The impact of feature selection on one and two-class classification performance for plant MicroRNAs

Waleed Khalifa, Malik Yousef, Müşerref Duygu Saçar Demirci, Jens Allmer

MicroRNAs (miRNAs) are short nucleotide sequences that form a typical hairpin structure which is recognized by a complex enzyme machinery. It ultimately leads to the incorporation of 18 – 24 nt long mature miRNAs into RISC where they act as recognition keys to aid in regulation of target mRNAs. It is involved to determine miRNAs experimentally and, therefore, machine learning is used to complement such endeavors. The success of machine learning mostly depends on proper input data and appropriate features for parameterization of the data. Although, in general, two-class classification (TCC) is used in the field; because negative examples are hard to come by, one-class classification (OCC) has been tried for pre-miRNA detection. Since both positive and negative examples are currently somewhat limited, feature selection can prove to be vital for furthering the field of pre-miRNA detection. In this study, we compare the performance of OCC and TCC using eight feature selection methods and seven different plant species providing positive pre-miRNA examples. Feature selection was very successful for OCC where the best feature selection method achieved an average accuracy of 95.6%, thereby being ~29% better than the worst method which achieved 66.9% accuracy. While the performance is comparable to TCC, which performs up to 3% better than OCC, TCC is much less affected by feature selection and its largest performance gap is ~13% which only occurs for two of the feature selection methodologies. We conclude that feature selection is crucially important for OCC and that it can perform on par with TCC given the proper set of features.

The impact of feature selection on one- and two-class classification performance for plant microRNAs

Waleed Khalifa[1,2], Malik Yousef[1,2], Müşerref Duygu Saçar Demirci[3], and Jens Allmer[3,4]

[1] Computer Science, The College of Sakhnin, Sakhnin, 30810, Israel
[2] The Institute of Applied Research- the Galilee Society, P.O. Box 437 Shefa Amr, 20200, Israel
[3] Molecular Biology and Genetics, Izmir Institute of Technology, 35430 Urla, Izmir, Turkey
[4] Bionia Incorporated, IZTEKGEB A8, 35430 Urla, Izmir, Turkey

Contact Emails: malik.yousef@gmail.com, duygusacar@gmail.com, khwalid@hotmail.com, jens@allmer.de

Abstract

MicroRNAs (miRNAs) are short nucleotide sequences that form a typical hairpin structure which is recognized by a complex enzyme machinery. It ultimately leads to the incorporation of 18 – 24 nt long mature miRNAs into RISC where they act as recognition keys to aid in regulation of target mRNAs. It is involved to determine miRNAs experimentally and, therefore, machine learning is used to complement such endeavors. The success of machine learning mostly depends on proper input data and appropriate features for parameterization of the data. Although, in general, two-class classification (TCC) is used in the field; because negative examples are hard to come by, one-class classification (OCC) has been tried for pre-miRNA detection. Since both positive and negative examples are currently somewhat limited, feature selection can prove to be vital for furthering the field of pre-miRNA detection. In this study, we compare the performance of OCC and TCC using eight feature selection methods and seven different plant species providing positive pre-miRNA examples. Feature selection was very successful for OCC where the best feature selection method achieved an average accuracy of 95.6%, thereby being ~29% better than the worst method which achieved 66.9% accuracy. While the performance is comparable to TCC, which performs up to 3% better than OCC, TCC is much less affected by feature selection and its largest performance gap is ~13% which only occurs for two of the feature selection methodologies. We conclude that feature selection is crucially important for OCC and that it can perform on *par* with TCC given the proper set of features.

Introduction

Gene regulation is of prime importance in all living organisms and there are multiple levels at which gene expression can be modulated. MicroRNAs (miRNAs) play a role in post-transcriptional gene regulation [1] and, among other functions, fine-tune the amount of translated protein product [2]. Mature miRNAs are short nucleotide sequences discovered about two decades ago [3]. From databases which host miRNAs like miRBase [4] it can be gleaned that miRNAs exist in a wide range of organisms ranging from viruses [5] to plants [6]. It has also been proposed that the plant miRNA system may have evolved independently [7] and some organisms like yeasts also display differences to the canonical pathway [8]. Any regulatory element itself may be miss-regulated and miRNAs are no exception, and therefore, have been implicated in, for example, human diseases [9], [10] and in plant stress response [11]. While miRNAs may lead to inter-kingdom communication in special cases [12], it is not likely that there is extensive communication among eukaryotes [13]. Experimentally detected and/or validated miRNAs are available in databases such as miRBase [14] and miRTarBase [15]. MicroRNAs' effect can only be established when it is co-expressed with its targets [2], which complicates experimental analysis since only a fraction of the genome is expressed at a given time, in a tissue, or in response to stress conditions; and testing all conditions experimentally is elusive. Additionally, such an analysis needs to be performed on transcript and protein level, concurrently, over multiple time points to establish a causative relationship. Therefore, it seems impossible to experimentally detect all possible miRNAs of any higher eukaryotic organism [16]–[19]. Moreover, it has become clear that even among the experimentally validated miRNAs in miRBase and mirTarBase, there may be dubious

