Table S1. Metrics used in binary classification. Adopted from [1]–[9]

ID	Metric	Representation	Observations	Used in TEs
1	Accuracy	$\frac{(TP + TN)}{(TP + FP + FN + TN)}$	Measures the percentage of samples that are correctly classified	[10]–[14]
2	Precision (Positive predictive value)	$\frac{TP}{(TP + FP)}$	Percentage of correctly classified positive samples among all positive-classified ones	[3]
3	Sensitivity (recall or true positive rate)	$\frac{TP}{(TP + FN)}$	Represents the proportion of positive samples that are correctly predicted	[3], [10], [12], [15]
4	Specificity	$\frac{TN}{(TN + FP)}$	Represents the proportion of negative samples that are correctly predicted	[15]
5	Matthews correlation coefficient	$\frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}$	The MCC can be seen as a discretization of the Pearson correlation for binary variables.	NO
6	Performance coefficient	$\frac{TP}{(TP + FN + FP)}$	Ratio of correct predictions belonging to the positive class and predictions belonging to the false class	NO
7	F1 score	$\frac{2 \times TP}{(2\ x\ TP + FP + FN)}$	Harmonic mean of precision and sensitivity	NO

8	Precision-recall curves	Graphics	Plots the precision of a model as a function of its recall	[3], [16]
9	Receiver Operating Characteristic curves (ROCs)	Graphics	Commonly used to evaluate the discriminative power of the classification model at different thresholds	[15]
10	Area under the ROC curve (AUC)a	$\frac{1}{2}(\frac{tp}{tp+fn}+\frac{tn}{tn+fp})$	Summary measure that indicates whether prediction performance is close to random (0:5) or perfect (1:0). Also describes the sensitivity versus the specificity of the prediction	[14], [16]
11	Area under the Precision Recall Curve (auPRC)b	$\frac{1}{2}(\frac{tp}{tp+fn}+\frac{tp}{tp+fp})$	Measures the fraction of negatives misclassified as positives and plots the precision vs. recall ratio	NO
12	False positive rate	1 − Specificity	Percentage of predictions marked as belonging to the positive class, but that are part of the negative class.	[14], [15]

Although [a] and [b] are areas under the curve, they can be viewed as a linear transformation of Youden Index [17].

Table S2. Metrics used in multiclass classification. Adopted from [1]–[9]

ID	Metric	Representation	Observations	Used in TEs
13	Average Accuracy	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$	The average per-class effectiveness of a classifier	[18]
14	Error Rate	$\frac{\sum_{i=1}^{l} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$	The average per-class classification error	NO
15	Precisionμ	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fp_i)}$	Agreement of the data class labels with those of a classifiers if calculated from sums of per-text decisions	NO
16	Recallμ	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l} (tp_i + fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions	NO
17	Fscoreμ	$\frac{(\beta^2 + 1)\,\mathrm{Pr}ecision_\mu\,\mathrm{Re}call_\mu}{\beta^2\,\mathrm{Pr}ecision_\mu + \mathrm{Re}call_\mu}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions	NO

18	PrecisionM	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$	An average per-class agreement of the data class labels with those of a classifiers	[3], [19]
19	RecallM	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$	An average per-class effectiveness of a classifier to identify class labels	NO
20	FscoreM	$\frac{(\beta^2+1)\,\mathrm{Pr}ecision_M\,\mathrm{Re}call_M}{\beta^2\,\mathrm{Pr}ecision_M + \mathrm{Re}call_M}$	Relations between data's positive labels and those given by a classifier based on a per-class average	NO

Table S3. Metrics used in hierarchical classification. Adopted from [1]–[9].

ID	Metric	Representation	Observations	Used in TEs
21	Precision↓	$\frac{\lvert C_{\downarrow}^{c} \cap C_{\downarrow}^{d} \rvert}{\lvert C_{\downarrow}^{c} \rvert}$	Positive agreement on subclass labels w.r.t. the subclass labels given by a classifier	[20]–[22]
22	Recall↓	$\frac{\lvert C_{\downarrow}^{c} \cap C_{\downarrow}^{d} \rvert}{\lvert C_{\downarrow}^{d} \rvert}$	Positive agreement on subclass labels w.r.t. the subclass labels given by data	[20]–[22]

23	Fscore↓	$\frac{(\beta^2+1)\,\mathrm{Pr}\,ecision_{\downarrow}\,\mathrm{Re}\,call_{\downarrow}}{\beta^2\,\mathrm{Pr}\,ecision_{\downarrow}+\mathrm{Re}\,call_{\downarrow}}$	Relations between data's positive subclass labels and those given by a classifier	[20]–[22]
24	Precision↑	$\frac{\lvert C_{\uparrow}^{c}\cap C_{\uparrow}^{d}\rvert}{\lvert C_{\uparrow}^{c}\rvert}$	Positive agreement on superclass labels w.r.t. the superclass labels given by a classifier	[20]–[23]
25	Recall↑	$\frac{\lvert C_{\uparrow}^{c}\cap C_{\uparrow}^{d}\rvert}{\lvert C_{\uparrow}^{d}\rvert}$	Positive agreement on superclass labels w.r.t. the superclass labels given by data	[20]–[23]
26	Fscore↑	$\frac{(\beta^2+1)\,\mathrm{Pr}\,ecision_{\uparrow}\,\mathrm{Re}\,call_{\uparrow}}{\beta^2\,\mathrm{Pr}\,ecision_{\uparrow}+\mathrm{Re}\,call_{\uparrow}}$	Relations between data's positive superclass labels and those given by a classifier	[20]–[23]

Table S4. Evaluation for metric collection.

ID	Metric	Level of applicability to TEs	Level of measured features	Observations
1	Accuracy	Low	Low	For unbalanced datasets (such as in TE classes), this metric is a meaningless performance measurement [74], because it does not reveal the true classification performance of the rare classes.
2	Precision (Positive predictive value)	Medium	Medium	In the TE detection and classification problem, the number of correct predictions should be maximized. This metric can be informative since Precision is the percentage of predictions that are correct [35].

