Supervised Rank aggregation (SRA): A novel rank aggregation approach for ensemble-based feature selection

Rahi Jain[1¶], Wei Xu[2*]

[1]Biostatistics Department, Princess Margaret Cancer Research Centre, Toronto, Ontario, Canada

[2]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[*] Corresponding author

Telephone Number: (+1) 416-946-4497

Email: wei.xu@uhnres.utoronto.ca (WX)

Rahi Jain is a postdoctoral fellow at the Biostatistics Department, Princess Margaret Cancer Research Centre, University Health Network. His research interests are primarily in feature selection in high dimensional data.

Wei Xu is an associate professor at the Dalla Lana School of Public Health, University of Toronto. His research interests focus on biostatistics and bioinformatics methodology and statistical genetics.

Abstract

Background: Feature selection (FS) is critical for high dimensional data analysis. Ensemble based feature selection (EFS) is a commonly used approach to develop FS techniques. Rank aggregation (RA) is an essential step of EFS where results from multiple models are pooled to estimate feature importance. However, the literature primarily relies on rule-based methods to perform this step which may not always provide an optimal feature set.

Method and Results: This study proposes a novel Supervised Rank Aggregation (SRA) approach to allow RA step to dynamically learn and adapt the model aggregation rules to obtain feature importance. The approach creates a performance matrix containing feature and model performance value from all models and prepares a supervised learning model to get the feature importance. Then, unsupervised learning is performed to select the features using their importance. We evaluate the performance of the algorithm using simulation studies and implement it into real research studies, and compare its performance with various existing RA methods. The proposed SRA method provides better or at par performance in terms of feature selection and predictive performance of the model compared to existing methods.

Conclusion: SRA method provides an alternative to the existing approaches of RA for EFS. While the current study is limited to the continuous cross-sectional outcome, other endpoints such as longitudinal, categorical, and time-to-event medical data could also be used.

Keywords

high dimensional data, supervised rank aggregation, artificial intelligence, machine learning, ensemble feature selection, random forest

Introduction

A high dimensional data has challenges associated with model fitting, generalizability [1], and computation complexity [2,3], which prevents modeling by many classic statistical techniques. Feature selection is an important component in high dimensional data analysis domains like genomics [4] and radiomics [5], as it helps reduce the dimensions of the dataset. Literature provides many techniques to perform feature selection. However, these techniques could be categorized based on their feature selection (FS) approach (Figure 1). One broad category of FS techniques uses only expert or domain knowledge to perform feature selection [6,7]. These techniques work in scenarios with few features without interaction among features and are well known in the research domain [8]. Another broad category of FS techniques combines expert or domain knowledge with data [9,10]. FS techniques designed in the Bayesian framework incorporate prior knowledge in the feature selection process [9].

The third and major category of FS techniques relies on the dataset to perform feature selection and is referred to as data-based FS techniques in this paper. These techniques are sub-categorized into Filter, Wrapper, and Embedded FS techniques [11,12]. Filter methods select features based on internal data structures like association [13] and information gain [14,15]. Wrapper methods evaluate multiple subsets of features iteratively by building models to get the feature subset, which achieves the best performance [16–18]. Embedded methods build the model that simultaneously performs features selection [19–22].

Literature suggests different approaches to use the FS techniques for FS. These approaches can be categorized into base, hybrid, and ensemble approaches. In the base approach, a single FS technique is used. In the hybrid approach, multiple FS techniques are used in a sequence to perform feature selection [10,23]. Commonly, a filter based FS technique is used as coarse FS followed by a wrapper or an embedded based FS technique for final FS [23]. Some approaches create a sequence by combining expert based FS with other FS techniques [10].

Figure 1: Different feature selection approaches

In an ensemble approach, instead of a single model, multiple models are created from the same dataset. The performance of features from these models is pooled and ranked based on their relevance. Finally, the relevant features are selected based on the cut-off of importance. Two approaches can be used to generate multiple models, namely homogenous ensemble approach and heterogeneous ensemble approach [24–26]. In a homogenous ensemble approach, multiple datasets are created from the same data by sub-setting the samples, features, or both followed by using a single technique to build the model on each of these datasets [25]. In a heterogeneous ensemble approach, a single dataset is modeled using different techniques to generate multiple models [26]. An ensemble approach could perform better than single model approaches [27].

In an ensemble approach, one of the essential steps is to pool together the performance of features obtained from different models and is referred to as rank aggregation (RA) in this study. The performance metric used for RA varies across the studies, like model estimates [8,21] and goodness of fit [8]. Literature provides various techniques to aggregate the feature performance obtained from different models, but these techniques mainly rely upon a pre-defined rule to aggregate the performance of features, i.e., rule-based rank aggregation approaches. Commonly used methods to aggregate the performance of features is to find the mean, median or Robust rank aggregation (RRA) performance of the feature across all models [28] [29]. However, they cannot learn from the data about the RA rule dynamically and may even be sensitive towards extreme values like mean values.

In high-dimensional data analysis, the performances of rule-based analysis have been challenged by machine learning (ML) based approaches like supervised learning. ML-based approaches are considered effective in the dynamic and complex environment as compared to rule-based approaches because ML creates dynamic rules by learning and adapting to the existing environment [30]. In the case of ensemble FS, the data structure is dynamic and varies across datasets, so it may not always be possible for a predefined rule to give optimal results for all the scenarios [30,31]. Thus, it is desirable to explore the application of ML in all steps of ensemble FS owing to its dynamic

learning characteristics. ML approaches like supervised learning are well established in the model building step [32], but no supervised learning approach is designed for the RA step.

