User documentation

EnsembleFS web-based tool for a filter ensemble feature selection of molecular data

Aneta Polewko-Klim, Paweł Grablis and Witold R. Rudnicki

Contents

1 Methods

1.1 Feature filters

The filters used in the ensemble were chosen based on the following criteria:

1. FS algorithms should be based on different assumptions to reduce the risk of omitting important biological variables;

2. each feature filter should generate the ranking of features with a statistically well-defined cutoff between informative and non-informative ones;

3. each FS algorithm should be suitable for high-dimensional and correlated biomedical data;

4. at least one of the FS methods should be sensitive to interactions between variables to consider potential interactions among biomarkers.

In the current version, the EnsembleFS tool offers five feature selection (FS) algorithms for removing irrelevant variables, namely the U-test [20], the minimum redundancy maximum relevance (MRMR) [7], the Monte Carlo feature selection (MCFS) [8], and the 1- and 2-dimensional versions of multidimensional feature selection (MDFS-1D and MDFS-2D respectively) [22]. The Mann-Whitney test is a non-parametric statistical test that assigns a probability to the hypothesis that two samples corresponding to two decision classes are drawn from populations with the same average value. The MDFS method measures the decrease of the information entropy of the decision variable due to knowledge of k-dimensional tuples of variables and measures the influence of each variable in the tuple [22]. It performs an exhaustive search over all possible k-tuples and assigns to each variable a maximal information gain due to a given variable that was achieved in any of the k-tuple that included this variable. The two-dimensional version of this algorithm (MDFS-2D) can capture the synergistic interactions between pairwise features. The MRMR method is based on mutual information as a measure of the relevancy and redundancy of features, where the redundancy of selected features is an aggregate mutual information measure between each pair of features in the selected feature subset, and the relevance to a class variable is an aggregate mutual information measure between each feature with respect to the class variable. The MCFS method relies on a Monte Carlo approach to select informative features. The MCFS algorithm is capable of incorporating inter-dependencies between features. This FS method offers several cut-off methods for discerning informative and non-informative features, such as the critical angle, k-means, and permutations. Due to short calculation time, the *EnsembleFS* uses the k-means method as the default cut-off method. The k-means method groups the relative importance values into two clusters and sets the cut-off border to separate the two clusters.

The U-test, MDFS-1D, and MDFS-2D methods compute the importance of features with p-values. To adjust p-values for multiple pairwise comparisons, the *EnsembleFS* tool offers five p-value correction methods, namely: the Bonferroni correction [9], the Benjamini & Hochberg correction [1], the Benjamini & Yekutieli correction (alias FDR)[2], the Holm correction [15], and the Hochberg correction [14].

For U-test, MDFS-1D, and MDFS-2D filters, the p-value of less than 0.05 indicates that the feature is a significant predictor.

1.2 Learning method

Predictive models are constructed with the random forest algorithm [3]. Random forest was selected because it works well on data sets with a small number of objects, has few tune-able parameters that don't relate directly to the data, very rarely fails and usually gives results that are often either best or very close to the best results achievable by any classification algorithm [11]. This algorithm is well-suited for imbalanced data sets. [16].

1.3 Predictive model evaluation metrics

In the EnsembleFS pipeline, the classification accuracy (ACC), the area under the receiver operator curve (AUC), and the Matthews correlation coefficient (MCC) [21] are used for the evaluation of the quality of predictive RF models. The ACC metric is defined as:

$$ACC = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{1}$$

where: TP is the number of true positives, TN the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. MCC is defined as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

AUC and MCC metrics are better suited to evaluate the quality of binary classifier for the unbalanced population than the simple classification accuracy [30, 4].

1.4 Feature stability measure

To measure the similarity between sets of the most informative features, the average of the pairwise similarity for all pairs of the feature sets with different cross-validation samples was calculated. For this, Lustgarten's stability measure (ASM) was used, which is described by the formula [19]:

$$ASM = \frac{2}{(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left(\frac{|S_i \cap S_j| - |S_i| * |S_j| / p}{min(|S_i|, |S_j|) - max(0, |S_i| + |S_j| - p)} \right) \tag{3}$$

where: S_i and S_j are patrainingf the most informative feature subsets fro; namelyns of a model in k-fold cross-validation, $m = n \cdot k$, and p is total feature number of dataset.

1.5 Feature selection process

Figure S1 presents the scheme of the proposed FS procedure for biomarker discovery with quantitative omics data. *EnsambleFS* uses the heterogeneous ensemble feature selection method to select the relevant features from the omic data [27]. In the current version, five different FS algorithms are offered to users to reduce the risk of omitting biologically relevant biomarkers. The feature selection and model-building procedure for omics data describes Algorithm 1.

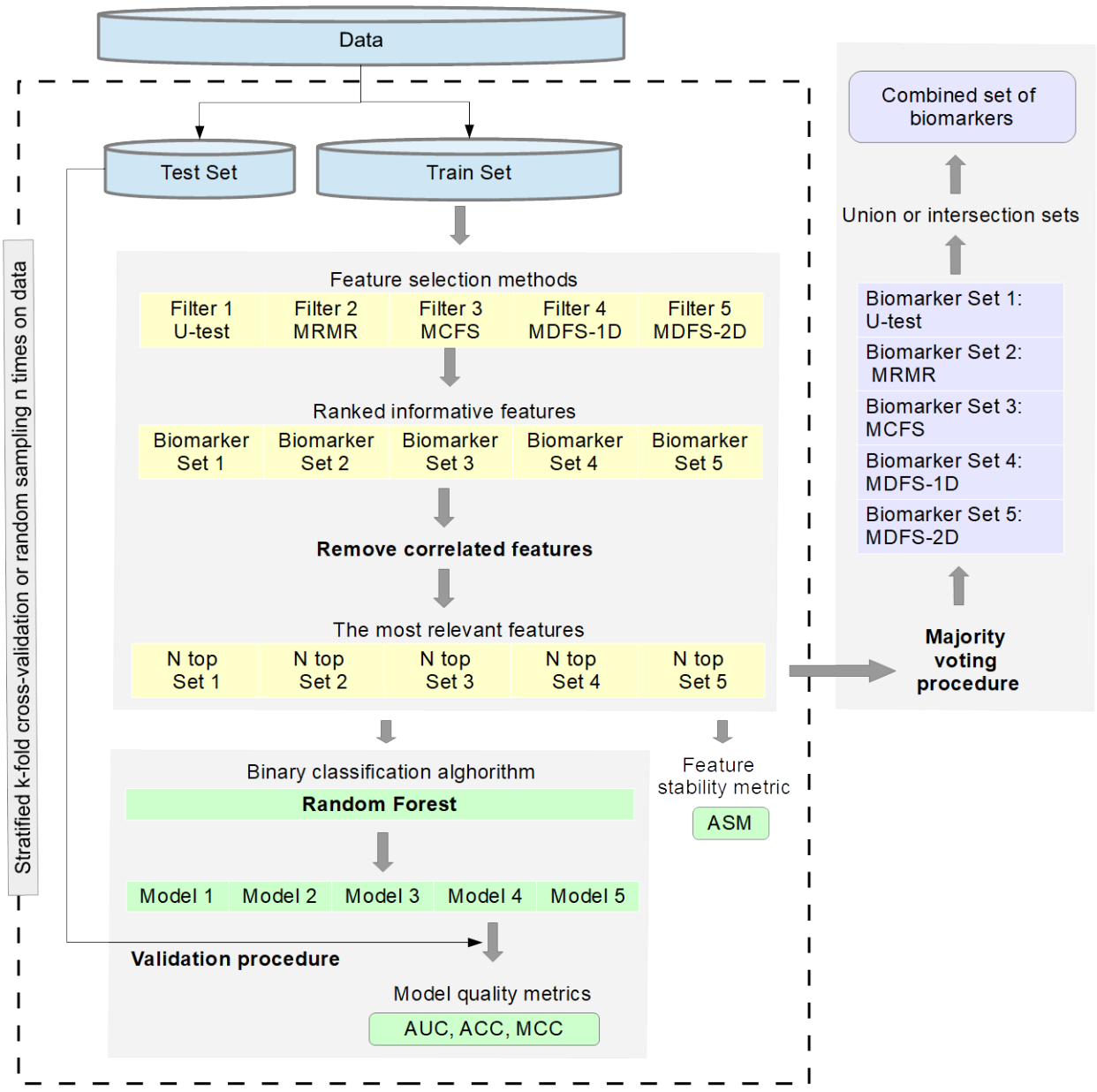

Figure 1: Scheme of the ensemble feature selection process with validation. See notation in text.

Algorithm 1: EFS$(l, f, S = \{P_1, \ldots, P_k\})$ the ensemble feature selection algorithm with binary classifier

input : Random Forest classifier l
Feature filters f_j, $j = 1, \ldots, m$
Dataset $S = \{(y, X)\}$ with n entries of p features $V = \{v_1, \ldots, v_p\}$
belonging to one of two classes, randomly split into k partitions P_i
output: Combined set of informative features F
Ranked informative feature set F_j, $j = 1, \ldots, m$
Performance estimation metric E_j, $j = 1, \ldots, m$
Feature selection stability measure A_j, $j = 1, \ldots, m$
repeat r **times**
  foreach S_i **do**
    Generate the training set $S_{\setminus i}(V) \leftarrow S(V) \setminus P_i(\mathrm{V})$
    foreach f_j **do**
      Perform feature selection on the training set $W_i \leftarrow f(S_{\setminus i}(V))$
      Collect the ranked informative feature set $W_i = \{v_1, \ldots, v_d\}$
      Remove highly correlated features with W_i
      Build the model on the training set $L_i \leftarrow l(S_{\setminus i}(U_i))$ using top N features U_i with W_i
      Performance estimation E_i: use the trained model L_i on a test set P_i
    end
  end
end
$E_j \leftarrow \frac{1}{r \cdot k} \Sigma E_i$, $i = 1, \ldots, r \cdot k$
Assess the feature selection stability A_j of $r \cdot k$ feature sets U_i
Collect the feature set F_j from $r \cdot k$ sets U_i by using the majority voting strategy, for each of m feature filters
Collect combined feature set $F = \bigcup_{j=1}^{m} F_j$ or $F = \bigcap_{j=1}^{m} F_j$

To evaluate the quality of the feature set, *EnsambleFS* applies the supervised ML procedure. The predictive models are built with top N features using the random forest algorithm, and the stratified k-fold cross-validation (cv) procedure or random sampling is used to evaluate classification models. The mean values of ACC, AUC, MCC, and ASM metrics are calculated for each FS method. This validation procedure provides an unbiased estimation of the quality of predictions for trained models.