3	Sensitivity (recall or true positive rate)	Medium	Medium	It can be a very informative measure for TEs since sensitivity is the percentage of true samples that are correctly detected [75].
4	Specificity	Low	Low	In TE identification or classification, true positive results are more important than the others. Thus, this metric might not be practical because sensitivity is the measure by which false samples are correctly rejected [75].
5	Matthews correlation coefficient	High	Low	MCC can be a key measurement because it is a balanced measurement, even if the sizes of positive and negative samples have great differences [76], such as in TE datasets. This constant indicates how much the predicted results agree (near to one), disagree (near to minus one) or are random (near to zero) compared to the observed data.
6	Performance coefficient	Low	Low	This metric could give an uninformative value due to class imbalance. The value will probably be low since positive results are much lower than negative ones given the size of the negative class (which would be much bigger than the positive).
7	F1 score	**High**	**High**	Since F1-score is the harmonic mean of precision and recall (it weighs both metrics equally [77]), it can be used to find an optimized threshold in a ML task [78], especially in TE problems.
8	Precision-recall curves	High	High	Since only a small genome fraction contains TE sequences from a specific class, it is most important to recognize positive samples [35]. This graphic plots a list of precision and recall values for all possible thresholds in a sorted order [32], providing a more informative picture than other metrics in skewed datasets [69].
9	Receiver Operating Characteristic curves (ROCs)	Low	Low	When the number of negative examples greatly exceeds the number of positives, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis [69]. Thus, in TE identification, this can be problematic because the negative dataset would be much bigger than the positive one. Also, in the classification process, one of the classes would be much smaller than the others.
10	Area under the ROC curve (AUC)a	Low	Low	AUC can be a wrong indicator of performance because, in heavily unbalanced classes, large values of AUC may not necessarily indicate good performance [74], [79].
11	Area under the Precision Recall Curve (auPRC)b	**High**	**High**	In contrast to AUC, this metric is more appropriate when class distribution is heavily unbalanced [74]. Also, since this metric is order-based, it is not influenced by the imbalanced-classes problem because the proportions of positive and negative classes are not considered [78]. A high value indicates that the model makes very few mistakes [74].

12	False positive rate	Medium	Low	This metric would be not interesting for the TE detection and classification problem since it represents the proportion of negative cases that are incorrectly identified as positive cases in the data. A higher FPR implies that more negative data points have been incorrectly classified. In TE detection, the negative class is much larger than the positive one; thus, this imbalance would increase its value. Also, in the classification process, the size of the classes differs, due to the dynamics of TEs.
13	Average Accuracy	Low	Low	As in its binary case, this metric is meaningless because TE datasets contain imbalanced classes [32]. This aspect is caused by the dynamics of TEs, where some lineages or superfamilies are more frequent than others, and this distribution changes depending on the organism.
14	Error Rate	Low	Low	Since this metric measures how many negative results (FP and FN) are obtained by the algorithm, it could be not very informative in imbalanced datasets, such as TEs classes.
15	Precisionμ	Medium	Low	Precision is a relevant metric for TE analysis, because it is the percentage of correct predictions [35]. Also, in TE detection and classification, it is more interesting to measure positive than negative results. However, the "micro" averaging strategy may not be the best option because we are not interested in treating each sample equally, due to the great diversity of TEs within classes.
16	Recallμ	Medium	Low	Similar to precision, this metric is based on positive results, which is the main interest when analyzing TEs. Like multi-class precision, the "micro" averaging strategy could be not correctly applicable to TEs.
17	Fscoreμ	**High**	**Low**	F1-score could be a metric that better measures the performance of ML algorithms applied to TEs since it considers precision and recall. Again, the "micro" averaging strategy could be not correctly applicable to TEs.
18	PrecisionM	**Medium**	**Medium**	The "macro" averaging strategy could be the best option to obtain the same weight for each class, because macro-averaging treats all classes equally while micro-averaging favors bigger classes [34].
19	RecallM	**Medium**	**Medium**	The "macro" averaging strategy could be the best option to obtain the same weight for each class, because macro-averaging treats all classes equally while micro-averaging favors bigger classes [34].
20	FscoreM	**High**	**High**	The "macro" averaging strategy could be the best option to obtain the same weight for each class, because macro-averaging treats all classes equally while micro-averaging favors bigger classes [34].
21	Precision↓	**Medium**	**Low**	This metric measures descendant performance in terms of the agreement between predicted and given labels. Given the hierarchical classification of TEs, ML algorithms could be measured by hierarchical metrics.
22	Recall↓	**Medium**	**Low**	This metric measures descendant performance in terms of the agreement between predicted and given labels. Given the hierarchical classification of TEs, ML algorithms could be measured by hierarchical metrics.
23	Fscore↓	**High**	**Low**	This metric measures descendant performance in terms of the agreement between predicted and given labels. Given the hierarchical classification of TEs, ML algorithms could be measured by hierarchical metrics.

24	Precision↑	Medium	Medium	This metric measures ancestor performance in terms of the agreement between predicted and given labels, which could be better for lineage classification. Given the hierarchical classification of TEs, ML algorithms could be measured by hierarchical metrics.
25	Recall↑	Medium	Medium	This metric measures ancestor performance in terms of the agreement between predicted and given labels, which could be better for lineage classification. Given the hierarchical classification of TEs, ML algorithms could be measured by hierarchical metrics.
26	Fscore↑	**High**	**High**	This metric measures ancestor performance in terms of the agreement between predicted and given labels, which could be better for lineage classification. Given the hierarchical classification of TEs, ML algorithms could be measured by hierarchical metrics.

Rows in bold were selected to perform invariance analyses.

Table S5. Results of experiment 1.