45 examples [20]. Therefore, carefully designed computational experiments are required to complement experimental
46 approaches for miRNA detection.
47
48 Many computational approaches to miRNA detection have been proposed and most of them derive numerical
49 features [21] to describe a pre-miRNA and then use machine learning to establish a model for miRNA identification
50 [22]–[26]. Of these approaches, most, with few exceptions [16], [23], [27], employ two class classification; the latter
51 has been compared previously [24], [28]. Classification in machine learning depends on positive examples for
52 training the classifier in case of one-class classification (OCC) and additionally on negative data in case of two-class
53 classification (TCC). The negative data, however, proves to be difficult to establish (if not impossible), so that all
54 negative datasets currently in use are based on arbitrary selection of examples from parts of a genome deemed not
55 miRNA genic or from randomly generated sequences. While both approaches are questionable, they present the only
56 alternative to using OCC and in the absence of proper benchmark data need to be used for TCC [29]. Since OCC
57 only needs examples for the target class (here positives), it can obliterate the need to define artificial negative
58 examples [30], [31] and can be used to differentiate between target and unknown class. We have recently analyzed
59 the use of OCC for miRNA detection in plants and found that it was competitive in comparison to TCC although the
60 analysis was unduly biased towards TCC [6]. Our previous study also showed that among the hundreds of features
61 proposed for miRNA parameterization [21] some are more discriminative than others. Since feature selection is NP-
62 hard [32], selecting the best subset from more than 1000 features on a per dataset basis is not achievable. Feature
63 selection has been investigated before, but mostly for TCC [33]–[35], while only little has been done for OCC [36]–
64 [38]. In this study, we used different feature selection approaches and compared their effectiveness for OCC and
65 TCC classification performance.
66 Both machine learning approaches, OCC and TCC, benefit from feature selection. While feature selection is
67 essential for OCC and a difference of about 30% accuracy can be observed, the maximum difference for TCC is
68 ~10%. Moreover, for TCC 7 out of 8 feature selection methods lead to accuracy greater than 90% whereas such high
69 accuracy was only achieved for two methods when using OCC. For the LIG feature selection method, intended as a
70 negative control, both classifiers display lowest performance but TCC is about 20% better than OCC. With
71 increasing accuracy (i.e. better feature selection for OCC), the accuracy for TCC also increases; except for the SFC
72 which is best for OCC but only third best for TCC. While the performance difference for LIG is large, it decreases
73 with the use of better feature selection methods. TCC is only 3% better when the SFC feature selection method is
74 considered, which provided the best performance for OCC. A difference in performance among plant species was
75 observed for both classifiers, but for TCC it was about 5% whereas for OCC it was 15%. In conclusion, feature
76 selection is essential for OCC, but does not affect TCC as much. We propose that due to the lack of true negative
77 data, more focus should be put on the further development of OCC approaches to pre-miRNA detection.

Materials and methods

Data

80 Positive examples for pre-miRNAs from selected plant species were downloaded from miRBase [14] (Releases 20
81 and 21). *Glycine max* (gma), *Zea mays* (zma), *Sorghum bicolor* (sbi), *Physcomitrella patens* (ppt), *Arabidopsis
82 thaliana* (ath), *Populus trichocarpa* (ptc), and *Oryza sativa* (osa) make up the positive dataset. Negative examples
83 for miRNAs consisted of 980 pseudo pre-miRNAs from the PlantMiRNAPred dataset [39]. For these data, all pre-
84 miRNA features were calculated as described previously [21], [23], [40]. We chose plant pre-miRNAs with large
85 amount of pre-miRNA examples and from different clades for this study. Additionally, plant miRNAs have not been
86 investigated as extensively as metazoan miRNAs which adds to the reason to choose plant pre-miRNAs.

One class classification

88 For one-class classification the DDtools [41] implementation of an OCC was utilized. 100 fold Monte Carlo cross
89 validation [42] was performed using randomly sampled 90% of the positive data for training and 10% for testing.
90 Moreover, the pseudo negative sequences were injected as unknown class during testing. We employed *k*-means in
91 this study as previously described [43] since it performed well in respect to OCC although it is a clustering
92 algorithm. During learning, labeled examples are clustered (miRNAs and unknown) and during testing and in
93 prediction, the label of the closest cluster is assigned to the sample.

Two class classification

Support Vector Machines (SVMs) are used for machine learning and were first proposed by [44]. In bioinformatics and in the field of pre-miRNA detection, SVMs have been used [17], [18], [39], [45]. Here, the WEKA library [46] SVM implementation which is based on LibSVM [47] was utilized. The radial basis function was set to a gamma value of 0.7 and the cost parameter was chosen to be 4.0 and the normalization option was set to true. Any machine learning algorithm needs initial training and we performed a 10 fold Monte Carlo cross validation [42] during learning, by employing random sampling using 90% of the data for training and 10% for testing.

Feature selection strategies

Feature selection has been shown to be an NP-hard problem and, therefore, other approximate feature selection strategies are being developed. In machine learning for pre-miRNAs, more than 1000 features have been proposed which makes feature selection especially hard. To investigate the impact of feature selection on model performance for OCC and TCC, four negative and four positive feature selection methods were designed. Previously, we found that a set of 50 to 100 features may be sufficient for successful pre-miRNA detection [21]. Using more than 50 features increases the likelihood that the feature set contains some features which may conceal differences among feature selection methods. Therefore a feature set size of 50 was selected for model training in this study.

We have previously performed feature selection for OCC [48] using similar feature selection methods as we propose here, but it is important to compare the impact between OCC and TCC.

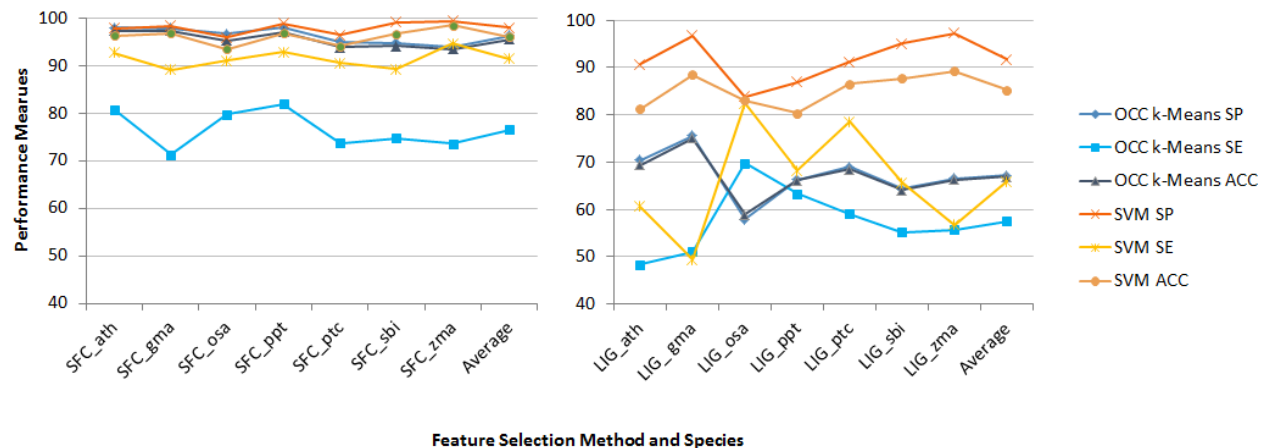
Eight feature selection methods were devised and four of them were expected to lead to low performance while the remaining methods were thought to perform well. The former were selecting features with low information gain (LIG), random feature selection (RFS), selecting random feature from feature clusters (RFC), and selecting features from clusters (SFC). The latter were selecting features with high information gain (HIG), selecting the highest information gain from feature clusters (HIC), zero-norm feature selection (ZNF), and Pearson correlation-based feature selection (PCF).

