ExhauFS: exhaustive search-based feature selection for classification and survival regression

Stepan Nersisyan [Corresp., 1], **Victor Novosad** [1, 2], **Alexei Galatenko** [3, 4], **Andrey Sokolov** [3, 4], **Grigoriy Bokov** [3, 4], **Alexander Konovalov** [3, 4], **Dmitry Alekseev** [3, 4], **Alexander Tonevitsky** [1, 2, 5]

[1] Faculty of Biology and Biotechnology, HSE University, Moscow, Russia

[2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia

[3] Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, Moscow, Russia

[4] Moscow Center for Fundamental and Applied Mathematics, Moscow, Russia

[5] Institute of Nanotechnologies of Microelectronics RAS, Moscow, Russia

Corresponding Author: Stepan Nersisyan
Email address: s.a.nersisyan@gmail.com

Feature selection is one of the main techniques used to prevent overfitting in machine learning applications. The most straightforward approach for feature selection is exhaustive search: one can go over all possible feature combinations and pick up the model with the highest accuracy. This method together with its optimizations were actively used in biomedical research, however, publicly available implementation is missing. We present ExhauFS – the user-friendly command-line implementation of the exhaustive search approach for classification and survival regression. Aside from tool description, we included three application examples in the manuscript to comprehensively review the implemented functionality. First, we executed ExhauFS on a toy cervical cancer dataset to illustrate basic concepts. Then, multi-cohort microarray breast cancer datasets were used to construct gene signatures for 5-year recurrence classification. The vast majority of signatures constructed by ExhauFS passed 0.65 threshold of sensitivity and specificity on all datasets, including the validation one. Moreover, a number of gene signatures demonstrated reliable performance on independent RNA-seq dataset without any coefficient re-tuning, i.e., turned out to be cross-platform. Finally, Cox survival regression models were used to fit isomiR signatures for overall survival prediction for patients with colorectal cancer. Similarly to the previous example, the major part of models passed the pre-defined concordance index threshold 0.65 on all datasets. In both real-world scenarios (breast and colorectal cancer datasets), ExhauFS was benchmarked against state-of-the-art feature selection models, including L_1-regularized sparse models. In case of breast cancer, we were unable to construct reliable cross-platform classifiers using alternative feature selection approaches. In case of colorectal cancer not a single model passed the same 0.65 threshold. Source codes and documentation of ExhauFS are available on

GitHub: https://github.com/s-a-nersisyan/ExhauFS .

1 ExhauFS: exhaustive search-based feature selection
2 for classification and survival regression

4 Stepan Nersisyan[1], Victor Novosad[1,2], Alexei Galatenko[3,4], Andrey Sokolov[3,4], Grigoriy
5 Bokov[3,4], Alexander Konovalov[3,4], Dmitry Alekseev[3,4], Alexander Tonevitsky[1,2,5]

7 [1] Faculty of Biology and Biotechnology, HSE University, Moscow, Russia
8 [2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia
9 [3] Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, Moscow,
10 Russia
11 [4] Moscow Center for Fundamental and Applied Mathematics, Moscow, Russia
12 [5] Institute of Nanotechnologies of Microelectronics RAS, Moscow, Russia

14 Corresponding Author:
15 Stepan Nersisyan[1]
16 Vavilova str. 7, Moscow, 1173112, Russia
17 Email address: snersisyan@hse.ru

19 Abstract

20 Feature selection is one of the main techniques used to prevent overfitting in machine learning
21 applications. The most straightforward approach for feature selection is exhaustive search: one
22 can go over all possible feature combinations and pick up the model with the highest accuracy.
23 This method together with its optimizations were actively used in biomedical research, however,
24 publicly available implementation is missing. We present ExhauFS – the user-friendly
25 command-line implementation of the exhaustive search approach for classification and survival
26 regression. Aside from tool description, we included three application examples in the
27 manuscript to comprehensively review the implemented functionality. First, we executed
28 ExhauFS on a toy cervical cancer dataset to illustrate basic concepts. Then, multi-cohort
29 microarray breast cancer datasets were used to construct gene signatures for 5-year recurrence
30 classification. The vast majority of signatures constructed by ExhauFS passed 0.65 threshold of
31 sensitivity and specificity on all datasets, including the validation one. Moreover, a number of
32 gene signatures demonstrated reliable performance on independent RNA-seq dataset without any
33 coefficient re-tuning, i.e., turned out to be cross-platform. Finally, Cox survival regression
34 models were used to fit isomiR signatures for overall survival prediction for patients with
35 colorectal cancer. Similarly to the previous example, the major part of models passed the pre-
36 defined concordance index threshold 0.65 on all datasets. In both real-world scenarios (breast
37 and colorectal cancer datasets), ExhauFS was benchmarked against state-of-the-art feature
38 selection models, including L_1-regularized sparse models. In case of breast cancer, we were
39 unable to construct reliable cross-platform classifiers using alternative feature selection

40 approaches. In case of colorectal cancer not a single model passed the same 0.65 threshold.
41 Source codes and documentation of ExhauFS are available on GitHub: https://github.com/s-a-
42 nersisyan/ExhauFS.
43

Introduction

45 Classification algorithms are widely used in biomedical applications, allowing one to construct a
46 rule which will separate data from different classes. For example, classification methods could
47 be applied to determine whether a patient has a particular disease based on gene expression
48 profile (diagnostic test), or to separate patients into groups of high and low risk (prognostic test)
49 (Cruz & Wishart, 2006; Kourou et al., 2015; Kang, Liu & Tian, 2016). In some clinical studies
50 time-to-event data are also available: this means that the data contain not only occurrence or not
51 occurrence of the event, but also the time of observation (e.g., overall, or recurrence-free survival
52 data). Survival regression models, such as Cox proportional hazards model, are used to analyze
53 this type of data (Zhang, 2002; Kleinbaum & Klein, 2012; Kamarudin, Cox & Kolamunnage-
54 Dona, 2017).
55

56 One of the main challenges related to machine learning applications in biomedical problems is
57 the so-called "curse of dimensionality": high-dimensional data with a low number of samples
58 often leads to overfitting and poor performance of learning algorithms (Salsburg, 1993; Asyali et
59 al., 2006; Sánchez & García, 2018; Mirza et al., 2019). One of the possible approaches to
60 overcome overfitting is called feature selection. Within this framework, some small subset of
61 features is being selected for further classification or regression. Existing feature selection
62 techniques include selection of the most relevant individual features (e.g., the most differentially
63 expressed genes between two classes of interest) and the "native" selection procedures of the
64 most important features from some models, such as random forests, regression techniques with
65 L_1-regularization and many others (Saeys, Inza & Larranaga, 2007; Chandrashekar & Sahin,
66 2014; Wang, Wang & Chang, 2016). Besides that, several approaches were designed specifically
67 for classification problems involving cancer transcriptomics data, including gene ranking,
68 filtration and combining the most relevant genes in a single model (Arakelyan, Aslanyan &
69 Boyajyan, 2013; Zhang et al., 2021; Rana et al., 2021).
70

71 Recently, Galatenko et al proposed to construct classifiers based on all possible feature
72 combinations, i.e., to perform an exhaustive search over feature subsets (Galatenko et al., 2015).
73 Specifically, the authors analyzed all possible gene pairs which allowed them to construct
74 reliable prognostic signatures for breast cancer. Further, the method was successfully applied to
75 the colorectal cancer datasets (Galatenko et al., 2018b,a). Unfortunately, computational
76 complexity of such an approach grows exponentially with the length of tested feature subsets.
77 This means that exhaustive search of all gene triples, quadruples and larger combinations is
78 computationally infeasible even for the most powerful supercomputers. A possible solution for
79 this problem was given in our recent report: the number of features (genes) was preliminarily

80 reduced to allow the search of all possible k-element gene subsets, which resulted in the
81 construction of highly reliable eight-gene prognostic signatures for breast cancer (Nersisyan et
82 al., 2021a).

83

84 In this paper we generalize the latter approach and propose ExhauFS – the user-friendly
85 command-line tool for exhaustive search-based feature selection for classification and survival
86 regression. ExhauFS allows one to vary many algorithm parameters, including classification
87 models, accuracy metrics, feature selection strategies and data pre-processing methods. Source
88 codes and documentation of ExhauFS are available on GitHub: https://github.com/s-a-
89 nersisyan/ExhauFS.

90

91 First, we executed ExhauFS on a "toy" cervical cancer dataset to illustrate the basic concepts
92 underlying the approach. Then, the tool was applied to the problem of 5-year breast cancer
93 recurrence classification based on gene expression profiles in patients' primary tumor. The
94 microarray-based dataset was composed of five independent patient cohorts, the constructed
95 models were also evaluated on the RNA sequencing data from The Cancer Genome Atlas Breast
96 Invasive Carcinoma (TCGA-BRCA) project (Koboldt et al., 2012). In order to justify our
97 approach, we tested two commonly used alternative classifier construction pipelines not
98 involving exhaustive search: they were based on univariate feature selection and L_1-penalized
99 logistic regression. Finally, we used expression profiles of miRNA isoforms (isomiRs) to build
100 signatures for predicting overall survival in colorectal cancer patients from The Cancer Genome
101 Atlas Colon Adenocarcinoma (TCGA-COAD) project (Muzny et al., 2012). As in case of breast
102 cancer, two alternative feature selection methods were benchmarked against ExhauFS: univariate
103 feature selection and sparse Cox model with L_1 penalty term.