The correlated variables can be removed from the training dataset to minimise the collinearity of features. To this end, Spearman's rank correlation coefficient (ρ) among features is determined, and the least informative features with the ρ values higher than some cut-off level are removed.

The most informative features, according to each individual FS algorithm, are identified as those that appear most consistently in the top N features selected in k resampling operations. The final feature set is determined by the majority voting method, i.e. only the features selected by more than half of the contributing filters are included in the final set. The full feature set for the final biological analysis includes all the best features or overlapping best features selected by basic feature filters.

2 Implementation details

To perform feature selection, the following R packages were used: *rmcfs* R package [8] for MCFS-ID method, *MDFS* R package [24] for MDFS-1D and MDFS-2D methods, and *mRMRe* R package [6] for MRMR method. To adjust p-values for multiple pairwise comparisons, the *stats* R package [26] are applied. The *caret* R package [18] is used to build random forest predictive models. For data visualization the following R packages are used: *ggplot2* [12], *plotly* [29], *venn* [10], and slickR [28]. The *gprofiler2* R package [17] is an interface to the g:Profiler web server for functional interpretation of gene lists. It is used to infer biological information from the list of relevant features.

3 EnsembleFS web app

3.1 Installation and launching the app

EnsambleFS R Shiny application is freely available online website at https://uco.uwb.edu.pl/apps/EnsembleFS (webserver demo), while the source code of the standalone version of *EnsambleFS* is hosted on the GitHub pages (https://github.com/biocsuwb/EnsembleFS). The EnsembleFS application can be directly launched from GitHub or cloned into a local system. In this latter case, the user should run the following command:

```
git clone https://github.com/biocsuwb/EnsembleFS
```

at the command line in git bash (Windows) or the terminal of bash shell in Linux or Mac.
RStudio users can easily download and run the *EnsambleFS* app as follows:

```
# set the working directory
setwd(dir = "path/folder")

# download a .zip file from the GitHub repository
download.file(
url = "https://github.com/biocsuwb/EnsembleFS/archive/refs/heads/main.zip",
destfile = "EnsembleFS-main.zip")

# unzip the .zip file
unzip(zipfile = "EnsembleFS-main.zip")

# run a Shiny app
library(shiny)
runApp("EnsembleFS-main/EnsembleFS")
```

3.2 Application tabs

Feature Selection tab consists of six sub-tabs dedicated to filtering and ranking informative biomarkers and comparative graphical analysis of the efficiency of used FS methods, namely, Data, Ranking list, FS Stability, Model Accuracy, Plots, and Download zip. The functionalities of the Feature Selection module are displayed in Figure 2. Here, the *EnsambleFS* application performs feature selection from data and builds a binary random forest model to predict the class for all samples. The user interface is clean and transparent. Users can freely choose the feature filters and

modify the default parameter values of the FS methods and other ML model options, such as the type of resampling method (stratified 3-fold cross-validation or 0.3 stratified random sampling), the optimum N number of variables for classifier building (N in the range from 2 to 100), the n number of iterations in the model (n in the range from 1 to 30), and the cut-off value of the Spearman correlation coefficient ρ (ρ in the range from 0 to 1). The default parameters for the module were chosen based on our experience in molecular data analysis and computer modelling. In its current version, the *EnsambleFS* demo web server accepts the input data limited to 5000 instances (30 MB limit on the total amount of data).

Data sub-tab allows the user to preview and control the quality of the tabular input data and perform the biomarker selection and classification. This module enables the selection of the subsets of relevant biomarkers using up to five FS methods, and ML algorithms. **Ranking list sub-tab** generates a interactive pivot table of the most informative biomarkers grouped by the FS filter names and the total number of occurrences of feature in all the final feature sets.

FS Stability sub-tab display summary information about the evaluation of feature stability of applied FS algorithms. The interactive table provides information about the mean ASM value for each FS method across the number of top-N features (N = 5, 10, 15, 20,..., 50, 75, 100).

Model Accuracy sub-tab compares the performance of the prediction RF models for each FS method across the N top features (N = 5, 10, 15, 20,..., 50, 75, 100). The interactive table includes the average values and error values (the standard deviation) of the ACC, AUC, and MCC metrics of predictive models built on N top features in repeated n times the k-fold cross-validation (or the random sampling).

Plots sub-tab generates comparative plots of the average values of ASM, ACC, AUC, and MCC metrics versus N top variables for all applied FS methods. The interactive plot allows the user to, for example, zoom in/out of any area of interest, select and crop a particular area, get the validation metric value of the selected point with the mouse, and filter results for each feature filter, and more.

Download zip sub-tab allows the user to download the EnsembleFS results in zip file format. The model results are output as an R object. The model parameters are saved in the text file for each FS method. All plots are saved in a PDF document.

Gene information tab is the second major functional module of EnsembleFS that supports the user in the biological analysis of molecular data. This tab allows the user access to information about the most relevant biomarkers from nine biological databases (the Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Reactome (REACT), the WikiPathways (WP), the Transfac (TF), the miRTarBase (MIRNA), the Human Protein Atlas (HPA), and the Human Phenotype Ontology (HPA)). As shown in Figure 1, user analysis may include the combined set of top biomarkers from up to five FS methods (the intersection or union of biomarker sets). Venn diagrams show the number of the most relevant biomarkers from each of the applied FS methods and the number of annotated biomarkers in the biological databases. The biological information assigned to the gene is presented in tabular format, wherein the type of display information can be set. The functionalities of the Gene Information module are shown in Figure 5.

Help tab includes the following groups: Terminology, Tutorial, and Example. Herein is demonstrated how to use the EnsembleFS web app to train ML models and search for information about key biomarkers in selected databases. The tutorial presents the short workflow and the video tutorial on YouTube (https://www.youtube.com/embed/ENf3LEMb56E). Example sub-tab

presents the capabilities and limitations of the EnsembleFS tool in a real case study. The RNA-seq data from The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) program [13], [5] is analyzed.

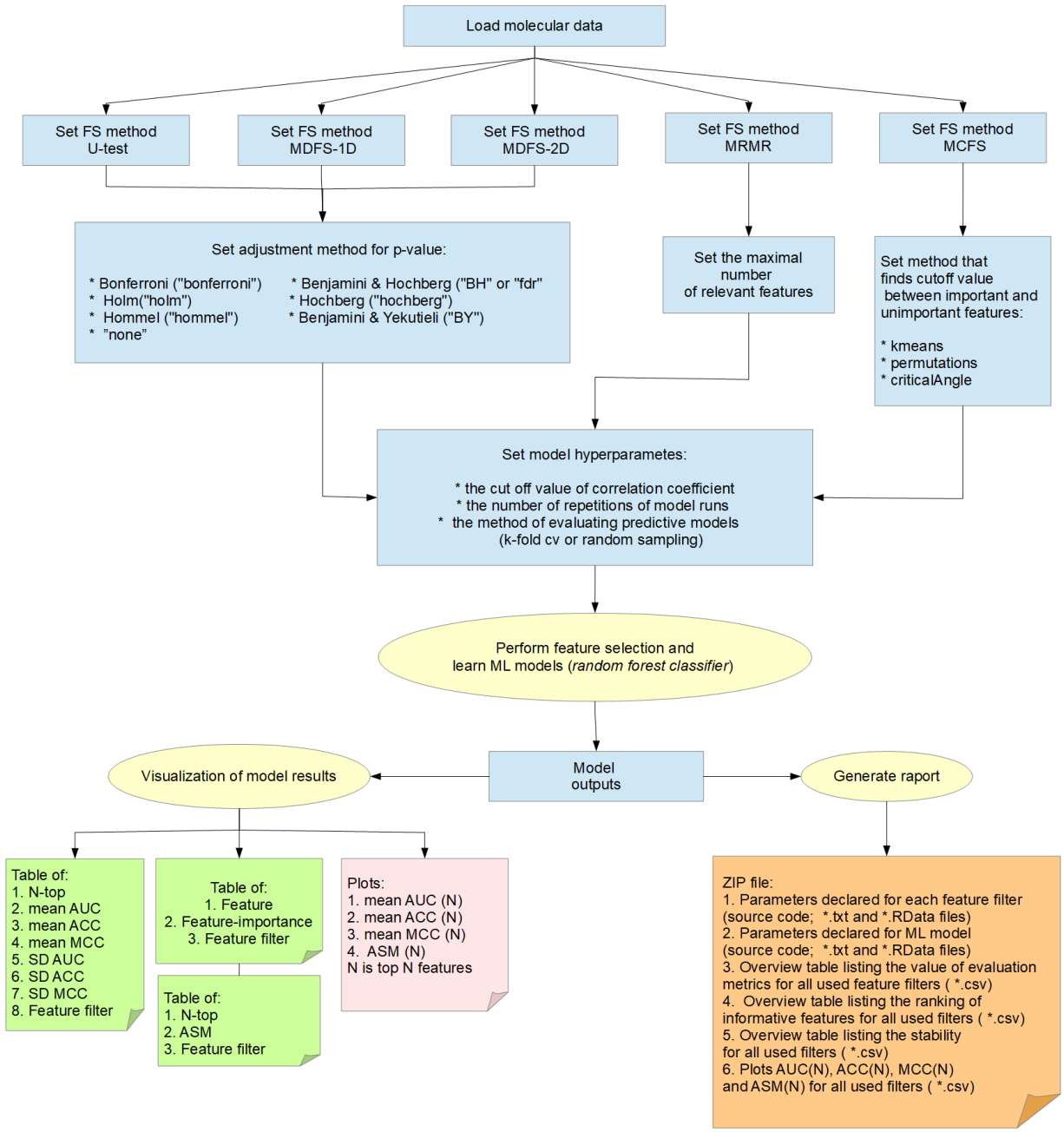

Figure 2: Main functionality of the Feature Selection module of EnsembleFS web app.