All feature selection methods except for the last two were performed using KNIME [49] and the selected features are available in Supplementary Table 1; information on how to calculate them are provided in Supplementary File 3. The workflows for our feature selection methods, developed in KNIME, are available for download from our website: http://bioinformatics.iyte.edu.tr/supplements/featsel.

In order to calculate LIG and HIG, for each dataset, the information gain (IG) among features was established (using KNIME's InformationGainCalculator node) and the 50 features with lowest IG (LIG) or highest IG (HIG) were selected. For RFS, 50 random features were selected using the Row Sampling node in KNIME. To establish RFC, features were clustered using WEKA k-Means implementation in KNIME ($k = 100$). From each cluster a random feature was selected and from the 100 random features the final set of 50 features was selected randomly (KNIME's Row Sampling approach). For SFC, clustering was performed as for RFC. Clusters were ordered by number of cluster members (largest to smallest) and the 50 features were chosen from the top. To derive HIC, the same clustering approach as for RFC and SFC was taken, but the features in each cluster were ranked according to IG and the best one was selected. The selected 100 features were again ranked using IG and the best 50 were selected. ZNF is defined to be the non-zero values for all feature vectors of positive examples. Among the non-zero ones, the 50 features with highest sum of values were selected. PCF was established according to Lorena et al. [36] and after Pearson clustering the features with lowest correlation score were retained. Feature selection was performed on a per species basis which led to the selection of different features (Supplementary Table 1). Combined feature selection uses the occurrence of selected features among the 7 selected species and 5 mixed datasets in respect to the top 100 features. The Features were ranked according to their frequency and top 50 were selected (Supplementary Table 1).

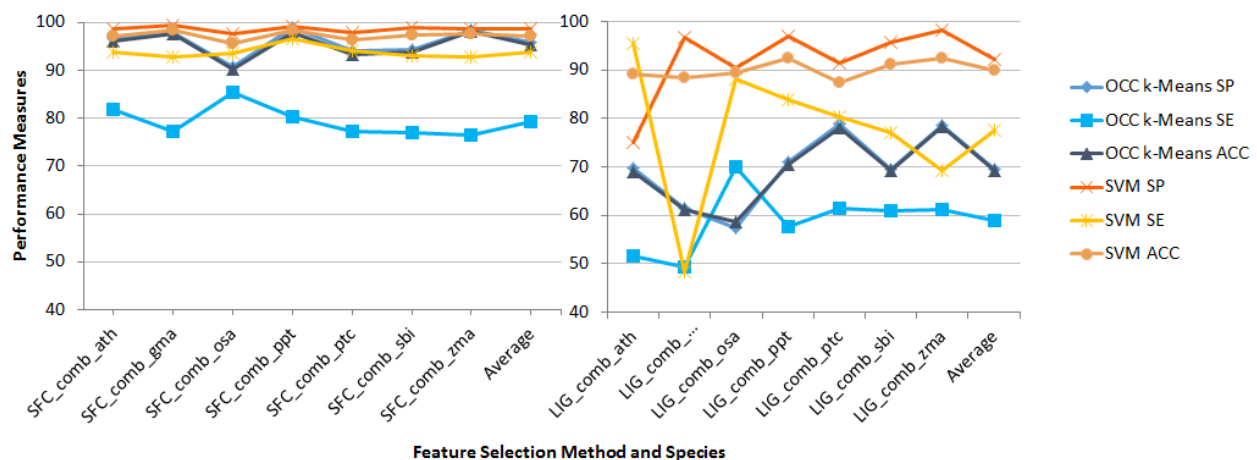
Results and discussion

Eight feature selection methods were designed and they were applied to seven plant datasets. For each dataset OCC (100) and TCC (10) models were established using Monte Carlo cross validation (MCCV). Feature selection was performed on a per plant dataset basis. The 50 features selected varied to some extend and, therefore, we defined another feature set (indicated by 'comb') which was created by selecting the features ordered by decreasing incidence based on the individual selections. The selected features are provided in Supplementary Table 1 by their acronyms which are explained in more detailed in our previous studies [21], [24].

We applied the 8 feature selection methods to the 7 plant species' datasets individually and recorded the model performance. Figure 1 shows the average model performance (OCC: 100, TCC: 10 fold cross validation) for the best feature selection method we found (SFC) and the worst one (LIG).

146

Feature Selection Method and Species

147
148 **Figure 1: Average model performance for SFC feature selection method for selected plant species (left) and LIG feature**
149 **selection model (right). OCC performance is in blue tone and SVM is toned orange. SP: specificity, SE: sensitivity, and**
150 **ACC: accuracy. Lines between points do not convey meaning, but were used to simplify visual tracking. Supplementary**
151 **Table 2 contains further information for all feature selection methods as well as standard deviations.**

152 Sensitivity was the performance measure most affected for both machine learning approaches (Figure 1). For TCC
153 the average accuracy among plant species dropped about 10% between SFC and LIG while it dropped about 30%
154 for OCC. The results for the remaining six feature selection methods are presented in Supplementary Table 2.
155 The impact on using combined feature selection for SFC and LIG is quite similar to individual feature selection
156 (Figure 2). The combined features were not calculated for PCF and ZNF since combination of features was not
157 supported by our workflow in this case. Overall accuracy is slightly reduced for the combined feature selection by
158 on average 1% (OCC) and 2% (TCC) when compared to individual feature selection.
159

Feature Selection Method and Species

160
161 **Figure 2: Average model performance for SFC feature selection method using combined feature set for selected plant**
162 **species (left) and LIG feature selection model using combined feature set (right). OCC performance is in blue tone and**
163 **SVM is toned orange. SP: specificity, SE: sensitivity, and ACC: accuracy. Lines between points do not convey meaning,**
164 **but were used to simplify visual tracking. Supplementary Table 2 contains further information for all feature selection**
165 **methods as well as standard deviations.**

166 The performance analysis of the remaining feature selection methods are presented in Supplementary Table 2. In
167 order to compare the performance of all feature selection methods for the two machine learning approaches, the
168 average model accuracy was plotted (Figure 3). It is striking that for most (6 out of 8) TCC performance results the
169 accuracy is above 95% for all plant species.
170

