CSC8398: Research Project for Bioinformatics

A Comparison of Various Machine Learning Approaches for Biomarker Identification in Hip/Knee Osteoarthritis

Raman Chahal

160067554

1 Introduction

Osteoarthritis (OA) has been suggested by The Centers of Disease Control and Prevention to be the fourth most frequent cause of hospitalisation (Centers for Disease Control and Prevention [Cdc], 2009). OA is the most widespread form of arthritis which, corresponding to Murphy and Helmick's 2012 study on OA, approximately 10% of adults are diagnosed with (n = 27 million) (Murphy and Helmick, 2012). OA is characterised as a degenerative joint condition which can cause damage to any joint, although high weight-bearing joints are typically most affected. These consist of joints in your hip, knees, and spine (Valdes and Spector, 2011; Ryd *et al*., 2015).

The progression of OA is a long complex process, typically spanning 10 - 20 years, during which time the protective cartilage covering the ends of bones breaks down. The mechanisms by which the cartilage deteriorates includes erosion, synovial inflammation, subchondral bone modifications and osteophyte formation. Once OA has developed to its final stage of progression, as recommended by Chu *et al*, whole joint replacement is the most effective treatment (Chu *et al*., 2014).

Biomarkers are naturally occurring molecules which can identify a biological process. These can be used for the diagnosis and prognosis of OA. Dry biomarkers, such as observing of radiographic and clinical data from structural changes to the joint are most frequently used to diagnose OA. These techniques, however, have been suggested by 2006 study by Wright to have low sensitivity as well as a large precision error, insinuating there is a need to identify more sensitive biomarkers, which can predict the initial pathogenesis of OA before major structural damage can take place (Wright *et al*., 2006).

Several studies have used machine learning methods for identifying biomarkers in OA (Lazzarini *et al*., 2017; Widera, P *et al*, 2020). However, these studies primarily focus on one feature selector and classifier method. Thus, it remains unclear as to which combination of feature selector and classifier, as well as their best settings, performs the best with biomedical data.

1.1 The Machine Learning Approach

To identify novel biomarkers in disease, machine learning (ML) methods can be employed. By using Feature Selection (FS) methods, a model's performance can be iteratively improved by identifying the poorest performing features and removing them. A feature is denoted as a property of a process which has been created from the initial input variables. Many features within the model are known to obscure the computational model. This is known was the curse of dimensionality problem, where the number of features is significantly larger than the sample count. Classifiers which fit over this data perform inadequately. Therefore, the use of feature selection techniques, such as RFE, methodically removes redundant features, whilst retaining a high predictive power for every model. This is achieved by lessening the raw high-dimensional data into distinct manageable features, hence alleviating the curse of dimensionality (Ceotto *et al*., 2011).

Predictive models can be made by using a feature selector coupled with a classifier, such as RFE with a random forest algorithm. Random forests comprise of a group of decision trees, each produced from using varying subsets of training and test data. Each tree votes for which it thinks the outcome is, from which the majority vote of the set of trees denotes the prediction.

Random forests have been shown to perform without issue with high-dimensional data and are able to identify prognosticators of the desired result (Breiman L *et al*., 2001). One frequent problem, which negatively impacts RF's capacity to distinguish the best predictors, is the presence of highly correlated predictors – resulting from high-dimensional data. Correlated predictors decrease the estimated importance scores for correlated variables (Gregorutti B *et al*., 2017). As such, RF will be used in conjunction with RFE to perform feature selection.

Preventing loss of information remains one of the biggest challenges for FS methods. RFE will be used to avoid removing features which contain crucial information about the analysed data. By iteratively training a model, ranking it's features and removing the lowest ranking features RFE can solve this issue (Guyon I *et al*., 2002). Studies by Jiang H, Svetnik V and Gregorutti B have demonstrated the use of RF-RFE to be positive when using data with correlated features (Gregorutti B *et al*., 2017; Jiang H *et al*., 2004; Svetnik V *et al*., 2004).

In the present study, a ML pipeline (a series of ML computational tasks stringed together) is used to analyse a cohort study of participants with high prognosis of developing OA, aiming to detect biomarkers which can aid in the identification of hip/knee OA. The pipeline consists of creating a valid pre-processing strategy and combining several feature selectors and classifiers within a K-fold cross-validation loop. The model will use feature selectors RFE and Relief with the classifiers RF, Logistic Regression (LR), Support Vector Classifier (SVC) and Xgboost (Xgb).

Comparing the different machine learning models' performance across several classifiers, the top features each model has selected can be evaluated. By exploring relevant literature, we can better understand what each model has learned, and how to improve these pipelines for further use.

Finally, although the ML methods used are concentrated on analysing data corresponding to hip/knee OA, the proposed pipeline for identifying biomarkers is sufficiently generic enough to be used in future work, for an array of biomedical data. The methodology has the capability to identify novel biomarkers for many different diseases and can help to better understand their mechanisms.

1.2 Aims and Objectives

The objectives of this study are to first outline a data pre-processing strategy, which tackles many of the problems machine learning methods have with biomedical data (missing values, data of mixed types). Secondary, to compare each model's performance and feature selection across classifiers. By analysing these metrics, behavior of the models can be interpreted to evaluate their clinical applicability. Finally, this study will also review different methods on evaluating model performance, as well as the limitations of this experiment design, and how it fits with relevant literature. Suggestions about further work will also be discussed. To reiterate, the overall objectives of this study are:

- Data exploration and remove variables which can be misrepresented
- Outline Pre-processing strategy
- Build models
- Compare each models' performance and features they selected
- Review these methods and compare against literary findings
- Further work

The general hypothesis for this study is that the tree-based classifiers, RF and Xgboost, are to perform better than the linear models, as the data each model is based on has many variables, which are not so clearly linear. Thus, the linear based classifiers (LR, SVC) are hypothesised to perform worse.

2 Methods

2.1 Dataset and Individuals

The data used for this study came from the Cohort Hip & Cohort Knee (CHECK) dataset, provided by our clinical collaborators at the University Medical Centre Utrecht (Janet Wesseling *et al*., 2016). The study was a population-based randomised cohort study of 1106 individuals demonstrating early symptoms of OA of the hip and/or knee. CHECK combines data from The Western Ontario and McMaster Universities Arthritis Index (WOMAC), a pain questionnaire evaluating hip and knee OA. The questionnaire aims to cover active and passive hip or knee pain experienced. The criteria involves stiffness: after waking up and later in the day; physical function: walking up and downstairs; and heavy/light domestic tasks (American College of Rheumatology WOMAC).

The general methodology for the proposed project consisted of 1) Data exploration; 2) Pre-processing and cleaning data; 3) Defining evaluation pipelines; 4) Implementing these pipelines in experiments; and 5) Data interpretation.

2.2 Data Exploration

An initial exploration of the data was performed using Pandas Profiling (McKinney, W *et al*., 2011), to visualise missing values within the dataset, and the variables types, be it categorical, Boolean, or numerical.

Furthermore, by making a correlation matrix, highly correlated values were identified. Although some variables showed correlation, it was suggested to keep them in the dataset as the values may represent a causal relationship when it comes to biomarker identification, especially if such biomarkers are important in noting progression of the disease.

2.3 Defining Evaluation Pipelines

K-Fold

As elucidated by Lazzarini *et al* (Lazzarini *et al*., 2017), identifying biomarkers for OA using ML methods involves four main steps: 1) Train the model on the training set; 2) calculate the ranking based on the feature importance; 3) Eliminate the features of lowest ranking; and 4) repeating this process until the top n features are selected.

Typical feature selection methods, such as RFE, are sensitive to minor disturbances between training and testing sets of the data. Therefore, by adopting K-fold cross validation – a leave-one-out strategy – which aims to improve the stability and robustness of each model. K-fold cross validation avoids over-fitting of the data, hence increasing the number of estimates and increasing the accuracy of the prediction model. The original training sample is stochastically split into "K" subsamples, where one partition is held as the validation data set to be used for testing the model's performance, while K-1 subsamples are retained as the training data. Cross-validation repeats "K" number of times so that each of the "K" validation subsamples are used just one time. A single estimation is achieved after averaging the K results from each fold. K-fold CV is advantageous to methods using random sub-sampling because it ensures all observations are used for training and validation a single time. This was achieved by using K-Fold package from Sklearn (Pedregosa, F. *et al*., 2011).

Data pre-processing

The CHECK data presented a difficult task from a ML perspective as there were a lot of missing values present in the dataset, along with the data being heterogeneous – having both discrete and continuous variables. First, attributes which could be misinterpreted by the model were removed, these included whether the patient was married or not, patient ID's and dates.

To tackle the problem of missing values, the discrete variables were imputed by using the most frequent value for each training and test set for each of the differing splits in K-Fold. The continuous values were imputed by the mean values using the same method. Normalisation of the continuous values was performed using Standard-Scalar from Sklearn (Pedregosa, F. *et al*., 2011). After missing values were imputed for both nominal and continuous variables, the discrete variables were one-hot encoded. This involves the removal and replacement of the variables with a new unique binary variable for every distinctive integer value. Parameter tuning is performed based on the nested cross-validation on the training data only, so that the test data is only used at the final steps to evaluate the final generated models.

This imputation and one-hot encoding method was achieved using Sklearn's Column Transformer (Pedregosa, F. *et al*., 2011) and implementing that into a pipeline, along with the respective feature selector and classifier. Using a pipeline ensured that the pre-processing steps occur within the cross-validation loop, so that the median and mean values can be imputed on the training set only, without access to the test data.