This study proposes a novel perspective to perform RA using the supervised learning approach of the ML called supervised rank aggregation (SRA). First, SRA creates a performance matrix that contains the performance of all features in all the models as the input and the performance of each model in achieving the final data analysis goal as the label. Then, supervised learning is used to find the relative rank or performance of features based on their potential to help achieve the best performance in the final data analysis.

SRA based ensemble feature selection (EFS) is highly innovative in many ways. Firstly, perspective is unique as it pools and ranks features dynamically rather than using fixed rules for EFS. Secondly, it provides a unique application of supervised learning models as they replace the static rule-based RA approach with a dynamic rule-based RA approach. Thirdly, it is versatile, which allows its integration with existing ensemble methods.

This paper provides the "Methodology" section to explain the SRA based EFS. Then, its performance is compared against existing rank aggregation methods used in EFS for simulations and real studies in the "Simulation Studies" and "Real Studies" sections. Finally, we summarize and provide future directions for research in the "Conclusion and Discussion" section.

Methodology

SRA methodology is developed to integrate the supervised learning in the rank aggregation step of the ensemble learning (Figure 2). A dataset of sample size, n, with given input feature space, p, and an outcome is fed into the EFS process, where multiple models are created either by creating multiple bootstrapped datasets from the original dataset (homogenous approach) or by using multiple modeling techniques (heterogeneous approach). Then a performance matrix is created from these multiple models by extracting feature performance and model performance. A

supervised learning algorithm is trained on this performance matrix, and feature importance obtained from the algorithm is used as the final feature ranking or importance. Finally, the features are selected based on an importance cut-off obtained from a predefined threshold or an unsupervised ML algorithm. The proposed methodology is discussed below in more detail.

Figure 2: Graphical representation of SRA methodology based Ensemble feature selection

Generate multiple models

From the original dataset D of feature space p, outcome y, and sample size n, k randomly sampled datasets are generated by randomly sampling features without repeats $q_i | \ i \in \{1, \dots, k\}, 1 < q \leq p$. All k sample datasets have a sample size of n by sampling with replacement from dataset D. A model m is created for every k sample dataset using any modeling technique.

$$m_i : y_i = f(q_i) | \ i \in \{1, \dots, k\} \#(1)$$

where, modeling technique used to prepare the i^{th} dataset model m_i will determine the function f. In this study, RIDGE regression is used as the modeling technique for building the models. Optimal hyperparameter values for each model are obtained from 10-fold cross-validation.

Create Performance Matrix

A performance matrix C is prepared from m models containing k rows and $p + 1$ columns. The matrix contains feature performance, FP as the input features and model performance for study objective, MP as the outcome or label for all m models.

$$C = |FP_{ij} \quad \mathrm{MP}_i | \ | \ i \in \{1, \dots, k\}, j \in \{1, \dots, p\} \#(2)$$

In the current study, model estimates are used as FP metric and predictive performance of a model on the left out bootstrap samples as MP. Accordingly, MP metric used in the study is inverse of root mean square error (RMSE).

Supervised Rank Aggregation

A supervised learning model (SLM) is created from the performance matrix with FP of p features as predictors and MP as the outcome.

$$SLM: MP = g(FP) \#(3)$$

where, machine learning technique used for SLM will determine the function g. Currently, only ML techniques like penalized regression and decision trees which could provide feature importance, $fimp$ in achieving the model performance could be used.

Feature Selection

The importance for each feature is used to select target features q_{best}. It is assumed that the features with more importance should be target features as they are more relevant in achieving higher model performance. In literature, the cut-off value for features is obtained by using a pre-defined threshold [8,21], rule-based threshold estimation [33], or unsupervised learning based threshold estimation [21]. A predefined threshold may require the tuning step to arrive at an appropriate cut-off value, which will give optimal results for a given scenario [8,21]. Rule-based methods may not always provide optimal results [30,31]. Thus, in this study, the K-means based unsupervised learning technique is used for obtaining the threshold cut-off as it will eliminate the need for tuning and dynamically adapt to the given scenario. Since clustering will be happening on a single dimension, hence high dimension limitation of K-mean clustering is avoided. K-means is used to cluster the features into two groups, and the features in the cluster with a higher mean $fimp$ value are selected as final features q_{best}. Pseudo Algorithm summarizes the complete SRA based ensemble feature selection algorithm.

Pseudo Algorithm: SRA based ensemble feature selection	
Input:	Feature data X (p × n) Target feature Y (1 × n) Number of sample datasets k Performance matrix $C = \{empty\}$
Output:	Final Feature set q_{best}

Begin:
 # Step I: Generate multiple models
for i=1 **to** k
  Generate q_i random features from p
  Generate samples $\left(X^i, Y^i \in R^{n \times (q_i+1)}\right)$
  Build embedded model m_i (like RIDGE) from $\left(X^i, Y^i\right)$
end for

 # Step II: Prepare Performance Matrix
for i=1 **to** k
  Compute feature performance estimate FP_i of the model
  Compute model performance estimate MP_i of the model
  Add (FP_i, MP_i) to C
end for

 # Step III: Supervised Rank Aggregation
Build a supervised learning model (like LASSO, RIDGE, Random Forest), SLM from C
Compute feature importance estimate (like feature coefficient, feature importance score) $fimp$ from SLM model for p

 # Step IV: Feature Selection
Generate two clusters (c_1, c_2) from $fimp$ of p features using unsupervised learning like K-means
Compute mean $fimp$ of c_1 $mc_1 = \mu(c_1)$
Compute mean $fimp$ of c_2 $mc_2 = \mu(c_2)$
if mc_1= MAXIMUM(mc_1, mc_2)
  Add c_1 features to q_{best} to get final feature selection
else if mc_2 = MAXIMUM(mc_1, mc_2)
  Add c_2 features to q_{best} to get final feature selection
else
  Add p features to q_{best} to get final feature selection
End

