of the S. oryzae. This experiment was conducted at Biological Research Laboratory, Universiti Teknologi MARA Cawangan Perlis, Malaysia during 2019.

2.2. Morphological analysis

Forty-five adults of S. oryzae was collected from the starter culture and been killed using warm water. The morphological features of S. oryzae i.e. rostrum length and width, prothorax length and width, club length and width, and abdomen width at third and last sternite were measured using Olympus SZX7 Zoom Stereomicroscope [17]. The measurement then been analysed using independent sample t-test in order to evaluate the statistical different between morphological measurement of S. oryzae gender.

2.3. Prediction of S. oryzae Gender using Discriminant Analysis

Discriminatory analysis (DA) is a linear combination of two or more independent variables which can discriminate between classes [18]. DA aims at identifying the relative contribution of p variables to group separation, finding the optimal plane on which points may better illustrate the group configuration
[19]. It can be used with raw data to create a discriminating function (DF) for of group [20]. This DF can be computed with Equation 1

\[ f(G_i) = k_i + \sum_{j=1}^{n} w_j P_j \]  

where; 
- \( i \) = the number of group \( G \) 
- \( k_j \) = constant inherent to each group 
- \( n \) = the number of parameters used to classify a set of data into a given group 
- \( w_j \) = the weight coefficient assigned by discriminant function analysis (DFA) to a given parameter \( P_j \).

DA displays the relations between two or more predefined classes of substances focusing on a set of two or more separating variables [20]. In either case, DA was conducted using the norm for the first time. These have been used to build DFs that classify \( S. oryzae \) sex. The gender (male and female) of \( S. oryzae \) are the grouping (dependent) variables, while all the numerical information of \( S. oryzae \) morphological feature are independents variables. This then was analyzed using DA in XLSTAT 2019 [21].

In this analysis a total of 45 \( S. oryzae \) were used, where 35 were used in model development and classification assessments. Another 10 does not appear in the model, but as a validation sample for classification. There were five models checked. Model 1 consists of all combination of eight (8) morphological features in creating the model. Then, one by one dataset had been eliminated to observe the performance in differentiating the gender of the \( S. oryzae \) using the lesser morphological feature measurement. Model 2 consists of only rostrum width and length, Model 3 consists of only prothorax length and width, Model 4 consists of only third and last abdomen length and Model 5 consists of club width and length.

Two approaches had been used in assessing the overall fit of the model was used in the study to identify the variables which discriminate between gender of \( S. oryzae \). The Wilks Lambda value test for the standard mode gives a \( P < 0.0001 \). The null hypothesis stated that the means of vectors of the 2 clusters are equal. The alternative hypothesis, alongside, states that at least one of means of vector is different from another. Since the computed \( P \)-value is lower than the significance level of alpha = 0.05, one should reject a null hypothesis and accept the alternative hypothesis. The risk of rejecting the null hypothesis while it is true is lower than 0.01%. Thus, the 2 clusters are indeed different from one another.

The second assessment of overall model fit has been done by assessing the classification accuracy of the model. This classification accuracy had been converted into the form of the percentage. The performance of each model in predicting the gender of \( S. oryzae \) was evaluated using the area under the receiver operating characteristic (ROC) curve or also called as the area under the curve (AUC) using the XLSTAT software. The AUC ranged from 0 to 1. A value of 1 indicated a perfect model agreement. A value of 0.5 indicated equal chance and value of 0 indicated complete disagreement of the model [22].

2.4. Prediction of \( S. oryzae \) Gender using Support Vector Machine

The similar dataset was used in model development and classification assessments. The gender (male and female) of \( S. oryzae \) are the response (dependent) variables, while all the numerical information of \( S. oryzae \) morphological feature are independents variables.