Database	Filling	Coding Scheme	Pre-processing	LR	LDA	KNN	SVM	MLP	RF	DT	NB
Repbase	Self	Complementary	None	0,483	0,436	0,318	0,506	0,443	0,367	0,390	0,531
Repbase	Self	DAX	None	0,395	0,417	0,359	0,494	0,367	0,350	0,415	0,485
Repbase	Self	EIIP	None	0,402	0,443	0,302	0,450	0,425	0,346	0,415	0,409
Repbase	Self	Enthalpy	None	0,339	0,329	0,297	0,329	0,341	0,343	0,408	0,367
Repbase	Self	Galois-4	None	0,390	0,422	0,348	0,492	0,334	0,357	0,402	0,496
Repbase	Self	Kmers	None	0,958	0,931	0,880	0,963	0,168	0,875	0,746	0,741
Repbase	Self	PC	None	0,380	0,383	0,420	0,172	0,381	0,438	0,420	0,362
Repbase	Self	Complementary	Scaling	0,225	0,436	0,316	0,501	0,467	0,344	0,390	0,531
Repbase	Self	DAX	Scaling	0,195	0,417	0,353	0,497	0,464	0,359	0,415	0,483
Repbase	Self	EIIP	Scaling	0,190	0,443	0,302	0,490	0,453	0,343	0,415	0,408
Repbase	Self	Enthalpy	Scaling	0,188	0,329	0,297	0,366	0,343	0,364	0,408	0,364
Repbase	Self	Galois-4	Scaling	0,360	0,422	0,348	0,478	0,464	0,350	0,395	0,489
Repbase	Self	Kmers	Scaling	0,961	0,931	0,877	0,963	0,974	0,882	0,746	0,737
Repbase	Self	PC	Scaling	0,380	0,383	0,450	0,193	0,476	0,438	0,418	0,351
Repbase	Self	Complementary	PCA	0,162	0,104	0,304	0,483	0,413	0,406	0,487	0,395
Repbase	Self	DAX	PCA	0,213	0,097	0,343	0,492	0,452	0,404	0,438	0,381
Repbase	Self	EIIP	PCA	0,473	0,086	0,295	0,443	0,476	0,392	0,417	0,346

Repbase	Self	Enthalpy	PCA	0,376	0,090	0,302	0,315	0,339	0,357	0,380	0,330
Repbase	Self	Galois-4	PCA	0,146	0,098	0,336	0,489	0,441	0,408	0,441	0,402
Repbase	Self	Kmers	PCA	0,633	0,628	0,765	0,764	0,596	0,785	0,727	0,674
Repbase	Self	PC	PCA	0,378	0,378	0,401	0,195	0,418	0,369	0,383	0,369
Repbase	Self	Complementary	Scaling + PCA	0,243	0,056	0,299	0,471	0,457	0,436	0,492	0,392
Repbase	Self	DAX	Scaling + PCA	0,232	0,098	0,337	0,487	0,455	0,373	0,432	0,395
Repbase	Self	EIIP	Scaling + PCA	0,246	0,077	0,293	0,459	0,411	0,387	0,401	0,360
Repbase	Self	Enthalpy	Scaling + PCA	0,243	0,102	0,301	0,339	0,327	0,350	0,383	0,323
Repbase	Self	Galois-4	Scaling + PCA	0,322	0,132	0,343	0,464	0,469	0,380	0,425	0,401
Repbase	Self	Kmers	Scaling + PCA	0,954	0,968	0,887	0,963	0,977	0,817	0,787	0,723
Repbase	Self	PC	Scaling + PCA	0,376	0,376	0,415	0,283	0,431	0,395	0,417	0,357

Table S6. Results of experiment 2.

Database	Filling	Coding Scheme	Pre-processing	LR	LDA	KNN	SVM	MLP	RF	DT	NB
Repbase	Self	Complementary	None	0,258	0,152	0,133	0,208	0,184	0,106	0,243	0,226
Repbase	Self	DAX	None	0,127	0,124	0,120	0,184	0,146	0,098	0,241	0,176
Repbase	Self	EIIP	None	0,120	0,135	0,053	0,185	0,168	0,100	0,241	0,129
Repbase	Self	Enthalpy	None	0,085	0,104	0,049	0,171	0,140	0,099	0,231	0,151
Repbase	Self	Galois-4	None	0,134	0,157	0,116	0,193	0,172	0,099	0,238	0,190
Repbase	Self	Kmers	None	0,952	0,912	0,784	0,957	0,050	0,786	0,616	0,772
Repbase	Self	PC	None	0,086	0,119	0,210	0,054	0,085	0,227	0,223	0,094
Repbase	Self	Complementary	Scaling	0,195	0,152	0,133	0,209	0,210	0,099	0,248	0,224
Repbase	Self	DAX	Scaling	0,156	0,124	0,118	0,189	0,190	0,092	0,241	0,174
Repbase	Self	EIIP	Scaling	0,144	0,135	0,053	0,176	0,188	0,093	0,241	0,128
Repbase	Self	Enthalpy	Scaling	0,144	0,104	0,050	0,162	0,187	0,086	0,231	0,151
Repbase	Self	Galois-4	Scaling	0,233	0,157	0,114	0,191	0,214	0,101	0,228	0,186
Repbase	Self	Kmers	Scaling	0,944	0,912	0,860	0,948	0,973	0,750	0,617	0,772
Repbase	Self	PC	Scaling	0,101	0,119	0,225	0,051	0,249	0,234	0,222	0,093

