Supplementary

Application of filter feature selection methods to the genetic data of complex diseases

1 Information theoretic methods

Dimensionality reduction is the process of reducing the number of features that need to take into account when making predictions. Dimensionality reduction methods can be classified as 'feature extraction' or 'feature selection' methods. Feature extraction methods transform a high-dimensional space of feature measurements to a space of fewer dimensions. Principal Component Analysis (PCA) can achieve such a transformation without reference to the output scores by exploiting dependencies between feature measurements, that may be assumed to be linear[1]. Many other transformations exist which can exploit nonlinear dependencies and the statistics of the output scores. 'Feature selection' approaches try to find a subset of the original variables that enable more accurate prediction by the elimination of irrelevant and confusing information. Filter methods select features based on a performance measure regardless of the employed data modeling algorithm and separate the classification and feature selection components. Filter methods are generally applied as pre-processing steps, with subset selection procedures that are independent of the learning algorithm and the defining component of filter based methods is scoring criterion, which is often as 'relevance index'[2]. The relevance index denotes how useful each feature is likely to be for the ML classification methods. Although this leads to a faster learning process, it is possible for the criterion used in the pre-processing step to result in a subset that may not work very well downstream in the learning algorithm. The information theoretic methods investigate the multivariate interaction within features and the scoring criterion is weighted sum of feature relevancy and redundancy. The main goal of feature selection is obtaining a subset of features that produces the highest 'Area under the ROC Curve' (AUROC) and the precsion-recall curve on the classification models[3]. It has been mathematically proven that the performance ranks of two models remain same in the ROC space and the PR space[4]. The classification performance is necessarily proportional to removal of redundant features. Wrapper methods[1,5] search the space of feature subset based on the accuracy of a particular classifier (e.g. LG or RF). Embedded methods perform feature selection in the process of training and during the modeling algorithm's execution. Hybrid methods were proposed to combine the best properties of filters and wrappers. Direct feature selection searches to identify, individually, the relevant features and discard the irrelevant ones. Such methods are instances of a wide range of general strategies for dimensionality reduction, which seek to map the input variables into a lower dimensional space prior to running the supervised learning algorithm. A learning algorithm is faced with the problem of selecting a relevant subset of features which makes the best prediction while ignoring the rest in the features. Since the usual goal of supervised learning algorithms is to minimise regression error on an unseen test set, we have adopted this as our goal in guiding the feature subset selection. Univariate and multivariate methods are two categories for all filter based methods. Univariate methods, the scoring criterion only consider the relevancy of features while ignoring the feature redundancy. Multivariate method investigates the multivariate interaction within features and the scoring criterion is a weighted sum of feature relevancy and redundancy.

Mutual information is univariate feature selection approach (Shannon, 1948) measures the amount of information shared by an input feature X and class label (target) Y. Where the lower case x or y is possible values that the variables X and Y can adopt from the alphabet X and Y respectively in (1). To obtain this, we need to estimate the distribution of p_x and p_y respectively[6].

$$I(X;Y) = \sum_{x \in x} \sum_{y \in y} p(xy) log \frac{p(xy)}{p(x)p(y)} \tag{1}$$

Mutual Information Maximization (MIM) method given by (2) examines the mutual information between a class label Y and a feature X_k, where K is the top features[6]. MIM assumes that all the features are independent and it does not account any dependencies between the features.

$$J_{\text{mim}}(X_k) = I(X_k;Y) \tag{2}$$

Multivariate feature selection methods are described as follows Joint Mutual Information (JMI) was proposed by Yang and Moody (1999)[7,8]. JMI is the information between the targets and a joint random variable defined by pairing the candidate X_n with each current feature. The redundancy term full captures by JMI.

$$J_{\text{jmi}} = \sum_{X_j \in S} I(X_k X_j;Y) \tag{3}$$

Minimal-Redundancy-Maximal-Relevance (mRMR) given by (4) was proposed by Peng et al[9]. This takes the mean of redundancy term and it elemenated the conditional term. In equation (4), n is size of a feature set.

$$J_{\text{mRMR}} = I(X_n;Y) - \frac{1}{n-1} \sum_{k=1}^{n-1} I(X_n;X_k) \tag{4}$$

Conditional Mutual Information Maximization (CMIM) given by (6) was proposed by Fleuret (2004)[10] and is probably the most-well known recent criterion. CMIM measures the information between a feature and the target and it is conditioned on each current feature. The interaction information is the term in square brackets which can be both negative and positive. A negative value indicate that the shared information between X_k and Y has decreased as the result of including X_n.

$$J_{\text{cmim}} = I(X_n;Y) - max_k[I(X_j;X_k,Y) - I(X_n;X_k|Y)] \tag{5}$$

Mutual information feature selection (MIFS) does not consider conditional redundancy (g = 0), but it does incorporate the redundancy penalty (Brown et al., 2012).

$$J_{\text{mifs}} = I(X_k;Y) - \beta(X_n;X_k) - I(X_n;X_k)] \tag{6}$$

Double input symmetrical relevance (DISR) aims to better include such complimentary features by expanding JMI[11]. Disr normalises the information provided by a feature by how well the given feature complements the other features.

$$J_{\text{disr}}(X_k) = \frac{I(X_nX_k;Y)}{H(X_nX_k;Y)} \tag{7}$$

The interaction capping (ICAP) approximated by following equation.

$$J_{\text{icap}}(X_k) = I(X_k;Y) - \sum_{X_j \in S} max_k[0, I(X_k;X_j) - I[(X_k;X_j|Y)] \tag{8}$$

2 Nested Cross Validation

Nested cross validation was used at model development stage to assure good generalisability when the models were tested with hold out data. Varma and Simon[12] report a bias in error estimation when using cross-validation for model selection; therefore, we used stratified nested cross-validation as an almost unbiased estimate of the true AUC. The validation data-set is often used to fine-tune models. For example, we try out various sets of K for a KNNC model by finding the AUC produced by each set of K for the validation data-set. This would allow us to choose among the competing sets of K. In such a case, the AUC with the validation data-set will be an optimistic estimate of how the fine-tuned model would perform with unseen data[1]. This is because the final K will has been chosen such that the AUC with the validation data-set is the highest possible.

3 Figure for orginal dataset with potential confounders

We added three potential confounders aao, PC1 and PC2 to the dataset and no mitigation is applied to the confounders in Figure 1. aao had the vote 100 for all features selection criteria and followed by HLA_B _*06 with the vote '100' in 'MIM' and JMI and HLA_DRB1_07 with the vote '97' in 'mim'and HLA_B_*27 with the vote '89' in 'disr. PC1 and PC2 had the average vote for the most FS criteria and there effect was mitigated post to the mitigation.

4 Figures for impact of original dataset confounding on feature selection

Figure 2 and Figure 3 show the average AUC in nested cross validation and AUC in hold out set respectively when CMIM, DISR, ICAP, JMI, MIFs, MRMR and overall ranking was applied to the original dataset.

5 Figures for impact of mitigated confounding on feature selection

Figure 4 and 5 illustrate the average AUC in nested cross validation and the AUC in hold out set respectively when CMIM, DISR, ICAP, JMI, MIFs, MRMR and overall ranking was applied to the mitigated dataset

6 Figures for evaluation metrics

Figures 6,7 show ROC curve and precision-recall for 6 ML models

Figures8 show the accuracy, precision, recall and F1 score for 448 generated different models. Figures 9 depict the 64 different combination for each ML model.