3. Results and Discussion

3.1 Morphological analysis

Figure 1 shows the morphological features, i.e. rostrum length and width, prothorax length and width, club length and width, and abdomen width at third and last sternite of male and female \( S. oryzae \). The sex was established by the division of the abdomen and the male was shown by the presence of aedeagus, while the lack of aedeagus showed female \( S. oryzae \). Figure 2 shows the aedeagus that resulted from
dissection of the abdominal part of *S. oryzae*. The male was confirmed with the aedagus present, the female was confirmed by the absence of aedeagus. An aedeagus is a reproductive organ of male insect.

Figure 1. Morphological features *S. oryzae*

Figure 2. Structure of aedagus

Independent sample t-test was used to observe the significant different of the rhostrum length, abdomen width at 3rd sternite, club width and club length between *S. oryzae* gender and the P-value less than 0.05. It show that there were significant difference of this rhostrum length, abdomen width at 3rd
sternite, club width and club length structure between gender of *S. oryzae*(Table 1). While Rhostrum width, prothorax length, prothorax width, and abdomen width at last sternite, P-value more than 0.05 show there were no significant difference between gender of *S. oryzae* based on this morphological feature only. Therefore, its difficult to conclude which gender of *S. oryzae* is larger numerically.

**Table 1.** Morphological analysis of morphological features of *S. oryzae*

| Morphological features          | Gender       |           | P-value |
|--------------------------------|--------------|-----------|---------|
|                               | Female (Mean±SD) | Male (Mean±SD) |         |
| Rhostrum length               | 0.93±0.64    | 0.48±0.25 | 0.048*  |
| Rhostrum width                | 0.14±0.1     | 0.11±0.054| 0.19    |
| Prothorax length              | 0.91±0.63    | 0.62±0.26 | 0.059   |
| Prothorax width               | 0.89±0.64    | 0.66±0.3  | 0.128   |
| Abdomen width at 3rd sternite | 0.94±0.65    | 0.56±0.24 | 0.016*  |
| Abdomen width at last sternite| 0.63±0.41    | 0.45±0.19 | 0.065   |
| Club width                    | 0.11±0.07    | 0.07±0.03 | 0.027*  |
| Club length                   | 0.18±0.1     | 0.12±0.05 | 0.023*  |

* The mean difference is significant at the 0.05 level

In many insect species, the females tended to be larger and live longer compared to male. However, for some insect taxa, the body size dimorphism varies substantially. This can be seen in a most seed beetle (Chrysomelidae, subfamily Bruchinae) where females are larger than males, but in the genus Stator, male are generally larger than females [23]. For *S. oryzae* there are limited studies conducted so far showing the differences in male or female body size. The morphological characteristic of male and female are very identical, and there were no apparent differences in the body size of male and female of *S. oryzae*.

According to Omar [17], the main differences between male and female of *S. oryzae* are the side view of the ventral surface; where male have curved posterior while the female has a straight line. Another difference is the rostrum of the male is smaller and thicker and less curved compared to female. This can be prove in the Table 1 that show female rostrum length are longer that male while rostrum width are not significantly different between *S. oryzae* gender.

### 3.2 Discriminant Analysis Model

The assessment of overall model fit has been done by assessing the classification accuracy of the model. This classification accuracy had been converted into the form of the percentage (Table 2). Model 1 that consisted of all morphological features shows 91% correct classification for the classification estimation and 90% classification validation. Model 2 that consisted rostra width and length show 91% correct classification for the classification estimation and 100% classification validation. However, the model that consisted only prothorax, abdomen and club part, there percentages are below 75% correct classification in the estimation and validation sample. Meaning that about 30% chances of wrong identification will be made by using this model.