104

105 While near-to-optimal solution can be easily obtained for the first "toy" dataset, the problems
106 solved in the second and the third example are believed to be hard. Specifically, the quality of
107 state-of-the-art solutions in terms of area under the receiver operating characteristic curve (ROC
108 AUC), sensitivity and specificity is well under 1.0 (Berg et al., 2009; Yang, Zhang & Yang,
109 2019).

110

Materials & Methods
Pipeline Workflow
113 The search for predictive models at the top level consists of the following phases. First, a fixed
114 set of features is selected for the downstream processing (feature pre-selection step). Then, n
115 most reliable features are picked according to the feature selection criteria. Next, all possible k-
116 element feature subsets are generated from n selected features. For each such a subset, a
117 predictive model (classifier or regressor) is tuned on a training set. Before fitting a model, one
118 can specify data pre-processing transformation. All constructed models are then evaluated on
119 training and filtration sets, user-defined quality metric functions are calculated, and the results

120 are compared with a predefined threshold parameter. In case of successful threshold passage, the
121 model is being saved and evaluated on the validation (test) set.
122
123 There is a special part of the tool dedicated to the estimation of running time and, hence,
124 choosing the right values for n and k (so that the pipeline execution takes an appropriate time).
125 The main idea behind the script is to calculate execution time on a limited number of feature
126 subsets and extrapolate this time according to the total number of k-element subsets composed
127 from a set of n genes equal to the binomial coefficient $\binom{n}{k}$.
128
Feature pre-selection and selection
130 The first stage of the pipeline is feature pre-selection. The aim of this step is to limit the
131 downstream analysis to a set of user-specified features. The most straightforward and common
132 way to enable pre-selection is to pass a list of features to ExhauFS through the file. Another
133 useful method is ANOVA F-test, which can be used to pre-select features with equal means
134 across training and filtration sets – this option can be beneficial to tackle batch effects (see
135 application example 2 in Results). Besides, pre-selection step can be performed based on some
136 domain knowledge or purpose of a study, e.g., limiting the set of all genes to the list of genes
137 encoding cell adhesion molecules in cancer studies. To validate benefits of using a certain pre-
138 selection method, one should compare reliability of models constructed with and without feature
139 pre-selection.
140
141 The next stage is feature selection: features are sorted by some user-specified function, and a
142 certain number of features with the highest scores are picked. Feature selection methods are
143 divided into three main groups: common methods, and methods specific for classification or
144 survival regression problems.
145
146 The following feature selection methods are common (i.e., can be used both for classification
147 and regression):
148   • Taking a ranked list of features from a user-specified file and selecting first n entries.
149   • Selecting n features with the highest median value across a dataset.
150 The latter could be potentially useful, for example, in miRNA expression research, where the
151 most expressed entries tend to be the most biologically active ones.
152
153 For binary classification problems, the basic feature selection method is Student's t-test, whose
154 objective is to score features with the highest differences between means in two considered
155 classes. For multiclass problems we also included ANOVA F-test (generalization of t-test on
156 three or more groups). Another useful univariate method for data with ordinal class labels
157 consists in calculation of Spearman's correlation coefficients between feature values and class

158 labels, with further selection of features with the highest absolute correlation values. Embedded
159 feature selection methods include L_1-regularized logistic regression.
160
161 Several available feature selection methods for the survival regression problems are based on the
162 single-factor Cox proportional hazard model. Concordance index and likelihood score can be
163 used to select the most individually informative features. Another group of methods is based on a
164 separation of samples into groups of low and high risk by the median risk score value. In this
165 case, one can use time-dependent AUC, hazard ratio or logrank test p-value for feature selection.
166 Similarly to the classification case, one can use sparse L_1-penalized Cox models to select
167 features with non-zero weights.
168
169 **Data pre-processing and model fitting**
170 Most machine learning methods require data normalization prior to model fitting
171 (Alexandropoulos, Kotsiantis & Vrahatis, 2019). A broad set of data pre-processing methods
172 from scikit-learn Python module (Pedregosa et al., 2011), including z-score transformation, min-
173 max scaler, binarization and others, can be natively used in ExhauFS.
174
175 A number of classification models from the scikit-learn module are also natively embedded in
176 ExhauFS interface. This includes support vector machine (SVM), logistic regression, random
177 forest, nearest neighbors and many other classifiers. In addition, one can use gradient boosting
178 models available in the xgboost package (Chen & Guestrin, 2016). It is possible to define both
179 lists of predefined model parameters and parameters which should be estimated through cross-
180 validation (for example, SVM penalty cost parameter or the number of decision trees in a
181 random forest). Survival regression models and metrics were imported from lifelines
182 (https://zenodo.org/record/4816284) and scikit-survival (Pölsterl, 2020) modules.
183
184 **Implementation**
185 ExhauFS was implemented using Python 3 programming language with an extensive use of
186 common pandas (McKinney, 2010), NumPy (Harris et al., 2020) and SciPy (Virtanen et al.,
187 2020) modules. Open-source code and the detailed documentation can be found on GitHub:
188 https://github.com/s-a-nersisyan/ExhauFS. To make the installation process more convenient, we
189 deposited ExhauFS to Python Package Index (PyPI, https://pypi.org/project/exhaufs/).
190
191 ExhauFS also supports parallel execution: each process performs a search on its own set of k-
192 element feature combinations. The number of parallel processes used for the pipeline execution
193 is specified by the user. Parallelization is based on the standard python multiprocessing module.
194
195 **Datasets used**

196   For the first "toy" example we used the small (n = 72) "Cervical cancer" dataset from UCI
197   Machine Learning Repository (Sobar, Machmud & Wijaya, 2016). The obtained data were
198   already divided into equal sized training and validation sets.

199

200   ER-positive breast cancer microarray dataset (Affymetrix Human Genome U133A Array) was
201   composed of five independent patient cohorts:
202   • Training set: the union of GSE3494 (Lundberg et al., 2017) and GSE6532 (Loi et al.,
203   2008) datasets.
204   • Filtration sets: GSE12093 (Zhang et al., 2009) and GSE17705 (Symmans et al., 2010).
205   • Validation set: GSE1456 (Hall et al., 2006).

206

207   Basic data processing was done like in our previous work (Nersisyan et al., 2021a). Briefly, raw
208   *.CEL files were normalized with the use of RMA algorithm available in affy R package
209   (Gautier et al., 2004), obtained intensity values were log_2-transformed. For each gene we
210   selected a probeset with the maximal median intensity across considered samples. Non-coding
211   and low-expressed genes (lower 25% according to the median intensity values) were discarded.
212   We also removed genes with near-constant expression (difference between 0.95 and 0.05
213   quantiles lower than two folds). To establish a binary classification problem, we separated all
214   patients into two groups: the ones with recurrence during the first five years after surgery or
215   recurrence-free with at least seven years follow-up. The remaining patients (early censored,
216   lately relapsed or from the grey 5-7 years zone) were included in the construction of Kaplan-
217   Meier plots. For the additional validation, TCGA-BRCA normalized RNA-seq data were
218   downloaded from GDC portal (https://portal.gdc.cancer.gov/) in the format of FPKM expression
219   tables. The log_2-transformed FPKM data were converted to the microarray units with the use of
220   quantile normalization, the useful technique for the gene expression distribution alignment across
221   data samples (Zhao, Wong & Goh, 2020). The number of patients in each dataset, as well as the
222   number of patients in each class are available in Table S1.

223

224   For the third ExhauFS application example we used isomiR expression data from TCGA-COAD
225   ($n = 413$ patients). MiRNA-seq read count tables were downloaded from GDC portal in the
226   form of *.isoforms.quantification.txt files. Library size normalization was performed with the
227   edgeR v3.30.3 package (Robinson, McCarthy & Smyth, 2009), TMM-normalized reads per
228   million mapped reads (TMM-RPM) matrices were generated. The default edgeR noise filtering
229   procedure was applied. We used conventional isomiR nomenclature introduced by Telonis et al
230   (Telonis et al., 2015). For example, hsa-miR-192-5p|+1 stands for the mature hsa-miR-192-5p
231   miRNA without the first nucleotide at the 5'-end (the number after the | sign stands for the 5'-
232   end offset in direction from 5'-end to 3'-end). The whole TCGA-COAD cohort was split into
233   three equal-sized sets (training, filtration and validation) with a stratification by outcome and
234   survival time (either date of death or date of the last follow-up). Differential expression analysis

235 for the comparison of tumor and normal samples was performed with the use of DESeq2 R
236 package (Love, Huber & Anders, 2014).

237

238 For both breast and colorectal cancer datasets, we generated grids of n and k values in such a
239 way that for a specific n, k pair pipeline execution time would not exceed 40 CPU-hours. Feature
240 combination lengths (k) were varied between 1 and 20.

241

242 All raw and processed data tables, configuration files and instructions for reproducing three
243 presented examples are available on the tool's GitHub page in the "Tutorials" section of the
244 README file.

245

246 **Results**
247 **Overview of the approach**
248 We developed an easy-to-use command line tool that combines all necessary parts needed to
249 perform an exhaustive feature selection. The first step of the pipeline is feature pre-selection: the
250 whole downstream analysis is performed for the fixed set of features determined during pre-
251 selection. For example, we recently pre-selected genes encoding extracellular matrix proteins
252 and their cellular receptors and constructed prognostic signatures for colorectal cancer
253 (Nersisyan et al., 2021b). The second step of the method is feature selection: features are sorted
254 according to some criteria and then "top n" features are selected (e.g., select n most differentially
255 expressed genes between normal and tumor samples).