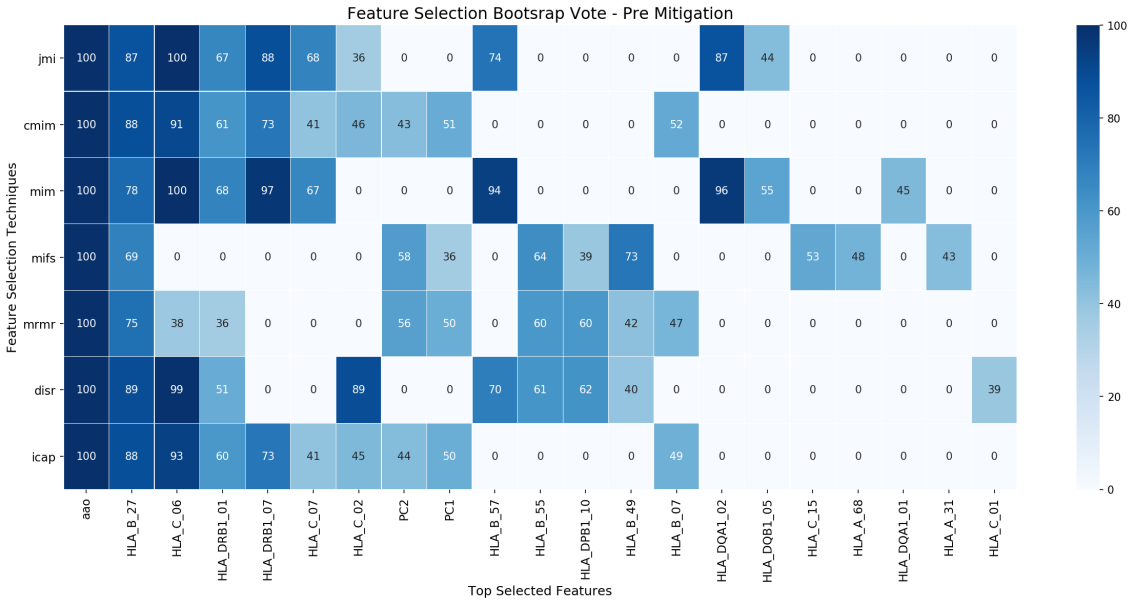


Figure 1. Heatmap (a) feature ranking original dataset with three confounders aao, PC1, PC2. The majority vote over 100 bootstrap for the top 10 selected features (in rows) and seven features selection techniques in (columns)

Model Num	The best models			Accuracy %			F1-score%		
	Model Name	Feature Selection	Top features	Cross validation	Hold out	External	Cross validation	Hold out	External
402	LG	disr	40	0.59	0.54	0.56	0.58	0.53	0.56
303	Adaboost	jmi	60	0.62	0.61	0.53	0.62	0.61	0.56
416	DT	disr	10	0.54	0.54	0.51	0.26	0.25	0.30
398	XGBoost	disr	40	0.58	0.55	0.55	0.58	0.55	0.55
232	KNNC	disr	60	0.73	0.76	0.53	0.76	0.53	0.74
39	NB Gaussain	mim	10	0.56	0.54	0.55	0.45	0.42	0.48
184	Random Forest	icap	20	0.58	0.54	0.57	0.56	0.49	0.58

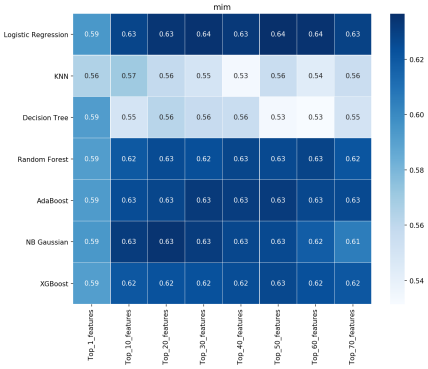
Table 1. The best generated models out of 448 generated models

References

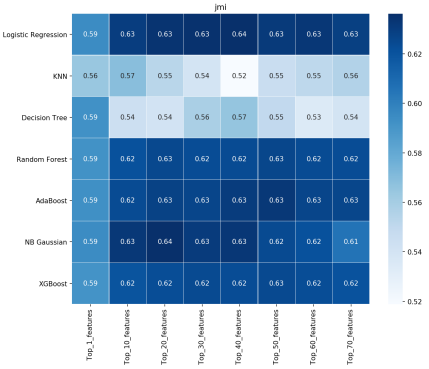
1. Jalalinajafabadi, F. *Computerised GRBAS Assessement of Voice Quality*. Ph.D. thesis, The University of Manchester (United Kingdom) (2016).
2. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. machine learning research* **3**, 1157–1182 (2003).
3. Das, S. Filters, wrappers and a boosting-based hybrid for feature selection. In *Icml*, vol. 1, 74–81 (2001).
4. Davis, J. & Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 233–240 (2006).
5. Kohavi, R., John, G. H. *et al.* Wrappers for feature subset selection. *Artif. intelligence* **97**, 273–324 (1997).
6. Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The journal machine learning research* **13**, 27–66 (2012).
7. Yang, H. & Moody, J. Feature selection based on joint mutual information. In *Proceedings of international ICSC symposium on advances in intelligent data analysis*, vol. 1999, 22–25 (Citeseer, 1999).
8. Brown, G. A new perspective for information theoretic feature selection. In *Artificial intelligence and statistics*, 49–56 (PMLR, 2009).

(a) CMIM

(b) DISR

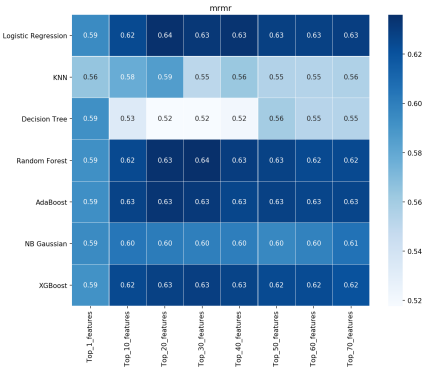


(c) mim



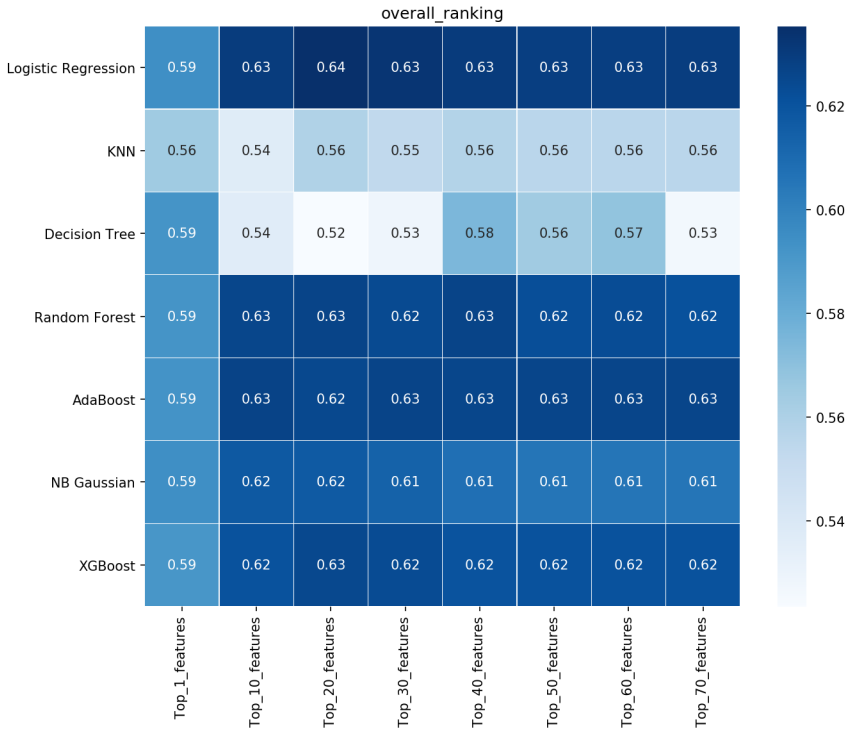
(d) JMI

(e) MIFS



(f) MRMR

Figure 2. Pre-Mitigated Feature Selection - Cross-validation.



(g) Overall Ranking

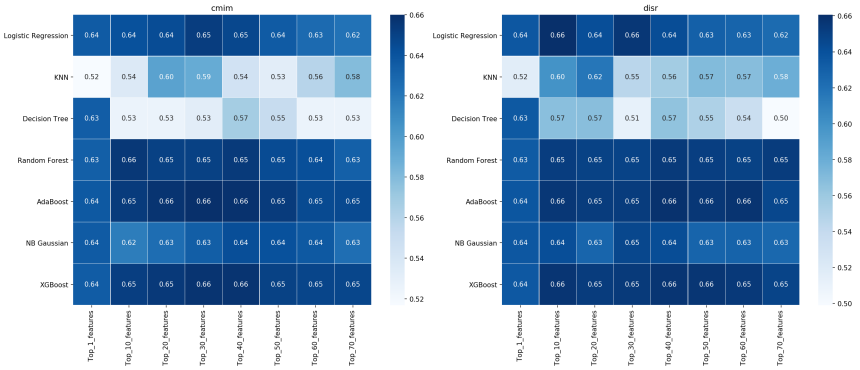
Figure 2. Pre-Mitigated Feature Selection - Cross-validation.

9. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis machine intelligence* **27**, 1226–1238 (2005).

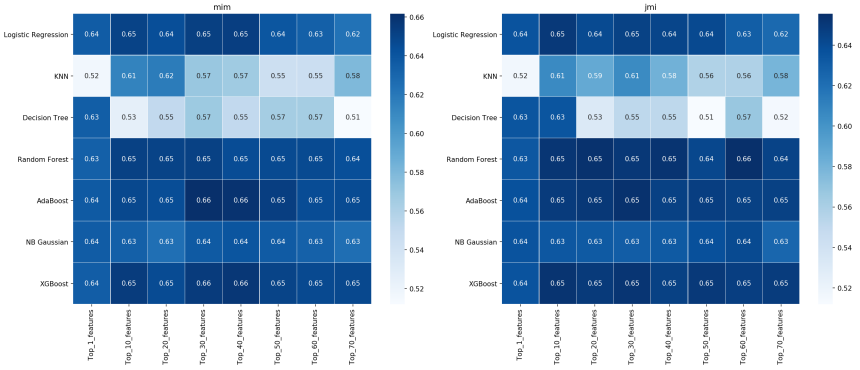
10. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. learning research* **5**, 1531–1555 (2004).

11. Bennasar, M., Setchi, R. & Hicks, Y. Feature interaction maximisation. *Pattern recognition letters* **34**, 1630–1635 (2013).

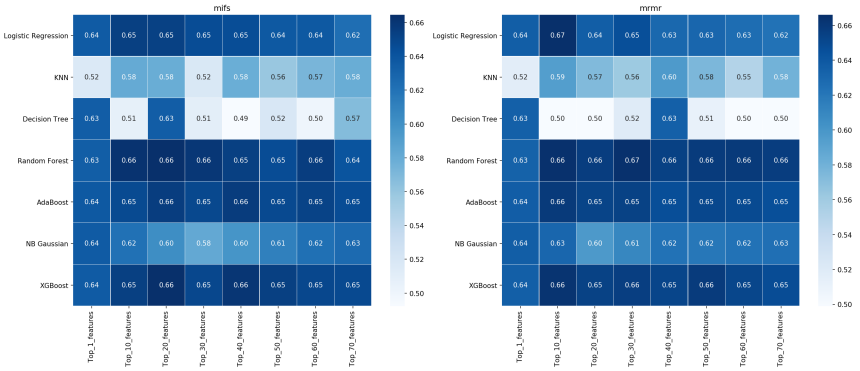
12. Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics* **7**, 91 (2006).



(a) CMIM **(b)** DISR

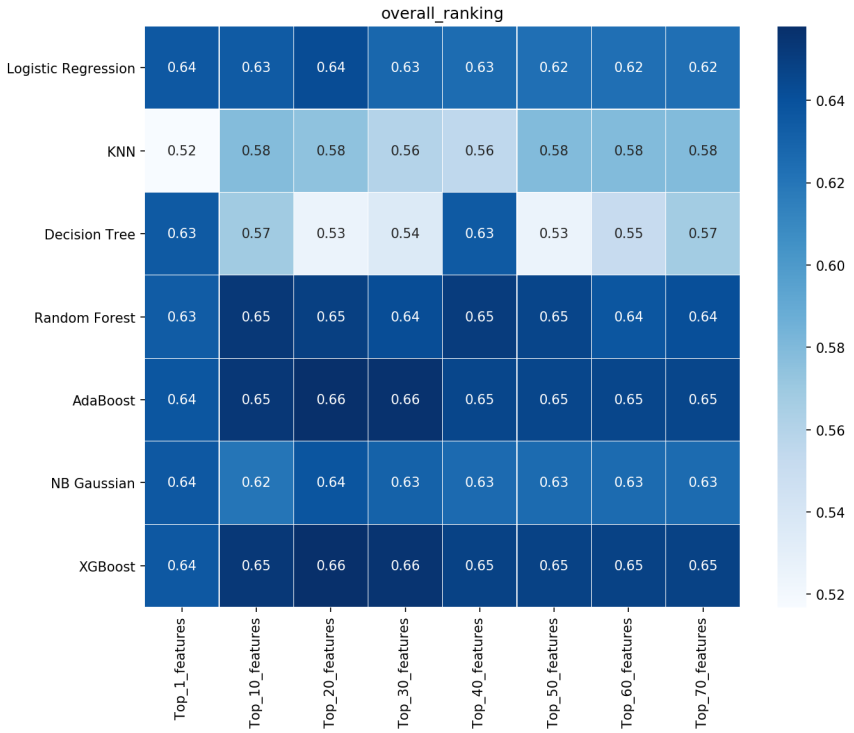


(c) MIM **(d)** JMI



(e) MIFS **(f)** MRMR

Figure 3. Pre-mitigated Feature Selection - hold out.



(g) Overall Ranking

Figure 3. Heatmap depicting the predictive performance (AUC for hold out set) for different number of HLA features(in rows) and different classification method in (columns). It can be observed for all feature selection all classifiers show relatively the same predictive performance in many cases. Pre-Mitigated Feature Selection -Hold-out.