Repbase	Self	Complementary	PCA	0,147	0,081	0,099	0,233	0,190	0,123	0,248	0,143
Repbase	Self	DAX	PCA	0,155	0,074	0,115	0,210	0,202	0,120	0,176	0,144
Repbase	Self	EIIP	PCA	0,150	0,068	0,042	0,157	0,156	0,115	0,201	0,131
Repbase	Self	Enthalpy	PCA	0,104	0,064	0,051	0,160	0,154	0,091	0,137	0,134
Repbase	Self	Galois-4	PCA	0,126	0,066	0,110	0,206	0,223	0,111	0,181	0,150
Repbase	Self	Kmers	PCA	0,357	0,407	0,582	0,601	0,371	0,620	0,563	0,524
Repbase	Self	PC	PCA	0,093	0,095	0,225	0,047	0,179	0,208	0,158	0,099
Repbase	Self	Complementary	Scaling + PCA	0,173	0,048	0,094	0,256	0,264	0,113	0,264	0,144
Repbase	Self	DAX	Scaling + PCA	0,140	0,084	0,108	0,208	0,204	0,118	0,203	0,155
Repbase	Self	EIIP	Scaling + PCA	0,148	0,053	0,038	0,185	0,217	0,105	0,142	0,133
Repbase	Self	Enthalpy	Scaling + PCA	0,141	0,089	0,050	0,162	0,198	0,086	0,145	0,130
Repbase	Self	Galois-4	Scaling + PCA	0,179	0,109	0,112	0,198	0,256	0,113	0,194	0,153
Repbase	Self	Kmers	Scaling + PCA	0,923	0,963	0,871	0,948	0,978	0,582	0,667	0,519
Repbase	Self	PC	Scaling + PCA	0,095	0,094	0,194	0,067	0,193	0,216	0,187	0,087

Table S7. Results of experiment 3.

Database	Filling	Coding Scheme	Pre-processing	LR	LDA	KNN	SVM	MLP	RF	DT	NB
PGSB	Self	Complementary	None	0,771	0,066	0,587	0,819	0,761	0,683	0,805	0,738
PGSB	Self	DAX	None	0,742	0,075	0,355	0,794	0,731	0,699	0,795	0,603
PGSB	Self	EIIP	None	0,718	0,089	0,324	0,637	0,724	0,695	0,795	0,620
PGSB	Self	Enthalpy	None	0,668	0,569	0,352	0,501	0,677	0,669	0,756	0,513
PGSB	Self	Galois-4	None	0,697	0,684	0,391	0,776	0,697	0,687	0,781	0,639
PGSB	Self	Kmers	None	0,993	0,988	0,974	0,989	0,360	0,977	0,895	0,865
PGSB	Self	PC	None	0,398	0,414	0,732	0,410	0,384	0,766	0,688	0,381

PGSB	Self	Complementary	Scaling	0,764	0,066	0,586	0,818	0,780	0,682	0,808	0,737
PGSB	Self	DAX	Scaling	0,735	0,075	0,358	0,801	0,768	0,697	0,793	0,603
PGSB	Self	EIIP	Scaling	0,699	0,089	0,323	0,764	0,750	0,750	0,795	0,620
PGSB	Self	Enthalpy	Scaling	0,646	0,569	0,353	0,700	0,710	0,664	0,756	0,513
PGSB	Self	Galois-4	Scaling	0,701	0,684	0,394	0,776	0,761	0,688	0,782	0,639
PGSB	Self	Kmers	Scaling	0,993	0,988	0,990	0,993	0,993	0,978	0,893	0,866
PGSB	Self	PC	Scaling	0,418	0,414	0,754	0,371	0,669	0,763	0,688	0,381
PGSB	Self	Complementary	PCA	0,712	0,669	0,584	0,817	0,763	0,549	0,670	0,528
PGSB	Self	DAX	PCA	0,706	0,595	0,356	0,793	0,757	0,512	0,584	0,460
PGSB	Self	EIIP	PCA	0,735	0,556	0,318	0,637	0,747	0,488	0,614	0,340
PGSB	Self	Enthalpy	PCA	0,688	0,608	0,342	0,499	0,686	0,522	0,660	0,443
PGSB	Self	Galois-4	PCA	0,634	0,699	0,389	0,767	0,702	0,526	0,597	0,471
PGSB	Self	Kmers	PCA	0,607	0,602	0,915	0,891	0,696	0,914	0,859	0,647
PGSB	Self	PC	PCA	0,382	0,348	0,678	0,381	0,614	0,661	0,642	0,393
PGSB	Self	Complementary	Scaling + PCA	0,746	0,742	0,582	0,817	0,783	0,571	0,669	0,533
PGSB	Self	DAX	Scaling + PCA	0,714	0,696	0,346	0,801	0,772	0,547	0,594	0,454
PGSB	Self	EIIP	Scaling + PCA	0,679	0,642	0,315	0,756	0,754	0,518	0,624	0,342
PGSB	Self	Enthalpy	Scaling + PCA	0,626	0,634	0,358	0,697	0,701	0,542	0,667	0,447
PGSB	Self	Galois-4	Scaling + PCA	0,677	0,730	0,391	0,768	0,758	0,551	0,595	0,478

PGSB	Self	Kmers	Scaling + PCA	0,993	0,986	0,991	0,991	0,994	0,981	0,862	0,742
PGSB	Self	PC	Scaling + PCA	0,370	0,348	0,666	0,370	0,611	0,642	0,638	0,391

Table S8. Results of experiment 4.

Database	Filling	Coding Scheme	Pre-processing	LR	LDA	KNN	SVM	MLP	RF	DT	NB
PGSB	Self	Complementary	None	0,621	0,053	0,385	0,694	0,555	0,441	0,668	0,620
PGSB	Self	DAX	None	0,551	0,069	0,348	0,653	0,524	0,479	0,643	0,510
PGSB	Self	EIIP	None	0,461	0,064	0,271	0,403	0,487	0,478	0,642	0,469
PGSB	Self	Enthalpy	None	0,381	0,364	0,278	0,261	0,412	0,388	0,540	0,385
PGSB	Self	Galois-4	None	0,063	0,062	0,061	0,610	0,078	0,061	0,085	0,545
PGSB	Self	Kmers	None	0,960	0,968	0,932	0,973	0,461	0,950	0,710	0,834
PGSB	Self	PC	None	0,069	0,111	0,562	0,075	0,077	0,615	0,442	0,102
PGSB	Self	Complementary	Scaling	0,616	0,053	0,384	0,667	0,610	0,432	0,667	0,620
PGSB	Self	DAX	Scaling	0,566	0,069	0,349	0,661	0,611	0,467	0,640	0,510
PGSB	Self	EIIP	Scaling	0,515	0,064	0,271	0,584	0,566	0,468	0,643	0,469
PGSB	Self	Enthalpy	Scaling	0,405	0,364	0,281	0,470	0,460	0,398	0,544	0,387
PGSB	Self	Galois-4	Scaling	0,537	0,548	0,344	0,639	0,624	0,461	0,579	0,545
PGSB	Self	Kmers	Scaling	0,968	0,968	0,970	0,981	0,975	0,946	0,711	0,835
PGSB	Self	PC	Scaling	0,083	0,111	0,598	0,079	0,415	0,605	0,442	0,102
PGSB	Self	Complementary	PCA	0,524	0,531	0,383	0,690	0,545	0,236	0,445	0,352