Simulation Studies

We perform simulation studies to evaluate the proposed RA method and compare its performance with multiple other RA methods for EFS. The study generates high-dimensional feature space for marginal models using multivariate normal distributions. The study uses regression model $y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon$ to provide a continuous outcome variable of simulated data with sample size, n for marginal models. β represents the effect of different features and intercept term on the outcome, $\varepsilon \sim N(0, \sigma^2)$ is the normally distributed error term and $x_i \sim N(0, 1)$ are normally distributed input features, p. Multi-collinearity is added between features using the covariance matrix as given below:

$$\begin{bmatrix} x_1x_1 & \cdot & x_1x_{15} & \cdot & \cdot & x_1x_p \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{15}x_1 & \cdot & x_{15}x_{15} & \cdot & \cdot & x_{15}x_p \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_px_1 & \cdot & x_px_{15} & \cdot & \cdot & x_px_p \end{bmatrix} = \begin{bmatrix} 1 & \cdot & 5 & \cdot & \cdot & 0 \\ \cdot & \cdot & 5 & \cdot & \cdot & \cdot \\ 5 & 5 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & \cdot & \cdot & 1 \end{bmatrix}$$

Multiple scenarios are simulated by changing p, n, β, the number of target features, and k (Table 1). Only true features are assigned a non-zero β value. We prepare homogenous ensemble models for feature selection. The dataset for each model is generated by randomly sampling two to p features from p feature space and sub-setting n samples from the original dataset with replacement. RIDGE is used to build models for each dataset. A penalized effect size of each feature obtained from the RIDGE models is scaled using the absolute maximum value, which is used as a feature performance metric.

Implementation of SRA is shown using three different supervised learning algorithms, namely LASSO (SRA-Lasso), RIDGE (SRA-Ridge), random forest (SRA-RF). While LASSO and RIDGE perform supervised learning using a linear combination of features, random forest performs supervised learning using a non-linear combination of features. A supervised learning model in each SRA is prepared using optimized hyperparameter values.

SRA performance is compared with existing rule-based RA methods, namely, mean based RA (MeRA), maximum based RA (MaRA), minimum based RA (MiRA), median based RA (MedRA), coefficient of variation based RA (CVRA), standard deviation based RA (SDRA), robust rank aggregation (RRA), t-test based RA (tRA), and Wilcoxon signed-rank test based RA (WRA). R 4.0.3 is used for the analysis. The study has used some inbuilt packages in statistical language R for the analysis like *glmnet* package [34] for LASSO and RIDGE, *randomForest* package [35] for random forest, and *RobustRankAggreg* package [29] for RRA.

The different RA methods are evaluated for their ability to select target features, discriminate between target and noise features, and predictive performance of the models built using selected features. We use the F1 score for feature discrimination ability evaluation and inverse RMSE for the

test data for the predictive performance evaluation. RIDGE is used to build the final model from the selected features for predictive performance evaluation. Ten trials are performed for each scenario.

Table 2 results suggest that all methods can select some target features under all scenarios, but SRA-Ridge consistently outperformed rule-based RA methods. SRA-Ridge selected almost all the target features in all scenarios. The performance of the other two SRA methods is at par with existing RA methods. Further, the results suggest that SRA-Ridge has a better or at par feature discriminative ability than other methods. Thus, SRA-Ridge not only selects target features but is also good in rejecting noise features as compared to other methods. The results suggest that SRA could be a good candidate to select target features.

Further, SRA-Ridge based selected features can build good predictive models and consistently outperformed rule-based RA methods (Table 2). These findings suggest that the SRA may provide better or at par prediction performance than existing methods. Further, SRA could enhance the performance of ensemble-based approaches in high-dimensional settings.

Real Studies

Three real studies are analyzed to compare the performance of SRA and existing RA methods. Study I is Community Health Status Indicators (CHSI) study that collected US county data (n=3141) containing 578 features to understand non-communicable diseases [36]. Study II is National Social Life, Health and Aging Project (NSHAP) study that collected aged Americans data (n=4377) containing 1470 features to understand their health and well-being [37]. Study III is the DNA methylation data (n=27578) containing 108 samples to understand its relationship with human age [38,39].

Table 3 shows the final cleaned dataset for these three studies used for analysis. Features and samples are filtered to remove highly correlated features, non-continuous features, missing values, and very low standard deviation. The final cleaned dataset is randomly split into training and test

208 dataset. The test dataset is used to evaluate the predictive performance of the features selected by
209 different RA methods. The study uses inverse RMSE as the predictive performance metric. The mean
210 performance of ten trials is used for comparison between RA methods. In the cases of Study I and
211 Study II, 100 ensemble models are created, while in Study III, 1000 ensemble models are created.

212 The results from Table 4 suggest that SRA methods provided better or at par predictive performance
213 than existing RA methods. The better performance of the SRA method suggests that it may be more
214 reliable than existing RA methods in identifying the target features. Further, unlike the simulated
215 data results, different SRA methods have shown different performances. In the case of Study I and
216 Study II, SRA-Ridge has the best predictive performance, but in Study III, SRA-Lasso has the best
217 predictive performance, which suggests that SRA methods performance may change with dataset
218 and ensemble models. In general, the variation in performance of feature selection techniques with
219 dataset has been reported in the literature and could be attributed to data characteristics [40].