(a) CMIM

(b) DISR

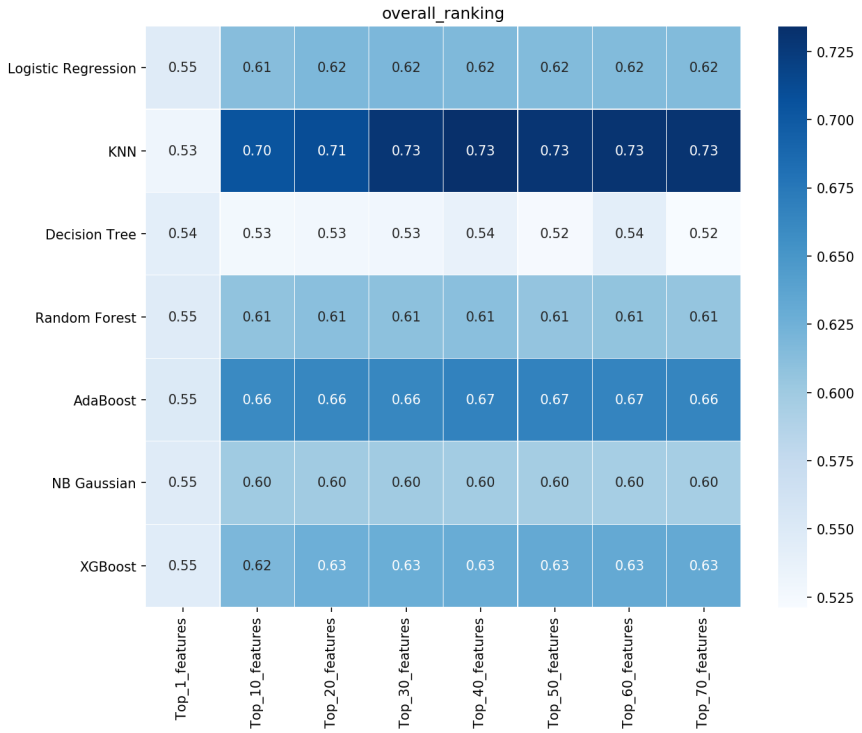
(c) MIM

(d) JMI

(e) MIFS

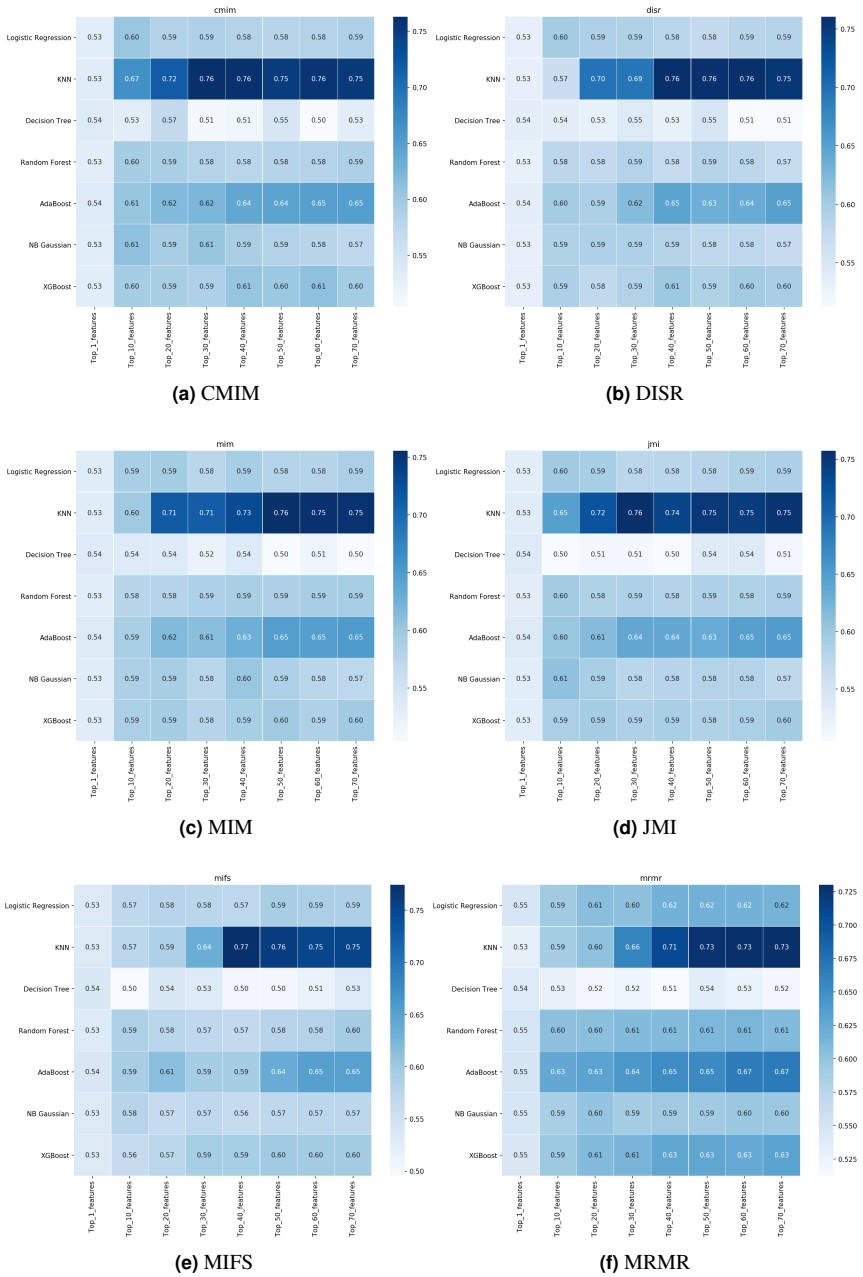
(f) MRMR

Figure 4. Heatmap depicting the predictive performance (AUC average over cross validation) for different number of HLA features(in rows) and different classification method in (columns). It can be observed for all feature selection all classifiers show relatively the same predictive performance in many cases. Mitigated Feature Selection - Cross-validation.



(g) Overall Ranking

Figure 4. Heatmap depicting the predictive performance (AUC average over cross validation) for different number of HLA features(in rows) and different classification method in (columns). It can be observed for all feature selection all classifiers show relatively the same predictive performance in many cases. Mitigated Feature Selection - Cross-validation.

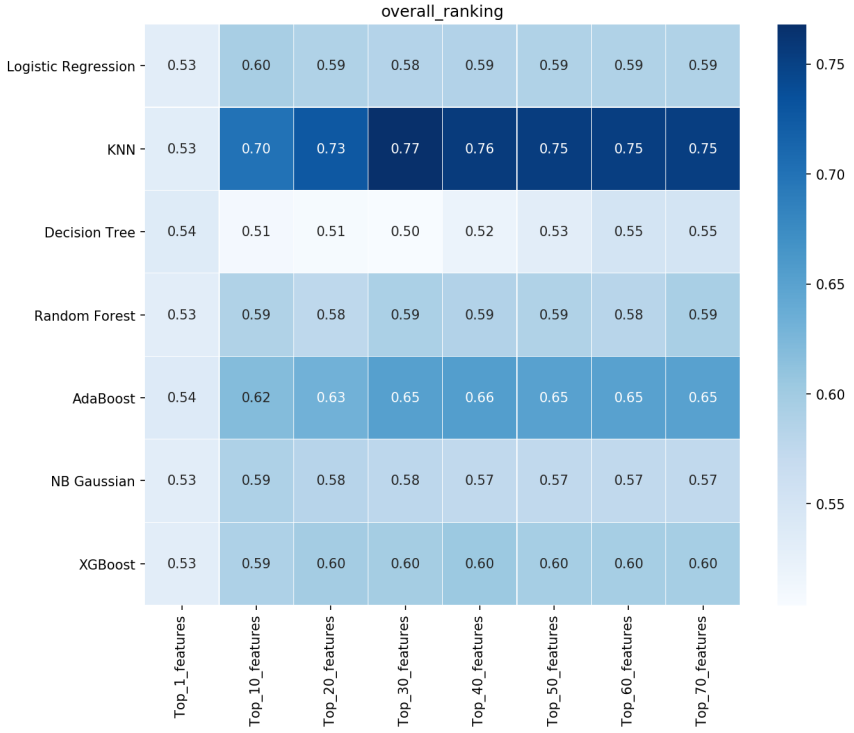


(a) CMIM **(b)** DISR

(c) MIM **(d)** JMI

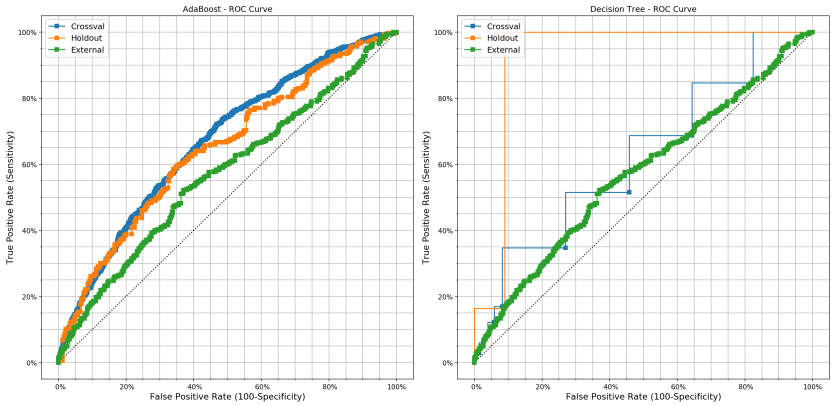
(e) MIFS **(f)** MRMR

Figure 5. Heatmap depicting the predictive performance (AUC hold out) for different number of HLA features(in rows) and different classification method in (columns). It can be observed for all feature selection all classifiers show relatively the same predictive performance in many cases.Mitigated Feature Selection -hold-out.



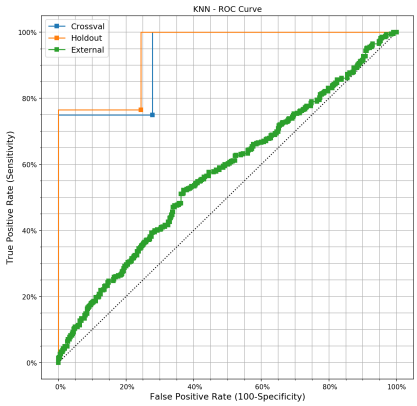
(g) Overall Ranking

Figure 5. Mitigated Feature Selection -hold-out.