PGSB	Self	DAX	PCA	0,546	0,474	0,346	0,659	0,566	0,228	0,327	0,329
PGSB	Self	EIIP	PCA	0,488	0,395	0,267	0,408	0,543	0,195	0,367	0,222
PGSB	Self	Enthalpy	PCA	0,392	0,392	0,280	0,260	0,444	0,221	0,381	0,274
PGSB	Self	Galois-4	PCA	0,440	0,565	0,342	0,606	0,499	0,236	0,318	0,368
PGSB	Self	Kmers	PCA	0,327	0,347	0,855	0,827	0,479	0,860	0,759	0,560
PGSB	Self	PC	PCA	0,061	0,063	0,489	0,074	0,343	0,479	0,435	0,115
PGSB	Self	Complementary	Scaling + PCA	0,581	0,613	0,382	0,705	0,595	0,257	0,434	0,369
PGSB	Self	DAX	Scaling + PCA	0,537	0,569	0,343	0,675	0,621	0,242	0,337	0,383
PGSB	Self	EIIP	Scaling + PCA	0,469	0,476	0,266	0,588	0,561	0,233	0,362	0,230
PGSB	Self	Enthalpy	Scaling + PCA	0,381	0,412	0,275	0,458	0,453	0,233	0,382	0,288
PGSB	Self	Galois-4	Scaling + PCA	0,504	0,608	0,337	0,616	0,616	0,259	0,347	0,406
PGSB	Self	Kmers	Scaling + PCA	0,972	0,964	0,970	0,978	0,969	0,960	0,675	0,832
PGSB	Self	PC	Scaling + PCA	0,058	0,062	0,480	0,064	0,357	0,455	0,449	0,107

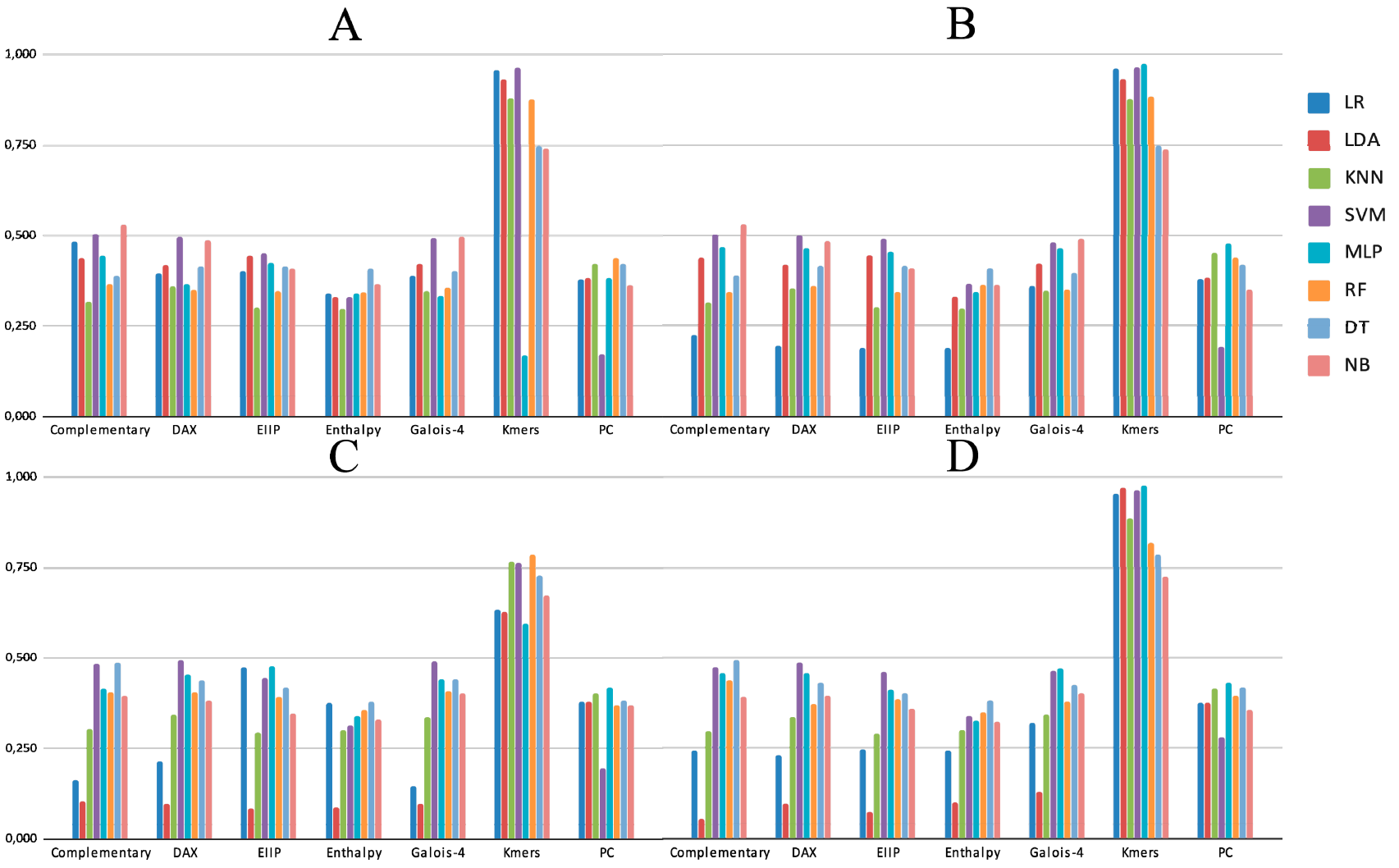

Figure S1. Performance of ML algorithms and Repbase using as main metric accuracy (experiment 1) and the following pre-processing techniques: a) none, b) scaling, c) PCA, d) PCA + scaling.

