PROJECT DELIVERABLE REPORT

Project Title:
Advanced personalised, multi-scale computer models preventing osteoarthritis
SC1-PM-17-2017 - Personalised computer models and in-silico systems for well-being

Deliverable number	D9.2
Deliverable title	Evaluation of OACTIVE in human population
Submission month of deliverable	M42
Issuing partner	HULAFE
Contributing partners	ALL
Dissemination Level (PU/PP/RE/CO):	PU
Project coordinator	University of Nicosia (UNIC)
Tel:	**+357 22 841 528**
Fax:	**+357 22 357481**
Email:	felekkis.k@unic.ac.cy **&** giannaki.c@unic.ac.cy
Project web site address	www.oactive.eu

Revision History

Version	Responsible	Description/Remarks/Reason for changes
1.0	CERTH, LJMU	First Draft
1.1	UNIC, HULAFE	Review of First Draft
2.0	CERTH, HULAFE	Second Draft
2.1	UNIC, LJMU, ANIMUS	Review of Second Draft
3	HULAFE	Final version

Contents

Abbreviations

CA	Consortium Agreement
Del	Deliverable
DoW	Description of Work
FS	Feature Selection
GA	Grant Agreement
ICT	Information and Communication Technologies
KL	Kellgren and Lawrence
KOA	Knee osteoarthritis
kNN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
ML	Machine learning
MDA	Mean Decrease in Accuracy
MLR	Multinomial Logistic Regression
OOB	Out-of-bag
PCA	Principal component analysis
RF	Random Forests Algorithm
vs	Versus
XGBoost	Extreme Gradient Boosting

1. Summary

The first aim of this deliverable (Task 9.1) is to present the results and outcomes of the long-term evaluation of OACTIVE using data from clinical studies.

This report refers to Deliverable 9.2, which relates to the OACTIVE WP 9, "Technology assessment and full system validation" led by HULAFE. The objective of WP9 is to validate the integrated OACTIVE system by employing a comprehensive methodology that involves: (i) Clinical studies in human populations and (ii) validation of the system using big data registries. The ethical, legal, and social challenges that need to be met in order for the scientific advances to be responsibly applied will be finally investigated.

2. Introduction

Osteoarthritis (OA) is a common joint disease, causing disability and reduction of quality of life. It is a leading cause of chronic pain, and health-care utilization (1). OA is characterized by cartilage loss, subchondral bone changes, synovial inflammation, and meniscus degeneration (2). Basic research approaches allowed the identification of pathophysiological factors that determine the existence of OA. However, the main part of this research is performed in the late stage of OA, and the pathological processes involved in the early stages of joint disease are not well understood (3), and current guidelines are not well suited for diagnosing patients in the early stages of disease and do not discriminate patients for whom the disease might progress rapidly. The most important challenge in OA management is identifying and classifying patients who will benefit most from treatment. Further efforts are needed in patient subgrouping and developing prediction models (4). For this purpose, we conducted clinical studies in 3 different countries and 3 different populations, and all the data collected has been combined to create big data models in order to predict outcomes for individual patients that could help prevent OA in a clinical context.

The main goal of this deliverable is to present the results and outcomes of the long-term evaluation of OACTIVE using data from clinical studies conducted through this project. The clinical studies collect data of 316 subjects of 3 different countries (Spain-HULAFE, Greece-ANIMUS and Cyprus-Apollonion Private Hospital). In each country, a different population was recruited (i. e., patients in risk to develop OA (HULAFE), athletes (ANIMUS), and patients with advanced OA (UNIC) in order to assess demographic, lifestyle, functional, and clinical risk factors that could be related to the onset and development of OA.

The data collected from each patient is divided in different data subsets: (i) training data: data that is being used for building the personalised models, (ii) fine-tuning data: data used for further optimising the models, and (iii) validation data: data used for testing the efficiency of the trained and fine-tuned patient-specific models. In OACTIVE, all the data collected for clinical studies has been combined to perform a detailed analysis about the risk factors involved in OA. This analysis is composed by thousands of individual data, that has served as input for the integrated computer hyper- models. With this big data models, it could be possible to predict outcomes for individual patients that could help to prevent OA in a clinical context.

Initially, we present an extensive analysis of OACTIVE's database. The database is separated into two datasets because a small number of Missing Not αt Random data exists. We start our analysis by presenting the main characteristics of both datasets, in terms of correlation, with a mixed association technique, in terms of the response variable's separation, by creating a heatmap of the feature's vs the samples, and finally, in terms of variable importance analysis, by utilizing the Random Forests Variable Importance. Afterwards, we visualise the data into two dimensions by utilising the Principal component

analysis (PCA) technique, and we colour each sample with the class it represents. Finally, in order to validate the datasets, we apply several classification models, such as Multinomial Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbours, Random Forests and XGBoost.

Classical statistical models were also applied, in particular logistic regression and induction trees with CART. Further, a self-explained neural network was used (https://arxiv.org/abs/1908.05978) which verified the dependence on covariates explicitly, showing that the dependence is almost exactly linear which justifies our preference for logistic regression. The interpretability of the risk score index provided by logistic regression is expressed in a form that most naturally integrates with clinical reasoning. The reason for this is that it gives a statistical assessment of the weight of evidence for making the diagnosis This analysis was presented at IDEAL (https://link.springer.com/chapter/10.1007/978-3-030-33617-2_13). This case study benchmarks a range of statistical and machine learning methods relevant to computer-based decision support in clinical medicine, focusing on the diagnosis of osteoarthritis of the knee at first presentation.

In addition, we present an extensive analysis of the OACTIVE MRI dataset. The dataset consists of two separate cases to be analysed. The first case based on Diagnosis task and the other case based on early diagnosis task. To perform, on each of the two cases, a complete analysis of this data we distinct it in 9 variables' categories dataset, and we analyse them separately. We start our analysis by presenting the main characteristics this database. In terms of correlation, with Spearman's correlation coefficient technique. In terms of variable importance analysis by utilizing the Random Forests Variable Importance. We do an extensive analysis of both the Principal Components Analysis and the Random Projections techniques to apply dimensionality reduction in our data. Afterwards, we visualise the data into two dimensions by utilising the PCA technique, and we colour each sample with the class it represents. Finally, in order to validate the datasets, we apply several classification models, such as Multinomial Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbours and Support Vector Machines.

Furthermore, we validated the robust data mining approach that could identify important risk factors which contribute to the diagnosis for different stages of KOA on OACTIVEs database. The first approach concerns the diagnosis and the second one the early diagnosis of KOA. The validation of the extracted factors was performed in subgroups employing seven well-known classifiers. We investigated the behavior of the best model, with respect to classification errors and the impact of used features, to confirm their clinical relevance.