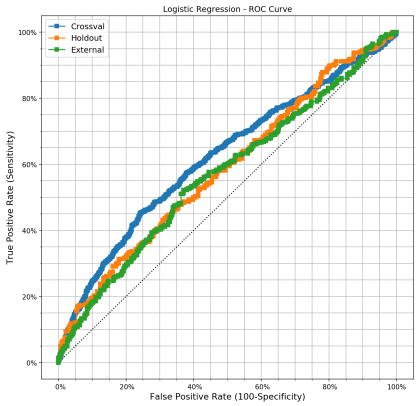


(a) AdaBoost

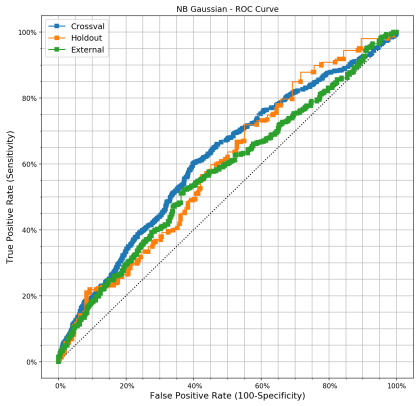
(b) Decision Tree



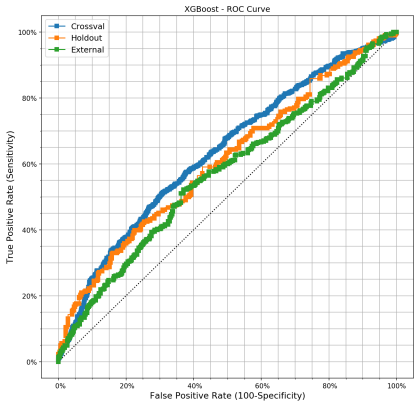
(c) KNN



(d) Logistic Regression

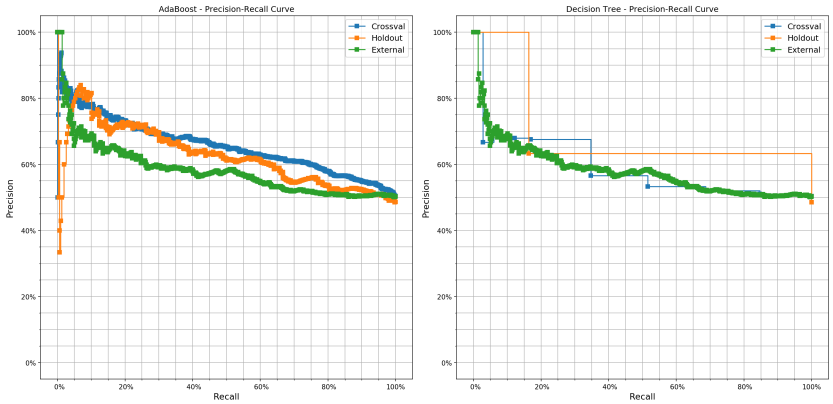


(e) NB Gussian



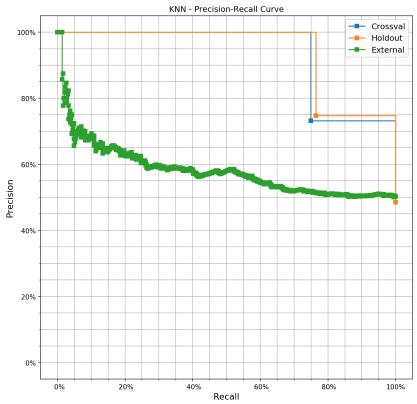
(f) XGBoost

Figure 6. Mitigated - ROC Curve.