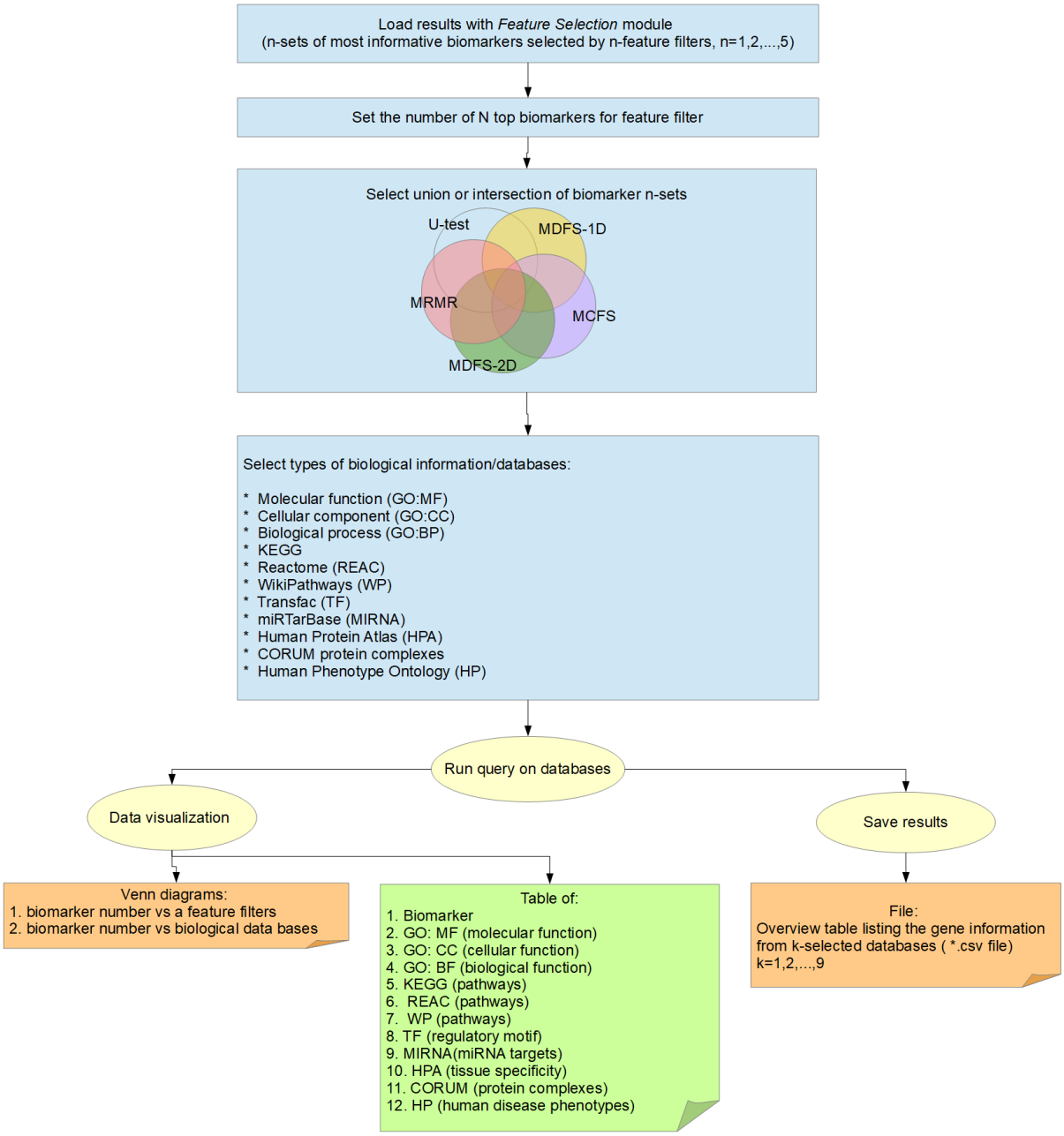

Figure 3: Main functionality of the Gene Information module of EnsembleFS web app.

3.3 Example application workflow and results

3.3.1 Data set

To demonstrate how to use the EnsembleFS web app to train ML models and search information in selected databases on genomics, transcriptomics, and proteomics, we used the RNA-seq data from TCGA-LUAD program [13], [5]. This dataset contains the differentially expressed genes (DEGs) set of tumor-adjacent normal tissues of lung adenocarcinoma cancer patients (LUAD). After the standard preprocessing procedure, the primary dataset contains 574 samples (59 normal and 515 tumours) described with 20171 DEGs. For testing purposes, we focused on 2000 of the 20171 biomarkers with the highest difference in the gene expression level between tumour and normal tissues for the above-described lung adenocarcinoma data. The LUAD data used in this study (exampleData_TCGA_LUAD_2000.zip file) can be downloaded using the GitHub link https://github.com/biocsuwb/EnsembleFS/tree/main/data_test.

3.3.2 Example workflow

Ensemble feature selection process. The process of selecting of the most informative biomarkers includes the following steps:

1. Navigate to the FEATURE SELECTION tab.
2. Load csv/txt file (separator = ";", decimal = ",") and enter the column number for the binary decision variable
3. Choose the following parameters:
    - Feature selection methods: U-test, MRMR, MCFS, MDFS-1D, MDFS-1D
    - Multitest correction: fdr
    - MRMR parameter number of significant features: 110
    - MCFS parameter cut-off method: k-means
    - Correlation coefficient: 0.75
    - Validation methods: 3-fold cross-validation
    - Number of repetitions: 10
4. Press RUN FEATURE SELECTION
5. Navigate to RANKING LIST tab for the set of most informative biomarkers
6. Navigate to FS STABILITY tab for stability calculation results
7. Navigate to MODEL ACCURACY tab for the model building results
8. Navigate to PLOT to visualize stability results and model building results
9. Navigate to DOWNLOAD ZIP tab to download all results as one archive

Searching biological information about biomarkers. The process of aggregating information about the most informative biomarkers includes the following steps:

1. Navigate to GENE INFORMATION tab for biological information on the top biomarkers
2. Number of relevant biomarkers:
   - Top N features with FS filter: 100
   - Combination of a set of biomarkers: union
   - Databases: all
3. Press GET ANALYSIS

Another example workflow for TCGA-LUAD data (500 DEGs) is described in the tab: Help → Example. Video tutorial is available at https://www.youtube.com/embed/ENf3LEMb56E.

3.3.3 Example results

Below we present a part of the results of the feature selection process, RF classification, and biological gene information collection for the LUAD-TCGA data (574 samples, 2000 biomarkers) described above.

Relevant biomarkers from the individual FS methods (RANKING LIST sub-tab).

DATA | RANKING LIST | FS STABILITY | MODEL ACCURACY | PLOTS | DOWNLOAD ZIP

Show 25 entries Search:

biomarker.name	frequency	method
AADAC	30	utest
AATK	30	utest
ABCA10	30	utest
ABCA12	30	utest
ABCA4	30	utest
ABCC13	30	utest
ABCC2	30	utest
ABCC3	30	utest
ABCG2	30	utest
ABLIM3	30	utest
ABP1	30	utest
ACMSD	30	utest
ACOXL	30	utest
ACSS3	30	utest
ACY3	30	utest

Figure 4: A part of the result of the rank list of relevant biomarkers from each of the feature selection methods, where a frequency is the number of occurrences of the biomarker in the 30 obtained feature subsets.

Stability of informative biomarker subsets (FS STABILITY sub-tab).

DATA | RANKING LIST | FS STABILITY | MODEL ACCURACY | PLOTS | DOWNLOAD ZIP

Show 25 entries Search:

N	stability.asm	method
5	0.4370067	utest
10	0.4930939	utest
15	0.5471888	utest
20	0.6364177	utest
30	0.6661783	utest
40	0.6808240	utest
50	0.6898834	utest
75	0.7041584	utest
100	0.6917898	utest
5	0.5790875	mrmr
10	0.5910486	mrmr
15	0.5794081	mrmr
20	0.5777293	mrmr
30	0.5765939	mrmr
40	0.5604011	mrmr

Figure 5: A part of the result of the stability of the 10 feature subsets composed of N-top uncorrelated features for all feature filters. The value of the stability of feature ranking method is expressed by the Lustgarten stability measure (ASM)

Model accuracy (MODEL ACCURACY sub-tab).

DATA | RANKING LIST | FS STABILITY | MODEL ACCURACY | PLOTS | DOWNLOAD ZIP

Show 25 entries Search:

N	mean.acc	mean.auc	mean.mcc	sd.acc	sd.auc	sd.mcc	method
5	0.9878070	0.9676896	0.9348178	0.008167782	0.02292876	0.04330351	utest
10	0.9904221	0.9743881	0.9491188	0.005986269	0.01653527	0.03047891	utest
15	0.9898977	0.9719303	0.9459845	0.006558149	0.01892269	0.03400362	utest
20	0.9902467	0.9735975	0.9479650	0.006532678	0.01782391	0.03370703	utest
30	0.9902485	0.9735981	0.9478493	0.005930441	0.01763384	0.03095354	utest
40	0.9897277	0.9718345	0.9451012	0.006197785	0.01772288	0.03187701	utest
50	0.9902522	0.9736425	0.9480075	0.006527504	0.01781246	0.03361428	utest
75	0.9907712	0.9776148	0.9509530	0.006379344	0.01527342	0.03270176	utest
100	0.9912957	0.9764338	0.9535080	0.006329094	0.01696770	0.03308630	utest
5	0.9874525	0.9583966	0.9316871	0.007101607	0.02357831	0.03873800	mrmr
10	0.9909393	0.9722952	0.9512085	0.005118920	0.01368568	0.02697868	mrmr
15	0.9916429	0.9742884	0.9553011	0.005234361	0.01596349	0.02728802	mrmr
20	0.9921637	0.9768317	0.9580562	0.004891170	0.01495917	0.02549865	mrmr
30	0.9921656	0.9783484	0.9584286	0.006091652	0.01482396	0.03107823	mrmr
40	0.9925128	0.9785422	0.9601179	0.005926926	0.01338576	0.03026225	mrmr

Figure 6: A part of the result of the predictive power of random forest models trained on top N features with different feature filters, where ACC is accuracy, AUC is the area under the ROC curve, and MCC is the Matthews correlation coefficient.