220 In the current study, Study III data is also used to compare the performance of SRA based selected
221 methylated features with state-of-art literature based selected features [41,42]. SRA-Lasso is used to
222 obtain the target features. The complete dataset is used for the FS step rather than the training
223 data. SRA-Lasso identified 484 methylation sites compared to 353 methylation sites identified by the
224 literature, but only ten methylation sites are shared between the two approaches (Supplementary
225 File 1). The selected methylation sites from the two approaches are compared for their predictive
226 performance on the test data. Accordingly, the Study III dataset is split into training (80%) and test
227 (20%) data. RIDGE model is prepared using training data followed by predictive performance
228 measurement on test data. It is found that SRA-Lasso based selected features provided a marginally
229 better predictive performance ($RMSE^{-1}$(95% CI): 0.06 (0.05-0.07)) compared to literature
230 recommended selected features ($RMSE^{-1}$(95% CI): 0.05 (0.04-0.05)).

231 Further, we identified the differentially expressed genes associated with selected methylated sites
232 using *BioMethyl* package [43]. SRA-Lasso based selected methylated sites are linked with 288 genes,

233 but literature based selected methylated sites are linked with only 136 genes (Supplementary File 2).
234 Only ten genes, namely SFRP1, STRA6, BNC1, CSPG5, DCHS1, DIRAS3, TCF15, ERG, PIPOX, and
235 MCAM, are shared between the two approaches. Literature also provides a database, *GenAge*, of
236 307 genes commonly associated with age [44]. Among the 136 genes linked with literature-based
237 methylation sites, only 1 out of 308 genes is found (Supplementary File 3). However, among the 288
238 genes linked with SRA-Lasso based methylation sites, 9 out of 308 genes are found (Supplementary
239 File 3). Thus, SRA-Lasso may be relevant in identifying target features that have both biological
240 importance and good predictive performance.

241 Conclusion and Discussion

242 This paper proposes SRA, an innovative approach, to perform rank aggregation in ensemble models
243 for feature selection. The approach allows dynamic learning of feature performance pooling
244 strategy, which current rule-based rank aggregation methods do not perform. The approach is
245 flexible and could be incorporated into any ensemble technique. The SRA could identify target
246 features while retaining very few noise features compared to other methods. The simulated data
247 studies showed that SRA outperforms existing methods in feature selection and prediction
248 performance. Similar performance in real datasets also demonstrates the practical relevance of SRA.

249 The proposed method has certain limitations. The scope of the current study is limited to concept
250 testing. Consequently, the robustness of the approach on different data types and modeling
251 techniques could be the focus of future research. The ensemble model used in the study assumes a
252 linear combination of features. Thus, future research could study SRA for algorithms designed to
253 explore the non-linear combinations of features.

254 Key Points

255 • Supervised Rank Aggregation (SRA) methods are better than rule-based rank aggregation
256 methods for ensemble-based feature selection (EFS).

- SRA Ridge could give much better discrimination between true and noise features as well as predictive performance than rule-based rank aggregation methods
- SRA could be useful in detecting the genomic features like methylation sites which could have biological relevance

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Availability of data and materials

All the datasets and code are in the github link: https://github.com/rahijaingithub/SRA.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada [NSERC Grant RGPIN-2017-06672 to W.X.]; and the Prostate Cancer Canada [Translation Acceleration Grant 2018 to R.J. and W.X.].

Author Contributions

ALL AUTHORS HAVE READ AND APPROVED THE MANUSCRIPT.

Conceptualisation: RJ, WX

Formal Analysis: RJ

Investigation: RJ

Methodology: RJ, WX

Software: RJ

Supervision: RJ, WX

Validation: RJ, WX

Writing-original draft: RJ

Writing-review & editing: RJ, WX

Acknowledgements

Not Applicable

Reference

1. Bellman R. Dynamic Programming. Math. Sci. Eng. 1967; 40:101–137

2. Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proc. Int. Congr. Math. Madrid, August 22–30, 2006 2007; 595–622

3. Ayesha S, Hanif MK, Talib R. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inf. Fusion 2020; 59:44–58

4. Piles M, Bergsma R, Gianola D, et al. Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. Front. Genet. 2021; 12:

5. Healy G, Salinas-Miranda E, Jain R, et al. Pre-operative radiomics model for prognostication in resectable pancreatic adenocarcinoma with external validation. Eur. Radiol. 2021; Online:

6. Walter S, Tiemeier H. Variable selection: Current practice in epidemiological studies. Eur. J. Epidemiol. 2009; 24:733–736

7. Heinze G, Wallisch C, Dunkler D. Variable selection – A review and recommendations for the practicing statistician. Biometrical J. 2018; 60:431–449

8. Jain R, Xu W. HDSI: High dimensional selection with interactions algorithm on feature selection and testing. PLoS One 2021; 16:1–17

9. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. J. Am. Stat. Assoc. 1988; 83:1023–1032

10. Zycinski G, Barla A, Squillario M, et al. Knowledge Driven Variable Selection (KDVS) - a new approach to enrichment analysis of gene signatures obtained from high-throughput data. Source

Code Biol. Med. 2013; 8:1–14

11. Yang P, Huang H, Liu C. Feature selection revisited in the single-cell era. Genome Biol. 2021; 22:1–17

12. Dhal P, Azad C. A comprehensive survey on feature selection in the various fields of machine learning. Appl. Intell. 2021; 51:1–39

13. Chormunge S, Jena S. Correlation based feature selection with clustering for high dimensional data. J. Electr. Syst. Inf. Technol. 2018; 5:542–549

14. Dash M, Liu H, Yao J. Dimensionality reduction of unsupervised data. Proc. Ninth IEEE Int. Conf. Tools with Artif. Intell. 1997; 532–539

15. Zhang R, Nie F, Li X, et al. Feature selection with multi-view data: A survey. Inf. Fusion 2019; 50:158–167

16. Kohavi R, John GH. Wrappers for feature subset seelction. Artif. Intell. 1997; 97:273–324

17. Tarkhaneh O, Nguyen TT, Mazaheri S. A novel wrapper-based feature subset selection method using modified binary differential evolution algorithm. Inf. Sci. (Ny). 2021; 565:278–305