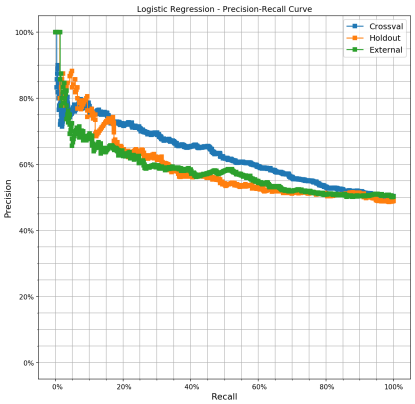


(a) AdaBoost

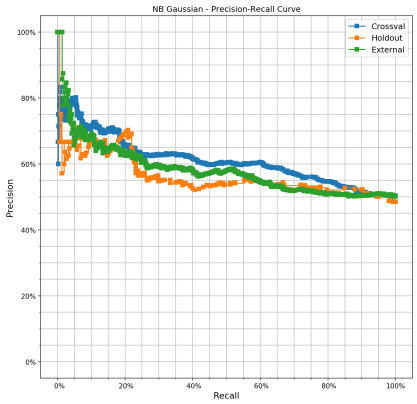
(b) Decision Tree



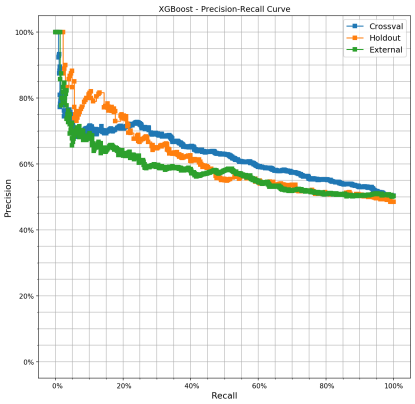
(c) KNN



(d) Logistic Regression

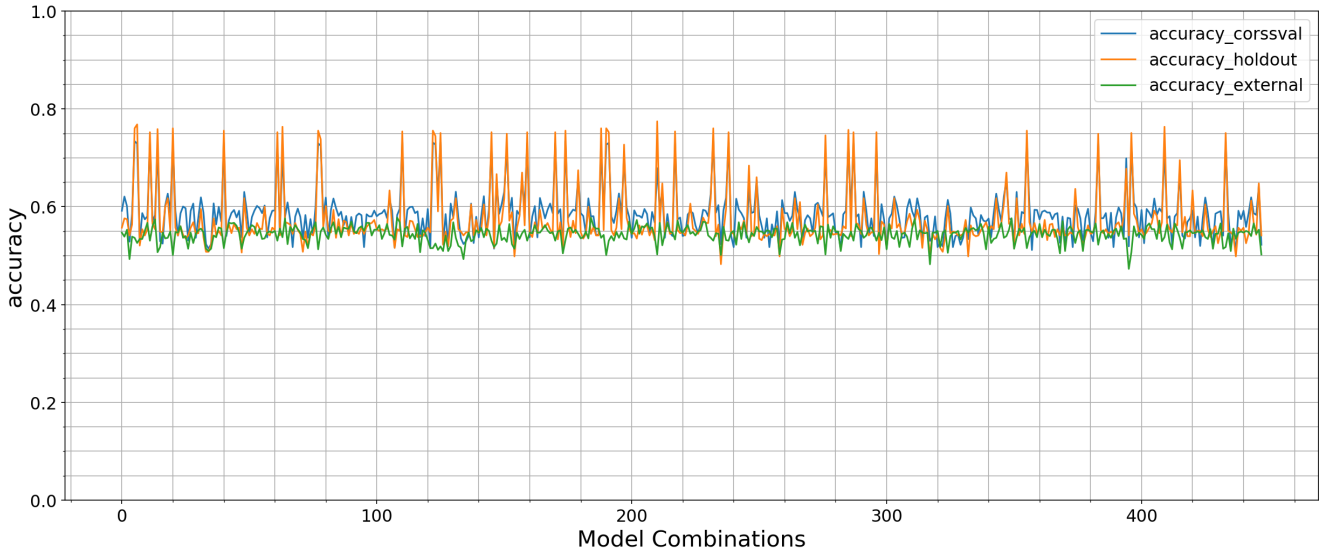


(e) NB Gaussian Forest

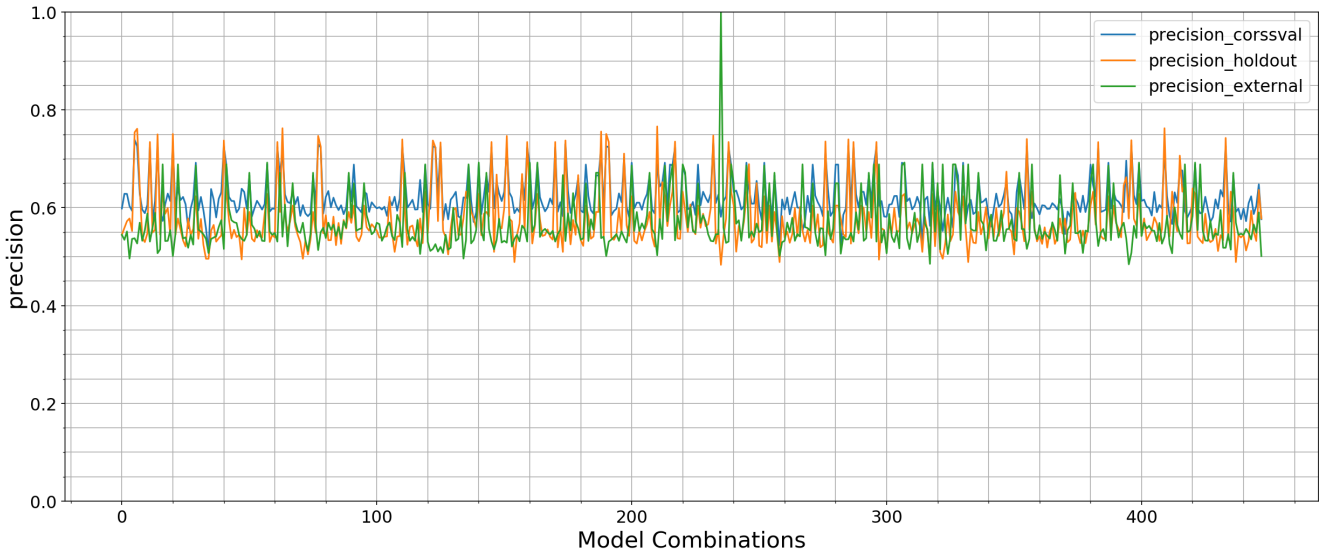


(f) XGBoost

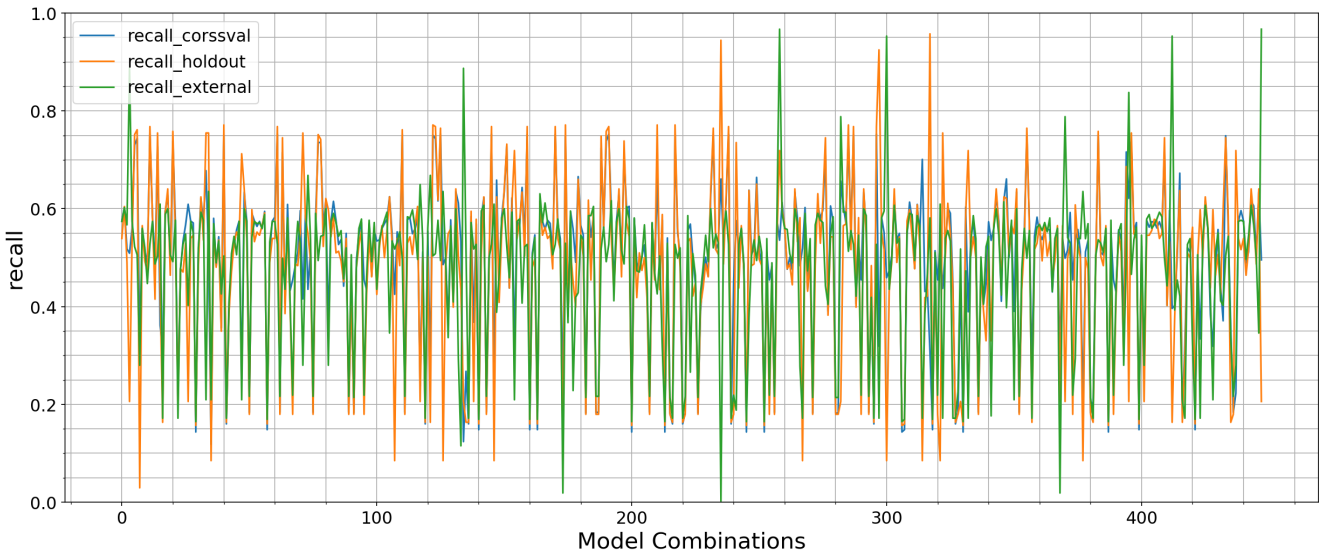
Figure 7. Mitigated - Precision & Recall Plots.



(a) accuracy

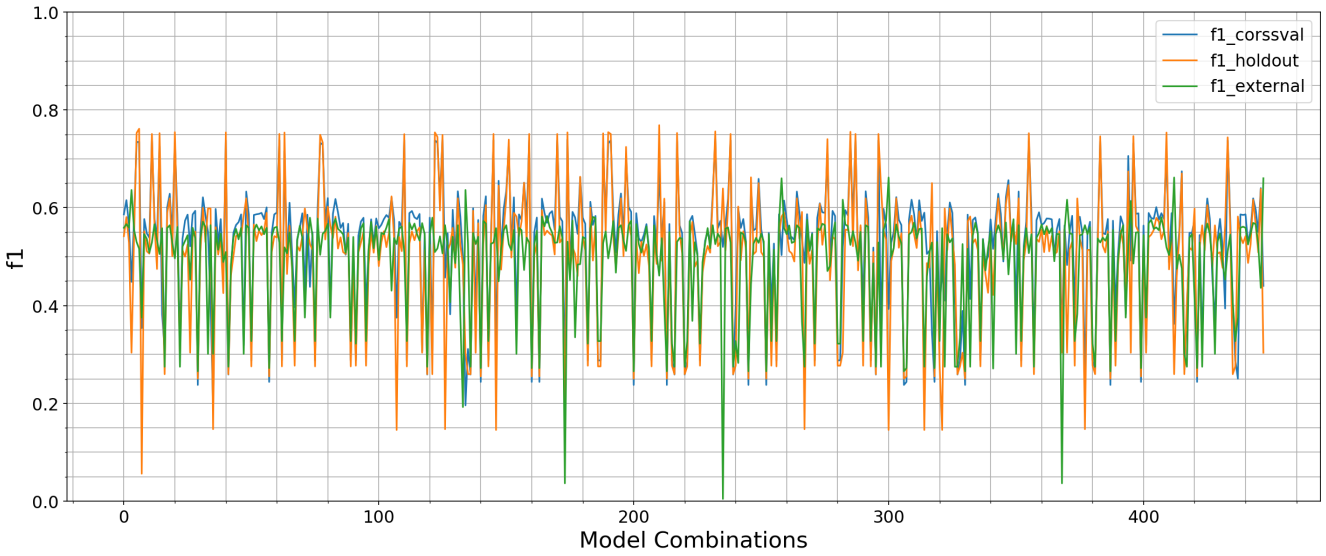


(b) precision



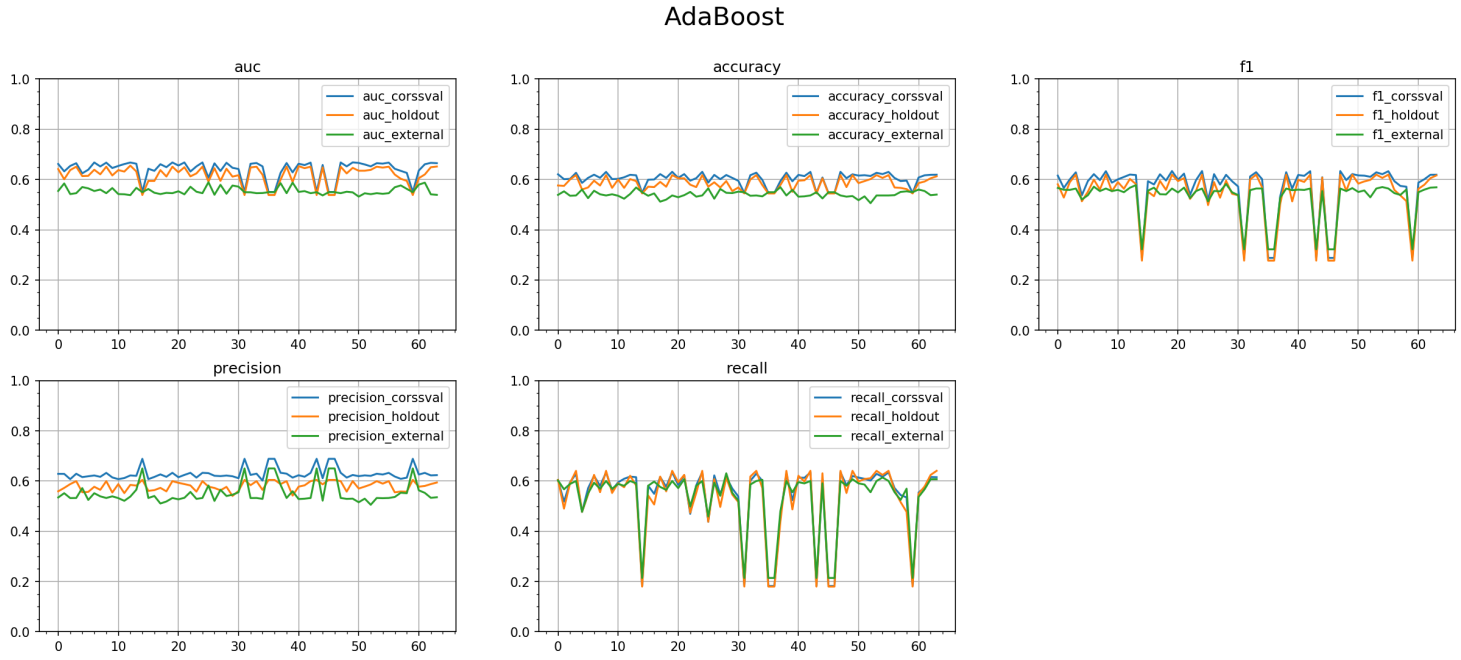
(c) recall

Figure 8. Comparison between accuracy, precision, recall and F1 score for cross validation, hold out and external set for 448 different generated ML models.There are 448 models trained using combination of "number of features", "Feature Selection",