Models and Parameters

In total 8-predicitive models were created using feature selectors RFE and Relief, along with classifiers RF, LR, SVC and Xgb:

1) RFE-RF, 2) Relief-RF, 3) RFE-LR, 4) Relief-LR, 5) RFE-SVC, 6) Relief-SVC, 7) RFE-Xgb, 8) Relief-Xgb

Each model involved the same pre-processing and train/test split methods (K-Fold). The number of features selected by RFE or Relief was kept the same for each model (10, 20, 30, 40, 50). Classifier parameters where kept broad and tuned. For the tree-based classifiers (RF and Xgb) max depth was set to [2, 3, 4, 5] and number of estimations set to [100, 200, 500, 1000]. It was important to keep the max depth to a maximum of 5 as to remove the danger of overfitting. For LR and SVC the "C" parameter was set to [0.001, 0.01, 0.1, 1, 10, 100, 1000]. Figure 1 Depicts the effect of overfitting on a models-performance.

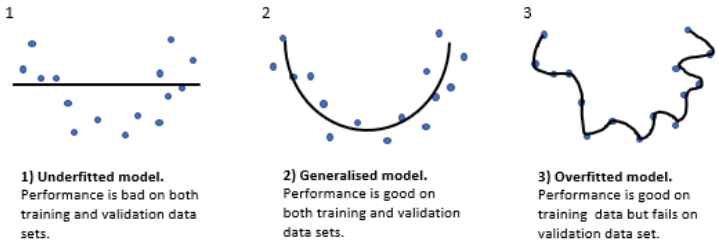


Figure 1. Visualisation of the effect underfitting/overfitting data on model performance.

Class imbalance

An interesting feature of the dataset provided by CHECK is that there is an imbalanced class distribution, such that there are more cases of "not OA" than "OA" (638 vs 468). Performing machine learning with imbalanced data sets has been explored numerous times and shown to make building robust models much harder (Al-Stouhi S *et al*., 2016). Each classifier used in this study was able to be made "cost-sensitive", so that class weights could be incorporated to punish misclassification by each classifier. Especially for random forest, a 2004 study by Chen *et al* (the creators of the classifier), demonstrated the advantages of using weighted variants with imbalanced datasets (Chen C *et al*., 2004).

Class weights were calculated using the in-build class weight calculator for the Sklearn classifiers. In this case classifiers RF, LR and SVC had their "class_weight" parameter set to "balanced". This calculates class weight based on the following equation (1):

$$w_j = \frac{n}{kn_j}$$

Equation 1: wj corresponds to the weight class j; n refers to the number of observations, where nj is the number of observations in class j; k is the total number of cases.

For the experiments using Xgboost, class weight was calculated using "compute_class_weight" Sklearn utilities. This method also uses equation 1 to compute class weight as "balanced" and was used since Xgboost does not have an in-built method to calculate class weight.

A visual representation of the full analytical pipeline is presented in Figure 2:

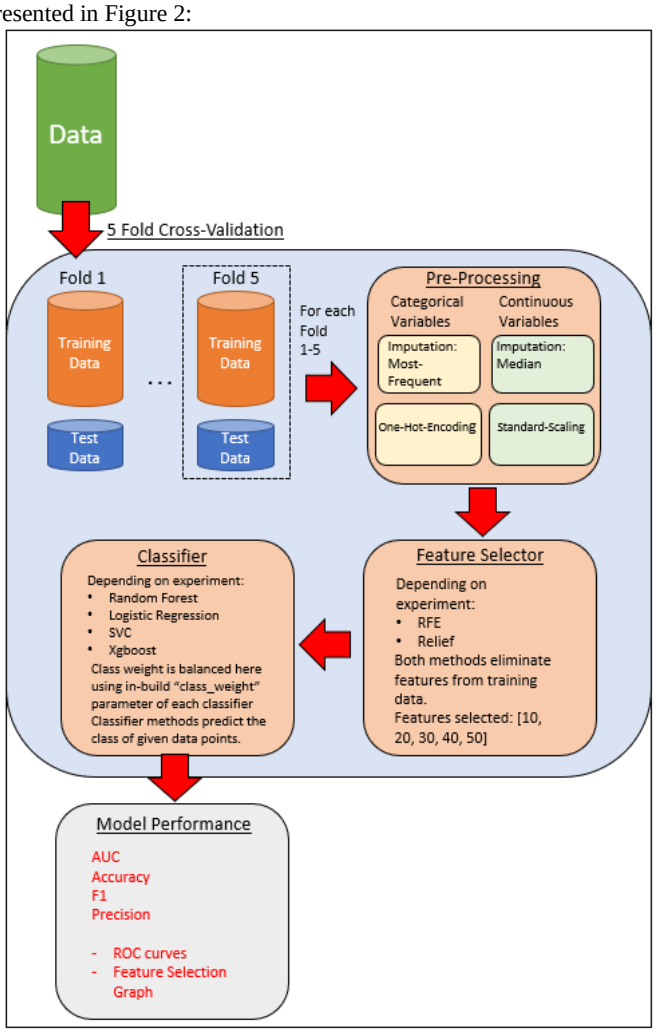


Figure 2. Analytical pipeline representation employed to identify the best predicative models. The dataset is first split into training and test sets by 5-Fold cross-validation. For each of the folds the data is pre-processed such that categorical variables are one-hot encoded after their missing values are imputed by the most-frequent occurring value in the respective column. Continuous variables have their missing values imputed by the median value of respective column before standard-scaling. Next, depending on experiment, feature selector is fitted, before the respective classifier is fitted. Model performance is assessed through generating scores of AUC, accuracy, F1 and precision. ROC plot and feature selection plots are generated using Matplotlib.

2.5 Experiments and Data Interpretation

The optimal model was found by examining the area under the curve (AUC) for Receiver Operating Characteristics (ROC). As shown by a study by Hajian-Tilaki, ROC curves are useful tools for diagnostic accuracy for biomarkers (Hajian-Tilaki, 2013). In this case, AUC provides an estimation of the accuracy for the diagnostic test of discrimination of incidence of OA in individuals. ROC curves were generated by plotting the false positive rate against the true positive rate, generated from Sklearn metrics packages. To overcome the issue of differences in true and false positive values between models, these were interpolated so a meaningful average could be obtained (Fawcett T. *et al*., 2005). By comparing the average AUC curves for each developed model, the optimal model can be discovered.

2.6 Additional model-performance metrics

To get a more rounded representation of each models' performance, the metrics: accuracy, F1 and precision were also gathered to be compared against each-other to determine the best performing model.

Model accuracy is defined as number of correct predictions divided by the total number of predictions made by the model. When working with an imbalanced dataset such as this, accuracy alone will not give a full picture on model performance when there is a substantial disparity between the amount of positive and negative labels.

Precision for a class is the ratio of true positives (number of correctly labelled predictions) to true + false positives (number of correctly labelled predictions + number of incorrectly labelled predictions). If a model has a precision score of 0.7 for example, it means when the model predicts cases of OA, its correct 70% of the time.

The F1 score is another measure for a models' accuracy. It computes a "harmonic mean" of precision and recall – where recall is the ratio of the number of correctly labelled positive results to all samples which should have been labelled as positive.

Although ROC curves are preferable, due to them being insensitive to changes in class distribution (Metz CE *et al*., 1978), using all the performance metrics in conjunction will give a clearer image on each model's performance.

The top features selected from each model were visualised in bar plots, generated from extracting the feature importance attributes after each model was fitted and using Matplotlib (Hunter, J *et al*., 2007). The most frequent – highest scoring features for the best models were noted and displayed in a table.

2.7 Software and packages used:

The ML pipeline was developed in PyCharm using Python 3.6 (Reference). All experiments were performed using libraries from scikit-learn (Pedregosa, F. *et al*., 2011), where the ML algorithms were derived. For the code involving data pre-processing and data analysis, the packages NumPy (Oliphant, T *et al*., 2007), pandas (McKinney, W., 2011), seaborn (Waskom, M., 2013) and Matplotlib (Hunter, J *et al*., 2007) were used.

3 Results

3.1 Data Exploration Findings

For all 1106 participants in the CHECK study, 994 individuals completed the full questionnaire – having follow-up data for 8-10 years. Baseline attributes of the 994 individuals are demonstrated in Table 1.

Variable	Mean value or Percentage	Missing Values (%)
Age (Years)	55.88	0
BMI (kg/m^2)	26.46	2
Menopausal Status	76.60%	13.7
Western Ethnicity	97.01%	0.2
Mild Symptoms (Knee Pain)	58.14%	3.7

Table 1. Baseline Characteristics of 994 individuals, along with their representative missing values, are displayed

The Pandas Profiling report showed a vast number of columns to have missing data, ranging from 2.0% missing values to 69.0% missing values. T5_wmtots, T5_wmpyns were displayed as having highly correlated values in Figure 3 along with T8_wmtots, T8_wmpyns; T10_wmtots, T10_wmpyns; T10_BMI, T8_BMI; T10_wmfuns, T10_wmtots; T0_Menopauze_01, T0_SEXE.