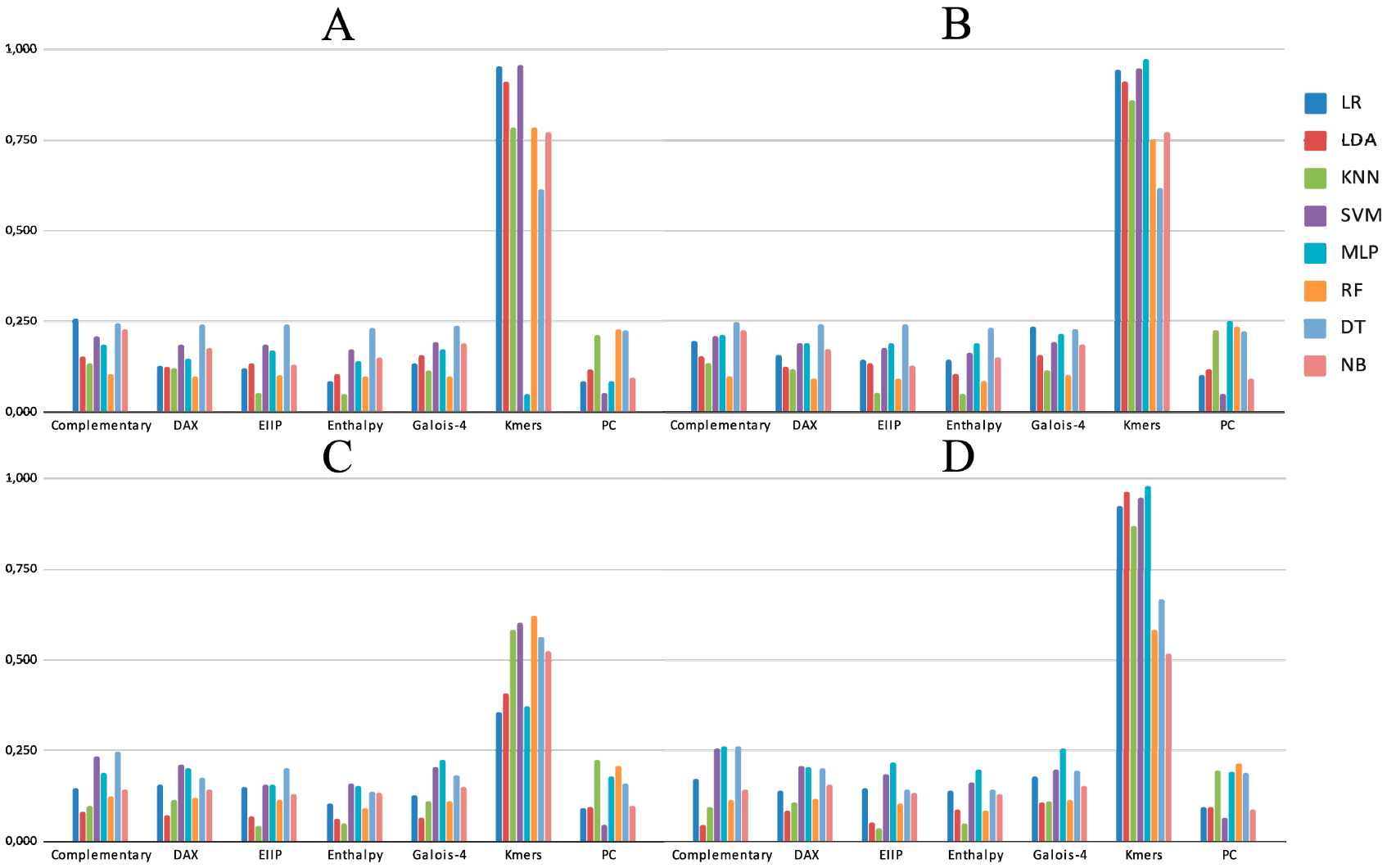

Figure S2. Performance of ML algorithms and Repbase using as main metric F1-score (experiment 2) and the following pre-processing techniques: a) none, b) scaling, c) PCA, d) PCA + scaling.