Model comparison plots (PLOTS sub-tab).

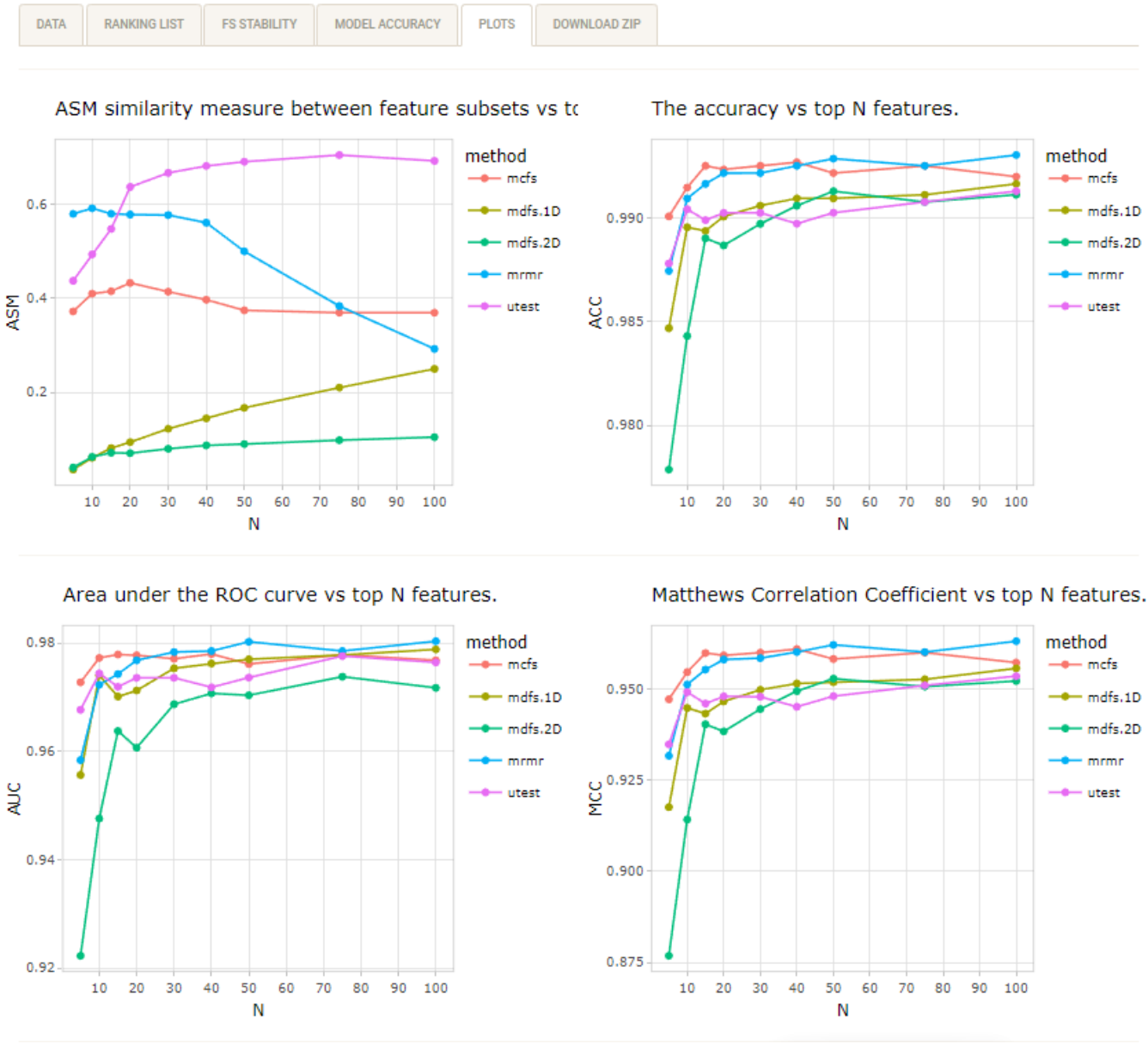

Figure 7: The average values for ACC, AUC, and MCC vs N top features for all features filters. ASM similarity measure between 10 feature subsets vs N top features.

Key biomarkers (GENE INFORMATION tab).

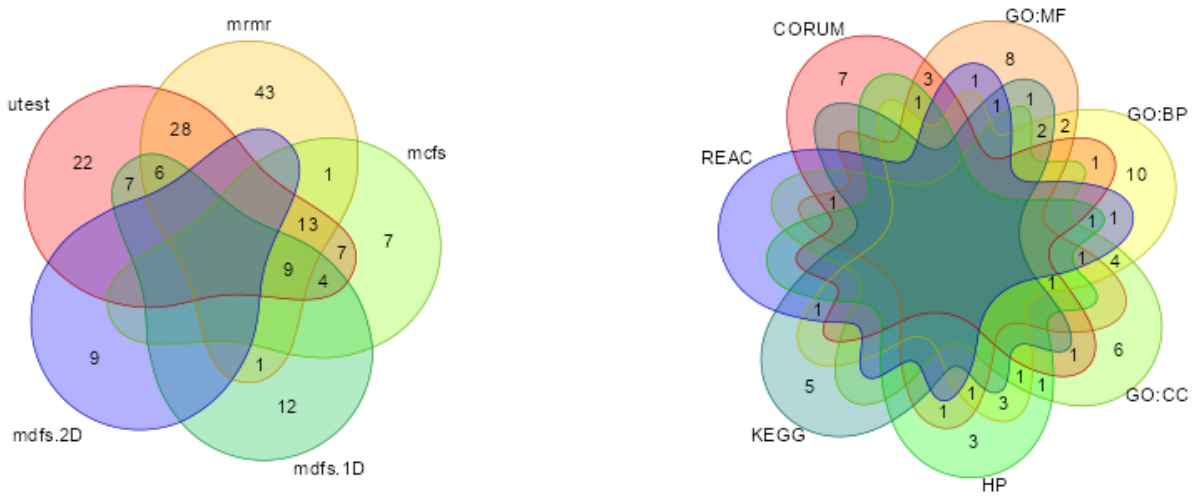

Figure 8: Left panel: the number of the most relevant biomarkers with all feature selection methods. Right panel: the number of annotated biomarkers in the biological databases

Biological gene information collection (GENE INFORMATION tab).

Show 25 entries Search:

term	source	term.ID	term.name
ADRB2	GO:MF	GO:0004941	beta2-adrenergic receptor activity
CAV1	GO:MF	GO:0070320	inward rectifier potassium channel inhibitor activity
GPT2	GO:MF	GO:0004021	L-alanine:2-oxoglutarate aminotransferase activity
GPT2	GO:MF	GO:0047635	alanine-oxo-acid transaminase activity
SLC39A8	GO:MF	GO:0097079	selenite:proton symporter activity
SLC39A8	GO:MF	GO:0140412	zinc:bicarbonate symporter activity
OTC	GO:MF	GO:0004585	ornithine carbamoyltransferase activity
OTC	GO:MF	GO:0016743	carboxyl- or carbamoyltransferase activity
STX1A	GO:MF	GO:0032028	myosin head/neck binding
DPYSL2	GO:MF	GO:0004157	dihydropyrimidinase activity
ADCY8	GO:MF	GO:0008294	calcium- and calmodulin-responsive adenylate cyclase activity
KCNT2	GO:MF	GO:0070089	chloride-activated potassium channel activity
KCNT2	GO:MF	GO:0005228	intracellular sodium activated potassium channel activity
PRKG2	GO:MF	GO:0004692	cGMP-dependent protein kinase activity
AGER	GO:MF	GO:0050785	advanced glycation end-product receptor activity

Figure 9: Molecular function annotation from the Gene Ontology database.

Show 25 entries Search:

term	source	term.ID	term.name
CAT	GO:CC	GO:0062151	catalase complex
CAV1	GO:CC	GO:0002095	caveolar macromolecular signaling complex
TNNC1	GO:CC	GO:1990584	cardiac Troponin complex
WNT3A	GO:CC	GO:1990851	Wnt-Frizzled-LRP5/6 complex
STX1A	GO:CC	GO:0070032	synaptobrevin 2-SNAP-25-syntaxin-1a-complexin I complex
STX1A	GO:CC	GO:0070033	synaptobrevin 2-SNAP-25-syntaxin-1a-complexin II complex
STX1A	GO:CC	GO:0070044	synaptobrevin 2-SNAP-25-syntaxin-1a complex
PAFAH1B3	GO:CC	GO:0008247	1-alkyl-2-acetylglycerophosphocholine esterase complex
HTR3C	GO:CC	GO:1904602	serotonin-activated cation-selective channel complex
MYO7A	GO:CC	GO:0120044	stereocilium base
COL10A1	GO:CC	GO:0005599	collagen type X trimer
COL10A1	GO:CC	GO:0005598	short-chain collagen trimer
COL10A1	GO:CC	GO:0030935	sheet-forming collagen trimer
COL10A1	GO:CC	GO:0098646	collagen sheet
SPTBN2	GO:CC	GO:0099189	postsynaptic spectrin-associated cytoskeleton

Figure 10: Cellular component from the Gene Ontology database.