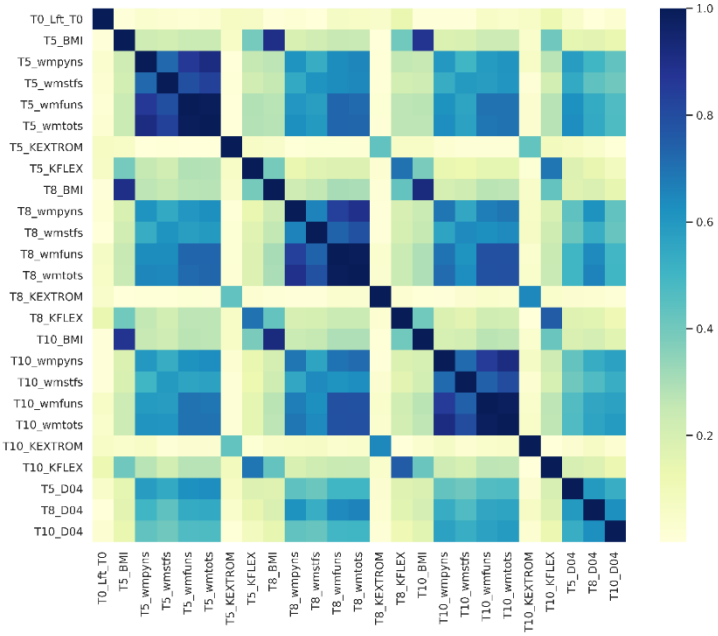


Figure 3. Correlation Matrix of the initial dataset before any pre-processing steps. The higher the correlation between two variables is shown by an increasing intensity of blue. Variables: T5_wmtots, T5_wmpyns; T8_wmtots, T8_wmpyns; T10_wmtots, T10_wmpyns; T10_BMI, T8_BMI; T10_wmfuns, T10_wmtots; T0_Menopauze_01, T0_SEXE are shown to have the highest correlation.

3.2 Model Interpretation

ROC

ROC curves for the 8-predicitive models are shown in Figure 4. Here, the models sorted by best AUC score are:

1. RFE-RF (AUC = 0.87)
2. Relief-RF (AUC =0.87)
3. Relief-Xgboost (AUC = 0.87)
4. Relief-LR (AUC = 0.86)
5. Relief-SVC (AUC = 0.86)
6. RFE-Xgboost (AUC = 0.86)
7. Relief-LR (AUC = 0.84)
8. RFE-SVC (AUC = 0.83)

Additional Metrics

For each model, the metrics: Accuracy, Precision and F1 score were obtained for each fold in the cross-validation loop, with their averages, along with AUC scores. These are presented in Figure 5. We can see, that by looking at the overall score (the mean of all scores), that the model Relief-Xgboost is the highest scoring. This model is followed by RFE-RF and Relief-RF as second and third highest scoring models.

Best Settings for each model

The settings used in the parameter grid for each model to achieve the highest scoring folds are described in Table 2. Selecting 30-50 features by the corresponding feature selector (RFE/Relief) resulted in best performance. For tree-based models (RFE-RF, Relief-RF, RFE-Xgboost and Relief-Xgboost) max depth was tuned to 5, with the number of trees ranging from 100-1000.

Model	Best Settings in Parameter Grid	AUC scores achieved
RFE-RF	N features to select -50, Max depth – 5, N estimators - 500	0.9
Relief-RF	N features to select -50, Max depth – 5, N estimators - 1000	0.89
RFE-LR	N features to select -40, C - 0.01	0.88
Relief-LR	N features to select -40, C - 0.01	0.89
RFE-SVC	N features to select -50, C - 0.01, Gamma - 1	0.86
Relief-SVC	N features to select -30, C - 0.01, Gamma - 1	0.89
RFE-Xgboost	N features to select -50, Max depth – 5, N estimators - 500	0.91
Relief-Xgboost	N features to select -50, Max depth – 5, N estimators - 100	0.9

Table 2. Best settings for each model are displayed, along with the respective AUC scores achieved using these settings.

Features Selected

Top features selected by the highest scoring models of each class is displayed in Figure 6. The common recurring features selected between each model includes: 10-year follow-up WOMAC standardized total sum score , 10-year follow-up WOMAC standardized physical functioning scale, 8-year follow-up WOMAC standardized total sum score, 8-year follow-up WOMAC standardized physical functioning scale, 5-year follow-up WOMAC standardized total sum score, Baseline age, 5-year follow-up BMI, 5-year follow-up knee extension active ROM (range of motion), 5-year follow-up WOMAC standardized total sum score, 10-year follow-up physical examination of pain present on knee extension, , 5-year follow-up WOMAC standardized physical functioning scale, 8-year follow-up BMI and 10-year follow-up BMI.

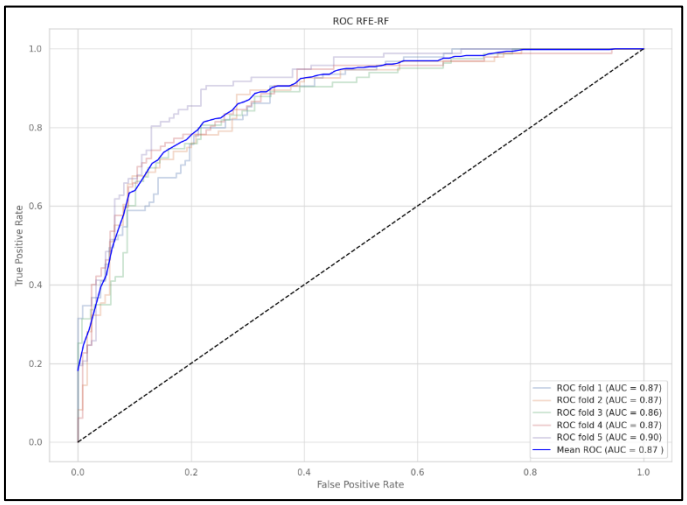
Without taking year into account, by abstraction we can see that the most reoccurring features are:
WOMAC standardized physical functioning scale (n = 8), WOMAC standardized total sum score (n = 7), BMI (n = 6), Physical examination of pain present on knee extension (n = 3), Baseline age (n = 2), knee extension active ROM (range of motion) (n = 2), WOMAC standardized pain scale (n = 2), WOMAC difficulty getting up from chair (n = 2) .

Table 3 shows a summary of the best features selected between the 8-predicitive models for OA outcome.

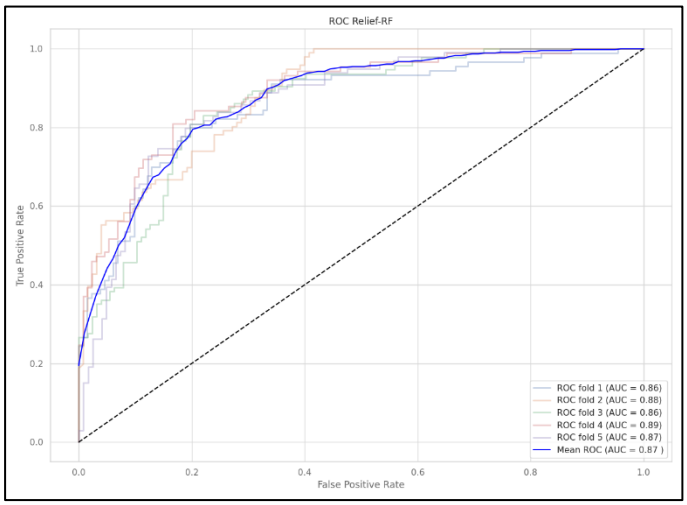
OA Measure	Biomarker
Knee Pain	WOMAC standardized physical functioning scale
	WOMAC standardized total sum score
	Knee extension active ROM
	WOMAC difficulty getting up from chair
	Physical examination of pain present on knee extension
Baseline Characteristics	Body Mass Index
	Age

Table 3. Summary of top features selected between each generated model

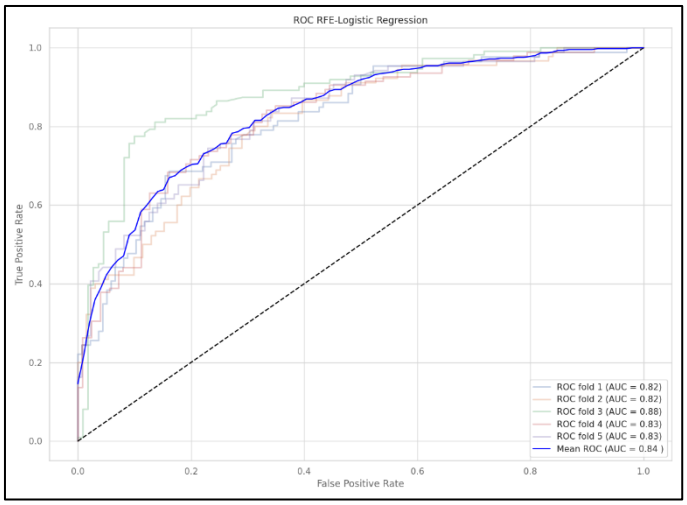
1


ROC RFE-RF
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.87)
ROC fold 2 (AUC = 0.87)
ROC fold 3 (AUC = 0.86)
ROC fold 4 (AUC = 0.87)
ROC fold 5 (AUC = 0.90)
Mean ROC (AUC = 0.87)