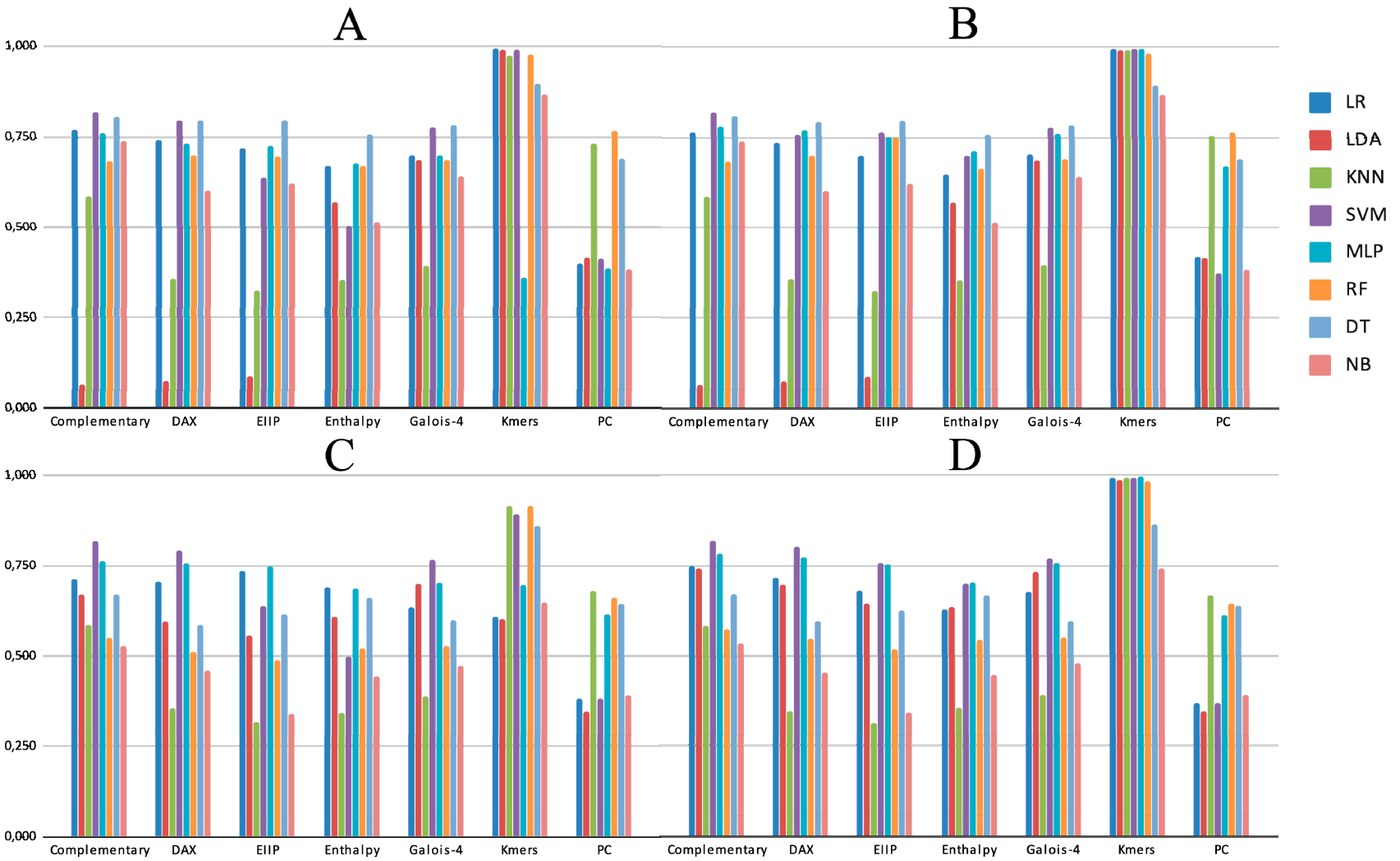

Figure S3. Performance of ML algorithms and PGSB using as main metric accuracy (experiment 3) and the following pre-processing techniques: a) none, b) scaling, c) PCA, d) PCA + scaling.