Show 25 entries Search:

term	source	term.ID	term.name
ANGPTL7	GO:BP	GO:0036331	avascular cornea development in camera-type eye
ANGPTL7	GO:BP	GO:1901346	negative regulation of vasculature development involved in avascular cornea development in camera-type eye
CAT	GO:BP	GO:0061691	detoxification of hydrogen peroxide
CAT	GO:BP	GO:0061692	cellular detoxification of hydrogen peroxide
CAV1	GO:BP	GO:1900085	negative regulation of peptidyl-tyrosine autophosphorylation
CAV1	GO:BP	GO:1903609	negative regulation of inward rectifier potassium channel activity
CD5L	GO:BP	GO:1903661	positive regulation of complement-dependent cytotoxicity
CYP1A2	GO:BP	GO:0009403	toxin biosynthetic process
PRKCE	GO:BP	GO:2001031	positive regulation of cellular glucuronidation
SLC39A8	GO:BP	GO:0030026	cellular manganese ion homeostasis
SLC39A8	GO:BP	GO:0097080	plasma membrane selenite transport
SLC39A8	GO:BP	GO:1990540	mitochondrial manganese ion transmembrane transport
WNT3A	GO:BP	GO:0003136	negative regulation of heart induction by canonical Wnt signaling pathway
WNT3A	GO:BP	GO:0009997	negative regulation of cardioblast cell fate specification
WNT3A	GO:BP	GO:0021874	Wnt signaling pathway involved in forebrain neuroblast division

Figure 11: Biological process annotation from the Gene Ontology database.

Show 50 entries Search:

term	source	term.ID	term.name
CYP1A2	KEGG	KEGG:00232	Caffeine metabolism
GPT2	KEGG	KEGG:01210	2-Oxocarboxylic acid metabolism
GPT2	KEGG	KEGG:00220	Arginine biosynthesis
OTC	KEGG	KEGG:00220	Arginine biosynthesis
SLC25A10	KEGG	KEGG:04964	Proximal tubule bicarbonate reclamation
FUT2	KEGG	KEGG:00603	Glycosphingolipid biosynthesis - globo and isoglobo series
CLEC4M	KEGG	KEGG:03264	Virion - Flavivirus
CLEC4M	KEGG	KEGG:03260	Virion - Human immunodeficiency virus
HS6ST2	KEGG	KEGG:00534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
RBP2	KEGG	KEGG:04977	Vitamin digestion and absorption
ST6GALNAC5	KEGG	KEGG:00604	Glycosphingolipid biosynthesis - ganglio series
IL4I1	KEGG	KEGG:00400	Phenylalanine, tyrosine and tryptophan biosynthesis
IL4I1	KEGG	KEGG:00360	Phenylalanine metabolism

Figure 12: KEGG.

Show 25 entries Search:

term	source	term.ID	term.name
CYP1A2	REAC	REAC:R-HSA-211957	Aromatic amines can be N-hydroxylated or N-dealkylated by CYP1A2
CYP1A2	REAC	REAC:R-HSA-9018681	Biosynthesis of protectins
STX1A	REAC	REAC:R-HSA-5250971	Toxicity of botulinum toxin type C (botC)
ACADL	REAC	REAC:R-HSA-77285	Beta oxidation of myristoyl-CoA to lauroyl-CoA
MGAT3	REAC	REAC:R-HSA-975574	Reactions specific to the hybrid N-glycan synthesis pathway
FUT2	REAC	REAC:R-HSA-9033807	ABO blood group biosynthesis
SLC4A1	REAC	REAC:R-HSA-5619050	Defective SLC4A1 causes hereditary spherocytosis type 4 (HSP4), distal renal tubular acidosis (dRTA) and dRTA with hemolytic anemia (dRTA-HA)
SLC6A4	REAC	REAC:R-HSA-380615	Serotonin clearance from the synaptic cleft

Figure 13: Reactome.

Show 25 entries Search:

term	source	term.ID	term.name
CYP1A2	WP	WP:WP2646	Lidocaine metabolism
CYP1A2	WP	WP:WP3633	Caffeine and theobromine metabolism
CYP1A2	WP	WP:WP694	Arylamine metabolism
CYP1A2	WP	WP:WP2542	Sulindac metabolic pathway
CYP1A2	WP	WP:WP699	Aflatoxin B1 metabolism
GPT2	WP	WP:WP4661	Amino acid metabolism pathway excerpt: histidine catabolism extension
OTC	WP	WP:WP4571	Urea cycle and related diseases
FABP4	WP	WP:WP4400	FABP4 in ovarian cancer
AGER	WP	WP:WP4479	Supression of HMGB1 mediated inflammation by THBD
ACADL	WP	WP:WP5241	Mitochondrial beta-oxidation
ALAS2	WP	WP:WP561	Heme biosynthesis
ALAS2	WP	WP:WP5169	Hemesynthesis defects and porphyrias
PCSK9	WP	WP:WP2846	PCSK9-mediated LDL receptor degradation
PCSK9	WP	WP:WP3408	Evolocumab mechanism to reduce LDL cholesterol

Figure 14: WikiPathways.

Show 25 entries Search:

term	source	term.ID	term.name
CYP1A2	TF	TF:M06228	Factor: ZNF543; motif: KGGWATRTGGGA
FAM189A2	TF	TF:M05901	Factor: ZNF561; motif: GATAGGGGNCGRTMGTCG
RTKN2	TF	TF:M06570	Factor: ZNF75CP; motif: KGTGTAGACATC
TACC1	TF	TF:M06150	Factor: ZNF81; motif: NTGGTTAAACGA
FABP4	TF	TF:M06478	Factor: ZNF433; motif: AGTCCAGATTAC
PECAM1	TF	TF:M06326	Factor: ZNF841; motif: NKGTCAAGAAAM
C20orf202	TF	TF:M11785_1	Factor: RORbeta; motif: NAWNTAGGTCRTGACCTANWTN; match class: 1
C20orf202	TF	TF:M09899_1	Factor: ehf; motif: NNANSAGGAAGTNNN; match class: 1
C20orf202	TF	TF:M11785	Factor: RORbeta; motif: NAWNTAGGTCRTGACCTANWTN
C20orf202	TF	TF:M11787	Factor: RORbeta; motif: NAWNTAGGTCATGACCTANWTN
FERMT1	TF	TF:M06794	Factor: ZNF107; motif: NATTAAAGCCGC
ALAS2	TF	TF:M06632_1	Factor: ZNF717; motif: GGGAAAAAGA; match class: 1
IGSF9	TF	TF:M06049_1	Factor: ZNF208; motif: NGNGGGAGTTCM; match class: 1
FOLR3	TF	TF:M06725	Factor: ZNF624; motif: NGGTCAATAYGA
PDPN	TF	TF:M12446_1	Factor: BHLHE22; motif: NNCAGCTGNN; match class: 1

Figure 15: Transfac.

Show 25 entries Search:

term	source	term.ID	term.name
GPT2	MIRNA	MIRNA:hsa-miR-4783-3p	hsa-miR-4783-3p
NCKAP5	MIRNA	MIRNA:hsa-miR-518e-3p	hsa-miR-518e-3p
STX11	MIRNA	MIRNA:hsa-miR-5188	hsa-miR-5188
ARHGAP31	MIRNA	MIRNA:hsa-miR-4730	hsa-miR-4730
FABP4	MIRNA	MIRNA:hsa-miR-369-5p	hsa-miR-369-5p
AMOTL1	MIRNA	MIRNA:hsa-miR-4479	hsa-miR-4479
KIF26B	MIRNA	MIRNA:hsa-miR-3944-3p	hsa-miR-3944-3p

Figure 16: miRTarBase.

Show 25 entries Search:

term	source	term.ID	term.name
PRKCE	HPA	HPA:0250152	hippocampus; synapses[≥Medium]
PRKCE	HPA	HPA:0250151	hippocampus; synapses[≥Low]
SCUBE1	HPA	HPA:0411243	retina; inner plexiform layer[High]
SCUBE1	HPA	HPA:0411242	retina; inner plexiform layer[≥Medium]
SCUBE1	HPA	HPA:0411253	retina; nerve fiber layer[High]
SCUBE1	HPA	HPA:0411263	retina; outer plexiform layer[High]
DPYSL2	HPA	HPA:0020033	adrenal gland; cells in zona glomerulosa[High]
DPYSL2	HPA	HPA:0020022	adrenal gland; cells in zona fasciculata[≥Medium]
DPYSL2	HPA	HPA:0020041	adrenal gland; cells in zona reticularis[≥Low]
DPYSL2	HPA	HPA:0020042	adrenal gland; cells in zona reticularis[≥Medium]
FGF11	HPA	HPA:0020043	adrenal gland; cells in zona reticularis[High]
FGF11	HPA	HPA:0020041	adrenal gland; cells in zona reticularis[≥Low]
FGF11	HPA	HPA:0020042	adrenal gland; cells in zona reticularis[≥Medium]
SLC6A4	HPA	HPA:0140141	dorsal raphe; neuronal projections[≥Low]
SLC6A4	HPA	HPA:0140142	dorsal raphe; neuronal projections[≥Medium]

Figure 17: The pathways annotation from the Human Protein Atlas database.