18. Alweshah M, Alkhalaileh S, Al-betar MA. Coronavirus herd immunity optimizer with greedy crossover for feature selection in medical diagnosis. Knowledge-Based Syst. 2020; 235:107629

19. Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. J. R. Stat. Soc. Ser. B Stat. Methodol. 2011; 73:273–282

20. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 2010; 72:3–25

21. Jain R, Xu W. RHDSI: A novel dimensionality reduction based algorithm on high dimensional feature selection with interactions. Inf. Sci. (Ny). 2021; 574:590–605

22. Lal TN, Chapelle O, Weston J. Embedded Methods. Featur. Extr. Found. Appl. 2006; 165:137–165

23. Hancer E, Xue B, Zhang M. A survey on feature selection approaches for clustering. Artif. Intell. Rev. 2020; 53:4519–4545

24. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, et al. Ensemble feature selection: Homogeneous and heterogeneous approaches. Knowledge-Based Syst. 2017; 118:124–139

25. Hosni M, Idri A, Abran A. On the value of filter feature selection techniques in homogeneous ensembles effort estimation. J. Softw. Evol. Process 2021; 33:e2343

26. Mera-Gaona M, López DM, Vargas-Canas R, et al. Framework for the ensemble of feature selection methods. Appl. Sci. 2021; 11:1–16

27. Tsai CF, Sung YT. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. Knowledge-Based Syst. 2020; 203:106097

28. Noureldien N, Mohmoud S. The Efficiency of Aggregation Methods in Ensemble Filter Feature Selection Models. Trans. Mach. Learn. Artif. Intell. 2021; 9:39–51

29. Kolde R, Laur S, Adler P, et al. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 2012; 28:573–580

30. van Ginneken B. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. Radiol. Phys. Technol. 2017; 10:23–32

347 31. Cronin RM, Fabbri D, Denny JC, et al. A comparison of rule-based and machine learning
348 approaches for classifying patient portal messages. Int. J. Med. Inform. 2017; 105:110–120

349 32. Lopez-Rincon A, Mendoza-Maldonado L, Martinez-Archundia M, et al. Machine learning-based
350 ensemble recursive feature selection of circulating mirnas for cancer tumor classification. Cancers
351 (Basel). 2020; 12:1–27

352 33. Seijo-Pardo B, Bolón-Canedo V, Alonso-Betanzos A. Testing Different Ensemble Configurations
353 for Feature Selection. Neural Process. Lett. 2017; 46:857–880

354 34. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via
355 Coordinate Descent. J. Stat. Softw. 2010; 33:1–22

356 35. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002; 2:18–22

357 36. [Dataset] Centers for Disease Control and Prevention. Community Health Status Indicators (CHSI)
358 to Combat Obesity, Heart Disease and Cancer. Healthdata.gov 2012;

359 37. [DATASET] Waite LJ, Laumann EO, Levinson WS, et al. National Social Life, Health, and Aging
360 Project (NSHAP): Wave 1, [United States], 2005-2006 (ICPSR 20541). Inter-university Consort. Polit.
361 Soc. Res. 2019;

362 38. Numata S, Ye T, Hyde TM, et al. DNA methylation signatures in development and aging of the
363 human prefrontal cortex. Am. J. Hum. Genet. 2012; 90:260–272

364 39. [Dataset] Akalin A. compGenomRData. Github 2019;

365 40. Parmezan ARS, Lee HD, Spolaôr N, et al. Automatic recommendation of feature selection
366 algorithms based on dataset characteristics. Expert Syst. Appl. 2021; 185:115589

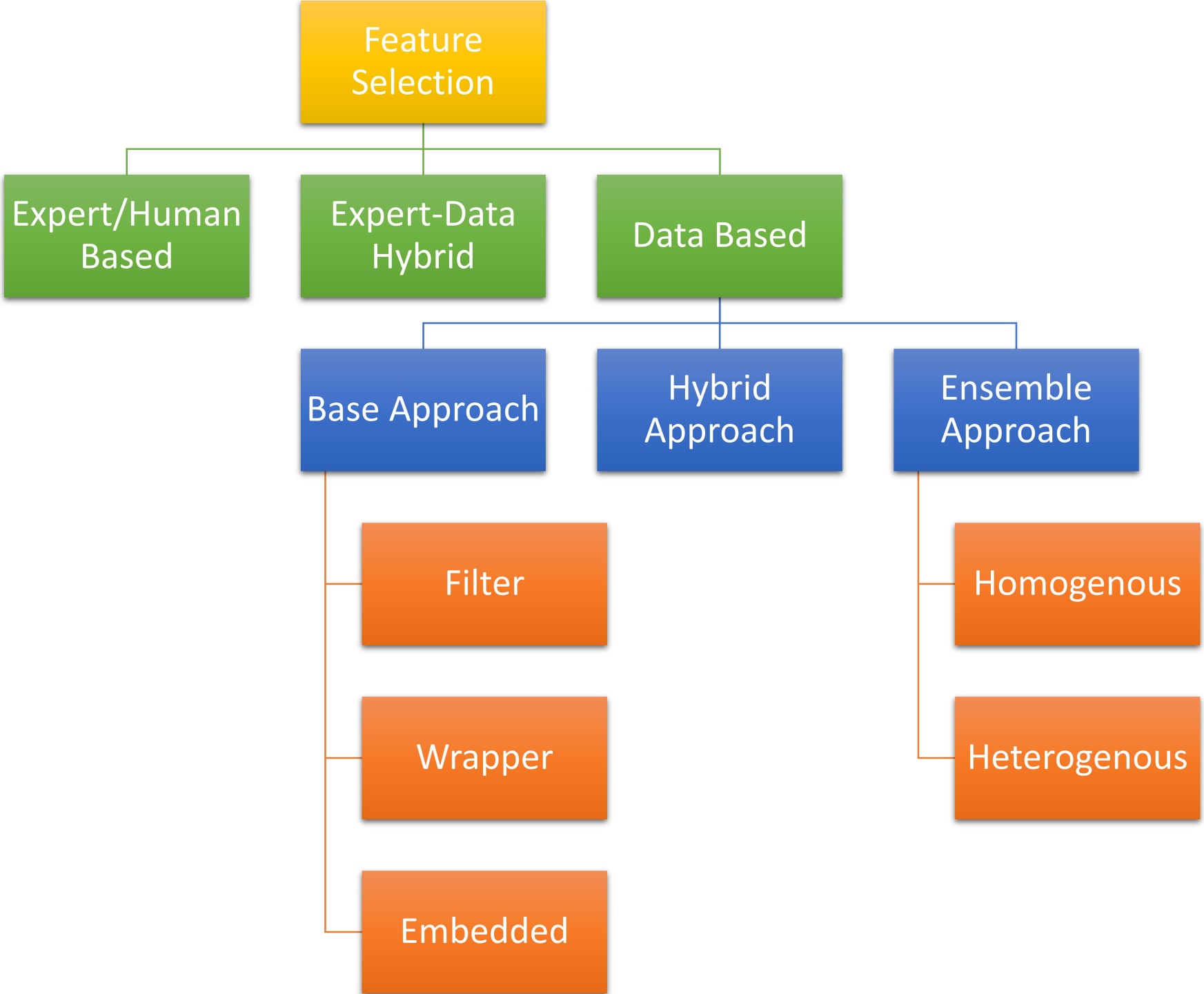
367 41. Pelegi-Siso D, De Prado P, Ronkainen J, et al. Methylclock: A Bioconductor package to estimate
368 DNA methylation age methylclock: A Bioconductor package to estimate DNA methylation age.
369 Bioinformatics 2021; 37:1759–1760