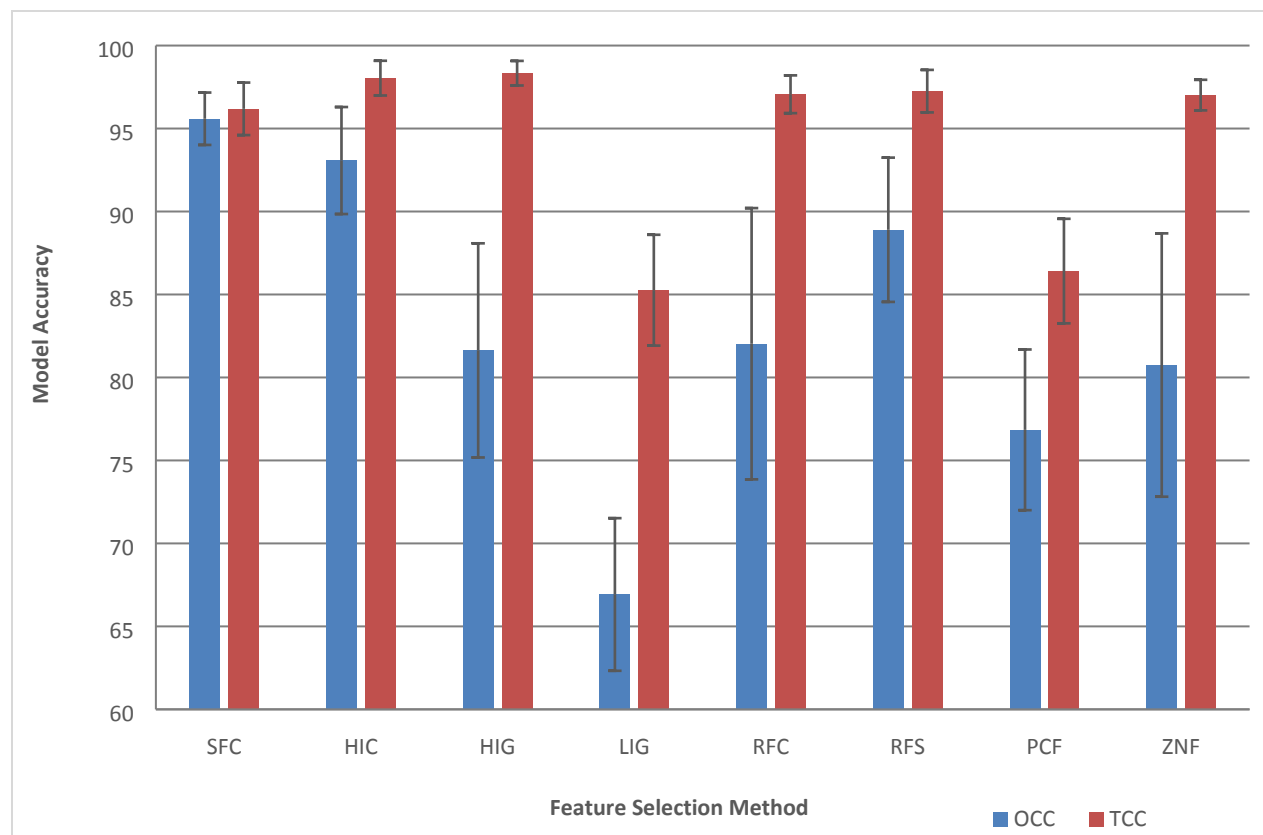
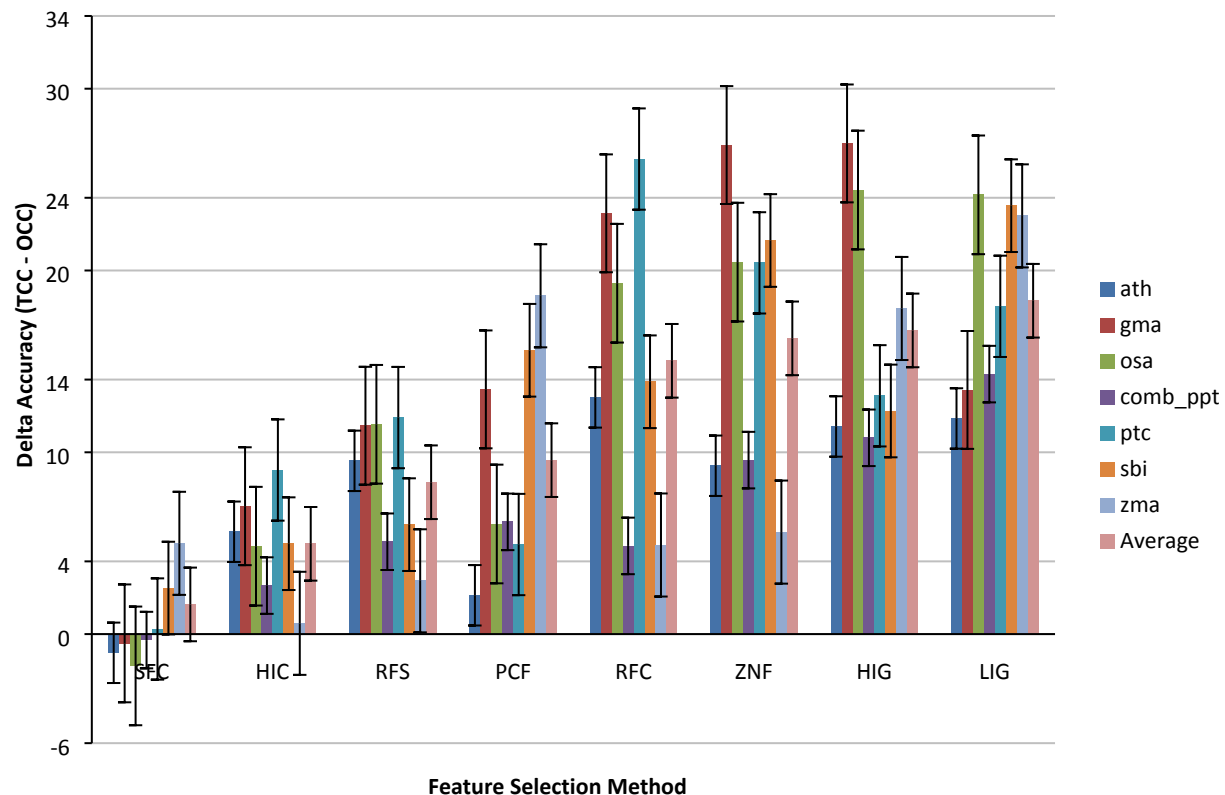



Figure 3: The average accuracy of OCC and TCC models created for selected plant species using eight different feature selection methods and their standard deviation. Data and figure are available for closer analysis in Supplementary Table 2.