Show 25 entries Search:

term	source	term.ID	term.name
CYP1A2	TF	TF:M06228	Factor: ZNF543; motif: KGGWATRTGGGA
FAM189A2	TF	TF:M05901	Factor: ZNF561; motif: GATAGGGGNCGRTMGTCG
RTKN2	TF	TF:M06570	Factor: ZNF75CP; motif: KGTGTAGACATC
TACC1	TF	TF:M06150	Factor: ZNF81; motif: NTGGTTAAACGA
FABP4	TF	TF:M06478	Factor: ZNF433; motif: AGTCCAGATTAC
PECAM1	TF	TF:M06326	Factor: ZNF841; motif: NKGTCAAGAAAM
C20orf202	TF	TF:M11785_1	Factor: RORbeta; motif: NAWNTAGGTCRTGACCTANWTN; match class: 1
C20orf202	TF	TF:M09899_1	Factor: ehf; motif: NNANSAGGAAGTNNN; match class: 1
C20orf202	TF	TF:M11785	Factor: RORbeta; motif: NAWNTAGGTCRTGACCTANWTN
C20orf202	TF	TF:M11787	Factor: RORbeta; motif: NAWNTAGGTCATGACCTANWTN
FERMT1	TF	TF:M06794	Factor: ZNF107; motif: NATTAAAGCCGC
ALAS2	TF	TF:M06632_1	Factor: ZNF717; motif: GGGAAAAAGA; match class: 1
IGSF9	TF	TF:M06049_1	Factor: ZNF208; motif: NGNGGGAGTTCM; match class: 1
FOLR3	TF	TF:M06725	Factor: ZNF624; motif: NGGTCAATAYGA
PDPN	TF	TF:M12446_1	Factor: BHLHE22; motif: NNCAGCTGNN; match class: 1

Figure 18: Transfac.

Show 25 entries Search:

term	source	term.ID	term.name
ADRB2	CORUM	CORUM:3830	ADRB2 homodimer complex
ADRB2	CORUM	CORUM:668	BKCA-beta2AR-AKAP79 signaling complex
ADRB2	CORUM	CORUM:672	BKCA-beta2AR complex
ADRB2	CORUM	CORUM:687	CFTR-NHERF-beta(2)AR signaling complex
CAV1	CORUM	CORUM:2462	Caveolin-1 homodimer complex
CAV1	CORUM	CORUM:5714	NOS3-CAV1 complex
CAV1	CORUM	CORUM:550	NOS3-CAV1-NOSTRIN complex
CAV1	CORUM	CORUM:5862	CAV1-VDAC1-ESR1 complex
CHRNA2	CORUM	CORUM:6440	CHRNA2-CHRNB4 complex
EPAS1	CORUM	CORUM:7274	ARNTL-EPAS1 complex
SH3GL3	CORUM	CORUM:2457	CIN85-SH3GL3 complex
STX11	CORUM	CORUM:3196	FHL4/STX11-ACT complex
STX1A	CORUM	CORUM:873	SNARE complex (STX1A, SNAP29)
STX1A	CORUM	CORUM:915	SNARE complex (SNAP23, STX1A)
ADRA1A	CORUM	CORUM:6807	ADRA1A-CXCR4 complex

Figure 19: CORUM protein complexes.

Show 25 entries Search:

term	source	term.ID	term.name
CAT	HP	HP:0012517	Reduced catalase level
CAT	HP	HP:0040113	Old-aged sensorineural hearing impairment
CAV1	HP	HP:0005320	Lack of facial subcutaneous fat
SLC39A8	HP	HP:0032098	Hypomanganesemia
LIMS2	HP	HP:0030284	Triangular tongue
ARHGAP31	HP	HP:0004476	Aplasia cutis congenita over parietal area
ARHGAP31	HP	HP:0007590	Aplasia cutis congenita over posterior parietal area
PRKG2	HP	HP:0003889	Abnormal deltoid tuberosity morphology
PRKG2	HP	HP:0003890	Prominent deltoid tuberosities
PRKG2	HP	HP:0003926	Abnormal humeral diaphysis morphology
FERMT1	HP	HP:0100517	Neoplasm of the urethra
COL10A1	HP	HP:0006414	Distal tibial bowing
COL10A1	HP	HP:0006634	Osteosclerosis of ribs
COL10A1	HP	HP:0045079	Distal femoral metaphyseal irregularity
COLEC10	HP	HP:0030025	Auricular pit

Figure 20: Human Phenotype Ontology.

Show 25 entries Search:

term	source	term.ID	term.name
ADRB2	CORUM	CORUM:3830	ADRB2 homodimer complex
ADRB2	CORUM	CORUM:668	BKCA-beta2AR-AKAP79 signaling complex
ADRB2	CORUM	CORUM:672	BKCA-beta2AR complex
ADRB2	CORUM	CORUM:687	CFTR-NHERF-beta(2)AR signaling complex
ADRB2	GO:MF	GO:0004941	beta2-adrenergic receptor activity
ANGPTL7	GO:BP	GO:0036331	avascular cornea development in camera-type eye
ANGPTL7	GO:BP	GO:1901346	negative regulation of vasculature development involved in avascular cornea development in camera-type eye
CAT	GO:BP	GO:0061691	detoxification of hydrogen peroxide
CAT	GO:BP	GO:0061692	cellular detoxification of hydrogen peroxide
CAT	GO:CC	GO:0062151	catalase complex
CAT	HP	HP:0012517	Reduced catalase level
CAT	HP	HP:0040113	Old-aged sensorineural hearing impairment
CAV1	CORUM	CORUM:2462	Caveolin-1 homodimer complex
CAV1	CORUM	CORUM:5714	NOS3-CAV1 complex
CAV1	CORUM	CORUM:550	NOS3-CAV1-NOSTRIN complex

Figure 21: All.

3.4 Report of feature selection and modelling results

3.4.1 Report files

Final report files with feature selection and ML tasks are compressed into ZIP file. The ZIP file contains the following:

- **ranking.csv** - set of most informative biomarkers;
- **model.csv** - model building results;
- **stability.csv** - feature selection stability measure;
- **result.pdf** - visualize stability results and model building results;
- **info.txt** - used parameters for the ensemble feature selection;
- **full_result.RData** - all results in one file (R's native file format);
- **utest.txt, mrmr.txt, mcfs.txt, mdfs1d.txt, mdfs2d.txt** - source code feature selection methods with used parameters.

An additional file **gene_information.csv** stores the crucial information about the most informative genes from the nine databases that support biological and biomedical research.

3.4.2 Sample report

Data set. Herein, we present the example of a generated report from the EnsembleFS web app for an exemplary METABRIC dataset [23]. The copy-number alterations data (CNA) was used (see Figure 22). The preprocessing procedure for this set is described in [25]. The clinical endpoints for 1394 breast cancer patients (781 survivors and 613 deceased) were predicted. For testing purposes, the set of biomarkers was limited to 2000 biomarkers with the highest difference in CNA levels between the two groups of patients. This dataset (exampleData_METABRIC_2000.csv file) can be downloaded using the GitHub link https://github.com/biocsuwb/EnsembleFS/tree/main/data_test.

DATA | RANKING LIST | FS STABILITY | MODEL ACCURACY | PLOTS | DOWNLOAD ZIP

Show 25 entries Search:

class	NOV	MTBP	MRPL13	DSCC1	TAF2	MAL2	SNTB1	COL14A1
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1
0	2	2	2	2	2	2	2	2
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Figure 22: View of the exemplary METABRIC dataset.

Model configuration parameters:

- feature selection methods (FS): U-test, MRMR, MCFS, MDFS-1D, MDFS-2D;
- multitest correction: fdr;
- MRMR parameter number of significant features: 200;
- MCFS parameter cut-off method: k-means;
- correlation coefficient: 0.75;
- validation methods: random sample (test set 30%);
- number iteration: 10;
- top N features with FS filter: 100;
- combination of a set of biomarkers: union.

The contents of outputs files:

1. The ranking list of the most informative biomarkers (ranking.csv file), see Figure 23.

	A	B	C	D	E	F	G	H	I	J	K
1		biomarker.name	frequency	method							
2	3	ARHGEF3	10	utest							
3	47	LINC01046	10	utest							
4	79	RNF168	10	utest							
5	7	C18orf42	9	utest							
6	...	...	...	...	etc.						
7	310	ARHGEF3	10	mrmr							
8	411	FAM117A	10	mrmr							
9	721	LINC01046	10	mrmr							
10	135	UBE2MP1	10	mrmr							
11	251	CSMD2	9	mrmr							
12	...	...	...	...	etc.						
13	232	SH2D4A	4	mcfs							
14	412	CDRT4	3	mcfs							
15	712	CTTN	2	mcfs							
16	...	...	...	...	etc.						
17	492	FAM117A	10	mdfs.1D							
18	262	CDRT4	8	mdfs.1D							
19	382	DBT	5	mdfs.1D							
20	...	...	...	...	etc.						
21	1432	RNF168	9	mdfs.2D							
22	493	FAM117A	8	mdfs.2D							
23	463	ERBB2	5	mdfs.2D							
24	...	...	...	...	etc.						

Figure 23: The frequency of occurrence of the most informative biomarkers in 10 feature subsets for each FS method (ranking.csv file).

2. The stability of feature selection methods (stability.csv file), see Figure 24,

	A	B	C	D	E	F	G	H	I	J	K
1		N	stability.asm	method							
2	1	5	0.0872727272727273	utest							
3	2	10	0.176767676767677	utest							
4	3	15	0.177373737373737	utest							
5	4	20	0.126868686868687	utest							
6	5	30	0.178401448133249	utest							
7	6	40	0.178401448133249	utest							
8	7	50	0.178401448133249	utest							
9	8	75	0.178401448133249	utest							
10	9	100	0.178401448133249	utest							
11	10	5	0.0375540641312453	mrmr							
12	11	10	0.068441461595824	mrmr							
13	12	15	0.112662192393736	mrmr							
14	...	...	...	...	etc.						

Figure 24: The Lustgarten adjusted stability measure (ASM) values for top N features for each filter.