3. Clinical studies

The clinical studies were conducted in 3 different countries with different populations:
- Spain- La Fe University Hospital (HULAFE): patients in risk to develop OA
- Cyprus-Apollonion Private Hospital (UNIC): patients with advanced OA
- Greece-ANIMUS: athletes post ACL injury, in risk for post-traumatic OA

3.1 Approval from Ethics Committees

The approval to perform the clinical studies was obtained from the local committees:

- Cyprus National Bioethics Committee at Apollonion Hospital, Nicosia, Cyprus (ΕΕΒΚ/ΕΠ/2018/19) (date: 03/05/2019)
- Ethics Committee on Drug Research of the University and Polytechnic Hospital La Fe, Valencia, Spain (CEIm La Fe 2017/0147) (date: 14/09/2018)
- Ethics and Deontology Committee of ANIMUS Rehabilitation Center, Larissa, Greece, 01/06/2018

3.2 Recruitment

The recruitment was slow at the beginning, due to logistic problems and the difficulty to find volunteers. Some actions were taken to speed-up the recruitment:

- Leaflet written in simple language (that outlined the benefits for the participation to the study) was disseminated to OA patients.
- UNIC: Workshop on Mar 16th, 2019 to increase the awareness of OA patients about OACTIVE and to inform them about the benefits they would have by participating in this study.
- HULAFE: An email account was provided for all the potential volunteers to be able to contact the researchers and ask any doubts, and to make communication easier.
- ANIMUS: An e-mail summarizing information and the benefits of the project was prepared and sent to orthopaedic surgeons, with great experience in treating ACL injury.
- To increase the number of participants a gait-analysis examination was offered free of charge to all volunteers from UNIC, HULAFE and ANIMUS.

Finally, a total of 130 patients (98 women, 32 men) were recruited from UNIC (patients with knee OA undergoing knee replacement surgery); 115 from HULAFE (51 Healthy subjects (HS), 55 subjects with Early OA (EOA) and 9 not meeting the criteria for HS or EOA); 113 participants from ANIMUS (athletes post ACL injury, 87 male, 26 female).

Patients were asked to participate in this research project with a voluntary decision and they should be competent to understand what is involved. To this end, a consent form was prepared; in ANIMUS group, consent for participation of participants < 18 year old was given by one of their parents. It should be pointed out that the anonymity of the patients was maintained.

The recruitment of patients with knee OA from UNIC was carried out by clinicians, and OA was defined according to the American College of Rheumatology criteria (https://www.rheumatology.org/) for the classification and reporting of osteoarthritis of the knee. The inclusion/exclusion criteria used are summarized in Table 3.1. In general, patients below the age of 50 were excluded from the study. Furthermore, patients with post-traumatic osteoarthritis, arthritis due to any autoimmune, infective, or inflammatory rheumatological conditions were also excluded for the study.

The recruitment of HS and EOA patients (HULAFE) was carried out by clinicians meeting the criteria in table 3.2.

The recruitment of athletes post ACL injury (ANIMUS) was carried out by clinicians and physiotherapists, according to the criteria shown in table 3.4. Participants above the age of 50, or with OA due to any autoimmune, infectious, inflammatory or rheumatological disease were excluded from the study.

A clinical evaluation meeting the Data Collection Protocol (DCP) elaborated in WP2 and detailed in Deliverable 2.2, consisting in physical examination, patient-based questionnaires, radiographs, MRI, and collection of biological samples (blood, urine, and faeces), was performed to each subject.

After the clinical evaluation, subjects from HULAFE were classified as follows, according to Luyten's proposal for EOA classification (Table 3.3):

- 51 Healthy subjects (35 women and 16 men),
- 55 with Early OA (35 women and 19 men),
- 9 (4 women and 5 men) not meeting criteria.

3.2.1. Tables

Table 3.1. The inclusion and exclusion criteria used for the recruitment of OA patients (UNIC)

Exclusion criteria	Inclusion criteria
1. Post-traumatic OA 2. Autoimmune OA 3. Infective/inflammatory OA 4. Rheumatologic conditions 5. Patient age <50 years	1. Knee pain 2. Radiological evidence of OA on plain film 3. Crepitus audible/ palpable 4. Stiffness lasting under 30 min 5. Patient age >50 years

Table 3.2. The inclusion and exclusion criteria used for the recruitment of HS/EOA patients (HULAFE)

Inclusion criteria		Exclusion criteria
Healthy subjects in risk of developing OA:	**Early OA:**	1. Autoinmune, infective or rheumatologic conditions 2. Athletes
1. Patient age greater than or equal to 40 years 2. BMI greater than or equal to 25 3. Kellgren & Lawrence (KL) 0-1	1. Patient age greater than or equal to 40 years 2. Kellgren & Lawrence 0-1	

Table 3.3. Luyten's proposal for classification criteria for early OA of the knee (5)

A. Patient-based questionnaires: Knee Injury and Osteoarthritis Outcome score (KOOS): 2 out of the 4 subscales need to score "positive" (≤85%) 1. Pain (9 items, including information on pain intensity, frequency, and duration) 2. Symptoms, stiffness (7 items) 3. Function, daily living (short version: 7 items) 4. Knee-related quality of life (4 items)
B. Clinical examination: at least 1 criterion needs to be present • Joint line tenderness • Crepitus
C. X-rays: Kellgren and Lawrence grade 0-1 standing, weight bearing (at least 2 projections: PA fixed flexion and skyline for patellofemoral OA)

Table 3.4. The inclusion and exclusion criteria used for the recruitment of athletes post ACL injury (ANIMUS)

Inclusion criteria	Exclusion criteria
1. Sports activity 2. ACL injury 3. ≥ 3 months post treatment 4. No other pathology affecting the gait/motor pattern 5. No history of OA	1. Autoimmune OA 2. Infectious/inflammatory OA 3. Rheumatic disease 4. Age > 50 years

3.3 Methodology of clinical studies

The subjects included on the clinical studies were subjected to the DCP, described in detail in Deliverable 2.2. Each group of subjects underwent a different set of tests and questionnaires, as described in table 3.5, but all of them were standardized among the three clinical partners.

Table 3.5. Tests performed on each group of subjects.

	Advanced OA patients (UNIC)	**HS/EOA patients (HULAFE)**	**Athletes post ACL injury (ANIMUS)**
Demographics			
Date	**X**	**X**	**X**
Sex	**X**	X	X
Age	X	X	X
Birth country	X	X	X
Ethnicity	X	X	X
Occupation	X	X	X
Socioeconomics			
Level of education (individual)	X	X	X
Level of education (parents)		X	X
Marital status	X	X	X
Residency		X	X
Household income		X	X
Housing status		X	X
Anamnesis			
Any current medication	X	X	X
XHigh blood pressure	X	X	X
Family OA history	X	X	X
Personal history of hip/hand OA	X	X	X
Do you have knee OA?	X	X	X
Have you ever been told that you have OA of your knee by a doctor?	X	X	X
Occupational risk	X	X	X
Smoking	X	X	X
Alcohol	X	X	X
Hormonal status (women)		X	X
Previous knee injuries	X	X	X
Regular sport leisure	X	X	X
Knee pain	X	X	X
Pain side	X	X	X
Time since pain start	X	X	X