256

257 During the main part, classification or survival regression models are constructed for all k-
258 element feature subsets from the set of features selected (n, k pairs are selected so that
259 computation time is practically acceptable). Concrete classifiers could be picked from the broad
260 list of available models; optimal user-specified parameters are automatically cross-validated. For
261 survival regression, Cox proportional hazards model is available. Additionally, a variety of data
262 pre-processing methods, such as z-score transformation or binarization, could be used.

263

264 To tackle batch effects and neutralize effects related to multiple model construction (high-quality
265 classifiers could appear by chance) we introduce filtration sets alongside training and validation
266 ones. First, model tuning (including cross-validation) is done on a training set for all possible
267 feature combinations. Then, each model is evaluated on one or more special filtration sets: if a
268 pre-defined quality threshold is not met on at least one set, then the model is discarded.
269 Otherwise, if the model demonstrates acceptable performance on both training and filtration sets,
270 it is further evaluated on a validation set. The workflow of the pipeline is presented in Fig. 1,
271 details can be found in Materials & Methods.

272

273 The output of the program consists of a table with quality metrics listed for all models which
274 passed the filtration step. To ensure a fair selection of the best model, the rows of the table are

275   sorted by the minimum of accuracy scores for the training and filtration datasets, and the first
276   model is considered as the best one. In case if several models pass the accuracy thresholds, some
277   of them may be chosen instead of the best model based on a relevance to a field of consideration
278   (e.g., combinations of genes with especially important biological roles). Besides, two summary
279   tables are generated. The first one contains the percentage of reliable models (with respect to a
280   validation set) out of all models which passed the filtration step. The second one summarizes the
281   most abundantly occurring features in a set of models which passed the filtration. The percentage
282   of models including a specific feature could be considered as an importance score of the feature.
283   Additionally, one can explore in detail a particular model: this includes construction of receiver
284   operating characteristic (ROC) curves for classification mode and Kaplan-Meier curves for the
285   survival regression mode.
286

287   **Application example 1: toy dataset (classification)**
288   As a simple example of how ExhauFS works, we tested it on a small dataset of cervical cancer
289   patients. We searched over all triples of 19 features and found that the best performing random
290   forest classifier was constructed on "perception_vulnerability", "socialSupport_instrumental" and
291   "empowerment_desires" features. This classifier had sensitivity (true positives rate, TPR) equal
292   to 0.9, specificity (true negatives rate, TNR) = 0.96, area under the receiver operating
293   characteristic curve (ROC AUC) = 0.96 (all presented metric were calculated on the validation
294   set). The ROC curve for the classifier is shown in Fig. S1A. In contrast, the performance of the
295   single feature classifiers did not exceed ROC AUC = 0.80 (Table S2).
296

297   The output of the program consists of a table with quality metrics listed for all models. To the
298   contrary, using a pure random forest classifier on all available features, we were able to achieve
299   much lower accuracy scores: TPR = 0.8, TNR = 0.92, ROC AUC = 0.92 (Fig. S1B).
300   Interestingly, the set of the most important features in this case ("behavior_personalHygine",
301   "empowerment_knowledge", "perception_severity") did not intersect with the previous one. The
302   model trained only on these three features led to even worse accuracy scores: TPR = 0.6, TNR =
303   0.88, ROC AUC = 0.83 (Fig. S1C). Thus, the exhaustive search-based approach allowed us to
304   select three features for a random forest classifier which could not be detected by standard
305   "native" feature selection available in the random forest model.
306

307   **Application example 2: prognostic gene signatures for breast cancer**
308   **(classification)**
309   **Experimental setup**
310   We applied ExhauFS to the real-world dataset of gene expression profiles in primary breast
311   tumors (transcriptome profiling was done by hybridization microarrays). The objective was to
312   classify patients into two groups: with recurrence within the first 5 year after surgery or without
313   recurrence with at least 7 years follow-up record. Training, filtration, and validation datasets
314   were composed of five independent patient cohorts (see Materials & Methods).
315

316   For feature selection we were picking "top n" most differentially expressed genes between
317   classes of patients with and without recurrence. Linear support vector machine (SVM) classifiers
318   were constructed for all k (length of gene signature) in the range from 1 to 20. Prior to model
319   fitting, we applied z-score transformation to all genes (i.e., subtracted means and normalized by
320   standard deviations, which were calculated on the training set). Before evaluation on the test set,
321   all trained models were filtered on training and two filtration sets with 0.65 threshold set on TPR,
322   TNR and ROC AUC. Our primary goal was to analyze the percentage of reliable classifiers in a
323   set of models which passed the filtration. For that, we calculated the percentage of classifiers
324   which passed the filtration step and passed the same 0.65 thresholds on the validation set
325   (separately for each n, k pair). The closer this value to 100%, the less overfitted the classifier is.
326

327   **Execution of ExhauFS on microarray data and pre-selection of "stable" genes**
328   First, we executed ExhauFS without any feature pre-selection. The results of analysis in such a
329   configuration were unsatisfactory. Namely, not a single classifier passed the filtration already for
330    $k \geq 6$ (Table S3). For shorter gene combinations, the maximum percentage of reliable
331   classifiers was only 53% ($k = 3, n = 130$). Interestingly, ROC AUC values calculated on the
332   validation set were high for all classifiers which passed the filtration: median ROC AUC = 0.76,
333   95% confidence interval (CI) 0.67-0.8. At the same time, we observed high disbalance between
334   sensitivity and specificity, which could be explained by a presence of batch effects in the data: a
335   decision threshold which is suitable for the training cohort is far from optimal for the validation
336   one. Notably, none of the individual-feature based classifiers ($k = 1$) passed the minimum
337   quality requirements on the training set.
338

339   To tackle batch effects and TPR/TNR disbalance, we added the feature pre-selector which was
340   used to put away genes with statistically significant differences in mean values between training
341   and filtration batches (ANOVA F-test, validation set was not included). In other words, we
342   preliminarily selected genes with similar expression patterns in training and filtration cohorts
343   ("stable" genes). Inclusion of such a gene pre-selection dramatically increased the quality of the
344   models: thousands of classifiers passed the filtration and more than 95% of them demonstrated
345   high TPR, TNR and ROC AUC values on the validation set for $k \geq 10$ (Table S4). Thus,
346   feature pre-selection is an effective strategy for reducing batch effects between training, filtration
347   and validation sets. Notably, some genes were highly overrepresented in the constructed gene
348   signatures (example data for $k = 10$ is presented in Table S5).
349

350   **Validation of cross-platform classifiers on RNA-seq data**
351   While the major part of existing transcriptomics data of clinical samples with sufficient follow-
352   up periods were generated with microarrays, the current standard transcriptome profiling
353   platform is RNA sequencing. Given that, we assessed whether the classifiers trained on
354   microarray data could demonstrate acceptable performance on RNA-seq data from TCGA-
355   BRCA project.

356
357   First, we transformed RNA-seq gene expression values to the microarray units with quantile
358   normalization. Then, the classifiers output by the previous run on microarray data were evaluated
359   on the transformed data without any changes in parameter values (including means and variances
360   of z-score scalers and weights/thresholds of SVM models). Eight gene signatures composed of
361   "stable" genes passed the 0.65 sensitivity/specificity threshold. One of the classifiers (a ten-gene
362   signature) had particularly high accuracy: ROC AUC = 0.68, TPR = 0.69, TNR = 0.69 (Fig. 2).
363   Another signature composed of only four genes (*TRIP13, ABAT, STC2, SIGLEC15*)
364   demonstrated slightly worse quality (ROC AUC = 0.66, TPR = 0.67, TNR = 0.69, Fig. S2).
365   Remarkably, sensitivity and specificity scores close to 0.7 are considered to be the current state
366   of the art for the breast cancer recurrence prediction problem (van 't Veer et al., 2002; Paik et al.,
367   2004; Nersisyan et al., 2021a). To the contrary, not a single classifier passed even 0.6 threshold
368   when no feature pre-selection was applied. Thereby, pre-selection of genes resistant to batch
369   effects allowed us to construct models which could be further generalized to other gene
370   expression profiling platforms.
371
372   In all aforementioned cases we used z-score transformation for data pre-processing. Though this
373   approach resulted in highly reliable and cross-platform classifiers, the tool allows one to easily
374   vary this step. For example, one can use binarization of gene expression profiles as a pre-
375   processing step: a continuous expression value becomes one if it is greater than a specific
376   threshold (e.g., the median gene expression level) and becomes zero otherwise. The use of such
377   an approach allowed us to achieve a near 100% passability of 0.65 threshold on the validation set
378   when no gene pre-selection was conducted. However, none of the models passed 0.65 accuracy
379   thresholds on the RNA-seq data.
380
381   **Comparisons with state-of-the-art feature selection methods justifies reliability of**
382   **exhaustive search**
383   To compare ExhauFS with alternative approaches, we implemented the canonical classifier
384   construction pipeline: a classifier is fitted directly on selected features (without exhaustive
385   search). Two ubiquitously used feature selection algorithms were tested:
386   • Univariate feature selection based on the rate of differential expression (Student's t-test)
387   (Samatov et al., 2017).
388   • Feature selection based on L_1-regularized linear models (logistic regression with L_1
389   penalty term) (Ma & Huang, 2008).
390
391   In both cases SVM model was fitted on a set of selected features to make an unbiased
392   comparison with ExhauFS. For the same comparison validity reasons, we imposed 0.65
393   threshold on TPR, TNR and ROC AUC calculated of training and filtration sets. Note that both
394   feature selection methods are available in ExhauFS package, so the proposed pipeline could be