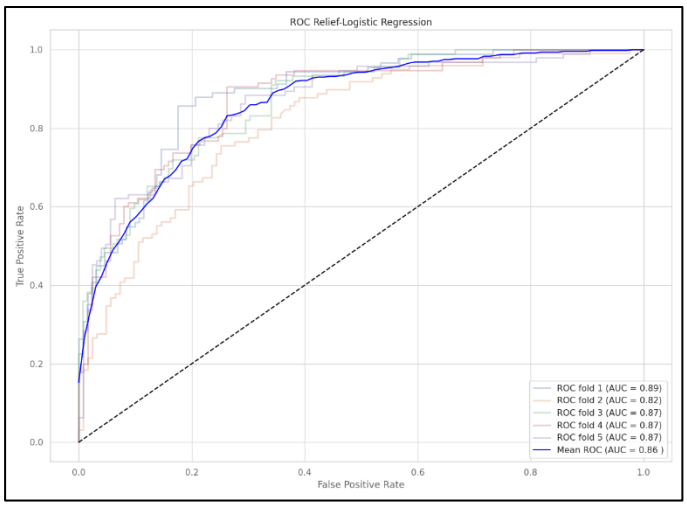
2


ROC Relief-RF
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.86)
ROC fold 2 (AUC = 0.88)
ROC fold 3 (AUC = 0.86)
ROC fold 4 (AUC = 0.89)
ROC fold 5 (AUC = 0.87)
Mean ROC (AUC = 0.87)

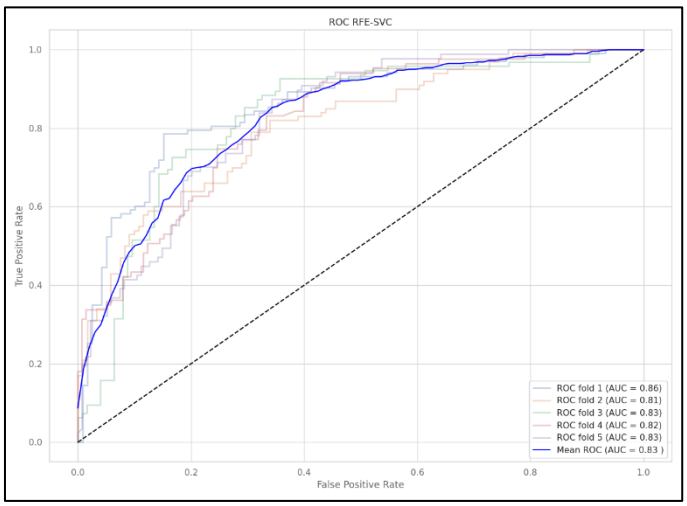
3


ROC RFE-Logistic Regression
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.82)
ROC fold 2 (AUC = 0.82)
ROC fold 3 (AUC = 0.88)
ROC fold 4 (AUC = 0.83)
ROC fold 5 (AUC = 0.83)
Mean ROC (AUC = 0.84)

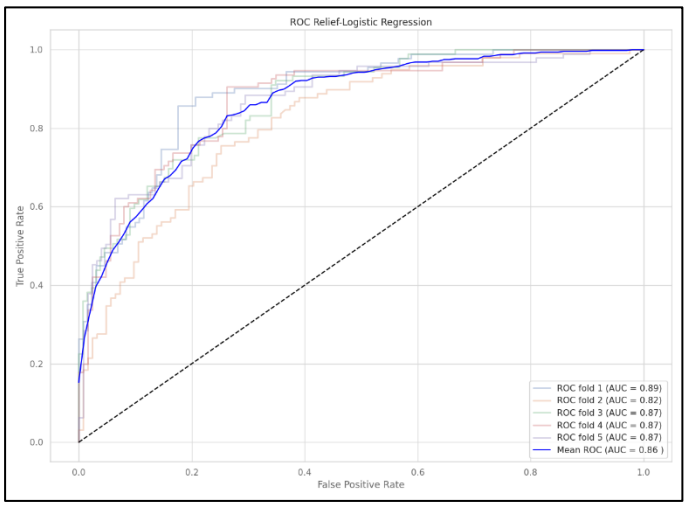
4


ROC Relief-Logistic Regression
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.89)
ROC fold 2 (AUC = 0.82)
ROC fold 3 (AUC = 0.87)
ROC fold 4 (AUC = 0.87)
ROC fold 5 (AUC = 0.87)
Mean ROC (AUC = 0.86)

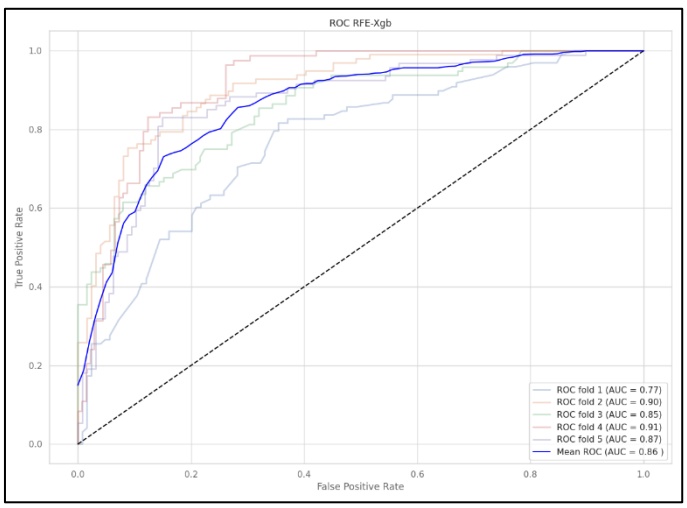
5


ROC RFE-SVC
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.86)
ROC fold 2 (AUC = 0.81)
ROC fold 3 (AUC = 0.83)
ROC fold 4 (AUC = 0.82)
ROC fold 5 (AUC = 0.83)
Mean ROC (AUC = 0.83)

6


ROC Relief-Logistic Regression
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.89)
ROC fold 2 (AUC = 0.82)
ROC fold 3 (AUC = 0.87)
ROC fold 4 (AUC = 0.87)
ROC fold 5 (AUC = 0.87)
Mean ROC (AUC = 0.86)

7


ROC RFE-Xgb
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.77)
ROC fold 2 (AUC = 0.90)
ROC fold 3 (AUC = 0.85)
ROC fold 4 (AUC = 0.91)
ROC fold 5 (AUC = 0.87)
Mean ROC (AUC = 0.86)

8

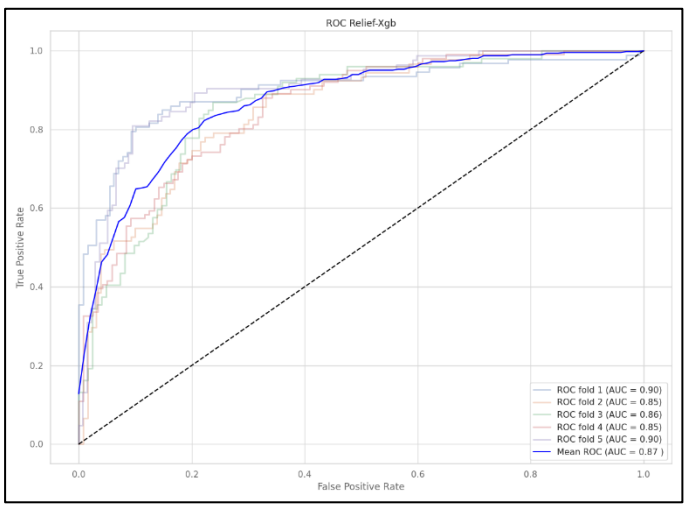

ROC Relief-Xgb
True Positive Rate
False Positive Rate
ROC fold 1 (AUC = 0.90)
ROC fold 2 (AUC = 0.85)
ROC fold 3 (AUC = 0.86)
ROC fold 4 (AUC = 0.85)
ROC fold 5 (AUC = 0.90)
Mean ROC (AUC = 0.87)

Figure 4. Receiver operating characteristic (ROC) curve for the top features selected by each model. Models 1-8 refer to: 1) RFE-Random Forest; 2) Relief-Random Forest; 3) RFE-Logistic Regression; 4) Relief-Logistic Regression; 5) RFE-Support Vector Classifier; 6) Relief-Support Vector Classifier; 7) RFE-Xgboost; 8) Relief-Xgboost.
The individual AUC scores for each of the 5 folds are shown in each of the plot legends, where the average AUC (3dp) between iterations are as follows:

1) 0.87; 2) 0.87; 3) 0.84; 4) 0.86; 5) 0.83; 6) 0.86; 7) 0.86; 8) 0.87

1

RFE-RF				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.783174472	0.783783784	0.755102	0.732673267
2	0.775375	0.78280543	0.741935	0.766666667
3	0.781691985	0.787330317	0.728324	0.7
4	0.790946126	0.791855204	0.767677	0.752475248
5	0.833970735	0.832579186	0.81592	0.788461538
	0.793031664	**0.795670784**	**0.761792**	**0.748055344**

2

Relief-RF				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.795959596	0.797297297	0.759358289	0.731958763
2	0.761791667	0.764705882	0.731958763	0.724489796
3	0.804657397	0.800904977	0.78	0.735849057
4	0.81541539	0.814479638	0.780748663	0.744897959
5	0.783573439	0.787330317	0.758974359	0.770833333
	0.792279498	**0.792943622**	**0.762208015**	**0.741605782**