Resting/Walking VAS		X	X
NHANES-type questions	X	**X**	X
Knee instability		X	X
Pain rhythm		X	X
Hip OA/surgery	X		
Physical examination			
Mass	X	X	X
Height	X	X	X
BMI	X	X	X
Joint line tenderness		X	
Patellofemoral pain		X	
Crepitus		X	
Flexion angle		X	
Extension angle		X	
Flexion deformity		X	
Muscle atrophy		X	
Knee laxity		X	X
Joint proprioception		X	X
Abdominal perimeter		X	
Dynamometric evaluation of knee extension strength		X	
Dynamometric evaluation of knee flexion strength		X	
5 sit to stand test		X	X
Walking speed 10 meter		X	
Knee morphology	X	X	X
Joint effusion	X	X	X
Increased local temperature		X	X
Local redness		X	X
Baker's cyst		X	X
Muscle strength (MRC)	X	X	X
Leg length discrepancy	X	X	X
Scales			
FACHS		X	
WOMAC		X	
KOOS		X	
HAD		X	
GADS		X	
Social participation questionnaire		X	
Radiographic data			
Leg-length inequality		X	
Knee alignment	X	X	X
KL	X	X	X
Patellofemoral lateral angle		X	
Lateral deviation patella		X	
Congruence angle		X	
MRI data	X	X	
Blood samples	X	X	X
Urine samples		X	
Faeces samples		X	
Biomechanical gait analysis data	X	X	

4. OACTIVEs database validation in diagnosis task

Data comprehension analysis

The datasets given to be analysed are two. The first dataset (**allData**) contains all the data collected by the OACTIVE clinical trials. It consists of 33 variables and 236 observations. In this data set there are incomplete/missing data obtained from the ANIMUS database.

The missingness mechanisms are three Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). MCAR mechanism represents a situation where missing data happen entirely by chance. MAR mechanism is considerably weaker than the MCAR. According to this assumption, the probability to be missing depends on the observed data only. More specifically, the observed data can describe the mechanism that leads to missing data. MNAR is a term used to describe a situation where the mechanism that causes missing data is related to unobserved measurements. In other words, the unobserved measurements influence the process governing missingness, even after controlling for observed measurements. The "allData" dataset consists generally of data belonging to the MNAR mechanism. There is a small insignificant amount of data belonging to the other two mechanisms.

For the "allData" dataset a complete pre-process analysis will be applied, including imputation of the missing values with the vtreat imputation methodology, significance pruning of the data and scaling of the data based on the response variable.

The second (**reducedData**) dataset contains all the data, except the observations of the Animus database that are missing, collected by the OACTIVE clinical trials. It consists of 33 variables and 211 observations. "reducedData" is a pre-processed dataset applying scaling and correlation analysis of the data.

From the initial analysis of the "allData" dataset, it was determined that from the 35 variables, 27 variables are categorical and the remaining 6 are numerical variables. Similarly, in the "reducedData" dataset, it was determined that from the 33 variables, 29 were categorical and the remaining 5 are numerical variables.

The response variable, for both datasets, is a two classes categorical variable. The first class, represented with "0", is the class of a person without osteoarthritis which will not present osteoarthritis in the near future. The second class, represented with "1", is the class with a person who shortly will present osteoarthritis in one or both of his knees. From now on the second class represented with "1" will be considered as the positive class.

The first step of our analysis is a correlation analysis of the variables, for both datasets. To produce correlation analysis the following process was implemented. A mixed association correlation was used for the calculation of the association of the variables. The strength of association is calculated for categorical vs categorical with a bias-corrected Cramer's V, numeric vs numeric with Spearman correlation, and categorical vs numeric with ANOVA. To apply the above process, it was ensured that the correlation of the variables with themselves is excluded.

Correlation Analysis

The "allData" dataset variables' correlation analysis showed that there is a small number of highly correlated variables inside the dataset. In Figure 1 we can see how many variable pairs are correlated (positively or negatively) and to what degree.

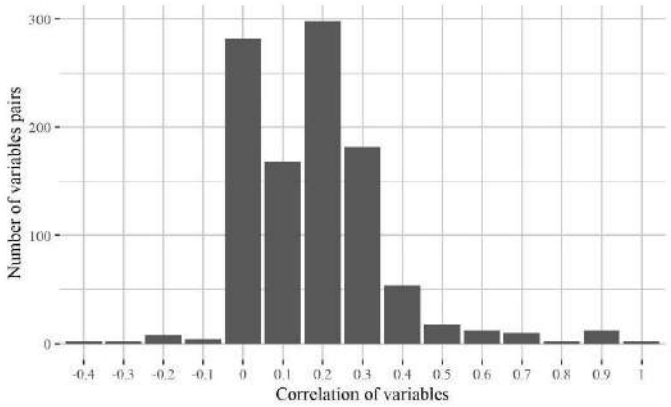


Figure 0. *allData correlation analysis.*

Figure 2 shows a graphical representation of the correlation for the aforementioned analysis. This graphical representation shows with blue colour the positive correlation of the variables (nodes), also the intensity of the colour depends on the degree of correlation between the two variables. Red colour, with the same logic as blue, represents the negative correlation.

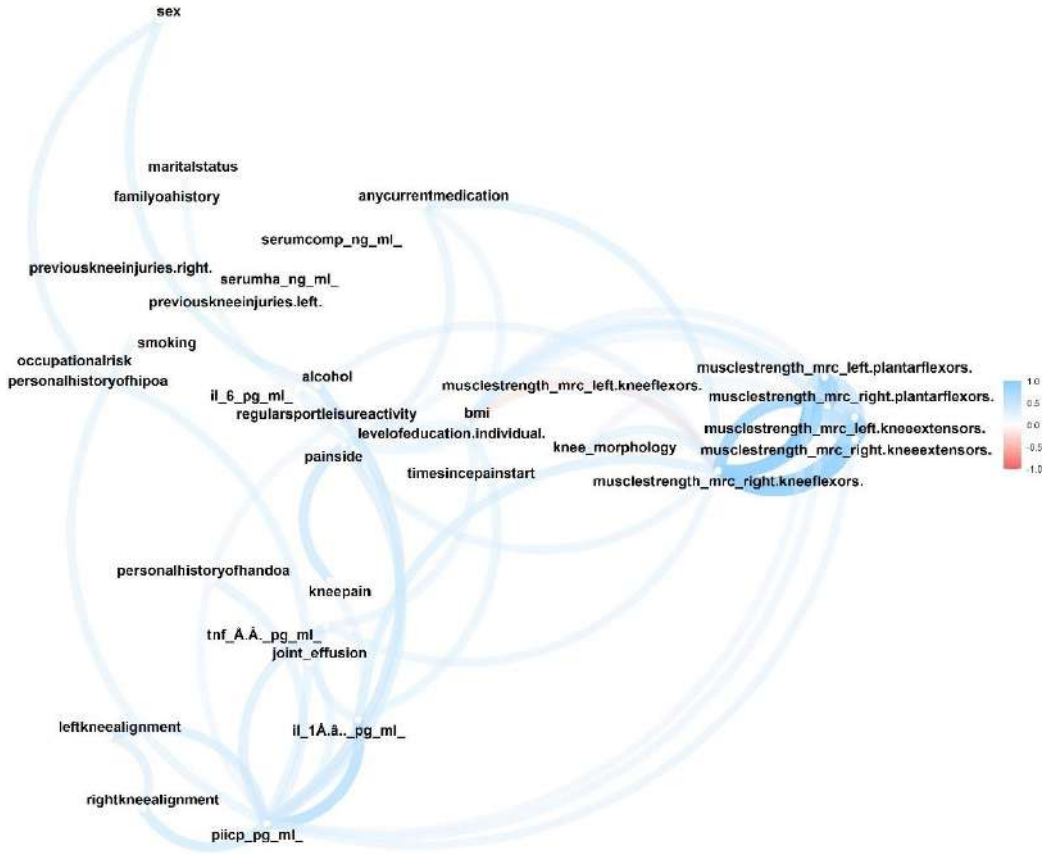


Figure 2. *allData correlation map.*

The "reducedData" dataset variables' correlation analysis, similarly, showed that there is a small number of highly correlated variables inside the dataset. In Figure 3 we can see how many variable pairs are correlated (positively or negatively) and to what degree.