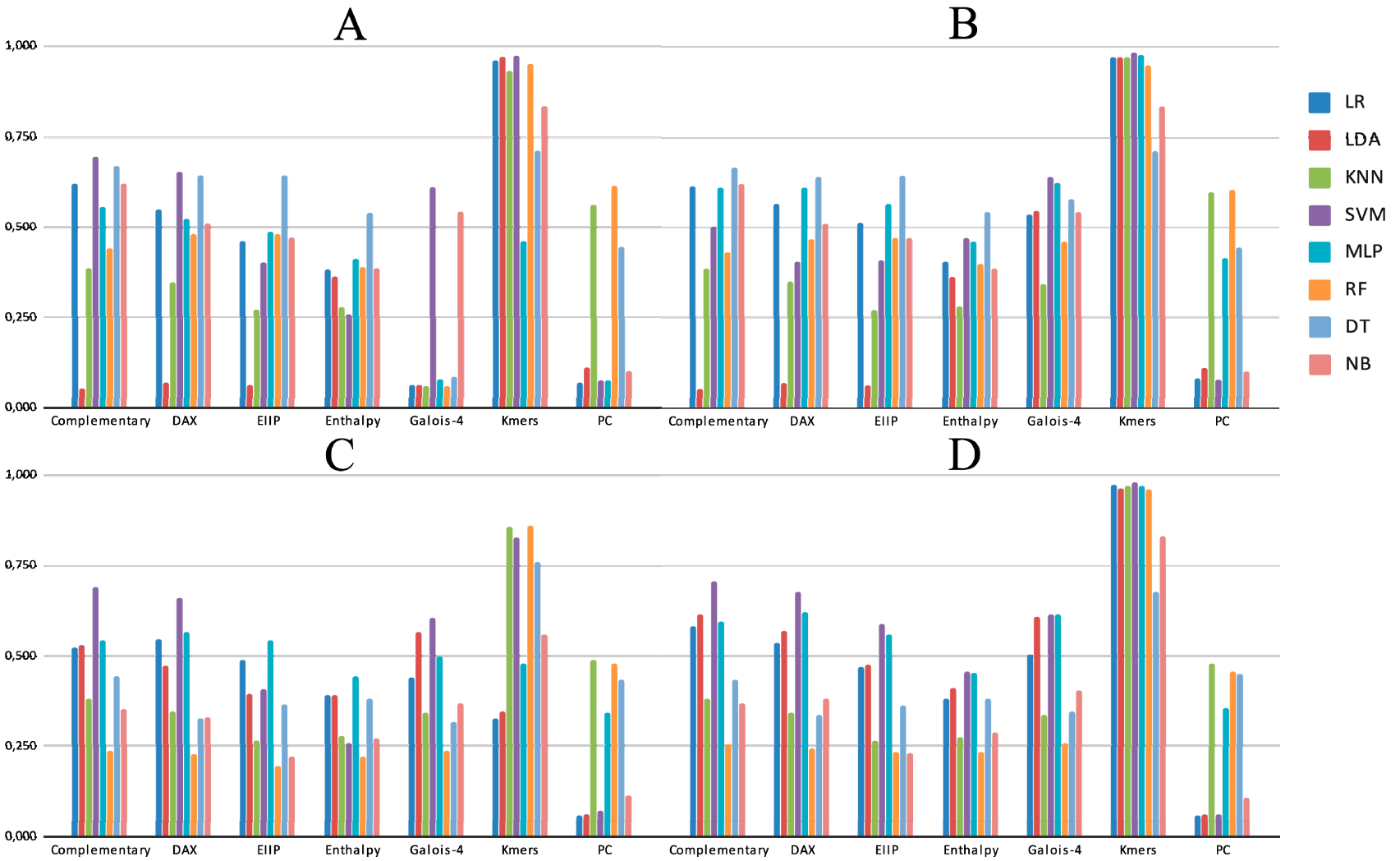

Figure S4. Performance of ML algorithms and PGSB using as main metric F1-score (experiment 4) and the following pre-processing techniques: a) none, b) scaling, c) PCA, d) PCA + scaling.

References

[1] L. Chen *et al.*, “Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection,” *Mol. Genet. Genomics*, vol. 293, no. 1, pp. 137–149, Feb. 2018.

[2] N. Yu, Z. Yu, and Y. Pan, “A deep learning method for lincRNA detection using auto-encoder algorithm,” *BMC Bioinformatics*, vol. 18, no. S15, p. 511, Dec. 2017.

[3] L. Schietgat *et al.*, “A machine learning based framework to identify and classify long terminal repeat retrotransposons.,” *PLoS Comput. Biol.*, vol. 14, no. 4, p. e1006097, Apr. 2018.

[4] M. A. Smith, S. E. Seemann, X. C. Quek, and J. S. Mattick, “DotAligner: identification and clustering of RNA structure motifs,” *Genome Biol.*, vol. 18, no. 1, p. 244, Dec. 2017.

[5] U. Kamath, K. De Jong, and A. Shehu, “Effective automated feature construction and selection for classification of biological sequences.,” *PLoS One*, vol. 9, no. 7, p. e99982, Jul. 2014.

[6] E. S. Segal *et al.*, “Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine Learning in a Stable Haploid Isolate of Candida albicans.,” *MBio*, vol. 9, no. 5, Oct. 2018.

[7] C. Ma, H. H. Zhang, and X. Wang, “Machine learning for Big Data analytics in plants,” *Trends Plant Sci.*, vol. 19, no. 12, pp. 798–808, Dec. 2014.

[8] J. Brayet, F. Zehraoui, L. Jeanson-Leh, D. Israeli, and F. Tahi, “Towards a piRNA prediction using multiple kernel fusion and support vector machine.,” *Bioinformatics*, vol. 30, no. 17, pp. i364-70, Sep. 2014.

[9] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[10] T. Loureiro, R. Camacho, J. Vieira, and N. A. Fonseca, “Boosting the Detection of Transposable Elements Using Machine Learning,” 2013, pp. 85–91.

[11] W. Ashlock and S. Datta, “Distinguishing endogenous retroviral LTRs from SINE elements using features extracted from evolved side effect machines,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 9, no. 6, pp. 1676–1689, 2012.

[12] T. Loureiro, R. Camacho, J. Vieira, and N. A. Fonseca, “Improving the performance of Transposable Elements detection tools.,” *J. Integr. Bioinform.*, vol. 10, no. 3, p. 231, Nov. 2013.

[13] G. Abrusan, N. Grundmann, L. DeMester, and W. Makalowski, “TEclass-a tool for automated classification of unknown eukaryotic transposable elements,” *BIOINFORMATICS*, vol. 25, no. 10, pp. 1329–1330, May 2009.

[14] Y. Zhang, A. Babaian, L. Gagnier, and D. L. Mager, “Visualized Computational Predictions of Transcriptional Effects by Intronic Endogenous Retroviruses,” *PLoS One*, vol. 8, no. 8, p. e71971, Aug. 2013.

[15] C. Douville *et al.*, “Detection of aneuploidy in patients with cancer through amplification of long interspersed nucleotide elements (LINEs).,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 8, pp. 1871–1876, Feb. 2018.

[16] G. M. M. Ventola, T. M. R. Noviello, S. D’Aniello, A. Spagnuolo, M. Ceccarelli, and L. Cerulo, “Identification of long non-coding transcripts with feature selection: a comparative study,” *BMC Bioinformatics*, vol. 18, no. 1, p. 187, Dec. 2017.

[17] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[18] W. Su, X. Gu, and T. Peterson, “TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome,” *Mol. Plant*, vol. 12, no. 3, pp. 447–460, Mar. 2019.

[19] J. Arango-López, S. Orozco-Arias, J. A. Salazar, and R. Guyot, “Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case,” 2017, pp. 156–170.

[20] F. K. Nakano, S. M. Mastelini, S. Barbon, and R. Cerri, “Improving Hierarchical Classification of Transposable Elements using Deep Neural Networks,” in *Proceedings of the International Joint Conference on Neural Networks*, 2018, vol. 2018–July.

[21] F. K. Nakano, S. Martiello Mastelini, S. Barbon, and R. Cerri, “Stacking methods for hierarchical classification,” in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, 2018, vol. 2018–Janua, pp. 289–296.

[22] F. K. Nakano, W. J. Pinto, G. L. Pappa, and R. Cerri, “Top-down strategies for hierarchical classification of transposable elements with neural networks,” in *Proceedings of the International Joint Conference on Neural Networks*, 2017, vol. 2017–May, pp. 2539–2546.

[23] B. Zamith Santos, G. Trindade Pereira, F. Kenji Nakano, and R. Cerri, “Strategies for selection of positive and negative instances in the hierarchical classification of transposable elements,” in *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, 2018, pp. 420–425.