395   reproduced by executing ExhauFS with $n = k$. Besides $k = 1,...,20$ (which was considered by
396   ExhauFS), we additionally examined values up to 100.
397
398   With univariate feature selection only four classifiers with 20, 21, 22 and 23 genes passed 0.65
399   quality threshold on training and filtration sets. These four models also demonstrated reliable
400   performance on the validation set (minimum of TPR and TNR 0.71). Since this experimental
401   setup was the special case of previously described ExhauFS run (with $n = k$), four identified
402   models were already found by the exhaustive search. Nevertheless, the use of univariate feature
403   selection without subsequent exhaustive search failed to construct shorter cross-platform gene
404   signatures.
405
406   With the second feature selection approach (sparse L_1-regularized logistic regression), a total of
407   31 classifiers were selected with 0.65 training and filtration thresholds. Only one model out of
408   them (3.2%) demonstrated acceptable performance on the validation set: 18-gene signature had
409   ROC AUC = 0.79, TPR = 0.67, TNR = 0.76, which was comparable with the best shorter
410   signatures found by ExhauFS. However, quality of predictions for TCGA RNA-seq data was
411   unsatisfactory as opposed to 4- and 10-gene signatures identified by ExhauFS. The detailed
412   information about quality scores for each dataset and gene signature length are presented in Fig.
413   S3.
414
415   **Application example 3: prognostic 5'-isomiR signatures for colorectal cancer**
416   **(survival regression)**
417   Survival regression module of the tool was applied to the problem of predicting overall survival
418   in colorectal cancer patients. For that we used miRNA expression data from TCGA-COAD
419   dataset; miRNA profiling was conducted with small RNA sequencing. Given the fact that even a
420   single nucleotide variation at 5'-end of a miRNA could significantly alter the targetome of the
421   molecule (van der Kwast et al., 2019; Zhiyanov, Nersisyan & Tonevitsky, 2021), we consider all
422   possible isoforms of a single miRNA with modified 5'-ends (5'-isomiRs, see Materials &
423   Methods for details about nomenclature).
424
425   To analyze the relationship between survival time and miRNA expression levels we used Cox
426   proportional hazard model and concordance index as a measure of model accuracy. Across
427   different feature selectors, the best results were obtained when concordance index was used to
428   select the most individually informative isomiRs. As a result, we identified thousands of isomiR
429   signatures with near-100% passability of 0.65 concordance index threshold on the validation set
430   (Table S6). For example, the seven-isomiR signature (hsa-miR-200a-5p|0, hsa-miR-26b-5p|0,
431   hsa-miR-21-3p|0, hsa-miR-126-3p|0, hsa-let-7e-5p|0, hsa-miR-374a-3p|0, hsa-miR-141-5p|0) had
432   concordance index equal to 0.71, hazard ratio (HR) = 2.97, 3-year ROC AUC = 0.67 and logrank
433   test $p = 8.70 \times 10^{-4}$ (Fig. 3A).
434

435 Another successful approach was to use a more biologically motivated feature selection method
436 (Lv et al., 2020): to pick the most differentially expressed isomiRs when comparing primary
437 tumors with adjacent normal tissues. While performance of the constructed models was on
438 average lower compared to the previous case (concordance index filtration threshold was
439 dropped to 0.63, Table S7), we were able to identify "short" reliable prognostic signatures. For
440 example, the four-isomiR signature (hsa-miR-126-3p|0, hsa-miR-374a-3p|0, hsa-miR-182-5p|0,
441 hsa-let-7e-5p|0) had concordance index equal to 0.70, HR = 3.43, 3-year ROC AUC = 0.67 and
442 logrank test $p = 6.51 \times 10^{-4}$ (Fig. 3B).

443

444 Benchmarking of exhaustive search with alternative methods was conducted in a similar manner
445 as for example 2:

- 446 • Cox model with univariate feature selection was directly applied to the data (i.e., $k = n$
  447 special case).
- 448 • L_1-regularized sparse Cox model was used to select features (Tibshirani, 1997),
  449 regression model was directly fitted using the selected isomiRs.
- 450 • Feature subset length was varied in the range from 1 to 50.

451 In both cases, not a single model passed 0.65 concordance index threshold on the filtration set,
452 (Fig. S4). Thus, exhaustive search was a necessary step to construct accurate survival regression
453 models based on 5'-isomiR expression data.

454

Discussion

456 With the use of ExhauFS, one can construct classification and survival regression models for all
457 subsets of selected features. In this manuscript we showed application examples of ExhauFS
458 both on toy and real-world biological datasets and made a comprehensive review on available
459 parameters. The main limitation of our approach consists in high computational complexity of
460 the pipeline: the number of feature subsets grows exponentially relative to the subset cardinality.
461 The proper use of the algorithm also requires a sufficient number of samples and presence of a
462 validation set. A promising future direction for ExhauFS development includes addition of
463 greedy techniques, which could be used in combination with exhaustive search to expand a set of
464 considered features: see, e.g., the report by Galatenko et al (Galatenko et al., 2015). Another
465 possible direction is inclusion of network analysis and clustering techniques (D'haeseleer, 2005;
466 Langfelder & Horvath, 2008) prior to feature selection, which will automatically eliminate
467 clusters of correlated features.

468

469 One of the tool application examples was construction of prognostic classification gene
470 signatures for breast cancer. The main challenge of this problem setting consisted in the
471 existence of strong batch effects and biases between independent training, filtration and
472 validation sets. Since tuned models should be further applied to data possibly coming from new
473 batches, conventional batch effect adjustment tools like ComBat (Johnson, Li & Rabinovic,
474 2007) are not applicable in this problem. Given that, we developed a special feature pre-selection

475   step which allowed us to discard genes which were prone to batch effects. As a result, multiple
476   reliable SVM classifiers were constructed. Moreover, some models showed high accuracy both
477   on microarray and RNA-seq data. Biological reliability of the prognostic signatures was
478   established by the analysis of the corresponding protein functions and previously published
479   cancer reports. For example, three out of four genes from the constructed short cross-platform
480   model (see Results and Fig. S2) were previously mentioned in the context of breast cancer
481   recurrence: *TRIP13* (Lu et al., 2019), *ABAT* (Budczies et al., 2013; Jansen et al., 2015), *STC2*
482   (Jansen et al., 2015). While we did not find reports connecting *SIGLEC15* expression with breast
483   cancer recurrence, it was previously shown to play a crucial role in a suppression of anti-tumor
484   T-cell immune responses (Hiruma, Hirai & Tsuda, 2011; Wang et al., 2019).

485

486   Another ExhauFS application example was the construction of survival regression models for
487   colorectal cancer using miRNA isoform (isomiR) expression data. Interestingly, the most reliable
488   isomiR signatures were composed only of canonical miRNA forms. These observations are in
489   agreement with our recent report, where we showed that the majority of 5'-isomiRs are tightly
490   co-expressed to their canonical forms (Zhiyanov, Nersisyan & Tonevitsky, 2021). This means
491   that despite functional differences between 5'-isomiRs of the same miRNA, it makes no sense to
492   include both molecules in the machine learning model. Additionally, numerous non-canonical
493   isomiRs were excluded from our analysis because of low expression levels.

494

Conclusions

496   ExhauFS is a user-friendly command-line tool which allows one to build classification and
497   survival regression models by using exhaustive search of features. A wide set of options can be
498   varied by user, including different algorithms for feature pre-selection, feature selection, data
499   pre-processing, classification and survival regression. Application of ExhauFS to real-world
500   datasets and comparison with alternative pipelines validated the proposed method and its
501   implementation. It is important to note that the scope of ExhauFS applications is not limited to
502   biology-related problems, and it can be used with any high-dimensional data (e.g., text
503   classification or machine learning analysis of financial data).

504

Acknowledgements

506   The authors thank Dr. Vladimir Galatenko, Dr. Mikhail Nosov, Yaromir Kobikov and
507   Mokhlaroy Akhmadjonova for valuable comments and discussions.

508

References

510   van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy
511   K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS,
512   Bernards R, Friend SH. 2002. Gene expression profiling predicts clinical outcome of breast
513   cancer. *Nature* 415:530–536. DOI: 10.1038/415530a.
514   Alexandropoulos S-AN, Kotsiantis SB, Vrahatis MN. 2019. Data preprocessing in predictive
515   data mining. *The Knowledge Engineering Review* 34:e1. DOI:

516   10.1017/S026988891800036X.

517   Arakelyan A, Aslanyan L, Boyajyan A. 2013. On knowledge-based gene expression data
518   analysis. In: *Ninth International Conference on Computer Science and Information*
519   *Technologies Revised Selected Papers*. IEEE, 1–6. DOI:
520   10.1109/CSITechnol.2013.6710349.

521   Asyali M, Colak D, Demirkaya O, Inan M. 2006. Gene Expression Profile Classification: A
522   Review. *Current Bioinformatics* 1:55–73. DOI: 10.2174/157489306775330615.

523   Berg AO, Armstrong K, Botkin J, Calonge N, Haddow J, Hayes M, Kaye C, Phillips KA, Piper
524   M, Richards CS, Scott JA, Strickland OL, Teutsch S. 2009. Recommendations from the
525   EGAPP Working Group: Can tumor gene expression profiling improve outcomes in
526   patients with breast cancer? *Genetics in Medicine* 11:66–73. DOI:
527   10.1097/GIM.0b013e3181928f56.