3

RFE-LR				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.732729138	0.743243243	0.674285714	0.662921348
2	0.720229008	0.7239819	0.673796791	0.649484536
3	0.832882883	0.832579186	0.821256039	0.885416667
4	0.753341688	0.755656109	0.721649485	0.707070707
5	0.748277347	0.746606335	0.698924731	0.65
	0.757492013	**0.760413355**	**0.717982552**	**0.710978652**

4

Relief-LR				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.829334787	0.824324324	0.8	0.75
2	0.732246557	0.728506787	0.714285714	0.669642857
3	0.762895812	0.769230769	0.718232044	0.706521739
4	0.777151211	0.78280543	0.744680851	0.752688172
5	0.778362573	0.778280543	0.751269036	0.725490196
	0.775998188	**0.776629571**	**0.745693529**	**0.720868593**

5

RFE-SVC				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.801419597	0.801801802	0.788461538	0.780952381
2	0.712561983	0.719457014	0.673684211	0.711111111
3	0.771888053	0.778280543	0.737967914	0.75
4	0.697310983	0.714932127	0.622754491	0.619047619
5	0.744081318	0.755656109	0.689655172	0.689655172
	0.745452387	**0.754025519**	**0.702504665**	**0.710153257**

6

Relief-SVC				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.75795082	0.756756757	0.740384615	0.712962963
2	0.803656869	0.809954751	0.764044944	0.755555556
3	0.774983148	0.78280543	0.736263736	0.744444444
4	0.759811166	0.760180995	0.748815166	0.745283019
5	0.78404924	0.778280543	0.732240437	0.67
	0.776090249	**0.777595695**	**0.74434978**	**0.725649196**

7

RFE-Xgboost				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.697251481	0.707207207	0.648648649	0.689655172
2	0.826779182	0.837104072	0.8	0.86746988
3	0.757708333	0.773755656	0.709302326	0.802631579
4	0.841976602	0.841628959	0.8	0.760869565
5	0.760093818	0.778280543	0.710059172	0.8
	0.776761883	**0.787595288**	**0.733602029**	**0.784125239**

8

Relief-Xgboost				
KFold	AUC Score	Accuracy Score	F1 Score	Precision Score
1	0.828332083	0.842342342	0.797687861	0.8625
2	0.723626374	0.746606335	0.658536585	0.739726027
3	0.784525584	0.787330317	0.76142132	0.765306122
4	0.748349835	0.755656109	0.712765957	0.770114943
5	0.819299618	0.841628959	0.777070064	0.835616438
	0.780826699	**0.794712812**	**0.741496358**	**0.794652706**

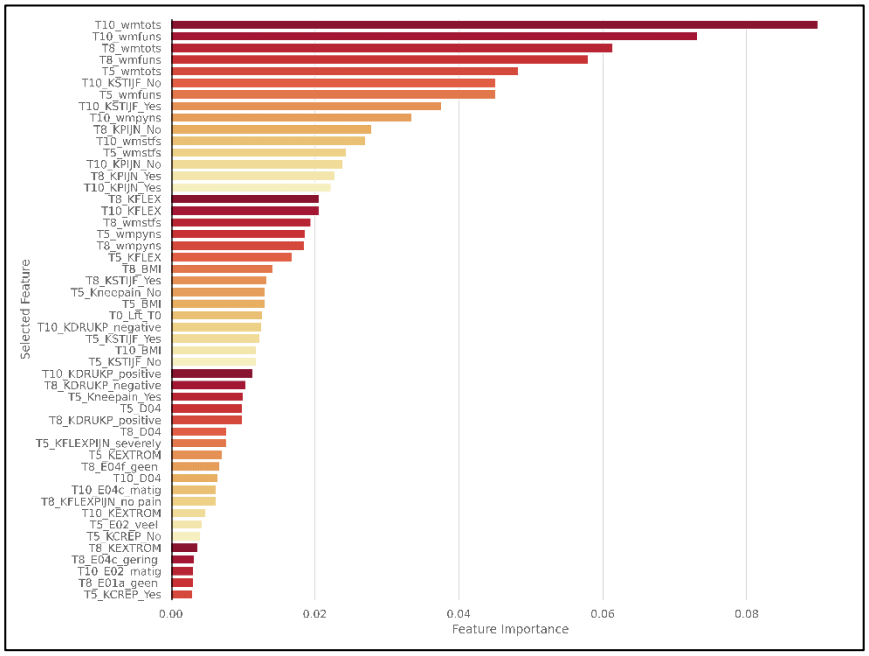
9

Model	Mean AUC Score	Mean Accuracy Score	Mean F1 Score	Mean Precision Score	Mean Overall Score
RFE-RF	0.793031664	0.795670784	0.761791678	0.748055344	0.774637368
Relief-RF	0.792279498	0.792943622	0.762208015	0.741605782	0.772259229
RFE-LR	0.757492013	0.760413355	0.717982552	0.710978652	0.736716643
Relief-LR	0.775998188	0.776629571	0.745693529	0.720868593	0.75479747
RFE-SVC	0.745452387	0.754025519	0.702504665	0.710153257	0.728033957
Relief-SVC	0.776090249	0.777595695	0.74434978	0.725649196	0.75592123
RFE-Xgboost	0.776761883	0.787595288	0.733602029	0.784125239	0.77052111
Relief-Xgboost	0.780826699	0.794712812	0.741496358	0.794652706	0.777922144

Figure 5. Complied tables (1-8) showing the performance metrics: AUC score, Accuracy score, F1 score and Precision score for each of the cross-validation folds in every predictive model.
Averages of these each of these scores for each model is also shown in panel 9. Here best performing model is highlighted in red, and second and third best models are noted in yellow and green, respectively.
Here, the model with the highest mean AUC score was RFE-RF; model with highest mean accuracy score was RFE-RF; highest mean F1 score was Relief-RF; and the model with the highest mean precision score was Relief-Xgboost.