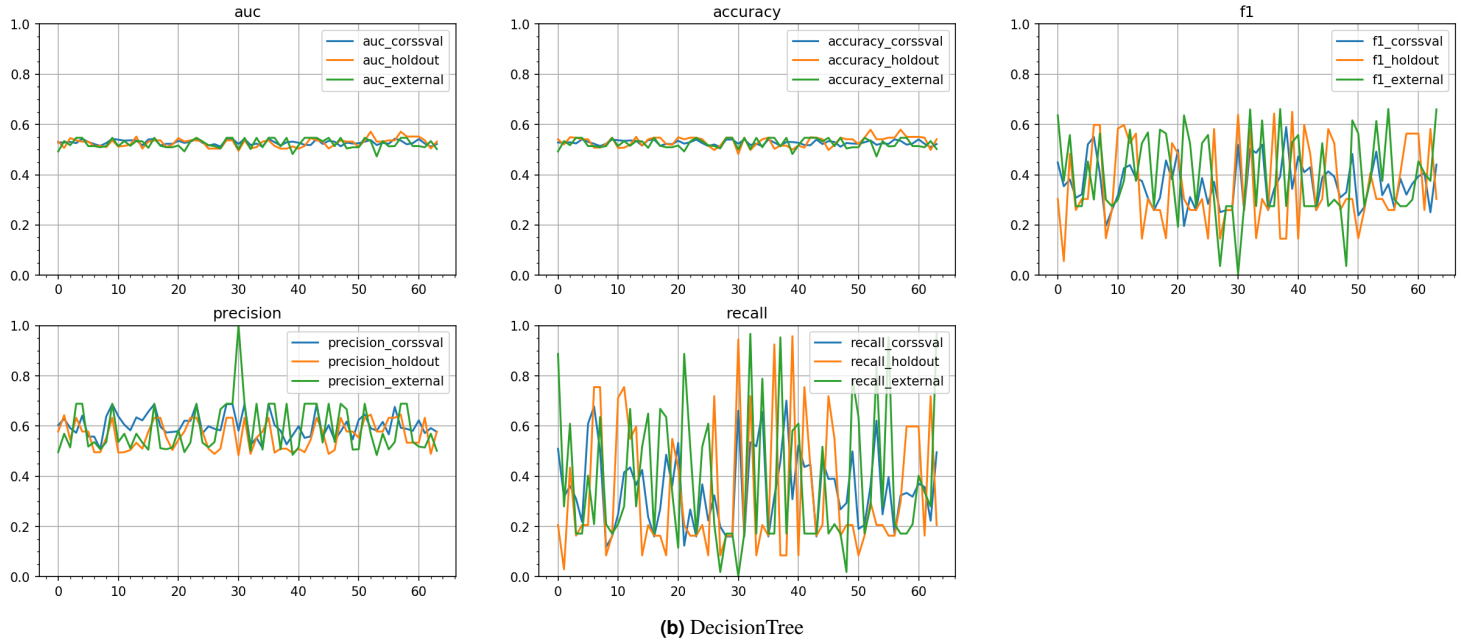


(d) F1-score

Figure 8. Mitigated Feature Selection -hold-out.