528   Budczies J, Brockmöller SF, Müller BM, Barupal DK, Richter-Ehrenstein C, Kleine-Tebbe A,
529   Griffin JL, Orešič M, Dietel M, Denkert C, Fiehn O. 2013. Comparative metabolomics of
530   estrogen receptor positive and estrogen receptor negative breast cancer: alterations in
531   glutamine and beta-alanine metabolism. *Journal of Proteomics* 94:279–288. DOI:
532   10.1016/j.jprot.2013.10.002.

533   Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. *Computers &*
534   *Electrical Engineering* 40:16–28. DOI: 10.1016/j.compeleceng.2013.11.024.

535   Chen T, Guestrin C. 2016. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International*
536   *Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 785–
537   794. DOI: 10.1145/2939672.2939785.

538   Cruz JA, Wishart DS. 2006. Applications of Machine Learning in Cancer Prediction and
539   Prognosis. *Cancer Informatics* 2:117693510600200. DOI: 10.1177/117693510600200030.

540   D'haeseleer P. 2005. How does gene expression clustering work? *Nature Biotechnology*
541   23:1499–1501. DOI: 10.1038/nbt1205-1499.

542   Galatenko V V., Galatenko A V., Samatov TR, Turchinovich AA, Shkurnikov MY, Makarova
543   JA, Tonevitsky AG. 2018a. Comprehensive network of miRNA-induced intergenic
544   interactions and a biological role of its core in cancer. *Scientific Reports* 8. DOI:
545   10.1038/s41598-018-20215-5.

546   Galatenko V V., Maltseva D V., Galatenko A V., Rodin S, Tonevitsky AG. 2018b. Cumulative
547   prognostic power of laminin genes in colorectal cancer. *BMC Medical Genomics* 11. DOI:
548   10.1186/s12920-018-0332-3.

549   Galatenko V V., Shkurnikov MYU, Samatov TR, Galatenko A V., Mityakina IA, Kaprin AD,
550   Schumacher U, Tonevitsky AG. 2015. Highly informative marker sets consisting of genes
551   with low individual degree of differential expression. *Scientific Reports* 5. DOI:
552   10.1038/srep14967.

553   Gautier L, Cope L, Bolstad BM, Irizarry RA. 2004. Affy - Analysis of Affymetrix GeneChip
554   data at the probe level. *Bioinformatics* 20:307–315. DOI: 10.1093/bioinformatics/btg405.

555   Hall P, Ploner A, Bjöhle J, Huang F, Lin CY, Liu ET, Miller LD, Nordgren H, Pawitan Y, Shaw
556   P, Skoog L, Smeds J, Wedrén S, Öhd J, Bergh J. 2006. Hormone-replacement therapy
557   influences gene expression profiles and is associated with breast-cancer prognosis: A cohort
558   study. *BMC Medicine* 4. DOI: 10.1186/1741-7015-4-16.

559   Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E,
560   Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M,
561   Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T,

562   Weckesser W, Abbasi H, Gohlke C, Oliphant TE. 2020. Array programming with NumPy.
563   *Nature* 585:357–362. DOI: 10.1038/s41586-020-2649-2.
564 Hiruma Y, Hirai T, Tsuda E. 2011. Siglec-15, a member of the sialic acid-binding lectin, is a
565   novel regulator for osteoclast differentiation. *Biochemical and Biophysical Research*
566   *Communications* 409:424–429. DOI: 10.1016/j.bbrc.2011.05.015.
567 Jansen MPHM, Sas L, Sieuwerts AM, Van Cauwenberghe C, Ramirez-Ardila D, Look M,
568   Ruigrok-Ritstier K, Finetti P, Bertucci F, Timmermans MM, van Deurzen CHM, Martens
569   JWM, Simon I, Roepman P, Linn SC, van Dam P, Kok M, Lardon F, Vermeulen PB,
570   Foekens JA, Dirix L, Berns EMJJ, Van Laere S. 2015. Decreased expression of ABAT and
571   STC2 hallmarks ER-positive inflammatory breast cancer and endocrine therapy resistance
572   in advanced disease. *Molecular Oncology* 9:1218–1233. DOI:
573   10.1016/j.molonc.2015.02.006.
574 Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data
575   using empirical Bayes methods. *Biostatistics* 8:118–127. DOI: 10.1093/biostatistics/kxj037.
576 Kamarudin AN, Cox T, Kolamunnage-Dona R. 2017. Time-dependent ROC curve analysis in
577   medical research: current methods and applications. *BMC Medical Research Methodology*
578   17:53. DOI: 10.1186/s12874-017-0332-6.
579 Kang L, Liu A, Tian L. 2016. Linear combination methods to improve diagnostic/prognostic
580   accuracy on future observations. *Statistical Methods in Medical Research* 25:1359–1380.
581   DOI: 10.1177/0962280213481053.
582 Kleinbaum DG, Klein M. 2012. The Cox Proportional Hazards Model and Its Characteristics. In:
583   97–159. DOI: 10.1007/978-1-4419-6646-9_3.
584 Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL,
585   Dooling DJ, Ding L, Mardis ER, Wilson RK, Ally A, Balasundaram M, Butterfield YSN,
586   Carlsen R, Carter C, Chu A, Chuah E, Chun HJE, Coope RJN, Dhalla N, Guin R, Hirst C,
587   Hirst M, Holt RA, Lee D, Li HI, Mayo M, Moore RA, Mungall AJ, Pleasance E, Robertson
588   AG, Schein JE, Shafiei A, Sipahimalani P, Slobodan JR, Stoll D, Tam A, Thiessen N,
589   Varhol RJ, Wye N, Zeng T, Zhao Y, Birol I, Jones SJM, Marra MA, Cherniack AD,
590   Saksena G, Onofrio RC, Pho NH, Carter SL, Schumacher SE, Tabak B, Hernandez B,
591   Gentry J, Nguyen H, Crenshaw A, Ardlie K, Beroukhim R, Winckler W, Getz G, Gabriel
592   SB, Meyerson M, Chin L, Kucherlapati R, Hoadley KA, Auman JT, Fan C, Turman YJ, Shi
593   Y, Li L, Topal MD, He X, Chao HH, Prat A, Silva GO, Iglesia MD, Zhao W, Usary J, Berg
594   JS, Adams M, Booker J, Wu J, Gulabani A, Bodenheimer T, Hoyle AP, Simons J V.,
595   Soloway MG, Mose LE, Jefferys SR, Balu S, Parker JS, Hayes DN, Perou CM, Malik S,
596   Mahurkar S, Shen H, Weisenberger DJ, Triche T, Lai PH, Bootwalla MS, Maglinte DT,
597   Berman BP, Van Den Berg DJ, Baylin SB, Laird PW, Creighton CJ, Donehower LA, Noble
598   M, Voet D, Gehlenborg N, Di Cara D, Zhang J, Zhang H, Wu CJ, Yingchun Liu S,
599   Lawrence MS, Zou L, Sivachenko A, Lin P, Stojanov P, Jing R, Cho J, Sinha R, Park RW,
600   Nazaire MD, Robinson J, Thorvaldsdottir H, Mesirov J, Park PJ, Reynolds S, Kreisberg RB,
601   Bernard B, Bressler R, Erkkila T, Lin J, Thorsson V, Zhang W, Shmulevich I, Ciriello G,
602   Weinhold N, Schultz N, Gao J, Cerami E, Gross B, Jacobsen A, Sinha R, Aksoy BA,
603   Antipin Y, Reva B, Shen R, Taylor BS, Ladanyi M, Sander C, Anur P, Spellman PT, Lu Y,
604   Liu W, Verhaak RRG, Mills GB, Akbani R, Zhang N, Broom BM, Casasent TD, Wakefield
605   C, Unruh AK, Baggerly K, Coombes K, Weinstein JN, Haussler D, Benz CC, Stuart JM,
606   Benz SC, Zhu J, Szeto CC, Scott GK, Yau C, Paull EO, Carlin D, Wong C, Sokolov A,
607   Thusberg J, Mooney S, Ng S, Goldstein TC, Ellrott K, Grifford M, Wilks C, Ma S, Craft B,