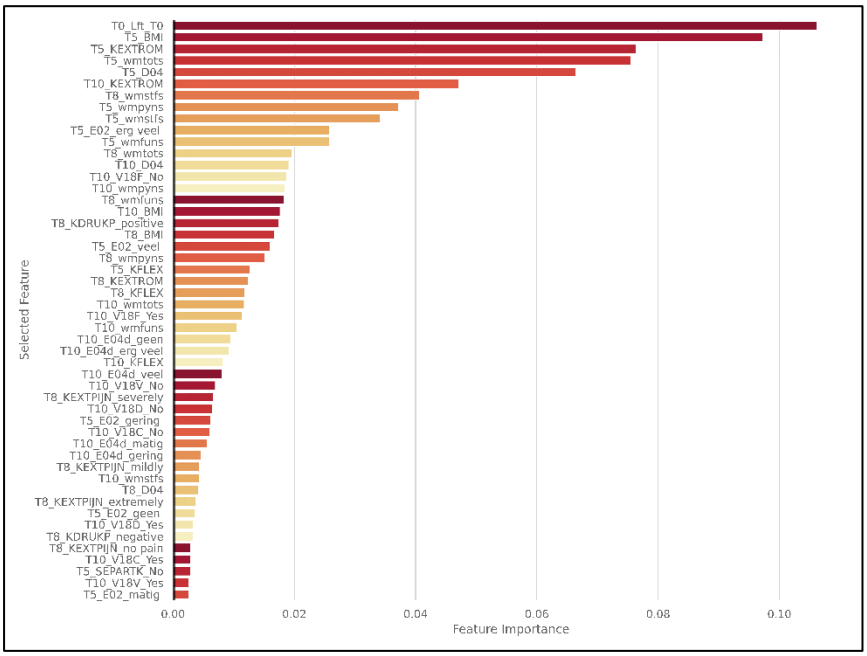
A

1



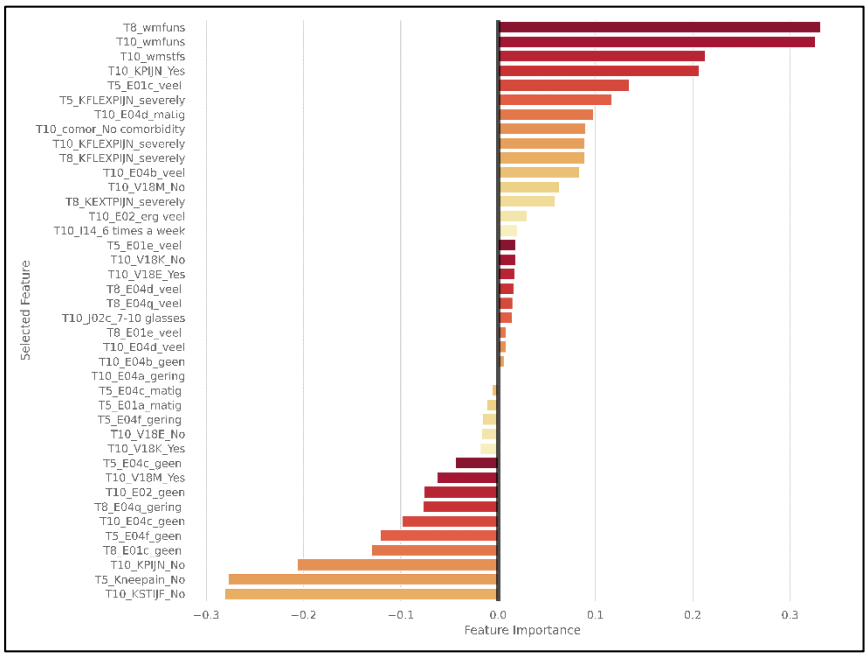
RFE-RF

2



Relief-RF

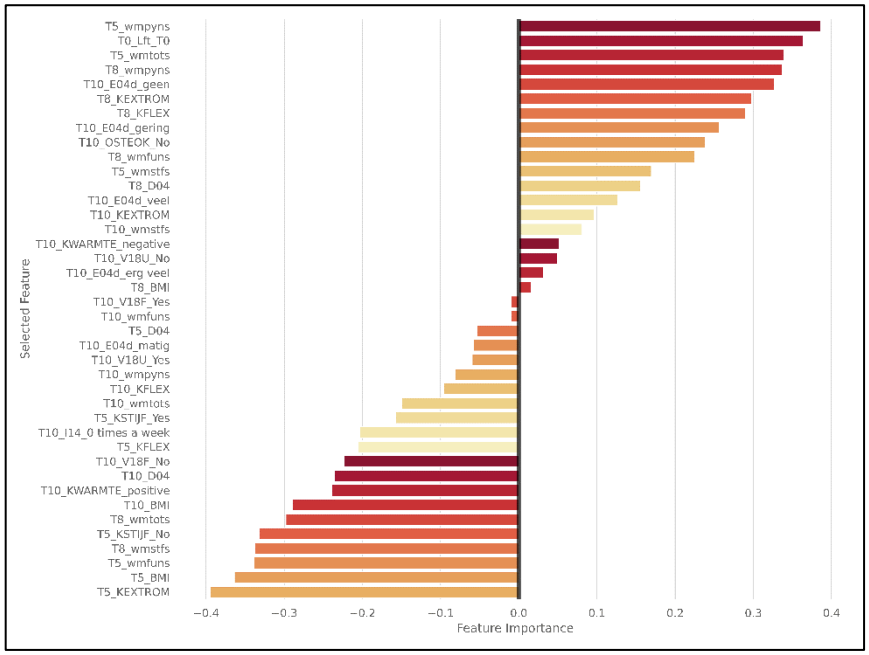
3



RFE-LR

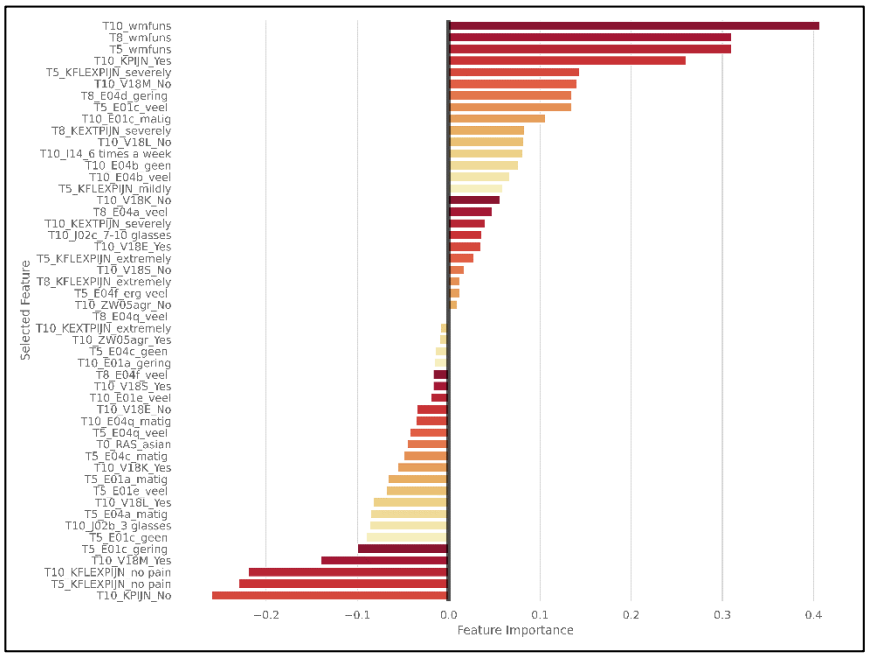
B

4



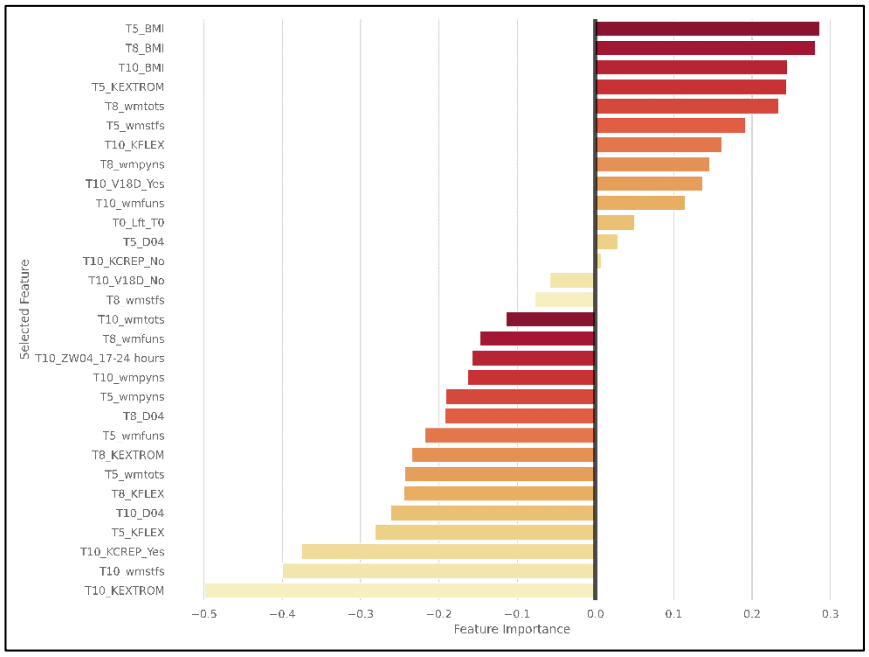
Relief-LR

5



RFE-SVC

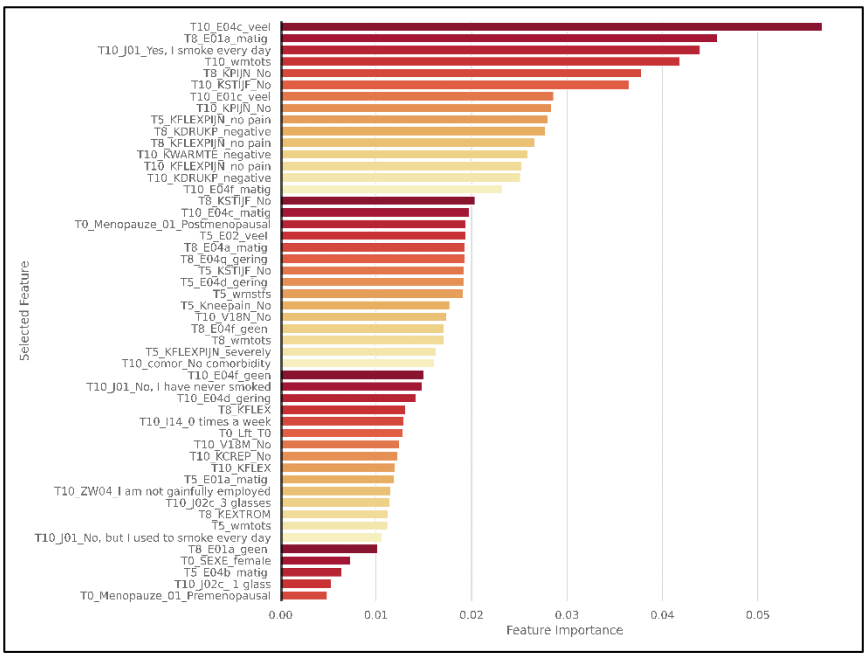
6



Relief-SVC

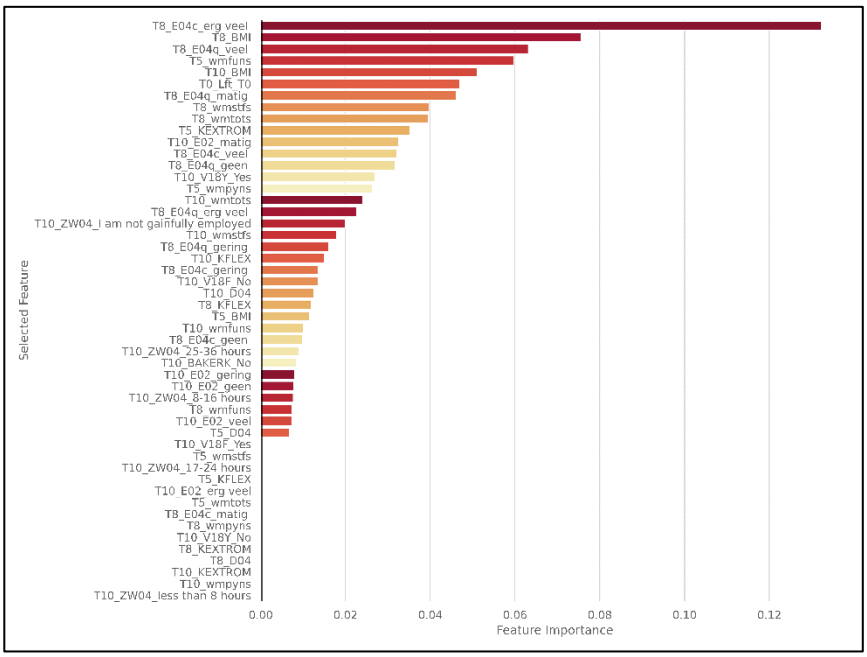
C

7



RFE-Xgboost

8



Relief-Xgboost

Figure 6.
Feature Importance graphs for each model, showing the relative feature importance of the selected features. The features are ordered by most impactful features. The top 5 selected features for models 1-8 (Panels A-C) are shown to be:

1. 10-year follow-up WOMAC standardized total sum score, 10-year follow-up WOMAC standardized physical functioning scale, 8-year follow-up WOMAC standardized total sum score, 8-year follow-up WOMAC standardized physical functioning scale, 5-year follow-up WOMAC standardized total sum score.
2. Baseline age, 5-year follow-up BMI, 5-year follow-up knee extension active ROM (range of motion), 5-year follow-up WOMAC standardized total sum score, 5-year follow-up NRS scores for pain intensity of knee the past week.
3. 8-year follow-up WOMAC standardized physical functioning scale, 10-year follow-up WOMAC standardized physical functioning scale, 10-year follow-up WOMAC standardized stiffness scale, 10-year follow-up physical examination of pain present on knee extension, 5-year follow-up WOMAC Night Pain.
4. 5-year follow-up WOMAC standardized pain scale, Baseline age, 5-year follow-up WOMAC standardized total sum score, 8-year follow-up WOMAC standardized pain scale, 10-year follow-up WOMAC difficulty standing.
5. 10-year follow-up, 8-year follow-up WOMAC standardized physical functioning scale, 5-year follow-up WOMAC standardized physical functioning scale, 10-year follow-up physical examination of pain present on knee extension, 5-year follow-up physical examination of pain present on knee flexion.
6. 5-year follow-up BMI, 8-year follow-up BMI, 10-year follow-up BMI, 5-year follow-up knee extension active ROM (range of motion), 8-year follow-up WOMAC standardized total sum score.
7. 10-year follow-up WOMAC difficulty getting up from chair, 8-year follow-up WOMAC pain while walking, Heavy smoker, 10-year follow-up WOMAC standardized total sum score, 8-year follow-up physical examination of pain present on knee extension.
8. 8-year follow-up WOMAC difficulty getting up from chair, 8-year follow-up BMI, 8-year follow-up WOMAC difficulty sitting, 5-year follow-up WOMAC standardized physical functioning scale, 10-year follow-up BMI

4 Discussion

In this work, 8 predictive models were developed to identify incidence of OA from the CHECK dataset. By evaluating the mean performance across metrics AUC score, accuracy, F1 and precision, the most performant model was identified. To minimize statistical bias, 5-fold cross validation was used, to train each model for every fold using randomised splits of the initial dataset. The model Relief-Xgboost was shown to have the best mean scores across classifiers, followed by RFE-RF and Relief-RF as second and third highest scoring models based on mean performance. The same models (RFE-RF, Relief-RF and Relief-Xgboost) were also shown to have the same interpolated AUC scores in Figure 4. This further supports that these are indeed the best models produced in this study and support our initial expectations that the tree-based models would perform best – due to the non-linearity of the CHECK data.

Interestingly, the performance of the model using RFE-Xgboost was worse than expected, however the best scoring fold for this model was 0.91, as shown in Table 2. Further examination of the performance of each fold in this model, Figure 4 (7) and Figure 5 (7), the AUC score in the first fold was below than average (0.697). This may be the cause of why the performance of RFE-Xgboost model was not in line with the other tree-based models.

4.1 ROC Interpolation

Examination of the AUC for ROC curves is informative of the performance of a model over an array of thresholds. However, when dealing with binary outcomes, such as in this study, there is but one threshold. Since comparing the AUC of each model is used to decide which of the model's perform best, it is imperative to consider how the AUC corresponds with the ROC curve in these such cases. A 2019 study by Muschelli, J demonstrates how, for binary predictors, the interpolation between thresholds affects the AUC score; that typical linear interpolation can produce distorted results. (Muschelli, J *et al*., 2019)

As the generated ROC plots for each model, and the averages among them, the specificity (x-axis) values differ from one another. As suggested by Tom Fawcettt, to overcome this, and generate a meaningful average, the false positive rate and true positive rates were interpolated (Fawcett T. *et al*., 2005). This may be the cause as to why the relative AUC scores in each ROC plot differs from the AUC score of each fold.

4.2 Features selected as potential biomarkers

The most impactful features selected by the highest scoring models, as demonstrated in Figure 5, of each type were WOMAC standardized physical functioning scale, WOMAC standardized total sum score, body mass index, physical examination of pain present on knee extension, baseline age, knee extension active range of motion, WOMAC standardized pain scale and WOMAC difficulty getting up from chair. These features may be candidates for potential biomarkers and make sense as so, as the WOMAC features (WOMAC standardized total sum score, WOMAC standardized physical functioning scale, WOMAC standardized stiffness scale, WOMAC standardized pain scale) were also the highest scoring features in a 2020 study by Pawel Widera *et al* (Widera, P *et al*, 2020) . The 2020 study used the same CHECK dataset, with an RFE-RF ML approach with the aim of predicting knee OA progression. The results obtained here are in-line with this study and thus the features have a high possibility to be biomarkers.

4.3 Biomarker Panels

It is preferred by clinicians for biomarker panels to be small, as assurance is needed that the selected biomarkers truly represent incidence of a disease. The predictive capabilities of a machine learning model are crucial in determining the prognosis of patients, and which would respond best to certain treatments. Therefore, biomarkers that describe risk of a disease should have widespread utility.

The European Society of Medical Oncology (ESMO) has suggested in its clinical guidelines that a biomarker panel should have no less than 20 molecular biomarkers in a panel (Besse B. *et al*., 2014; Eberhardt WE *et al*., 2015; Parker C *et al*., 2015; Van Cutsem E *et al*., 2016; Cardoso F *et al*., 2017). Although the biomarker panel selected in this project is on the larger side, it was done so to benefit model performance. One of the drawbacks of selecting a high number of features is speed but selecting fewer than 50 features with RFE or Relief resulted in lower scores across the board (accuracy, F1, precision, ROC-AUC), with scores below ~0.7. As each feature selector was given the option to select a range of number of features (10, 20, 30, 40, 50 features in the parameter grid), it would seem that selecting 50 features was the best trade-off between model performance and number of retained features.

4.4 Proposed Pipeline

The ML pipeline used here, especially the pre-processing strategy, is generic enough so that it can be applied to most biomedical datasets. Adapting these models to simulate patient selection has many clinical implications. For example, in clinical trials for drug discovery could use these methods to suggest which patients are preferable to be included into the study. This may improve efficiency of such clinical trials as patients most likely to show disease progression may allow for greater treatment appraisals.

4.5 Further Work

4.5.1 Correlated Features

As shown, the features selected by most of the models are highly correlated, as demonstrated in Figure 3. It is for this reason that if this experiment were to be repeated, removing the highly correlated values may improve the developed pipelines by making the learning algorithm faster, decreasing harmful bias, and decreasing the curse of dimensionality problem. By implementing this, and noting the top features selected for each model, it would be intriguing to see if and how they differ in future work.

4.5.2 Evaluation of ML Methods with Imbalanced Data and Choice of Performance Measure

As shown in the pandas profiling report, the dataset was imbalanced – having 638 cases of "Not OA" and 468 cases of "OA". The class weights were computed using an in-build method to balance them, but as elucidated by Samir *et al* (Al-Stouhi S *et al*., 2016), methods such as these for tackling imbalanced data inherently assume there is abundance of training examples - which are used to construct effective computation of class imbalance. Namely, only the problem of "relative imbalance" is addressed, where the number of training examples are sufficiently abundant and the sample number in one class is much larger than the other. Nevertheless, when the training data is not incredibly large, developing a representative model becomes much harder to accomplish because of the imbalanced class distribution. Samir *et al*, hence has proposed a novel classifier method which compensated for class imbalance thus rectifying the inherent skew, and offsets the lack of instances in training data by using data from an auxiliary domain. Perhaps repeating this experiment, incorporating the novel algorithm proposed by Samir *et al*, would yield more representative model performance.

From a recent 2020 study by Chang Cao *et al*, the use of ROC curves for evaluating binary classification methods was shown to exaggerate performance scores, especially if the dataset is imbalanced (Cao, Chang *et al*., 2020). The study suggested the use of an MCC-F1 performance curve to address the shortcomings of ROC curves. The novel MCC-F1 curve plots the MCC score against the F1 score and was shown to differentiate between highest and lowest scoring classifiers, even when the base dataset is imbalanced. Since time was limited for the present study, this was unable to be implemented. Repeating this experiment with the MCC-F1 performance curve, along with ROC curves and the additional metrics may provide further insight into which models perform the best.

4.5.3 Use of Archived Data

A fundamental component for biomarker discovery involves data acquisition. For the most part, participants involved in gathering data is limited by specimen availability instead of what would be best for a particular study. When developing predictive models retrospectively on these datasets, there may be biases where insufficient patient data may be unable to fulfil the necessary calculations centered on patient inclusion criteria, thus raising concerns on the true performance of the model.