370 42. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2015; 16:1–19

371 43. Wang Y, Franks JM, Whitfield ML, et al. BioMethyl: An R package for biological interpretation of
372 DNA methylation data. Bioinformatics 2019; 35:3635–3641

373 44. Tacutu R, Thornton D, Johnson E, et al. Human Ageing Genomic Resources: New and updated
374 databases. Nucleic Acids Res. 2018; 46:D1083–D1090

375


Feature Selection
Expert/Human Based
Expert-Data Hybrid
Data Based
Base Approach
Hybrid Approach
Ensemble Approach
Filter
Wrapper
Embedded
Homogenous
Heterogenous

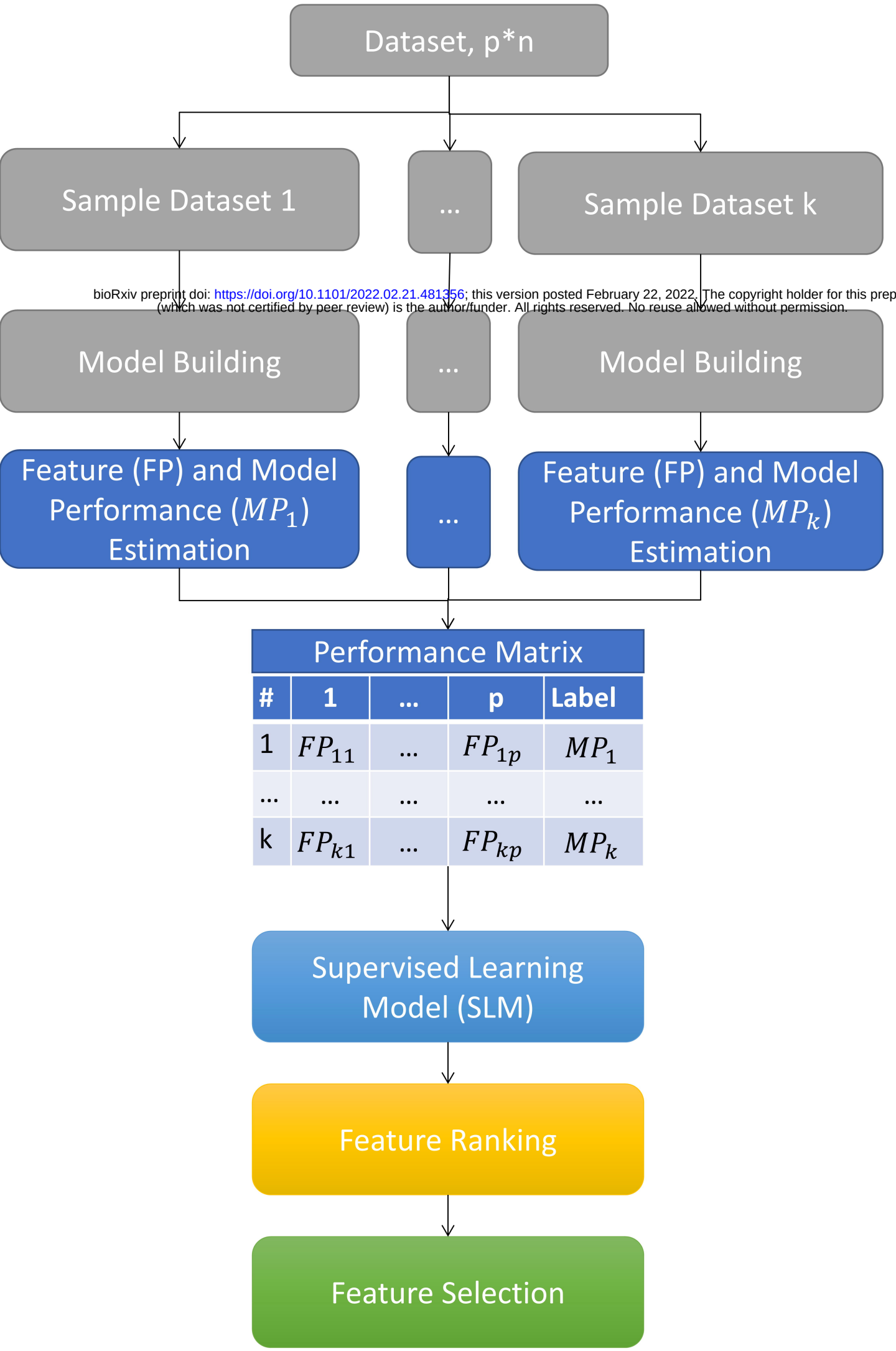


Table 1: Description of the scenarios of simulation studies

Scenario	β (Non-Zero coefficients)	p	Sample Size		σ	k
			Train (n)	Test		
A	$\{\beta_i \mid i = \{1, \ldots, 10\}\} = \{0.9, \ldots, 0.9\}$	75	100	500	0.25	300
B	$\{\beta_i \mid i = \{1, \ldots, 10\}\} = \{0.5, \ldots, 0.5\}$	100	100	500	0.25	100
C	$\{\beta_i \mid i = \{1, \ldots, 15\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, 0.4\}$	175	275	500	0.25	100
D	$\{\beta_i \mid i = \{1, \ldots, 15\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, 0.4\}$	75	275	500	0.25	100
E	$\{\beta_i \mid i = \{1, \ldots, 15\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, 0.4\}$	75	225	500	0.25	200
F	$\{\beta_i \mid i = \{1, \ldots, 20\}\} = \{0.4, -0.8, 0.4, -0.8, \ldots, -0.8\}$	125	225	500	0.25	200

Table 1: Comparison of model performance between SRA methods and Existing methods under six scenarios in terms of target feature selection, feature discrimination ability (F1 Score) and outcome prediction (1/RMSE)

RA technique		Scenarios					
		A	***B***	***C***	***D***	***E***	***F***
		Target Features (%) [μ(95%CI)]					
Existing	CVRA	**100 (100-100)**	**100 (100-100)**	46 (45-47)	46 (45-47)	47 (47-47)	51 (48-53)
	MARA	**100 (100-100)**	**100 (100-100)**	87 (81-93)	**97 (95-99)**	85 (78-93)	66 (56-76)