608 Yan C, Hu Y, Meerzaman D, Gastier-Foster JM, Bowen J, Ramirez NC, Black AD, Pyatt
609 RE, White P, Zmuda EJ, Frick J, Lichtenberg TM, Brookens R, George MM, Gerken MA,
610 Harper HA, Leraas KM, Wise LJ, Tabler TR, McAllister C, Barr T, Hart-Kothari M, Tarvin
611 K, Saller C, Sandusky G, Mitchell C, Iacocca M V., Brown J, Rabeno B, Czerwinski C,
612 Petrelli N, Dolzhansky O, Abramov M, Voronina O, Potapova O, Marks JR, Suchorska
613 WM, Murawa D, Kycler W, Ibbs M, Korski K, Spychała A, Murawa P, Brzeziński JJ, Perz
614 H, Łaźniak R, Teresiak M, Tatka H, Leporowska E, Bogusz-Czerniewicz M, Malicki J,
615 Mackiewicz A, Wiznerowicz M, Van Le X, Kohl B, Viet Tien N, Thorp R, Van Bang N,
616 Sussman H, Phu BD, Hajek R, Hung NP, Phuong TVT, Thang HQ, Khan KZ, Penny R,
617 Mallery D, Curley E, Shelton C, Yena P, Ingle JN, Couch FJ, Lingle WL, King TA,
618 Gonzalez-Angulo AM, Dyer MD, Liu S, Meng X, Patangan M, Waldman F, Stöppler H,
619 Rathmell WK, Thorne L, Huang M, Boice L, Hill A, Morrison C, Gaudioso C, Bshara W,
620 Daily K, Egea SC, Pegram MD, Gomez-Fernandez C, Dhir R, Bhargava R, Brufsky A,
621 Shriver CD, Hooke JA, Campbell JL, Mural RJ, Hu H, Somiari S, Larson C, Deyarmin B,
622 Kvecher L, Kovatich AJ, Ellis MJ, Stricker T, White K, Olopade O, Luo C, Chen Y, Bose
623 R, Chang LW, Beck AH, Pihl T, Jensen M, Sfeir R, Kahn A, Chu A, Kothiyal P, Wang Z,
624 Snyder E, Pontius J, Ayala B, Backus M, Walton J, Baboud J, Berton D, Nicholls M,
625 Srinivasan D, Raman R, Girshik S, Kigonya P, Alonso S, Sanbhadti R, Barletta S, Pot D,
626 Sheth M, Demchok JA, Shaw KRM, Yang L, Eley G, Ferguson ML, Tarnuzzer RW, Zhang
627 J, Dillon LAL, Buetow K, Fielding P, Ozenberger BA, Guyer MS, Sofia HJ, Palchik JD.
628 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70.
629 DOI: 10.1038/nature11412.
630 Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V., Fotiadis DI. 2015. Machine learning
631 applications in cancer prognosis and prediction. *Computational and Structural*
632 *Biotechnology Journal* 13:8–17. DOI: 10.1016/j.csbj.2014.11.005.
633 van der Kwast RVCT, Woudenberg T, Quax PHA, Nossent AY. 2019. MicroRNA-411 and Its
634 5′-IsomiR Have Distinct Targets and Functions and Are Differentially Regulated in the
635 Vasculature under Ischemia. *Molecular Therapy* 28:157–170. DOI:
636 10.1016/j.ymthe.2019.10.002.
637 Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network
638 analysis. *BMC Bioinformatics* 9:559. DOI: 10.1186/1471-2105-9-559.
639 Loi S, Haibe-Kains B, Desmedt C, Wirapati P, Lallemand F, Tutt AM, Gillet C, Ellis P, Ryder
640 K, Reid JF, Daidone MG, Pierotti MA, Berns EMJJ, Jansen MPHM, Foekens JA, Delorenzi
641 M, Bontempi G, Piccart MJ, Sotiriou C. 2008. Predicting prognosis using molecular
642 profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*
643 9. DOI: 10.1186/1471-2164-9-239.
644 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
645 RNA-seq data with DESeq2. *Genome Biology* 15:550. DOI: 10.1186/s13059-014-0550-8.
646 Lu S, Qian J, Guo M, Gu C, Yang Y. 2019. Insights into a Crucial Role of TRIP13 in Human
647 Cancer. *Computational and Structural Biotechnology Journal* 17:854–861. DOI:
648 10.1016/j.csbj.2019.06.005.
649 Lundberg A, Lindstrom LS, Harrell JC, Falato C, Carlson JW, Wright PK, Foukakis T, Perou
650 CM, Czene K, Bergh J, Tobin NP. 2017. Gene expression signatures and
651 immunohistochemical subtypes add prognostic value to each other in breast cancer cohorts.
652 *Clinical Cancer Research* 23:7512–7520. DOI: 10.1158/1078-0432.CCR-17-1535.
653 Lv Y, Duanmu J, Fu X, Li T, Jiang Q. 2020. Identifying a new microRNA signature as a

654   prognostic biomarker in colon cancer. *PLOS ONE* 15:e0228575. DOI:
655   10.1371/journal.pone.0228575.
656   Ma S, Huang J. 2008. Penalized feature selection and classification in bioinformatics. *Briefings*
657   *in Bioinformatics* 9:392–403. DOI: 10.1093/bib/bbn027.
658   McKinney W. 2010. Data Structures for Statistical Computing in Python. In: van der Walt S,
659   Millman J eds. *Proceedings of the 9th Python in Science Conference*. 56–61. DOI:
660   10.25080/Majora-92bf1922-00a.
661   Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. 2019. Machine Learning and Integrative
662   Analysis of Biomedical Big Data. *Genes* 10:87. DOI: 10.3390/genes10020087.
663   Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, Kovar CL, Lewis
664   LR, Morgan MB, Newsham IF, Reid JG, Santibanez J, Shinbrot E, Trevino LR, Wu Y-Q,
665   Wang M, Gunaratne P, Donehower LA, Creighton CJ, Wheeler DA, Gibbs RA, Lawrence
666   MS, Voet D, Jing R, Cibulskis K, Sivachenko A, Stojanov P, McKenna A, Lander ES,
667   Gabriel S, Getz G, Ding L, Fulton RS, Koboldt DC, Wylie T, Walker J, Dooling DJ, Fulton
668   L, Delehaunty KD, Fronick CC, Demeter R, Mardis ER, Wilson RK, Chu A, Chun H-JE,
669   Mungall AJ, Pleasance E, Gordon Robertson A, Stoll D, Balasundaram M, Birol I,
670   Butterfield YSN, Chuah E, Coope RJN, Dhalla N, Guin R, Hirst C, Hirst M, Holt RA, Lee
671   D, Li HI, Mayo M, Moore RA, Schein JE, Slobodan JR, Tam A, Thiessen N, Varhol R,
672   Zeng T, Zhao Y, Jones SJM, Marra MA, Bass AJ, Ramos AH, Saksena G, Cherniack AD,
673   Schumacher SE, Tabak B, Carter SL, Pho NH, Nguyen H, Onofrio RC, Crenshaw A, Ardlie
674   K, Beroukhim R, Winckler W, Getz G, Meyerson M, Protopopov A, Zhang J, Hadjipanayis
675   A, Lee E, Xi R, Yang L, Ren X, Zhang H, Sathiamoorthy N, Shukla S, Chen P-C, Haseley
676   P, Xiao Y, Lee S, Seidman J, Chin L, Park PJ, Kucherlapati R, Todd Auman J, Hoadley
677   KA, Du Y, Wilkerson MD, Shi Y, Liquori C, Meng S, Li L, Turman YJ, Topal MD, Tan D,
678   Waring S, Buda E, Walsh J, Jones CD, Mieczkowski PA, Singh D, Wu J, Gulabani A,
679   Dolina P, Bodenheimer T, Hoyle AP, Simons J V, Soloway M, Mose LE, Jefferys SR, Balu
680   S, O'Connor BD, Prins JF, Chiang DY, Neil Hayes D, Perou CM, Hinoue T, Weisenberger
681   DJ, Maglinte DT, Pan F, Berman BP, Van Den Berg DJ, Shen H, Triche Jr T, Baylin SB,
682   Laird PW, Getz G, Noble M, Voet D, Saksena G, Gehlenborg N, DiCara D, Zhang J, Zhang
683   H, Wu C-J, Yingchun Liu S, Shukla S, Lawrence MS, Zhou L, Sivachenko A, Lin P,
684   Stojanov P, Jing R, Park RW, Nazaire M-D, Robinson J, Thorvaldsdottir H, Mesirov J, Park
685   PJ, Chin L, Thorsson V, Reynolds SM, Bernard B, Kreisberg R, Lin J, Iype L, Bressler R,
686   Erkkilä T, Gundapuneni M, Liu Y, Norberg A, Robinson T, Yang D, Zhang W, Shmulevich
687   I, de Ronde JJ, Schultz N, Cerami E, Ciriello G, Goldberg AP, Gross B, Jacobsen A, Gao J,
688   Kaczkowski B, Sinha R, Arman Aksoy B, Antipin Y, Reva B, Shen R, Taylor BS, Chan
689   TA, Ladanyi M, Sander C, Akbani R, Zhang N, Broom BM, Casasent T, Unruh A,
690   Wakefield C, Hamilton SR, Craig Cason R, Baggerly KA, Weinstein JN, Haussler D, Benz
691   CC, Stuart JM, Benz SC, Zachary Sanborn J, Vaske CJ, Zhu J, Szeto C, Scott GK, Yau C,
692   Ng S, Goldstein T, Ellrott K, Collisson E, Cozen AE, Zerbino D, Wilks C, Craft B,
693   Spellman P, Penny R, Shelton T, Hatfield M, Morris S, Yena P, Shelton C, Sherman M,
694   Paulauskis J, Gastier-Foster JM, Bowen J, Ramirez NC, Black A, Pyatt R, Wise L, White P,
695   Bertagnolli M, Brown J, Chan TA, Chu GC, Czerwinski C, Denstman F, Dhir R, Dörner A,
696   Fuchs CS, Guillem JG, Iacocca M, Juhl H, Kaufman A, Kohl III B, Van Le X, Mariano
697   MC, Medina EN, Meyers M, Nash GM, Paty PB, Petrelli N, Rabeno B, Richards WG, Solit
698   D, Swanson P, Temple L, Tepper JE, Thorp R, Vakiani E, Weiser MR, Willis JE, Witkin G,
699   Zeng Z, Zinner MJ, Network TCGA, Medicine GSCBC of, Institute GSCB, Louis GSCWU