3. The results of predictive models (model.csv file), see Figure 25.

A	B	C	D	E	F	G	H	I
	N	mean.acc	mean.auc	mean.mcc	sd.acc	sd.auc	sd.mcc	method
1	5	0.562877385652728	0.547480919500909	0.0989436615140948	0.0245864866117616	0.0234998447332198	0.0490923751544072	utest
2	10	0.573086018617928	0.554301715626862	0.114763391227313	0.0228037615077848	0.022171383340788	0.0470579717544721	utest
3	15	0.581553967956734	0.565570526232469	0.136425554292445	0.0170833493263786	0.0178119393686916	0.0388428845661148	utest
4	20	0.586629980239546	0.573906479570657	0.150993576218337	0.024217786172024	0.0248970633626587	0.05192223740291	utest
5	30	0.597907779428847	0.587641565014439	0.177712836643514	0.0185036034223721	0.0195404917826178	0.0405643434299255	utest
6	40	0.59404092085772	0.583569738616687	0.169392729228894	0.0200020838389806	0.0215873526575906	0.0445044768159888	utest
7	50	0.591810799510339	0.581745528056807	0.165284240294732	0.0191329278812697	0.020740871745519	0.0417711576526861	utest
8	75	0.592056000001825	0.581838566576728	0.165794725427549	0.0172492050209299	0.0189150357859744	0.0394626929545466	utest
9	100	0.595417395137236	0.584468647786626	0.171635061315498	0.0145742147258853	0.015254610899241	0.0316553596867807	utest
10	5	0.56999587159005	0.551560764337663	0.109601658541233	0.0241913431273658	0.0234124418023388	0.050878898832222	mrmr
11	10	0.589812641792948	0.575675431327348	0.155130528342418	0.0187493360291753	0.0184614621783327	0.0362684106405233	mrmr
12	15	0.587705890769634	0.575396578192959	0.153363283894822	0.0188363295602237	0.0166153122368639	0.0338507702691667	mrmr
13	20	0.591604882145172	0.581084661134259	0.164037486907819	0.0214537812049896	0.0233267572776642	0.0463642553766794	mrmr
14	30	0.59771467812255	0.587446424035102	0.177378073100681	0.0219687591269373	0.0221500516515061	0.0458434888673972	mrmr
15	40	0.59605798356403	0.586324011963211	0.175026521927913	0.0298308720255611	0.0274565425090753	0.0565198790851156	mrmr
16	50	0.597113166682141	0.58761656592534	0.177531548293272	0.0239257099001541	0.0226065073827527	0.0465307522483843	mrmr
17	75	0.596906871468446	0.587666233328439	0.177217496527974	0.0269864828508867	0.0258819083609028	0.0527780401970378	mrmr
18	100	0.597192580737265	0.587372201737839	0.17708039112518	0.0282445422644443	0.0278280988942414	0.0570554842881587	mrmr
19	5	0.571142563239938	0.552363629150504	0.111467257692329	0.0173559179704809	0.0188036418321288	0.0356156796934172	mcfs
20	10	0.574871390074439	0.554691625199472	0.116660600605028	0.014781608713917	0.0167346660356388	0.032462046753666	mcfs
21	15	0.576717651780557	0.556506767420643	0.120563831494633	0.0139495260801689	0.0164472462410258	0.0317554198699697	mcfs
...	...	...	...	...	...	...	...	etc.

Figure 25: The mean values of evaluation metrics: the accuracy (ACC), the Matthews correlation coefficient (MCC), and the area under the receiver operator curve (AUC), for predictive models constructed using top N features returned by each feature filter.

4. The graphs for model performance comparison (result.pdf file), see Figure 26,

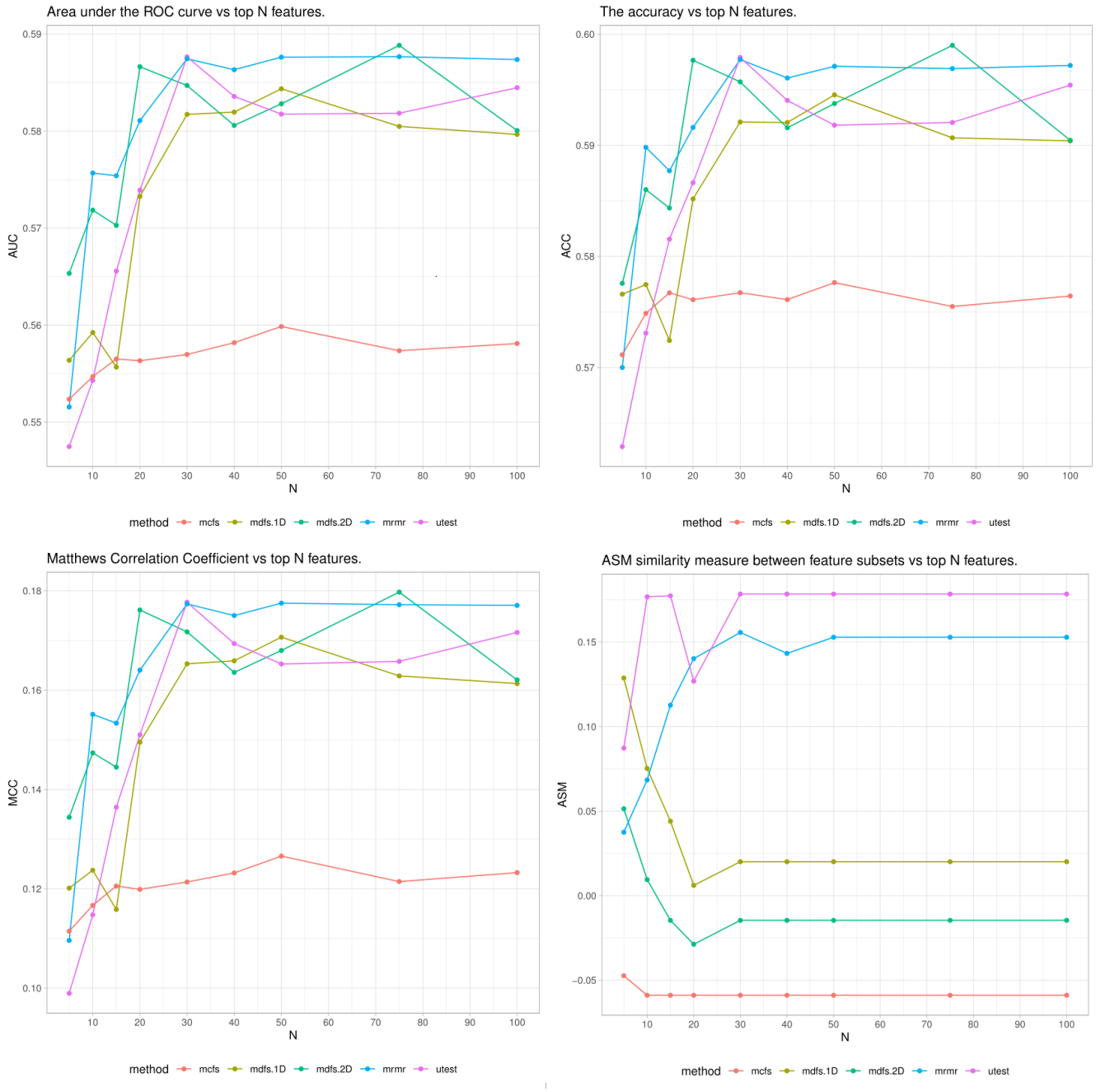

Figure 26: The average values for the accuracy (ACC), the Matthews correlation coefficient (MCC), the area under receiver operator curve (AUC), and ASM similarity measure between 10 feature subsets as a function of N top features for all filters.

5. The list of feature selection parameters and predictive model parameters (info.txt file):

```
list(methods = c("fs.utest",
                 "fs.mrmr",
                 "fs.mcfs",
                 "fs.mdfs.1D",
                 "fs.mdfs.2D"),
     p.adjust = "fdr,"
     level.correlation = 0.75,
     validation = "rsampling,"
     gene.info = NULL,
     number.repeats = 10)
```

6. The code of function that calls feature filter and return the ranking of relevant variables with analyzed data (utest.txt, mrmr.txt, mcfs.txt mdfs1D.txt, and mdfs2D.txt files), e.g., for MDFS-1D method:

```
function (x, y, params = list(adjust = "holm", alpha = 0.05))
{
 if(!is.data.frame(x))
    data = as.data.frame(x)
    dim0 = 1
    div0 = 3
    adjust = params$adjust
    alpha = params$alpha
    result = MDFS(data = x,
                  decision = y,
                  dimensions = dim0,
                  divisions = div0,
                  use.CUDA = FALSE,
                  p.adjust.method = adjust)
    var.names = names(x)
    index.imp = RelevantVariables(result$MDFS,
                                  level = alpha,
                                  p.adjust.method = adjust)
    var.imp.frame = data.frame(name = var.names,
                               Pvalue = result$p.value,
                               adjustPvalue = result$adjusted.p.value
                               )[index.imp,]
    var.imp = var.imp.frame[order(var.imp.frame$adjustPvalue,
                                  decreasing = F),]
    return(var.imp)
}
```

7. Information collected about the most informative biomarkers (gene.information.csv), see Figure 27.

Full report. The full report is available on GitHub repository
https:/github.com/biocsuwb/EnsembleFS/tree/main/data_test
in the exampleRaport_METABRIC_2000 folder.