	MeRA	**100 (100-100)**	**100 (100-100)**	47 (47-47)	47 (47-47)	47 (47-47)	53 (51-55)
	MedRA	**100 (100-100)**	**100 (100-100)**	47 (47-47)	47 (46-49)	47 (47-47)	53 (50-56)
	MIRA	62 (48-76)	94 (89-99)	**95 (91-98)**	77 (72-82)	85 (82-89)	88 (86-89)
	RRA	99 (97-100)	99 (97-101)	47 (47-47)	48 (46-50)	47 (46-49)	52 (50-54)
	SDRA	78 (67-89)	71 (67-75)	34 (29-39)	40 (35-45)	35 (27-42)	39 (32-46)
	tRA	**100 (100-100)**	**100 (100-100)**	46 (45-47)	45 (44-47)	47 (47-47)	51 (48-53)
	WRA	**100 (100-100)**	**100 (100-100)**	49 (45-53)	53 (53-53)	53 (53-53)	37 (34-40)
SRA	Lasso	92 (87-97)	37 (18-56)	41 (37-45)	58 (51-65)	63 (57-69)	46 (38-54)
	RF	98 (95-100)	53 (43-63)	63 (54-73)	67 (57-76)	65 (56-75)	61 (53-69)
	Ridge	**100 (100-100)**	99 (97-100)	**95 (89-100)**	95 (92-99)	**100 (100-100)**	**92 (88-96)**
RA technique		**F1 Score[μ(95%CI)]**					
Existing	CVRA	**1.00 (1.00-1.00)**	0.93 (0.89-0.97)	0.63 (0.62-0.64)	0.63 (0.62-0.64)	0.64 (0.64-0.64)	0.67 (0.65-0.69)
	MARA	0.58 (0.55-0.61)	0.70 (0.68-0.72)	0.60 (0.55-0.64)	0.83 (0.8-0.86)	0.65 (0.61-0.69)	0.44 (0.39-0.49)
	MeRA	0.83 (0.8-0.86)	0.81 (0.80-0.82)	0.64 (0.64-0.64)	0.64 (0.64-0.64)	0.64 (0.64-0.64)	0.69 (0.67-0.71)
	MedRA	0.85 (0.82-0.87)	0.81 (0.80-0.82)	0.64 (0.64-0.64)	0.64 (0.63-0.65)	0.64 (0.64-0.64)	0.69 (0.67-0.71)
	MIRA	0.41 (0.33-0.49)	0.60 (0.53-0.67)	0.79 (0.73-0.85)	0.75 (0.72-0.78)	0.70 (0.67-0.73)	0.76 (0.74-0.79)
	RRA	0.90 (0.87-0.93)	0.82 (0.80-0.83)	0.64 (0.64-0.64)	0.65 (0.63-0.66)	0.64 (0.63-0.65)	0.68 (0.67-0.7)
	SDRA	0.30 (0.27-0.32)	0.22 (0.20-0.24)	0.07 (0.06-0.07)	0.16 (0.15-0.18)	0.16 (0.14-0.17)	0.12 (0.10-0.14)
	tRA	**1.00 (1.00-1.00)**	0.92 (0.88-0.96)	0.63 (0.62-0.64)	0.62 (0.61-0.64)	0.64 (0.64-0.64)	0.67 (0.65-0.69)
	WRA	0.40 (0.38-0.42)	0.33 (0.32-0.34)	0.14 (0.13-0.16)	0.29 (0.27-0.3)	0.30 (0.28-0.32)	0.18 (0.17-0.2)
SRA	Lasso	0.96	0.33	0.58	0.73	0.77	0.62