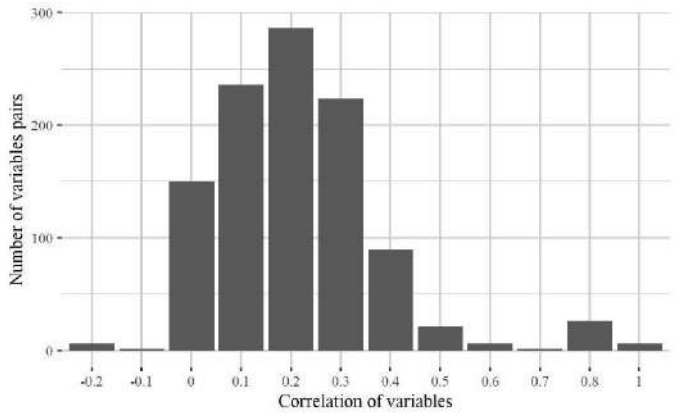


Figure 3. *reducedData correlation analysis.*

Figure 4 shows a graphical representation of the correlation for the aforementioned analysis.

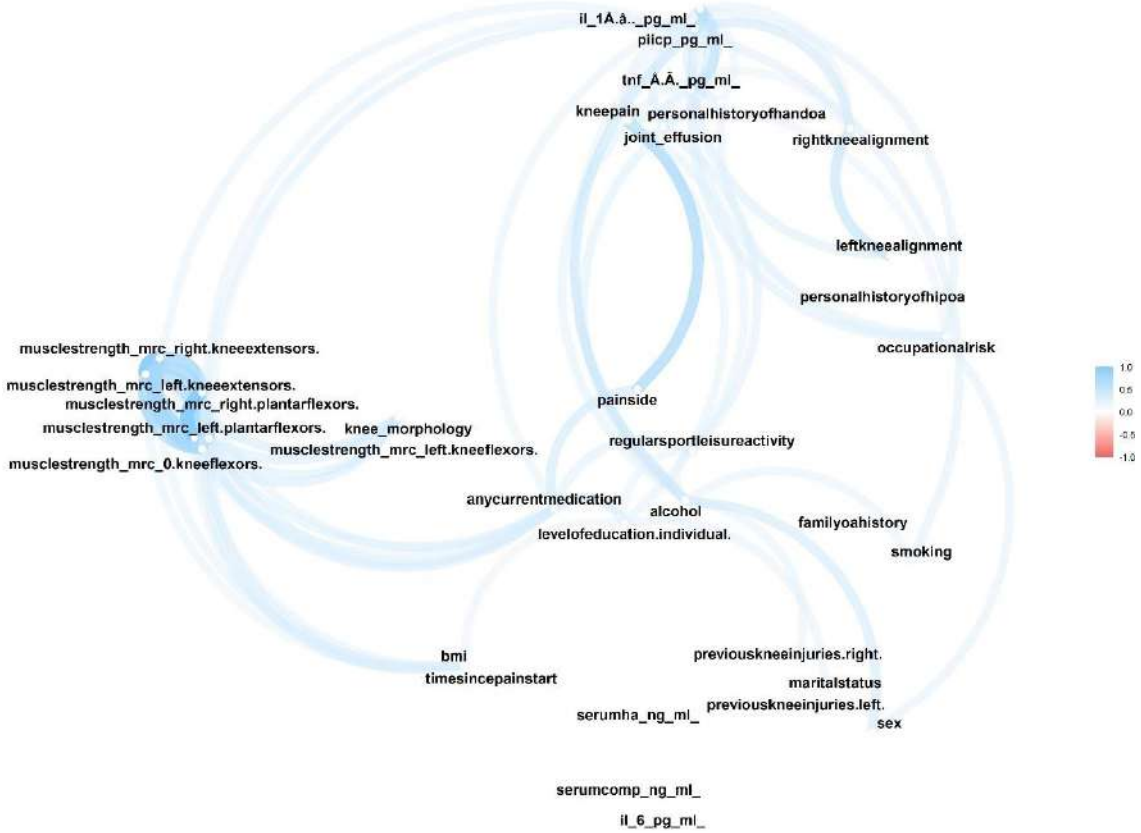


Figure 4. *reducedData correlation map.*

Heatmap analysis

One important step in the data comprehension for both datasets is the visualization of the data in a heatmap. The goal behind this analysis is the comprehension of the unique characteristics of each variable with respect to the response variable.

Figures 5 and 6 represents the heatmap of the "allData" and "reducedData" datasets, respectively. The lowest values are represented with blue and the highest value is represented with the colour red. The horizontal line in the middle of the heatmap separates the response variable's two classes. Above is represented the class "0" and below is represented the class "1". The conclusion from this heatmap is that the variables can separate well enough the two classes of the response variable.