Another analysis done to check the performance of each model is by observing the area under the curve (AUC). The model that consisted of the only rostrum and all morphological feature had the highest area under the curve. This showed a perfect model agreement. For the model that consists prothorax, abdomen and club the area under the curve less than 0.8.
Table 2. Overall statistical measure of DA models predicting *S. oryzae* gender.

| Type of Statistical Measure | Model 1 (All) | Model 2 (Rostrum width & length) | Model 3 (Prothorax width & length) | Model 4 (3rd and last sternite of abdomen) | Model 5 (Club width and length) |
|-----------------------------|---------------|---------------------------------|-----------------------------------|------------------------------------------|--------------------------------|
| Wilks’ Lambda test         | 0*            | 0*                              | 0.223                             | 0.060                                    | 0.103                          |
| Correct classification (%)  |                |                                 |                                   |                                          |                                |
| Female                     | 84            | 84                              | 58                                | 74                                       | 68                             |
| Male                       | 100           | 100                             | 67                                | 60                                       | 80                             |
| Total                      | 91            | 91                              | 62                                | 68                                       | 74                             |
| Correct classification (%)  |                |                                 |                                   |                                          |                                |
| Female                     | 100           | 100                             | 40                                | 80                                       | 60                             |
| Male                       | 80            | 100                             | 60                                | 20                                       | 80                             |
| Total                      | 90            | 100                             | 50                                | 50                                       | 70                             |
| Performance of Model        | ROC (AUC)     | 1                               | 0.611                             | 0.709                                    | 0.779                          |

* The mean difference is significant at the 0.05 level

Based on the overall performance on each statistical measure, Model 1 and 2 that consist of all morphological features and rostrum, respectively, perform well in Wilks’ Lambda test with P-value less than 0.05, has the highest percentage of correct classification and performance of model based on the area under ROC curve. Model 3, 4, and 5 that consist of the prothorax, abdomen and club does not perform well. It also shows that Model 1 that consists of all morphological features, greatly influenced by the rostrum structure. By using elimination techniques, all other morphological features (prothorax, abdomen, and club) can be eliminated. Rostrum measurement is enough to be used to differentiate the gender of *S. oryzae*. Therefore, Model 2, which modelled using rostrum length and width, was chosen as the best model due to practicality and ease of use.

3.3 Support Vector Machine
The assessment of overall model fit has been done by assessing the classification accuracy of the model. This classification accuracy had been converted into the form of the percentage (Table 3). Model 1, 2, 4 and 5 shows 62% correct classification for the classification training sample and 20% classification validation. While Model 3 that consisted prothorax width and length show 59% correct classification for the classification training sample and 20% classification validation.

Table 3. Overall statistical measure of SVM models predicting *S. oryzae* gender.

| Type of Statistical Measure | Model 1 (All) | Model 2 (Rostrum width & length) | Model 3 (Prothorax width & length) | Model 4 (3rd and last) | Model 5 (Club width and length) |
|-----------------------------|---------------|---------------------------------|-----------------------------------|------------------------|--------------------------------|

6
sternite of abdomen)

| Correct classification (%) | Female | 32  | 32  | 26  | 26  | 32 |
|----------------------------|--------|-----|-----|-----|-----|-----|
| Male                       |        | 90  | 90  | 90  | 95  | 90  |
| **Total**                  |        | **62**| **62**| **59**| **62**| **62**|

| Correct classification (%) | Female | 20  | 20  | 20  | 20  | 20  |
|----------------------------|--------|-----|-----|-----|-----|-----|
| Male                       |        | 0   | 0   | 0   | 0   | 0   |
| **Total**                  |        | **20**| **20**| **20**| **20**| **20**|

3.4 Comparison between DA and SVM

The predicted model developed using DA is in line with results reported by Hagstrum & Subramanyam [3] and Rees [11] found that the gender of *S. oryzae* can be identified using rostrum. The male of *S. oryzae* is wider and shorter, while the female has thinner and longer rostrum. However, the differences in the rostrum size are tiny and not distinct. Therefore, dissection of the abdomen of the insect is needed to confirm its gender. Dissection is a destructive method and not suggested and practical in determining the population assessment of *S. oryzae*. Thus, this developed model which based on the rostrum image of *S. oryzae* are more convenient in predicting its gender with 91% correct classification. In comparison, SVM model performed poorly in the classification process. Other than using the DA, Logistic Regression is another modelling technique that suitable in predicting the gender of *S. oryzae* [25].