For OCC, the performance is best for SFC where, on average, for plant species it achieves more than 95% accuracy (Figure 3). All other feature selection methods do not lead to high performing models with HIC being the second best, followed by RFS and PCF. For most feature selection methodologies the accuracy among plant species is quite similar for TCC, but for OCC the differences are much larger.

In order to compare the variance in performance between OCC and TCC, the difference between TCC and OCC accuracy was calculated ($TCC_{ACC} - OCC_{ACC}$) and is presented in Figure 4. Positive values signify better performance of TCC.

182
Figure 4: Eight feature selection methodologies applied to OCC and TCC. The difference in accuracy between TCC and OCC is presented. The groups are sorted by increasing average difference. Results are presented on a per species basis. The 'Average' averages the OCC or TCC performance among species. Data and figure are available for closer analysis in Supplementary Table 2.

The most accurate OCC model is on the left (SFC) and it is seen that TCC is outperformed by OCC on several plant species (ath, gma, osa, and ppt). Figure 4 shows that OCC is more affected by feature selection than TCC and, therefore, with increasing effectiveness of the feature selection methodology, the difference between classifiers diminishes. For improper feature selection it can reach up to about 30%; whereas it drops to almost similar performance for the best feature selection method in this study (SFC, ~0.6% on average).

Conclusions

Many general purpose feature selection methods have been described or used in bioinformatics [50]. For OCC feature selection nothing has been done in the area of pre-miRNA detection while one study investigated feature selection based on OCC for mature miRNA prediction [37]. When considering two class classification of pre-miRNAs SVM recursive feature elimination (RFE) has been used [51]. Meng et al. also used RFE, but modified it and compared to principal component analysis (PCA), correlation-based feature subset selection (CFS), and not using any filtering [52]. They report the best accuracy for SVM using their back SVM-RFE FS with 97.2% closely followed by PCA using SVM with 97.0% accuracy. One approach used genetic algorithm in combination with information gain and also taking into account feature redundancy for FS and achieved almost 99.5% accuracy, alas on a limited dataset [53]. These competing methods using different strategies for FS in pre-miRNA detection do not refer to OCC. However, they clearly show that feature selection has a large impact on model performance. The previous methodologies used correlation among features or feature redundancy for FS but did not put a clear focus on the correlation issue. We, therefore, devised eight feature selection methodologies with a focus on feature correlation and applied them to several plant miRNA datasets. Feature selection was performed on a per plant species basis, but we also investigated the combined feature set using the features shared among species; both of which were not done in previous approaches. Our SFC feature selection methodology was particularly successful

208 and there was no great difference for feature selection on a per plant basis or when combined (Figures 1, 2; left
209 panels). As expected, the LIG methodology did not perform well at all and was intended as a negative control.
210 However, the SVM learner was not nearly as much affected as the OCC one (Figures 1, 2; right panels) although
211 sensitivity was strongly affected for both learners.
212 Of the eight feature selection methods tested in this study, only 3 show good performance for OCC (SFC, HIC, and
213 RFS; Figure 3) while only two did not seem applicable for SVM (LIG and PCF; Figure 3). For most feature
214 selection methods average SVM performance is above 95% while OCC performance is generally below 90%
215 (Figure 3).
216 It is instructive to analyze the performance difference between SVM and OCC. Figure 4 shows the performance
217 difference and it can be seen that for most feature selection methods SVM performs better than OCC (positive
218 values; Figure 3). However, the SFC feature selection method which is among the best for SVM clearly performs
219 best for OCC and the latter can surpass the SVM performance for several of the selected plant species.
220 From this study it can be concluded that the more successful the feature selection the less difference between OCC
221 and TCC model performance and the better the overall model performance. Thus we conclude, that in the absence of
222 missing negative data OCC should be used and, therefore, additional feature selection strategies should be tried to
223 improve its performance.

References

225 [1] A. E. Erson-Bensan, "Introduction to microRNAs in biological systems.," *Methods Mol. Biol.*, vol. 1107, pp.
226   1–14, Jan. 2014.
227 [2] M. D. Saçar and J. Allmer, "Current Limitations for Computational Analysis of miRNAs in Cancer,"
228   *Pakistan J. Clin. Biomed. Res.*, vol. 1, no. 2, pp. 3–5, 2013.
229 [3] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs
230   with antisense complementarity to lin-14," *Cell*, vol. 75, no. 5, pp. 843–854, Dec. 1993.
231 [4] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing
232   data.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D152–7, Jan. 2011.
233 [5] F. Grey, "Role of microRNAs in herpesvirus latency and persistence.," *J. Gen. Virol.*, vol. 96, no. Pt 4, pp.
234   739–51, Apr. 2015.
235 [6] M. Yousef, J. Allmer, and W. Khalifaa, "Plant MicroRNA Prediction employing Sequence Motifs Achieves
236   High Accuracy," 2015.
237 [7] E. J. Chapman and J. C. Carrington, "Specialization and evolution of endogenous small RNA pathways.,"
238   *Nat. Rev. Genet.*, vol. 8, no. 11, pp. 884–896, Nov. 2007.
239 [8] C. Ender and G. Meister, "Argonaute proteins at a glance," *J. Cell Sci.*, vol. 123, no. 11, pp. 1819–1823,
240   Jun. 2010.
241 [9] B. Alural, G. A. Duran, K. U. Tufekci, J. Allmer, Z. Onkal, D. Tunali, K. Genc, and S. Genc, "EPO
242   Mediates Neurotrophic, Neuroprotective, Anti-Oxidant, and Anti-Apoptotic Effects via Downregulation of
243   miR-451 and miR-885-5p in SH-SY5Y Neuron-Like Cells.," *Front. Immunol.*, vol. 5, no. September, p. 475,
244   Sep. 2014.
245 [10] B. Alural, A. Ozerdem, J. Allmer, K. Genc, and S. Genc, "Lithium protects against paraquat neurotoxicity
246   by NRF2 activation and miR-34a inhibition in SH-SY5Y cells.," *Front. Cell. Neurosci.*, vol. 9, p. 209, May
247   2015.
248 [11] Z. Zhang, J. Yu, D. Li, Z. Zhang, F. Liu, X. Zhou, T. Wang, Y. Ling, and Z. Su, "PMRD: plant microRNA
249   database.," *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D806–13, Jan. 2010.
250 [12] C. Bağcı and J. Allmer, "Removing contamination from genomic sequences based on vector reference
251   libraries," in *2012 7th International Symposium on Health Informatics and Bioinformatics*, 2012, pp. 118–
252   122.
253 [13] C. Bağcı and J. Allmer, "One Step Forward, Two Steps Back; Xeno-MicroRNAs Reported in Breast Milk
254   Are Artifacts," *PLoS One*, vol. 11, no. 1, p. e0145065, 2016.
255 [14] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: tools for microRNA
256   genomics," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D154–8, Jan. 2008.
257 [15] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee,
258   C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang, "miRTarBase: a database
259   curates experimentally validated microRNA-target interactions.," *Nucleic Acids Res.*, vol. 39, no. Database
260   issue, pp. D163–9, Jan. 2011.