		(0.93-0.98)	(0.16-0.50)	(0.53-0.62)	(0.67-0.79)	(0.72-0.81)	(0.55-0.7)
	RF	0.75 (0.69-0.81)	0.29 (0.24-0.33)	0.64 (0.57-0.71)	0.73 (0.66-0.8)	0.77 (0.71-0.83)	0.70 (0.64-0.76)
	Ridge	**1.00 (1.00-1.00)**	**0.99 (0.98-1.00)**	**0.97 (0.94-1.00)**	**0.98 (0.96-0.99)**	**1.00 (1.00-1.00)**	**0.95 (0.93-0.98)**
RA technique		**Predictive Performance (1/RMSE) [μ(95%CI)]**					
Existing	CVRA	**3.50 (3.29-3.71)**	3.82 (2.90-4.75)	0.81 (0.79-0.84)	0.84 (0.81-0.86)	0.83 (0.80-0.85)	0.75 (0.72-0.77)
	MARA	2.67 (2.43-2.90)	3.65 (2.77-4.54)	1.73 (1.43-2.03)	2.42 (1.97-2.86)	1.43 (1.01-1.85)	0.58 (0.51-0.65)
	MeRA	2.94 (2.56-3.31)	3.67 (2.83-4.51)	0.82 (0.80-0.84)	0.84 (0.82-0.87)	0.83 (0.80-0.85)	0.76 (0.73-0.79)
	MedRA	2.96 (2.55-3.36)	3.67 (2.83-4.51)	0.82 (0.80-0.84)	0.85 (0.82-0.89)	0.83 (0.80-0.85)	0.76 (0.73-0.79)
	MIRA	0.80 (0.27-1.34)	2.58 (2.00-3.17)	2.45 (1.97-2.93)	1.45 (1.29-1.61)	1.74 (1.57-1.91)	1.31 (1.25-1.37)
	RRA	2.92 (2.42-3.42)	3.54 (2.61-4.46)	0.82 (0.80-0.84)	0.87 (0.83-0.91)	0.84 (0.81-0.87)	0.75 (0.73-0.77)
	SDRA	1.10 (0.53-1.68)	1.03 (0.96-1.10)	0.68 (0.66-0.7)	0.77 (0.74-0.8)	0.71 (0.66-0.76)	0.53 (0.47-0.58)
	tRA	**3.50 (3.29-3.71)**	3.79 (2.91-4.67)	0.81 (0.79-0.83)	0.84 (0.81-0.86)	0.83 (0.80-0.85)	0.75 (0.72-0.77)
	WRA	2.18 (1.96-2.39)	2.62 (2.39-2.85)	0.45 (0.44-0.45)	0.47 (0.45-0.48)	0.46 (0.45-0.46)	0.37 (0.36-0.38)
SRA	Lasso	2.17 (1.32-3.02)	0.77 (0.51-1.03)	0.79 (0.76-0.82)	1.00 (0.85-1.16)	1.09 (0.96-1.21)	0.72 (0.66-0.79)
	RF	2.62 (2.07-3.17)	0.88 (0.77-0.99)	0.70 (0.62-0.78)	0.87 (0.46-1.27)	0.73 (0.55-0.9)	0.55 (0.50-0.59)
	Ridge	**3.50 (3.29-3.71)**	**3.83 (2.91-4.75)**	**2.58 (2.07-3.08)**	**2.65 (2.02-3.28)**	**2.98 (2.51-3.44)**	**1.87 (1.46-2.28)**

4

Table 1: Summary of the real datasets

Real Studies	Marginal Features (p)	Outcome feature	Sample size (n)			k
			Total	Train	Test	
Study I	45	Height	1035	207	828	100
Study II	74	Number of unhealthy days	177	141	36	100
Study III	2289	Age	108	86	22	1000

Table 1: Comparison of SRA methods with Existing methods for three real studies in terms of outcome prediction (1/RMSE)

RA technique		Study		
		I	***II***	***III***
		Predictive Performance (1/RMSE) [μ(95%CI)]		
Existing	CVRA	1.08(1.07-1.1)	**1.28(1.22-1.35)**	2.36(2.15-2.56)
	MARA	1.14(1.12-1.16)	1.25(1.2-1.31)	2.14(1.95-2.34)
	MeRA	3.14(2.96-3.31)	**1.28(1.23-1.33)**	2.35(2.17-2.54)
	MedRA	3.16(2.98-3.35)	**1.28(1.23-1.33)**	2.38(2.19-2.57)
	MIRA	2.96(2.76-3.17)	1.22(1.16-1.27)	1.76(1.67-1.86)
	RRA	3.13(2.96-3.3)	**1.28(1.19-1.36)**	2.39(2.18-2.61)
	SDRA	1.06(1.03-1.09)	1.16(1.09-1.23)	1.05(1.02-1.09)
	tRA	1.08(1.07-1.1)	**1.28(1.2-1.36)**	2.35(2.15-2.55)
	WRA	1.13(1.11-1.14)	1.25(1.19-1.3)	2.05(1.87-2.23)
SRA	Lasso	1(0.99-1.01)	1.23(1.14-1.31)	**2.72(2.25-3.19)**
	RF	2.51(1.86-3.16)	1.23(1.14-1.32)	1.88(1.75-2)
	Ridge	**3.21(3.03-3.39)**	**1.28(1.22-1.34)**	2.30(2.12-2.47)