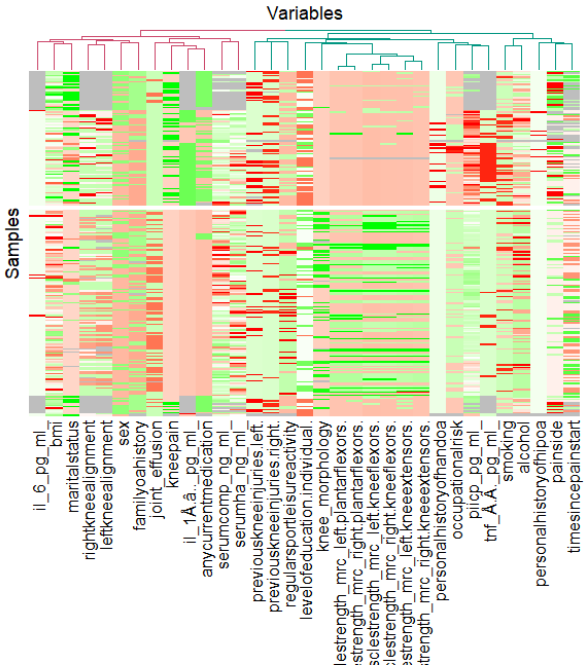


Figure 5. *allData heatmap.*

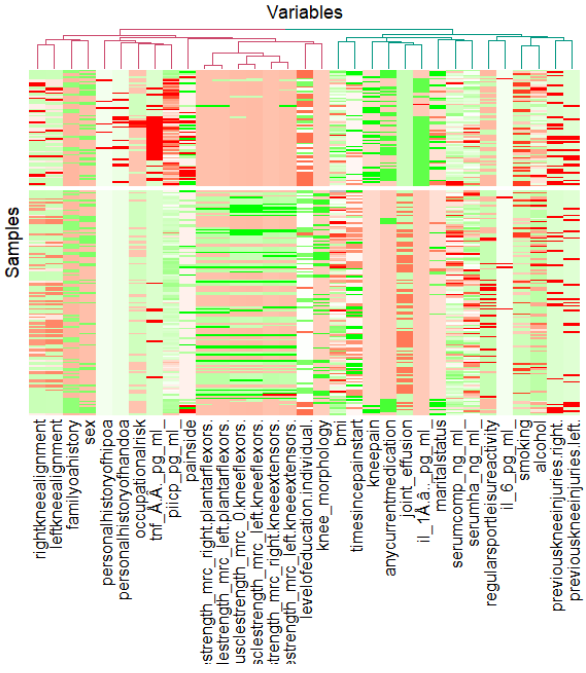


Figure 6. *reducedData heatmap.*

Variable Importance

The last step in the data comprehension analysis is the Variable Importance Analysis with the use of the Random Forests Algorithm (RF) and the Mean Decrease in Accuracy (MDA) measure. The resulting variables' selection was NOT used in the analysis. Its purpose is a better understanding of the data set.

To calculate the MDA with the RF algorithm the permuted out-of-bag (OOB) data where used. Specifically, by recording the prediction error on the OOB portion of data, for each tree. The same

process is repeated after permuting each predictor variable. The difference between the two (Decreases in Accuracy of Trees) is then averaged over all trees and normalized by the standard deviation of the differences.

$$MDA = \frac{Mean(Decreases\ in\ Accuracy\ of\ Trees)}{StandardDeviation(Decreases\ in\ Accuracy\ of\ Trees)}$$

For the calculation of the importance for the "allData" and "reducedData" datasets' variables, a RF model was created with the use of each datasets' raw variables. This model creation was possible because RF is a decision tree-based algorithm. This means that there is no need to convert the categorical variables to numeric with the use of dummy variables.

That way, the following figure (Figure 7A for "allData" and 7B for "reducedData" datasets) was created. In this figure, the variables with the highest MDA score, are presented in descending order, from the most important to the less important.

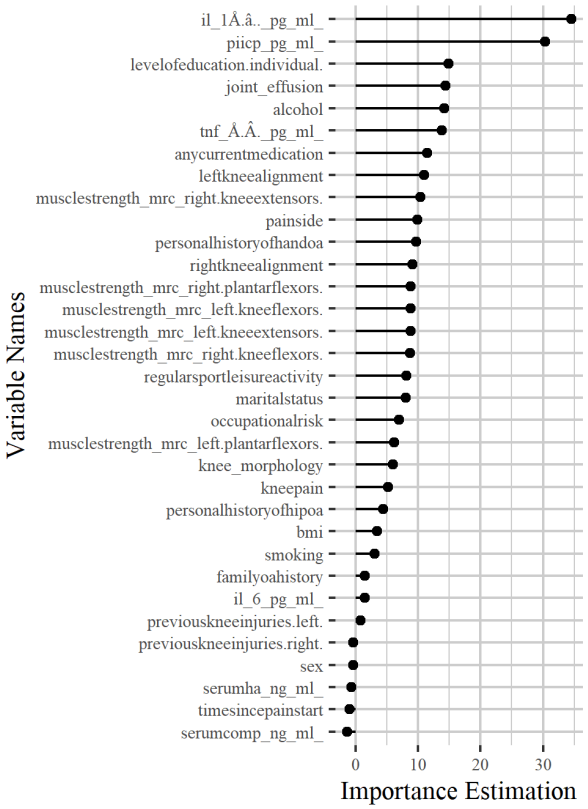


Figure 7A. *allData variable importance.*

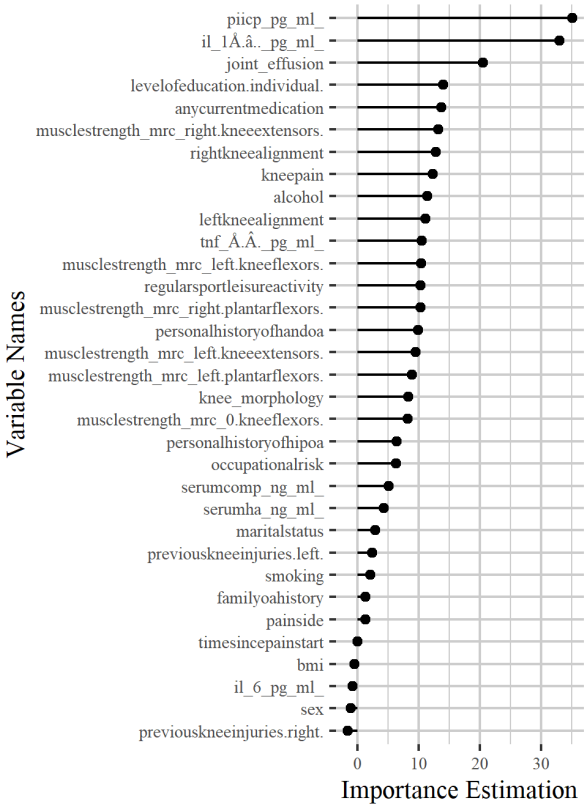


Figure 7B. *reducedData variable importance.*

Data pre-process

Imputation and Dummy Variables Process

The first step in every pre-process procedure is the treatment of the missing values. There are several methodologies for this procedure. The process selected in this case came in the form of the vtreat package. Vtreat is a methodology specialized for missing values under the missingness mechanism MNAR.

For the "allData" dataset the imputation method is the following. For the categorical variables, the well-known dummy variables process was applied to treat the missing values. With this process the categorical variables are converted to numerical, so they can be processed correctly in the classification tasks below. In this procedure, the missing values of the categorical variables were treated as part of the levels of the categorical variable. As a result, one additional level for each categorical variable was created. For the numerical variables, the classical mean substitution was applied. Additional to this procedure vtreat creates a further indicator for missing values, highlighting with 1 or 0 the occurrence or non-occurrence of missing values, a feature that enhances the knowledge discovery on ultra-noisy data. Concerning the categorical variables, vtreat creates further variables along with the well-known dummy variables process. More specifically, one more dummy variable is created to address the high cardinality categorical variables and two dummy variables are created to cope with the novel levels found in each categorical variable of the dataset.

Contrary, the "reducedData" dataset is already treated and contains no missing values. So, the procedure of the dummy variable creation and the process vtreat that was applied had the result to ignore some of the extra indicators created through the following process of significance pruning.

Significance Pruning

After imputation, each variable is evaluated based on its correlation with the response variable, for the isolation of the most significant variables. The idea behind this step is based on the fact that machine learning methods may be adversely affected by large numbers of irrelevant variables. More specific, we applied the treatment plan of vtreat package by decoding and removing the noisy variables. For this purpose, we estimated their correlation with the response variable using a X^2 Test. Then, we calculated the significance of each predictor variable using X^2-statistic on the logistic model between the response variable and each predictor variable.

Secondly, it prunes the variables with a higher significance value than the threshold $1/n\_vars$ in both cases, where n_vars indicates the resulted variable obtained from the matrix imputation step.