261 [16] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Learning from positive examples when the negative
262 class is undetermined--microRNA gene identification.," *Algorithms Mol. Biol.*, vol. 3, p. 2, Jan. 2008.
263 [17] J. Ding, S. Zhou, and J. Guan, "MiRenSVM: towards better prediction of microRNA precursors using an
264 ensemble SVM classifier with multi-loop features.," *BMC Bioinformatics*, vol. 11 Suppl 1, no. Suppl 11, p.
265 S11, Jan. 2010.
266 [18] Y. Wu, B. Wei, H. Liu, T. Li, and S. Rayner, "MiRPara: a SVM-based software tool for prediction of most
267 probable microRNA coding regions in genome scale sequences.," *BMC Bioinformatics*, vol. 12, no. 1, p.
268 107, Jan. 2011.
269 [19] W. Ritchie, D. Gao, and J. E. J. Rasko, "Defining and providing robust controls for microRNA prediction.,"
270 *Bioinformatics*, vol. 28, no. 8, pp. 1058–61, Apr. 2012.
271 [20] M. D. Saçar, H. Hamzeiy, and J. Allmer, "Can MiRBase provide positive data for machine learning for the
272 detection of MiRNA hairpins?," *J. Integr. Bioinform.*, vol. 10, no. 2, p. 215, Jan. 2013.
273 [21] M. D. Sacar and J. Allmer, "Data mining for microrna gene prediction: On the impact of class imbalance
274 and feature number for microrna gene prediction," in *2013 8th International Symposium on Health
275 Informatics and Bioinformatics*, 2013, pp. 1–6.
276 [22] J. Allmer and M. Yousef, "Computational methods for ab initio detection of microRNAs.," *Front. Genet.*,
277 vol. 3, p. 209, Jan. 2012.
278 [23] M. Yousef, J. Allmer, and W. Khalifa, "Sequence Motif-Based One-Class Classifiers Can Achieve
279 Comparable Accuracy to Two-Class Learners for Plant microRNA Detection," *J. Biomed. Sci. Eng.*, vol. 08,
280 no. 10, pp. 684–694, 2015.
281 [24] M. D. Saçar and J. Allmer, "Comparison of Four Ab Initio MicroRNA Prediction Tools," in *Proceedings of
282 the International Conference on Bioinformatics Models, Methods and Algorithms*, 2013, pp. 190–195.
283 [25] M. D. Saçar and J. Allmer, "Machine learning methods for microRNA gene prediction.," *Methods Mol.
284 Biol.*, vol. 1107, pp. 177–87, Jan. 2014.
285 [26] J. Allmer, "Computational and bioinformatics methods for microRNA gene prediction.," *Methods Mol.
286 Biol.*, vol. 1107, pp. 157–75, Jan. 2014.
287 [27] L. B. Koski, M. W. Gray, B. F. Lang, and G. Burger, "AutoFACT: an automatic functional annotation and
288 classification tool," *BMC Bioinformatics*, vol. 6, p. 151, 2005.
289 [28] I. de On Lopes, A. Schliep, and A. C. de Lf de Carvalho, "The discriminant power of RNA features for pre-
290 miRNA recognition.," *BMC Bioinformatics*, vol. 15, no. 1, p. 124, Jan. 2014.
291 [29] J. Allmer, "A Call for Benchmark Data in Mass Spectrometry-Based Proteomics," *J. Integr. OMICS*, vol. 2,
292 no. 2, Dec. 2012.
293 [30] L. M. Manevitz and M. Yousef, "One-Class SVMs for Document Classification," *J. Mach. Learn. Res.*, vol.
294 2, pp. 139–154, 2002.
295 [31] L. Manevitz and M. Yousef, "One-Class Document Classification via Neural Networks," *Neurocomputing*,
296 vol. 70, no. 7–9, pp. 1466–1481, Mar. 2007.
297 [32] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in
298 linear systems," *Theor. Comput. Sci.*, vol. 209, no. 1–2, pp. 237–260, 1998.
299 [33] S. Paul, M. Magdon-Ismail, and P. Drineas, "Feature selection for linear SVM with provable guarantees," *J.
300 Mach. Learn. Res.*, vol. 38, pp. 735–743, 2015.
301 [34] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support
302 Vector Machines," *Mach. Learn.*, no. 46, pp. 389–422, 2002.
303 [35] M. E. Ahsen, N. K. Singh, T. Boren, M. Vidyasagar, and M. A. White, "A new feature selection algorithm
304 for two-class classification problems and application to endometrial cancer," in *2012 IEEE 51st IEEE
305 Conference on Decision and Control (CDC)*, 2012, pp. 2976–2982.
306 [36] L. H. N. Lorena, A. C. P. L. F. Carvalho, and A. C. Lorena, "Filter Feature Selection for One-Class
307 Classification," *J. Intell. Robot. Syst.*, pp. 1–17, Sep. 2014.
308 [37] P. Xuan, M. Guo, Y. Huang, W. Li, and Y. Huang, "MaturePred: efficient identification of microRNAs
309 within novel plant pre-miRNAs.," *PLoS One*, vol. 6, no. 11, p. e27422, Jan. 2011.
310 [38] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining
311 software," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, Nov. 2009.
312 [39] P. Xuan, M. Guo, X. Liu, Y. Huang, W. Li, and Y. Huang, "PlantMiRNAPred: efficient classification of real
313 and pseudo plant pre-miRNAs.," *Bioinformatics*, vol. 27, no. 10, pp. 1368–76, May 2011.
314 [40] M. D. Saçar, C. Bağcı, and J. Allmer, "Computational Prediction of MicroRNAs from Toxoplasma gondii
315 Potentially Regulating the Hosts' Gene Expression.," *Genomics. Proteomics Bioinformatics*, vol. 12, no. 5,
316 pp. 228–238, Oct. 2014.