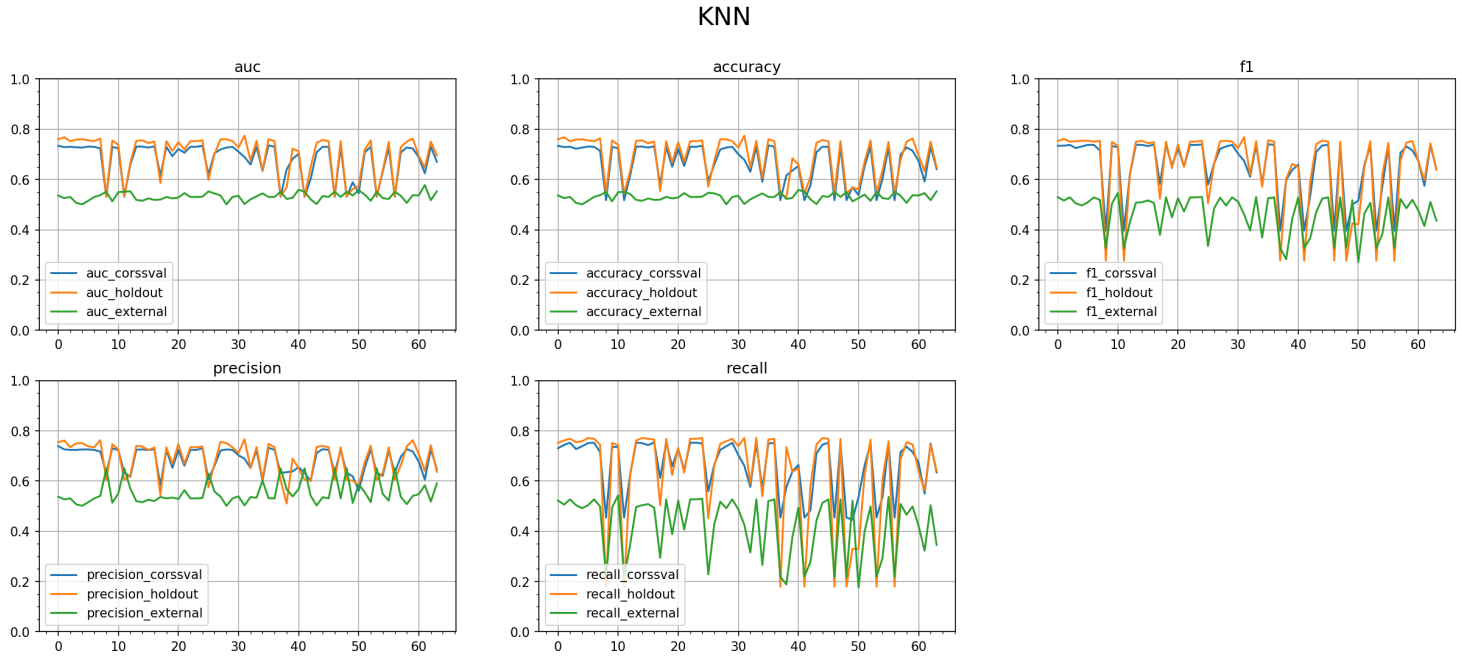


(a) AdaBoost

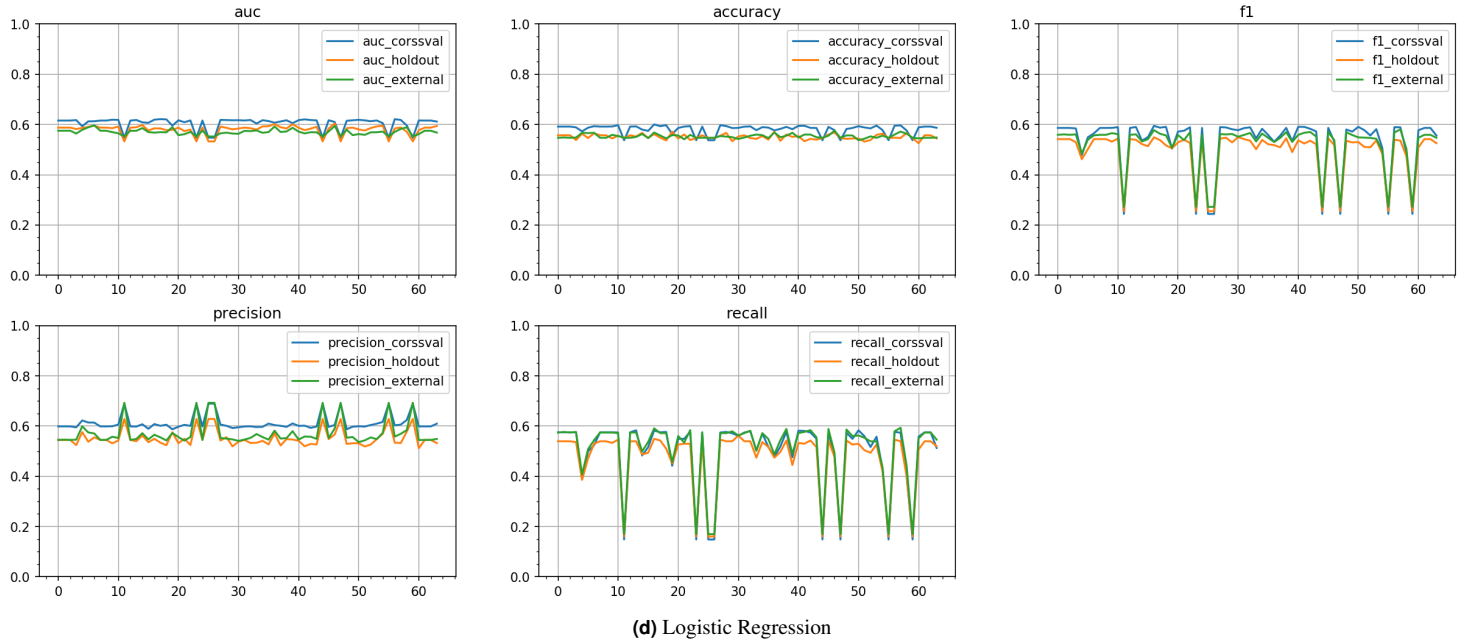


(b) DecisionTree

Figure 9. Accuracy,Precsion, Recall and F1 score curve of each models with internal(hold out and cross validation) and external model performance

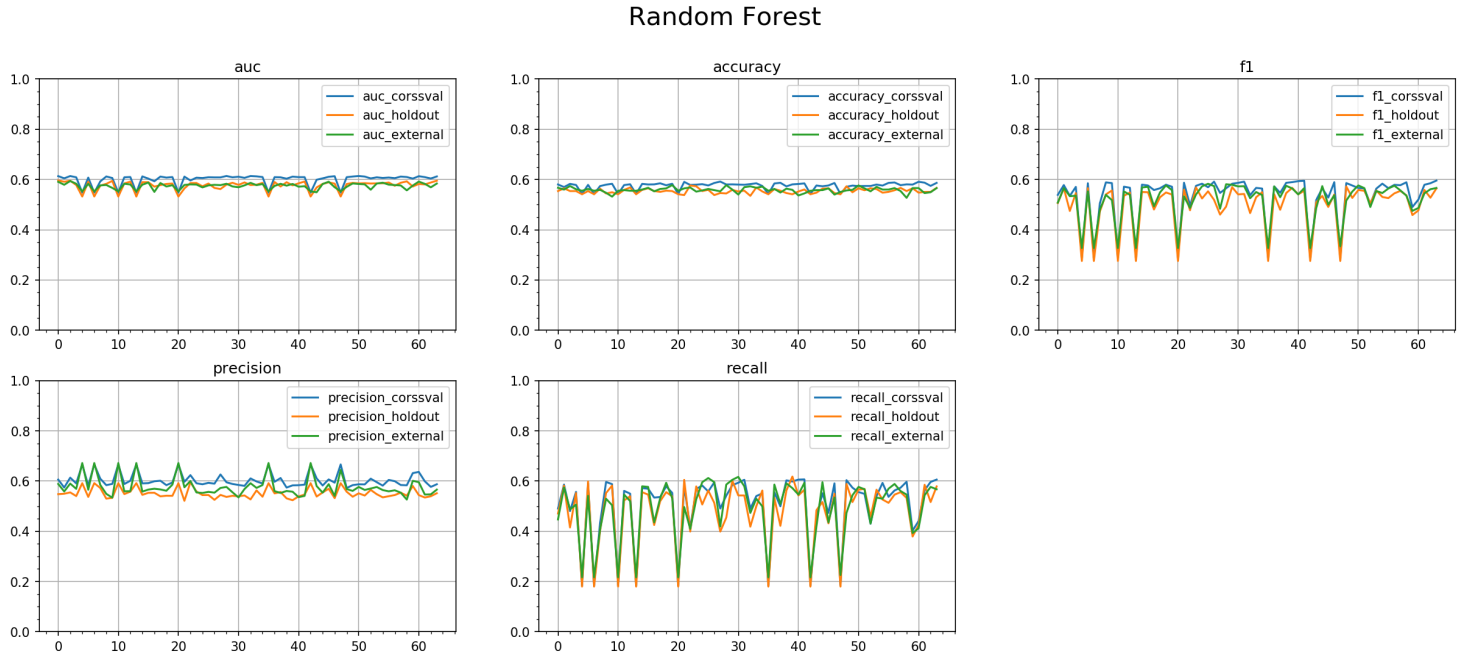


(c) KNNC

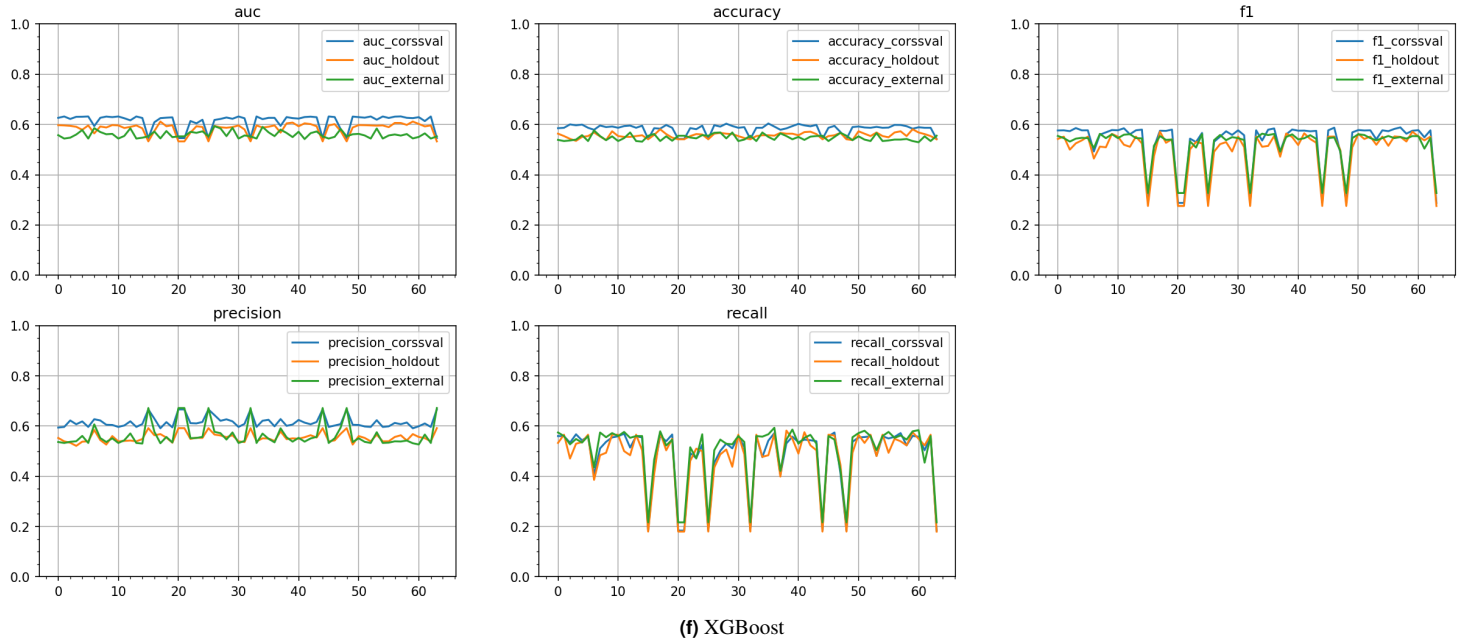


(d) Logistic Regression

Figure 9. Accuracy,Precsion, Recall and F1 score curve of the each models with internal(hold out and cross validation) and external model performance



(e) Random Forest



(f) XGBoost

Figure 9. Accuracy,Precsion, Recall and F1 score curve of the each models with internal(hold out and cross validation) and external model performance