Data normalization

For the normalization of each dataset the following logistic regression model was created for each predictor variable:

$$\log_b \frac{p}{1-p} = mx + b$$

where $p = P[y = TRUE]$ and y is the response variable, x is a predictor variable, m is the logistic regression coefficient and b is the intercept. Afterwards, normalization of each predictor variable was done with the following procedure:

$$x' = mx - mean(mx)$$

That way each time we express the predictor variables in the response variable's units (scaling) and centered at zero mean value.

Visualization of the data

For the visualization of the data the Principal Components Analysis was utilized. The projection of the data on the first 2 principal components is given in the figure bellow for both datasets (Figure 8 for "allData" and Figure 9**Figure** for "reducedData" datasets).

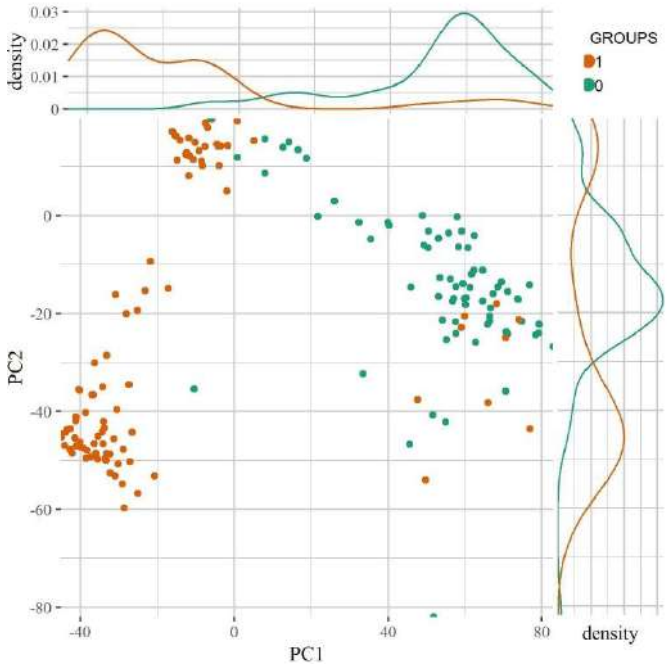


Figure 8. *allData PCA visualization in 2 principal components.*

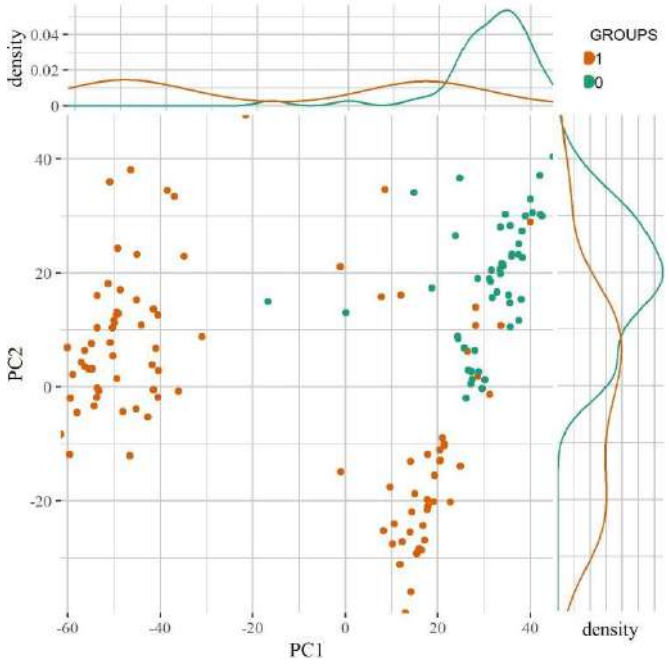


Figure 9. *reducedData PCA visualization in 2 principal components.*

The visualization of the data was applied on the data treated by the vtreat package. As it is immediately apparent from the representation of the data, the Classes "0" and "1" are segregated enough visually, especially in the axis of the first principal component in both cases. The interesting thing in these two visualizations is the presence of some outliers in both classes. Also, it is noteworthy that with the naked

eye one more class in the response of the data appeared, in both cases. This fact is interesting for a more comprehensive analysis.

Classification tasks

The classification algorithms used for this data set are the Multinomial Logistic Regression (MLR), the Linear Discriminant Analysis (LDA), the k-Nearest Neighbours (kNN), the Random Forests (RF) and finally the XGBoost. The results of the classifications can be seen on Table 1 for "allData" and Table 2 for "reducedData" datasets. The measures used are accuracy, sensitivity and specificity.

Table 1. *allData results table.*

	Accuracy	sensitivity	specificity
MLR	0.878	0.889	0.857
LDA	**0.975**	0.962	**1**
kNN	0.902	0.889	0.928
RF	0.658	**1**	0
XGBoost	0.902	0.889	0.928

Table 2. *resucedData results table.*

	Accuracy	sensitivity	specificity
MLR	0.889	0.909	0.857
LDA	**1**	**1**	**1**
kNN	0.972	0.954	**1**
RF	0.611	**1**	0
XGBoost	0.944	0.909	**1**

As it becomes apparent from the results of all the classification algorithms except for RF, which is the worst by far for this classification task, there is a very high classification accuracy. Second, to last is the kNN algorithm and the best classification of this task is implemented by the model of the LDA. The Ro Curves are presented below illustrating the results:

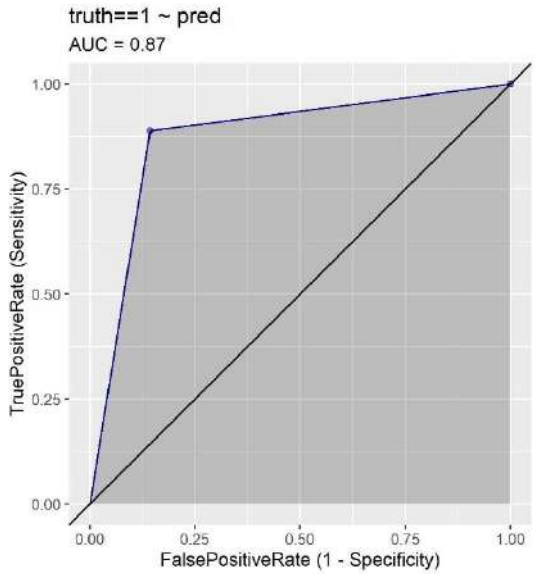


Figure 10. *allData Ro Curve.*

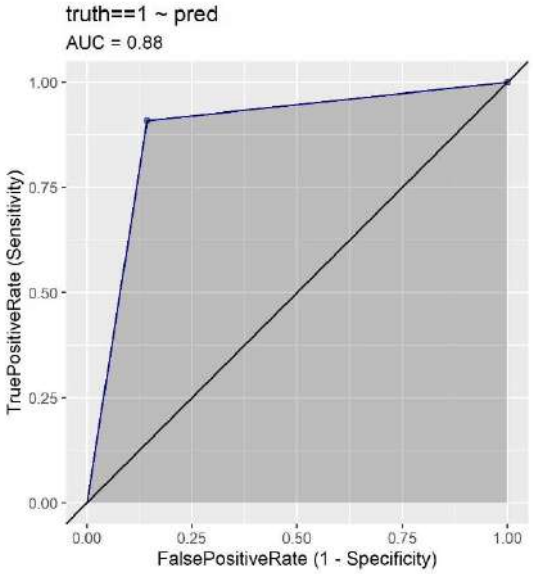


Figure 11. *reducedData Ro Curve.*

- **MLR**

- **LDA**

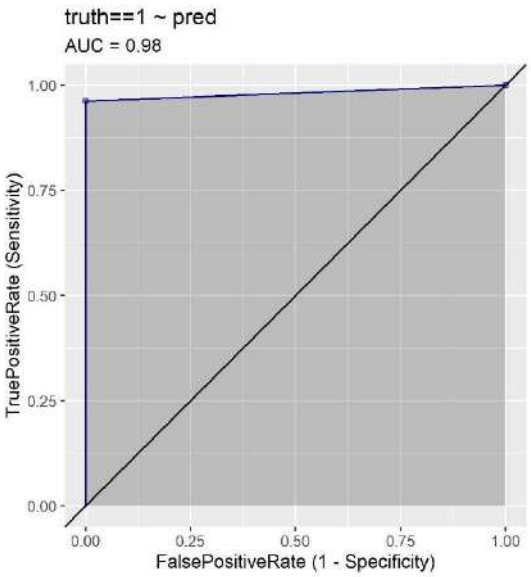


Figure 12. *allData Ro Curve.*

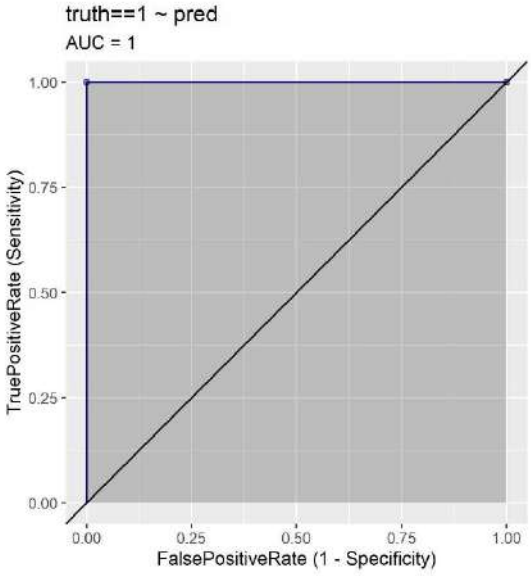


Figure 13. *reducedData Ro Curve.*

- **KNN**

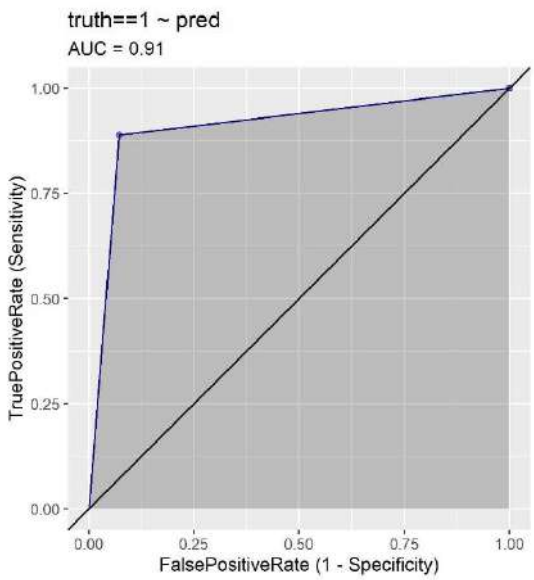


Figure 14. *allData Ro Curve.*

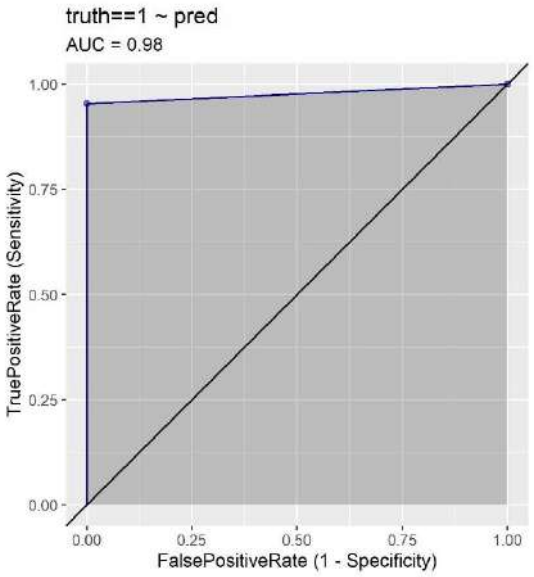


Figure 15. *reducedData Ro Curve.*

- **RF**

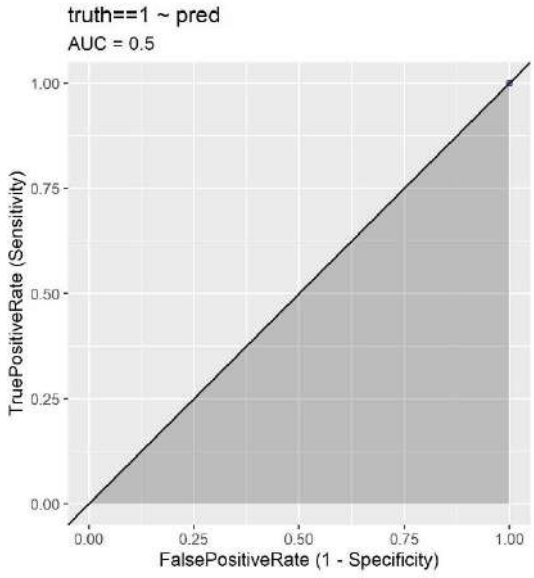


Figure 16. *allData Ro Curve.*

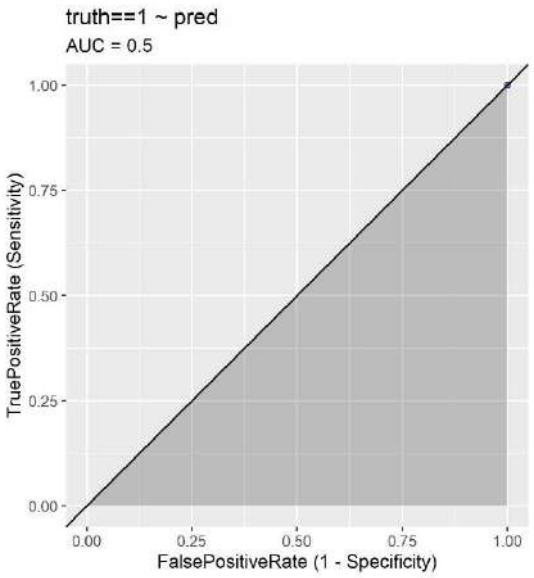


Figure 17. *reducedData Ro Curve.*

- **XGBoost**

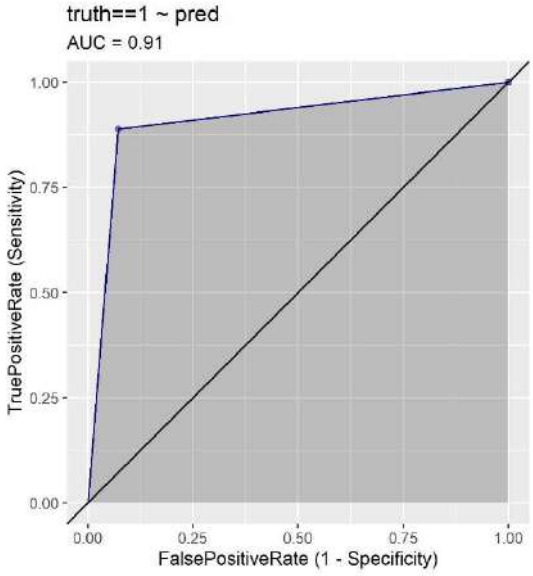


Figure 18. *allData Ro Curve.*

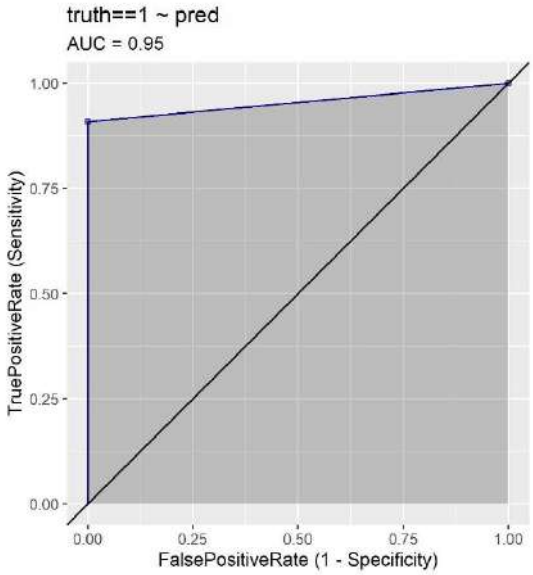


Figure 19. *reducedData Ro Curve.*

The extensive analysis presented in this report shows us the high quality of the OACTIVE's database. The findings of this analysis gave us valuable insight in this database. We know that it is a highly separable dataset even in the presence of missing data. The previous claim can be supported also from the classification models application, which has the result of classification accuracy 1 in the case of the "reducedData" and 0.975 in the case of the "allData". Finally, it is interesting for future investigation the reason behind the difference in all the examined metrics of the Random Forests and the XGBoost algorithms.

5. Interpretable models based on OACTIVE data

The analysis in this summary initially covers the OAI model validated on OACTIVE data, in work during the current review period. The next step in the analysis is OACTIVE model with OAI validation. As the OACTIVE dataset contains information on different variable cohorts, it is logical to try to leverage the information in these groups. Each variable group is considered separately for its own power to determine the presence of KOA. Following from individual analysis, the next step is to combine the variable groups, and perform feature selection on the variables to determine the most influential factors in disease presence at baseline.

Due to the way the data was collected in the cohorts of varying inclusion criteria, the OACTIVE data contains specificity to particular cohort definitions, which have skewed some of the validation results.

Variable Cohort Development

Both internal and external features contribute to the development of KOA. To have a model that can make use of the information about a person, their life, movements and biochemistry would give new insights into ways to earlier identify the onset or presence of disease. The OACTIVE dataset contains information on these variable sets.

The first step is to analyse each variable set individually, and then to incorporate all features into a model and use feature selection to establish the most influential features in identifying the presence of KOA.

This will then give a 'hybrid' model that we will compare the performance to each set individually to determine if the model performance is improved by the addition of the additional variable sets.

Demographic Variables

Usable cohort size: n = 206

Table 3. *Demographic variables*

Variable	*Definition*
Knee_swell	In the past 7 days, has the subject had knee swelling? Yes/ No
Gender	What is the subject's gender? Male/ Female