317 [41] D. M. J. Tax, "DDtools, the Data Description Toolbox for Matlab." 2015.

318 [42] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo cross validation," *Chemom. Intell. Lab. Syst.*, vol. 56, no. 1, pp. 1–
319 11, Apr. 2001.

320 [43] M. Yousef, J. Allmer, and W. Khalifa, "Feature Selection for MicroRNA Target Prediction Comparison of
321 One-Class Feature Selection Methodologies," in *Proceedings of the 9th International Joint Conference on*
322 *Biomedical Engineering Systems and Technologies*, 2016, pp. 219–225.

323 [44] V. N. Vapnik, *The nature of statistical learning theory*. New York, New York, USA: Springer-Verlag, 1995.

324 [45] K. L. S. Ng and S. K. Mishra, "De novo SVM classification of precursor microRNAs from genomic pseudo
325 hairpins using global and intrinsic folding measures.," *Bioinformatics*, vol. 23, no. 11, pp. 1321–30, Jun.
326 2007.

327 [46] J. E. Gewehr, M. Szugat, and R. Zimmer, "BioWeka--extending the Weka framework for bioinformatics.,"
328 *Bioinformatics*, vol. 23, no. 5, pp. 651–3, Mar. 2007.

329 [47] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr.
330 2011.

331 [48] M. Yousef, M. D. Saçar Demirci, W. Khalifa, and J. Allmer, "Feature Selection Has a Large Impact on One-
332 Class Classification Accuracy for MicroRNAs in Plants," *Adv. Bioinformatics*, vol. 2016, pp. 1–6, 2016.

333 [49] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel,
334 "KNIME - The Konstanz Information Miner," *SIGKDD Explor.*, vol. 11, no. 1, 2009.

335 [50] Y. Saeys, I. Inza, and P. Larra??aga, "A review of feature selection techniques in bioinformatics,"
336 *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

337 [51] J. Shu, K. Chiang, J. Zempleni, and J. Cui, "Computational Characterization of Exogenous MicroRNAs that
338 Can Be Transferred into Human Circulation.," *PLoS One*, vol. 10, no. 11, p. e0140587, 2015.

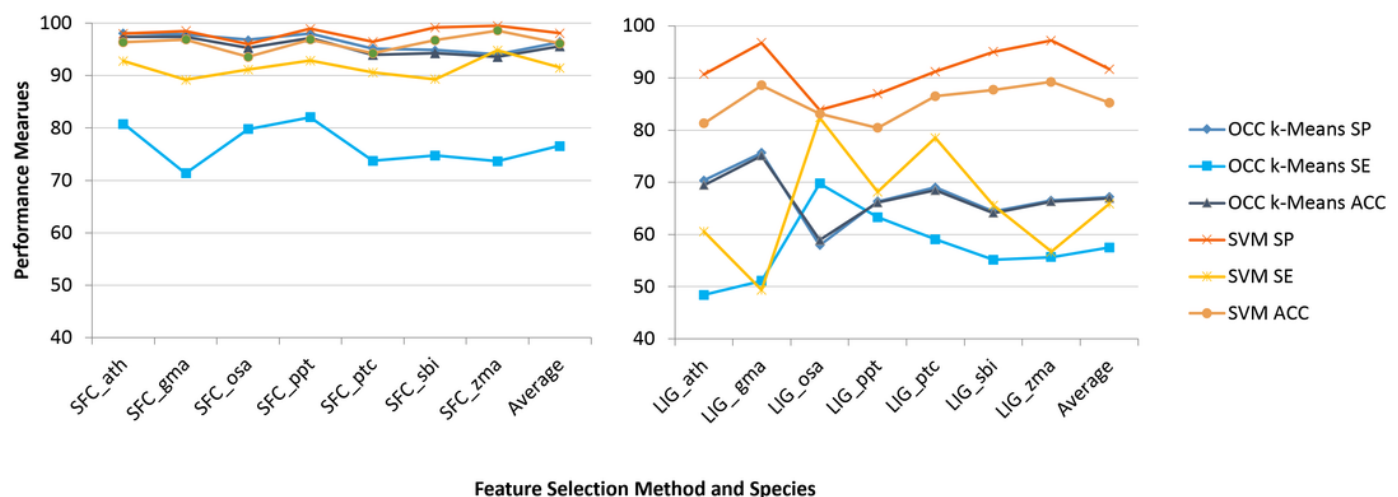
339 [52] J. Meng, D. Liu, C. Sun, and Y. Luan, "Prediction of plant pre-microRNAs and their microRNAs in
340 genome-scale sequences using structure-sequence features and support vector machine.," *BMC*
341 *Bioinformatics*, vol. 15, no. 1, p. 6595, 2014.

342 [53] P. Xuan, M. Z. Guo, J. Wang, C. Y. Wang, X. Y. Liu, and Y. Liu, "Genetic algorithm-based efficient feature
343 selection for classification of pre-miRNAs," *Genet. Mol. Res.*, vol. 10, no. 2, pp. 588–603, 2011.