	A	B	C	D	E	F
1	term	source	term.ID	term.name		
2	RNF168	GO:BP	GO:0036351	histone H2A-K13 ubiquitination		
3	RNF168	GO:BP	GO:0036352	histone H2A-K15 ubiquitination		
4	RNF168	WP	WP:WP5119	NIPBL role in DNA damage - Cornelia de Lange syndrome		
5	DBT	GO:MF	GO:0031405	lipoic acid binding		
6	DBT	GO:MF	GO:0043754	dihydrolipoyllysine-residue (2-methylpropanoyl)transferase activity		
7	TPK1	GO:BP	GO:0009229	thiamine diphosphate biosynthetic process		
8	TPK1	GO:BP	GO:0042724	thiamine-containing compound biosynthetic process		
9	TPK1	GO:MF	GO:0004788	thiamine diphosphokinase activity		
10	TPK1	GO:MF	GO:0030975	thiamine binding		
11	TPK1	KEGG	KEGG:00730	Thiamine metabolism		
12	TPK1	WP	WP:WP4297	Thiamine metabolic pathways		
13	PLCH2	HPA	HPA:0450622	skin		
14	PLCH2	HPA	HPA:0450621	skin		
15	PLCH2	HPA	HPA:0450000	skin		
16	PLCH2	TF	TF:M11790_1	Factor: REVERB-beta		
17	RELN	GO:BP	GO:1902076	regulation of lateral motor column neuron migration		
18	RELN	GO:BP	GO:1902078	positive regulation of lateral motor column neuron migration		
19	RELN	GO:BP	GO:1905483	regulation of motor neuron migration		
20	RELN	GO:BP	GO:1905485	positive regulation of motor neuron migration		
21	RELN	GO:CC	GO:0110157	reelin complex		
22	CDKN2B	MIRNA	MIRNA:hsa-miR-4790-5p	hsa-miR-4790-5p		
23	ROCK1	CORUM	CORUM:7424	ISLR2-ROCK1 complex		
24	ROCK1	GO:CC	GO:0106003	amyloid-beta complex		
25	ROCK1	MIRNA	MIRNA:hsa-miR-1280	hsa-miR-1280		
26	FHOD3	HP	HP:0031992	Apical hypertrophic cardiomyopathy		
27	FHOD3	HP	HP:4000001	Abnormal cardiac magnetic resonance imaging finding		
28	FHOD3	HP	HP:4000004	Myocardial late gadolinium enhancement		
29	FHOD3	TF	TF:M12407_1	Factor: PLAGL2		
30	ERBB2	CORUM	CORUM:6506	ERBB2-SPG1 complex		
31	ERBB2	CORUM	CORUM:2528	ERBB2-MEMO-SHC complex		
32	ERBB2	CORUM	CORUM:7351	GDF15-ERBB2 complex		
33	ERBB2	GO:CC	GO:0038143	ERBB3:ERBB2 complex		
34	...	...	...	...	etc.	

Figure 27: Information about the most informative biomarkers harvested from nine databases.

3.5 Computational aspects

Data set. We used the RNA-seq data of tumor-adjacent normal tissues of lung adenocarcinoma cancer patients from TCGA-LUAD project that was previously described in Section 3.3.1. For testing purposes, the number of biomarkers was limited to 2000 DEGs, with the highest difference in the gene expression level between tumour and normal tissues.
The test data (exampleData_TCGA_LUAD_2000.zip file) can be downloaded using the GitHub link https://github.com/biocsuwb/EnsembleFS-package/tree/main/data.

Computation time. To test the speed efficiency of the developed web application, we reviewed its performance for various data sizes. The following values of parameters were set up: multitest correction = fdr (U-test, MDFS-1D, MDFS-2D), number of relevant variables = 100 (MRMR), cut-off method = kmeans (MCFS), and correlation coefficient = 0.75. Table 1 presents the example execution times of one iteration of the feature selection and classification algorithm for TCGA-LUAD data limited to p = 100, 200, 1000 random biomarkers for EnsembleFS running on a cloud server (https://uco.uwb.edu.pl/apps/EnsembleFS), and the Table 2 for EnsembleFS running on the local

machine.

The one run of the algorithm involved the following steps: calling individual or all FS algorithms, removing correlated features, estimating a ranking of the features, and calling random forest classification algorithm for 0.3 resampling method or a one-time for 3-fold cross-validation method. The above procedure was executed as in Figure 1. Additionally, both tables include the execution time of the procedure of information searching about the most informative biomarkers from the ensemble FS in nine biological databases. An analysis of the data presented in Table 1 shows that the time of algorithm performance strongly depends on the FS method and hyper-parameter tuning when the feature number is increased. Among FS algorithms used in this study, the U-test and MDFS methods are the fastest for the initial 100 features, while the MRMR method is for 1000 features. FS methods that use p-values of the test to rank features (U-test, MDFS-1D and MDFS-2D) are the most time-consuming for 1000 features. However, it should be noted that the execution time of the MRMR and MCFS algorithms increases if their default parameters are changed, namely a more significant number of relevant features for the MRMR method and other cut-off methods for the MCFS method. The execution time of the MDFS-2D algorithm depends on using the processor's architecture (CPU or GPU).

Table 1: Execution times (hh:mm:ss) for a single iteration of the feature selection algorithm and random forest classification for the TCGA lung cancer dataset with 574 samples and p biomarkers. The execution time of the information searches in biological databases (DB) for the m-number of the most informative biomarkers with the ensemble FS method (union of top biomarkers with five FS methods). Two resampling procedures, namely 3-fold cross-validation (cv) and 0.3 random sampling (rs), were used for model validation (Val). Computations were performed on a GPU-accelerated version of MDFS on NVIDIA Tesla K80 co-processor, 14 GB RAM; see notes in the text.

No	p	m	Val	U-test	MDFS-1D	MDFS-2D	MRMR	MCFS	Ensemble	DB query
1	100	80	cv	00:00:21	00:00:21	00:00:18	00:00:22	00:00:36	00:01:57	00:04:06
		79	rs	00:00:06	00:00:06	00:00:06	00:00:06	00:00:10	00:00:35	00:03:36
2	200	83	cv	00:00:34	00:00:32	00:00:32	00:00:22	00:00:42	00:02:32	00:04:08
		131	rs	00:00:12	00:00:09	00:00:12	00:00:07	00:00:15	00:00:54	00:06:21
3	1000	204	cv	00:05:09	00:06:03	00:08:28	00:00:26	00:02:01	00:19:30	00:08:46
		256	rs	00:02:31	00:01:56	00:03:26	00:00:09	00:00:44	00:09:36	00:11:25

Table 2: Execution times for a single iteration of the feature selection algorithm and random forest classification for the TCGA lung cancer dataset with 574 samples and p biomarkers. Computations were performed on the Intel Core i5-12400 CPU using 32 GB RAM. For notes, see Table 1.

No	p	m	Val	U-test	MDFS-1D	MDFS-2D	MRMR	MCFS	Ensemble	DB query
1	100	80	cv	00:00:05	00:00:04	00:00:04	00:00:04	00:00:13	00:00:31	00:05:44
		82	rs	00:00:02	00:00:01	00:00:02	00:00:02	00:00:05	00:00:11	00:05:26
2	200	140	cv	00:00:09	00:00:09	00:00:08	00:00:05	00:00:20	00:00:52	00:09:51
		149	rs	00:00:03	00:00:03	00:00:03	00:00:02	00:00:07	00:00:19	00:09:38
3	1000	205	cv	00:02:18	00:02:16	00:02:19	00:00:07	00:01:14	00:08:17	00:14:23
		265	rs	00:00:49	00:00:47	00:00:49	00:00:02	00:00:27	00:02:57	00:13:59

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995.

[2] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[3] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[4] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6), 2020. doi: 10.1186/s12864-019-6413-7.

[5] E. Collisson, J. Campbell, A. Brooks, and et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511:543–550, 2014.

[6] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains. mrmre: an r package for parallelized mrmr ensemble feature selection. *Bioinformatics*, 29(18): 2365–2368, 2013.

[7] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, 2005.

[8] M. Draminski and J. Koronacki. rmcfs: An r package for monte carlo feature selection and interdependency discovery. *Journal of Statistical Software*, 85(12):1–28, 2018.

[9] O. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[10] A. Dussa. *venn: Draw Venn Diagrams*. 2021. URL https://CRAN.R-project.org/package=venn.

[11] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014. doi: 10.1117/1.JRS.11.015020.

[12] W. Hadley. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[13] P. Hammerman, M. Lawrence, D. Voet, and et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–525, 2012.

[14] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75 (4):800–802, 1988.

[15] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[16] A. H. M. Kamal, X. Zhu, A. S. Pandya, S. Hsu, and M. Shoaib. The impact of gene selection on imbalanced microarray expression data. In S. Rajasekaran, editor, *Bioinformatics and Computational Biology*, Lecture Notes in Computer Science, pages 259–269. Springer. ISBN 978-3-642-00727-9.

[17] L. Kolberg and U. Raudvere. *gprofiler2: Interface to the 'g:Profiler' Toolset*. 2021. URL https://CRAN.R-project.org/package=gprofiler2.

[18] M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26, 2008.

[19] J. L. Lustgarten, V. Gopalakrishnan, and S. Visweswaran. Measuring stability of feature selection in biomedical datasets. In *AMIA Annual Symposium Proceedings*, page 406–410, 2009. PMID: 20351889, PMCID: PMC2815476.

[20] H. B. Mann and D. R. Whitney. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Ann. Math. Statist.*, 18(1):50–60, 1947.

[21] B. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975.

[22] K. Mnich and W. R. Rudnicki. All-relevant feature selection using multidimensional filters with exhaustive search. *Inf. Sci.*, 524:277–297, 2020.

[23] B. Pereira, S. Chin, and O. e. a. Rueda. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7(11479), 2016.

[24] R. Piliszek, K. Mnich, S. Migacz, P. Tabaszewski, A. Sułecki, A. Polewko-Klim, and W. Rudnicki. Mdfs: Multidimensional feature selection in r. *The R Journal*, 11(1), 2019.

[25] A. Polewko-Klim, K. Mnich, and W. Rudnicki. Robust data integration method for classification of biomedical data. *Journal of Medical Systems*, 45(45), 2021. doi: 10.1007/s10916-021-01718-7.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[27] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, and A. Alonso-Betanzos. Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowledge-Based Systems*, 118: 124–139, 2017.

[28] J. Sidi. *slickR: Create Interactive Carousels with the 'JavaScript' 'Slick' Library*. 2020. URL https://CRAN.R-project.org/package=slickR.

[29] C. Sievert, C. Parmer, and e. a. Hocking, T. *plotly: Create Interactive Web Graphics via 'plotly.js'*. 2016. URL https://CRAN.R-project.org/package=plotly.

[30] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, and et. al. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. In *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 14–18, 2019.