700   in S, Agency GCCBCC, Institute G-CCB, School G-CCB and WH and HM, Genome-
701   Characterization Center University of North Carolina CH, University G-CCU of SC and JH,
702   Institute GDACB, Biology GDACI for S, Center GDACMS-KC, Center GDACU of
703   TMDAC, Genome Data Analysis Centers Santa Cruz and the Buck Institute U of C,
704   Consortium BCRIG, Resource NCHBC, group T source sites and disease working. 2012.
705   Comprehensive molecular characterization of human colon and rectal cancer. *Nature*
706   487:330–337. DOI: 10.1038/nature11252.
707 Nersisyan S, Galatenko A, Galatenko V, Shkurnikov M, Tonevitsky A. 2021a. miRGTF-net:
708   Integrative miRNA-gene-TF network analysis reveals key drivers of breast cancer
709   recurrence. *PLOS ONE* 16:e0249424. DOI: 10.1371/journal.pone.0249424.
710 Nersisyan S, Novosad V, Engibaryan N, Ushkaryov Y, Nikulin S, Tonevitsky A. 2021b. ECM–
711   Receptor Regulatory Network and Its Prognostic Role in Colorectal Cancer. *Frontiers in*
712   *Genetics* 12. DOI: 10.3389/fgene.2021.782699.
713 Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park
714   T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N. 2004. A Multigene Assay to
715   Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England*
716   *Journal of Medicine* 351:2817–2826. DOI: 10.1056/NEJMoa041588.
717 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer
718   P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M,
719   Duchesnay É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine*
720   *Learning Research* 12:2825–2830.
721 Pölsterl S. 2020. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-
722   learn. *Journal of Machine Learning Research* 21:1–6.
723 Rana P, Thai P, Dinh T, Ghosh P. 2021. Relevant and Non-Redundant Feature Selection for
724   Cancer Classification and Subtype Detection. *Cancers* 13:4297. DOI:
725   10.3390/cancers13174297.
726 Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: A Bioconductor package for differential
727   expression analysis of digital gene expression data. *Bioinformatics* 26:139–140. DOI:
728   10.1093/bioinformatics/btp616.
729 Saeys Y, Inza I, Larranaga P. 2007. A review of feature selection techniques in bioinformatics.
730   *Bioinformatics* 23:2507–2517. DOI: 10.1093/bioinformatics/btm344.
731 Salsburg D. 1993. Hundreds of Patients, Thousands of Observations: The Curse of
732   Dimensionality in Clinical Research. *Drug Information Journal* 27:597–609. DOI:
733   10.1177/009286159302700303.
734 Samatov TR, Galatenko V V., Block A, Shkurnikov MY, Tonevitsky AG, Schumacher U. 2017.
735   Novel biomarkers in cancer: The whole is greater than the sum of its parts. *Seminars in*
736   *Cancer Biology* 45:50–57. DOI: 10.1016/j.semcancer.2016.09.002.
737 Sánchez JS, García V. 2018. Addressing the Links Between Dimensionality and Data
738   Characteristics in Gene-Expression Microarrays. In: *Proceedings of the International*
739   *Conference on Learning and Optimization Algorithms: Theory and Applications - LOPAL*
740   *'18*. New York, New York, USA: ACM Press, 1–6. DOI: 10.1145/3230905.3230909.
741 Sobar ., Machmud R, Wijaya A. 2016. Behavior Determinant Based Cervical Cancer Early
742   Detection with Machine Learning Algorithm. *Advanced Science Letters* 22:3120–3123.
743   DOI: 10.1166/asl.2016.7980.
744 Symmans WF, Hatzis C, Sotiriou C, Andre F, Peintinger F, Regitnig P, Daxenbichler G,
745   Desmedt C, Domont J, Marth C, Delaloge S, Bauernhofer T, Valero V, Booser DJ,

746   Hortobagyi GN, Pusztai L. 2010. Genomic index of sensitivity to endocrine therapy for
747   breast cancer. *Journal of Clinical Oncology* 28:4111–4119. DOI:
748   10.1200/JCO.2010.28.4273.
749 Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I. 2015. Beyond the one-locus-one-miRNA
750   paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity.
751   *Nucleic Acids Research* 43:9158–9175. DOI: 10.1093/nar/gkv922.
752 Tibshirani R. 1997. The lasso method for variable selection in the Cox model. *Statistics in*
753   *medicine* 16:385–95. DOI: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-
754   sim380>3.0.co;2-3.
755 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E,
756   Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ,
757   Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore
758   EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris
759   CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A
760   Pietro, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN,
761   Fulton C, Masson C, Häggström C, Fitzgerald C, Nicholson DA, Hagen DR, Pasechnik D
762   V., Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA,
763   Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT,
764   Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer
765   J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M,
766   Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N,
767   Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier
768   S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP,
769   Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO,
770   Vázquez-Baeza Y. 2020. SciPy 1.0: fundamental algorithms for scientific computing in
771   Python. *Nature Methods* 17:261–272. DOI: 10.1038/s41592-019-0686-2.
772 Wang J, Sun J, Liu LN, Flies DB, Nie X, Toki M, Zhang J, Song C, Zarr M, Zhou X, Han X,
773   Archer KA, O'Neill T, Herbst RS, Boto AN, Sanmamed MF, Langermann S, Rimm DL,
774   Chen L. 2019. Siglec-15 as an immune suppressor and potential target for normalization
775   cancer immunotherapy. *Nature Medicine* 25:656–666. DOI: 10.1038/s41591-019-0374-x.
776 Wang L, Wang Y, Chang Q. 2016. Feature selection methods for big data bioinformatics: A
777   survey from the search perspective. *Methods* 111:21–31. DOI:
778   10.1016/j.ymeth.2016.08.014.
779 Yang G, Zhang Y, Yang J. 2019. A Five-microRNA Signature as Prognostic Biomarker in
780   Colorectal Cancer by Bioinformatics Analysis. *Frontiers in Oncology* 9. DOI:
781   10.3389/fonc.2019.01207.
782 Zhang M-J. 2002. Cox Proportional Hazards Regression Models for Survival Data in Cancer
783   Research. In: 59–70. DOI: 10.1007/978-1-4757-3571-0_4.
784 Zhang Y, Sieuwerts AM, McGreevy M, Casey G, Cufer T, Paradiso A, Harbeck N, Span PN,
785   Hicks DG, Crowe J, Tubbs RR, Budd GT, Lyons J, Sweep FCGJ, Schmitt M, Schittulli F,
786   Golouh R, Talantov D, Wang Y, Foekens JA. 2009. The 76-gene signature defines high-risk
787   patients that benefit from adjuvant tamoxifen therapy. *Breast Cancer Research and*
788   *Treatment* 116:303–309. DOI: 10.1007/s10549-008-0183-2.
789 Zhang J, Xu D, Hao K, Zhang Y, Chen W, Liu J, Gao R, Wu C, De Marinis Y. 2021. FS–GBDT:
790   identification multicancer-risk module via a feature selection algorithm by integrating
791   Fisher score and GBDT. *Briefings in Bioinformatics* 22. DOI: 10.1093/bib/bbaa189.

792   Zhao Y, Wong L, Goh WW Bin. 2020. How to do quantile normalization correctly for gene
793   expression data analyses. *Scientific Reports* 10:15534. DOI: 10.1038/s41598-020-72664-6.
794   Zhiyanov A, Nersisyan S, Tonevitsky A. 2021. Hairpin sequence and structure is associated with
795   features of isomiR biogenesis. *RNA Biology* 18:430–438. DOI:
796   10.1080/15476286.2021.1952759.
797

Figure 1

ExhauFS workflow.

Dashed lines represent optional relations (e.g., training/filtration datasets can be used in some feature pre-selection methods).