4. Conclusion

Based on the overall performance on every statistical test, DA Model 2 was selected as the best model to identify *S. oryzae* gender by using the rostrum length and width. The gender of the *S. oryzae* was able to distinguish with 91% correct classification. This model is consistently reliable, as the identification is based in comparison with previous numerical results, which have to be verified by an internal reproductive organ. The developed model not only help to identify and predict the *S. oryzae* gender, but it also enhance the process to measure fertility rate and gender ratio in order to understand the population assessment of *S. oryzae*.

References

[1] Longstaff B C 1981 *Prot. Ecol.* 2 83–130
[2] Hagstrum D, Phillips T W and Cuperus G 2012 Stored Product Protection (Kansas: Kansas State University)
[3] Hagstrum D, Klejdysz T, Subramanyam B and Nawrot J 2017 *Atlas of Stored-Product Insects and Mites* (Minnesota: AACC International)
[4] Swamy K C N, Mutthuraju G P, Jagadeesh E and Thirumalaraju G T 2014 *Curr. Biot.* 8 76–81
[5] Campbell J F 2002 *J. Insect Behav.* 15 429–45
[6] Hagstrum and Subramanyam 2006 *Fundamentals of Stored-Product Entomology* David (Minnesota: AACC International)
[7] Germinara G S, Cristofaro D, Rotundo, De Cristofaro A and Rotundo G 2008 *J. Chem. Ecol.* 34 523–9
[8] Germinara G S, Rotundo G and De Cristofaro A 2007 *J. Stored Prod. Res.* 43 229–33
[9] Dent D 2000 *Insect pest management* (New York: CABI Publishing)
[10] United States Environmental Protection Agency 2017 *Integrated Pest Management (IPM) Principles* (Environmental Protection Agency United State Press)
[11] Rees D 2004 *Insects of Stored Products* (London: CSIRO)
[12] Baker J E 1988 *J. Stored Prod. Res.* 24 193–8
[13] Arnqvist G and Nilsson T 2000 *Anim. Behav.* 60 145–64
[14] Flay C, Wang Q and He X 2014 *J. Insect Behav.* 27 268–78
[15] Johnson S D 1994 Oikos 69 172-176
[16] Rankin D J and Kokko H 2007 Oikos 116 335–48
[17] Omar Y 2012 Egypt J. Plant Prot. Pathol. 3 843–63
[18] Hair J F, Black W C, Babin B J and Anderson R E 2014 *Multivariate Data Analysis* (Essex: Pearson)
[19] Rencher 1998 Multivariate Statistical Inference and Applications (New York: Wiley-Interscience)
[20] Ibrahim A, Juahir H, Toriman M E, Mustapha A, Azid A and Isiyaka H A 2015 *J. Anal. Sci.* 19 338–48
[21] XLSTAT Discriminant Analysis (DA) XLSTAT Your data Anal. Solut.
[22] Khatchikian C, Sangermano F, Kendell D and Livdahl T 2011 *Med. Vet. Entomol.* 25 268–75
[23] Fox C W, Dublin L and Pollitt S J 2003 *Funct. Ecol.* 17 619–26
[24] Garikapati P, Balamurugan K, Latchoumi T P and Malkapuram R 2020 Silicon 63
[25] Azis T M F and Kasim K F 2020 *ARPN J. Eng. Appl. Sci.* 15 1006–10

**Acknowledgements**
The authors would like to thank Ministry of Higher Education (MOHE) (FRGS/1/2019/WAB01/UNIMAP/02/3) for providing financial support to this work and Universiti Malaysia Perlis (UniMAP) and UiTM Cawangan Perlis for providing the facilities and technical support.