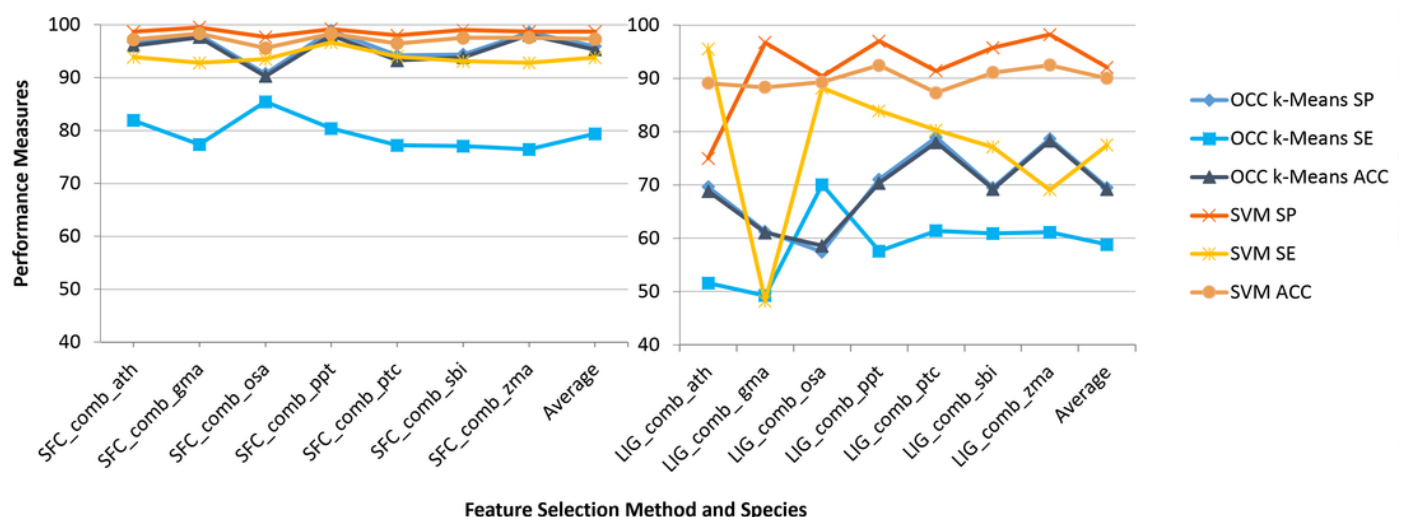
344

1

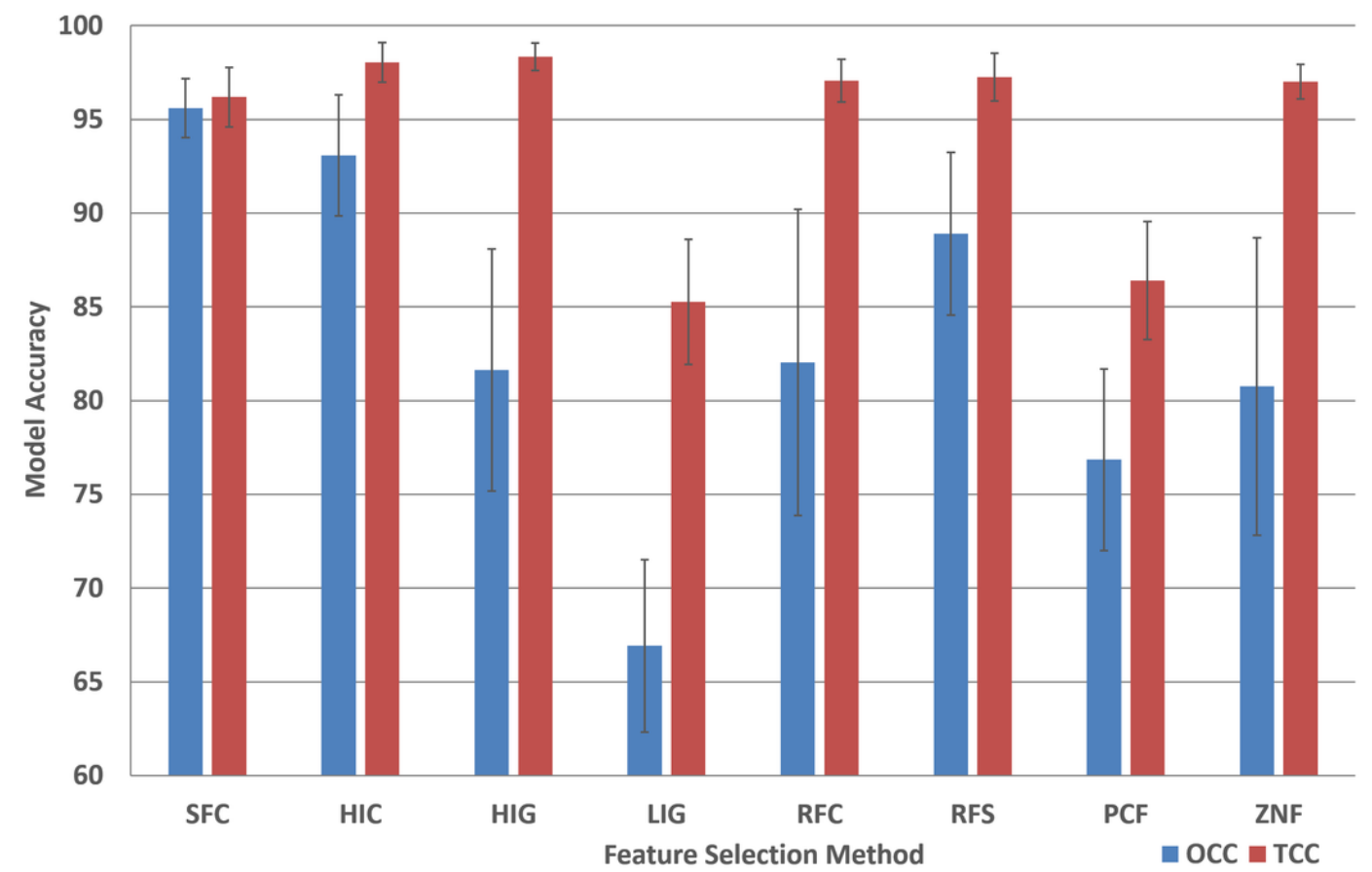
Best (SFC) versus worst (LIG) feature selection method on per species feature selection

Figure 1: Average model performance for SFC feature selection method for selected plant species (left) and LIG feature selection model (right). OCC performance is in blue tone and SVM is toned orange. SP: specificity, SE: sensitivity, and ACC: accuracy. Lines between points do not convey meaning, but were used to simplify visual tracking. Supplementary Table 2 contains further information for all feature selection methods as well as standard deviations.

2

Best (SFC) versus worst (LIG) feature selection method on consensus feature selection

Figure 2: Average model performance for SFC feature selection method using combined feature set for selected plant species (left) and LIG feature selection model using combined feature set (right). OCC performance is in blue tone and SVM is toned orange. SP: specificity, SE: sensitivity, and ACC: accuracy. Lines between points do not convey meaning, but were used to simplify visual tracking. Supplementary Table 2 contains further information for all feature selection methods as well as standard deviations.

3

Model accuracy comparison between OCC and TCC in respect to feature selection method

Figure 3: The average accuracy of OCC and TCC models created for selected plant species using eight different feature selection methods and their standard deviation. Data and figure are available for closer analysis in Supplementary Table 2.

4

Comparison of the effect of feature selection on two-class versus one-class classification

Figure 4: Eight feature selection methodologies applied to OCC and TCC. The difference in accuracy between TCC and OCC is presented. The groups are sorted by increasing average difference. Results are presented on a per species basis. The 'Average' averages the OCC or TCC performance among species. Data and figure are available for closer analysis in Supplementary Table 2.