All features

Pre-select features

Select n features

Training set

Filtration set

Generate all
k-element feature
subsets

Fit
classification/survival
regression models

Filter
classifiers/regressors

Validation set

Evaluate models

Generate reports
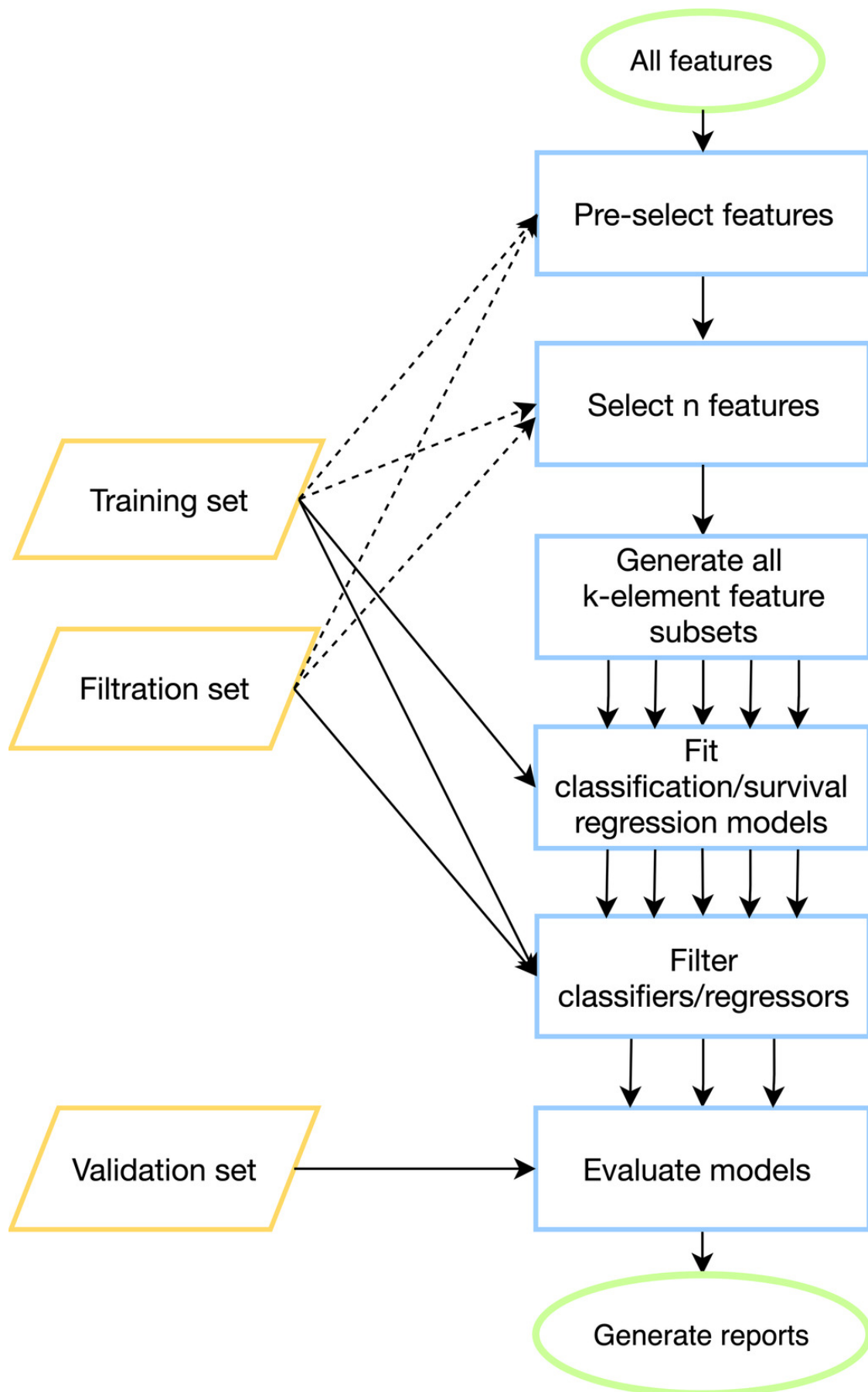
Figure 2

ROC and Kaplan-Meier plots for the ten-gene signature (TRIP13, ZWINT, EPN3, ECHDC2, CX3CR1, STARD13, MB, SLC7A5, ABAT, CCNL2) evaluated on GSE1456 (microarray) and TCGA-BRCA (RNA-seq) datasets.

Red points on ROC curves stand for the actual SVM threshold values.
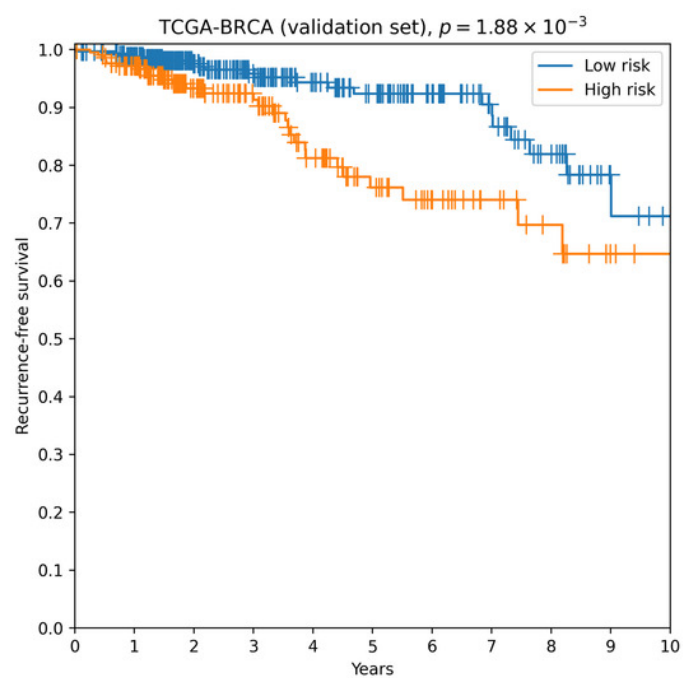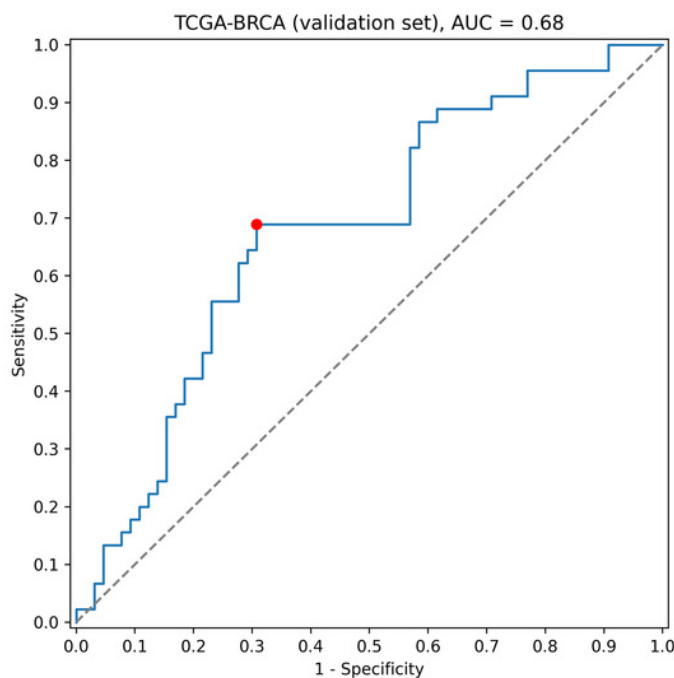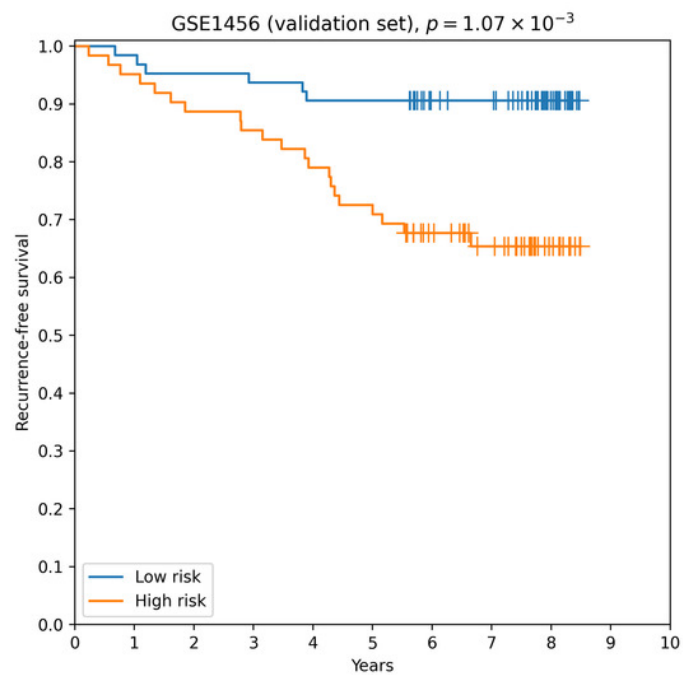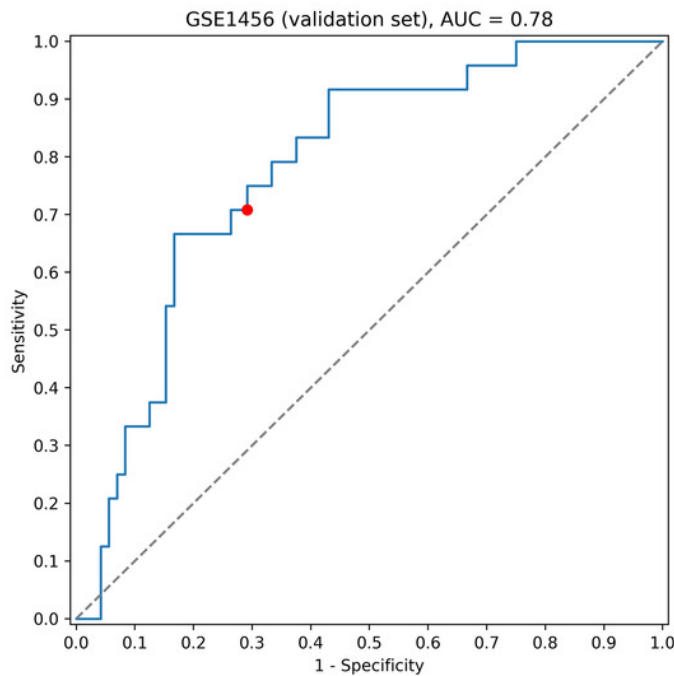
Figure 3

Kaplan-Meier plots for 5'-isomiR signatures in colorectal cancer.

(A) The seven-isomiR signature (hsa-miR-200a-5p|0, hsa-miR-26b-5p|0, hsa-miR-21-3p|0, hsa-miR-126-3p|0, hsa-let-7e-5p|0, hsa-miR-374a-3p|0, hsa-miR-141-5p|0). (B) The four-gene isomiR signature (hsa-miR-126-3p|0, hsa-miR-374a-3p|0, hsa-miR-182-5p|0, hsa-let-7e-5p|0).