A 2017 study by Selleck *et al* has proposed the use of archived datasets in conjunction with original datasets to validate biomarkers (Selleck MJ *et al*., 2017). Although this is not conventionally regarded as high-quality data, it would be interesting to repeat this experiment, using a combination of CHECK data and data used previously by the Lazzarini group (PROOF study) (Lazzarini *et al*., 2017) and note the effect on model performance.

4.5.4 New Feature Selector Heuristic

For the models RFE-LR, Relief-LR, RFE-SVC, RFE-Xgb, Relief-Xgb, some of the top selected features were categorical values, as shown in Figure X. With RFE and Relief from Sklearn/Skrebate, dummied variables are not considered as part of a block with its parent variables. Instead, each column is considered independent from one another.

Therefore, it would be reasonable to suggest that further work would include creating a new feature selector heuristic which would incorporate grouping each dummied variable to the original variable, so that if one is removed by RFE/Relief, then so are all the variables coming from the same original variable.

The proposed solution to this would involve 1) training the classifier, for example Random Forest, on the version of the dataset including all dummied variables; 2) Extracting the full feature importance rank from the trained RF model; 3) Collapsing the ranking, by taking the average rank for all dummy variables coming from the same original variable; 4) Create an elimination rule based on the collapsed ranking, such as if the lowest ranking variable is of a discrete class, then all dummied variables are removed at once; 5) Repeat this process until the specified n_features_to_select is reached in RFE.

It would be interesting to see how models which include the proposed feature selector heuristic would perform in line with the models generated here, and if the top features selected would be any different.

4.5.5 Visualising Feature Importance

Furthermore, evaluation of feature importance for the tree-based models (RF and Xgboost) was performed by sorting feature by their feature importance attribute. This may have led to some misleading results in relative importance of the selected features. Since feature importance is determined, for these models, by counting the number of times a tree has voted towards the final decision. Each split is therefore weighted equally. This may be problematic as splits closest to root splits influence the model the most.

Perhaps using an alternate feature importance method would be preferential for tree-based models. SHAP (Shapley additive explanations) combines a game theory approach to provide how each feature influences the outcome of a model (Lundbeg, S *et al*., 2018; Lundbeg, S *et al*., 2017; Å trumbelj *et al*., 2014). It would be interesting to incorporate this in a repeat of this experiment and see how the selected features differ from the features presented here.

5 Conclusion

The aim of this study was to compare the performance of 8 predictive ML models for detecting incidence of OA, and to suggest a suitable pre-processing strategy for many biomedical datasets. The data used was provided by a CHECK study, identifying WOMAC questionnaire data (standardized physical functioning scale, standardized total sum score, standardized pain scale and difficulty getting up from chair), BMI, pain present on knee extension, baseline age, knee extension active range of motion as potential biomarkers. Furthermore, the models Relief-Xgboost, RFE-RF and Relief-RF are suggested to perform the best for identifying biomarkers using the CHECK dataset, as shown by an overall mean score of 0.778, 0.775 and 0.772 respectively (mean score of mean AUC, mean accuracy, mean precision and mean F1 score).

Further work is required to fully realise the capabilities of the developed predictive models, involving review and evaluation of the methodology and feature selection process.

Acknowledgments

Code snippet used to extract features from column transformer was provided by answer on Stack Overflow: https://stackoverflow.com/questions/57528350/can-you-consistently-keep-track-of-column-labels-using-sklearns-transformer-api
Machine learning models are adapted from the Machine Learning Tutorial provided by Jaume Bacardit on GitHub: https://github.com/jaumebp/ML-tutorial

References

Centers for Disease Control and Prevention [Cdc]. (2009). Prevalence and most common causes of disability among adults–United States, 2005. MMWR Morb. Mortal. Wkly. Rep. 58, 421–426.

Murphy, L., and Helmick, C. G. (2012). The impact of osteoarthritis in the United States: a population-health perspective: a population-based review of the fourth most common cause of hospitalization in U.S. adults. Orthopedic Nurs. 31, 85–91. doi: 10.1097/NOR.0b013e31824fcd42

Valdes, A. M., and Spector, T. D. (2011). Genetic epidemiology of hip and knee osteoarthritis. Nat. Rev. Rheumatol. 7, 23–32. doi: 10.1038/nrrheum.2010.191

Chu, C. R., Millis, M. B., and Olson, S. A. (2014). Osteoarthritis: from palliation to prevention: aoa critical issues. J. Bone Joint Surg. Am. 96:e130. doi: 10.2106/JBJS.M.01209

Wright, Rick W., *et al*. "Radiographs are not useful in detecting arthroscopically confirmed mild chondral damage." Clinical Orthopaedics and Related Research 442 (2006): 245-251.

Ceotto, Michele, Gian Franco Tantardini, and Alán Aspuru-Guzik. 2011. Fighting the curse of dimensionality in first-principles semiclassical calculations: Non-local reference states for large number of dimensions. Journal of Chemical Physics 135(21): 214108.

Breiman L. Random forests. Mach Learn. 2001;45:5–32. doi: 10.1023/A:1010933404324.

Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2017;27:659–678. doi: 10.1007/s11222-016-9646-1.

Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422. doi: 10.1023/A:1012487302797.

Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, Tsai CJ, Zhang S. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics. 2004;5:81. doi: 10.1186/1471-2105-5-81.

Svetnik V, Liaw A, Tong C, Wang T. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T, editors. Multiple classifier systems. Berlin: Springer; 2004.

Lazzarini, N., Bacardit, J. RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers. BMC Bioinformatics 18, 322 (2017). https://doi.org/10.1186/s12859-017-1729-2

Janet Wesseling, Maarten Boers, Max A Viergever, Wim KHA Hilberdink, Floris PJG Lafeber, Joost Dekker, Johannes WJ Bijlsma, Cohort Profile: Cohort Hip and Cohort Knee (CHECK) study, International Journal of Epidemiology, Volume 45, Issue 1, February 2016, Pages 36–44, https://doi.org/10.1093/ije/dyu177

American College of Rheumatology. Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). http://www.rheumatology.org/practice/clinical/clinicianresearchers/outcomes-instrumentation/WOMAC.asp.

McKinney, W. pandas: a foundational Python library for data analysis and statistics. In Workshop on Python for High-Performance and Scientific Computing (PyHPC 2011) (Seattle, USA, 2011), https://www.dlr.de/sc/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.

Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830, http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (2011).

Al-Stouhi S, Reddy CK. Transfer Learning for Class Imbalance Problems with Inadequate Data. Knowl Inf Syst. 2016;48(1):201-228. doi:10.1007/s10115-015-0870-3

Chen, C., Liaw, A. & Breiman, L. Using random forest to learn imbalanced data. Tech. Rep., University of California, Berkeley (2004).

Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013;4(2):627-635.

Metz CE. Basic principles of ROC analysis. Seminars in Nuclear Medicine. 1978;8(4):283–98.

McKinney, W. pandas: a foundational Python library for data analysis and statistics. In Workshop on Python for High-Performance and Scientific Computing (PyHPC 2011) (Seattle, USA, 2011), https://www.dlr.de/sc/Portaldata/15/Resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.

Oliphant, T. E. Python for Scientific Computing. Computing in Science and Engineering 9, 10–20, https://doi.org/10.1109/MCSE.2007.58 (2007).

Waskom, M. seaborn: statistical data visualization (2013–), http://seaborn.pydata.org/.

Muschelli, J. ROC and AUC with a Binary Predictor: a Potentially Misleading Metric. J Classif (2019). https://doi.org/10.1007/s00357-019-09345-1

Tom Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, Volume 27, Issue 8, 2006, Pages 861-874, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2005.10.010.

Widera, P., Welsing, P.M.J., Ladel, C. *et al*. Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data. Sci Rep 10, 8427 (2020). https://doi.org/10.1038/s41598-020-64643-8

Besse B, Adjei A, Baas P, *et al*. 2nd ESMO Consensus Conference on Lung Cancer: non-small-cell lung cancer first-line/second and further lines of treatment in advanced disease. Ann Oncol. 2014;25:1475–1484.

Eberhardt WE, De Ruysscher D, Weder W, *et al*. 2nd ESMO Consensus Conference in Lung Cancer: locally advanced stage III non-small-cell lung cancer. Ann Oncol. 2015;26:1573–1588.

Parker C, Gillessen S, Heidenreich A, Horwich A, ESMO Guidelines Committee Cancer of the prostate: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2015;26:v69–v77.

Van Cutsem E, Cervantes A, Adam R, *et al*. ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. Ann Oncol. 2016;27:1386–1422.

Cardoso F, Costa A, Senkus E, *et al*. 3rd ESO-ESMO international consensus guidelines for Advanced Breast Cancer (ABC 3) Breast. 2017;31:244–259.

Cao, Chang & Chicco, Davide & Hoffman, Michael. (2020). The MCC-F1 curve: a performance evaluation technique for binary classification.

Selleck MJ, Senthil M, Wall NR. Making Meaningful Clinical Use of Biomarkers. Biomark Insights. 2017;12:1177271917715236. Published 2017 Jun 19. doi:10.1177/1177271917715236

Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent individualized feature attribution for tree ensembles. Computing Research Repository arXiv:1802.03888v2 https://arxiv.org/abs/1802.03888 (2018).

Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In I., Guyon *et al*. (eds.) Advances in Neural Information Processing Systems (NIPS 2017), 4765–4774 (Long Beach, CA, USA, 2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Å trumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41, 647–665, https://doi.org/10.1007/s10115-013-0679-x (2014).