BMI overweight	What is the subject's BMI? Less than 25/ 25 or more
AGE	How old is the subject at the assessment? 45-50/ 50-55/ 55-60/ 60-65/ 65+

Outcome:

No KOA: n = 72

KOA: n = 134

Biochemical Variables

Blood

Usable cohort size: n = 217

Table 4. *Biochemical variables (Blood)*

Variable	*Definition*
HA	Hyaluronic Acid
COMP	Cartilage oligomeric matrix protein
PIICP	Procollagen type II C-terminal propeptide

Outcome:

No KOA: n = 81

KOA: n = 136

Urine

Usable cohort size: n = 73

Table 5. *Biochemical variables (Urine)*

Variable	*Definition*
HA	Hyaluronic Acid
COMP	Cartilage oligomeric matrix protein
PIICP	Procollagen type II C-terminal propeptide
IL-1β	Interleukin 1 beta
IL- 6	Interleukin 6
TNF-a	Tumour Necrosis Factor Alpha

Outcome:

No KOA: n = 71

KOA: n = 2

Due to the small incidence of KOA in the dataset any model built with these features would naturally tend to the majority class and would therefore not be clinically useful for diagnosing KOA.

Biomechanical Variables

Usable cohort size: n = 84

Table 6. *Biomechanical variables*

Variable	*Definition*
KCF1_Walk	1st peak knee contact force – Walking
KCF2_Walk	2nd peak knee contact force – Walking
KAM1_Walk	1st peak knee adduction moment – Walking
KAM2_Walk	2nd peak knee adduction moment – Walking
KCF1_Step	1st peak knee contact force – Stepping
KCF2_Step	2nd peak knee contact force – Stepping
KAM1_Step	1st peak knee adduction moment – Stepping
KAM2_Step	2nd peak knee adduction moment – Stepping
KF_Walk	Peak knee flexion – Walking
KF_Step	Peak knee flexion – Stepping

Outcome:
No KOA: n = 64
KOA: n = 20

Socioeconomic Variables

Usable cohort size: n = 126

Table 7. *Socioeconomic variables*

Variable	*Definition*
Marital Status	Status of subject relationship
Residency	Status of living situation of subject
Household Income	How easily the subject gets by
Housing Status	Whether the subject owns or rents their property
Subject level of Education	The highest level of education a subject reached
Parent level of education	The highest level of education a subject's parents reached

Outcome:
No KOA: n = 107
KOA: n = 19

All Variable Cohort Hybrid

This cohort contains data across demographic, biochemical, biomechanical and socioeconomic.

Data size after all cohorts merged: n = 33

All subjects have no KOA at baseline. This dataset is not usable. To create a diagnostic model both outcomes need to be present in the data, and due to the usable cohort only containing subjects in one class no diagnostic modelling can take place.

When considering the isolated predictions of these 33 subjects, the biomechanical model correctly classified all subjects as no KOA, the socioeconomic model was correct on 32 subjects whilst the biochemical model was correct on 27 subjects. The variation between the actual and predicted presence

of KOA in the socioeconomic and biochemical models could be due to other features that are not considered in these models.

Hybrid Model 2

This cohort contains data from demographic and biochemical variable cohorts.

Data size after selected cohorts are merged: n = 194

Outcomes:
No KOA: n = 64
KOA: n = 130

Hybrid Model 3

This cohort contains data from demographic and biomechanical variable cohorts.

Data size after selected cohorts are merged: n = 45

Outcomes:
No KOA: n = 8
KOA: n = 37

Demographic Variable Model

This model was developed for use in a clinical setting as a point to aid clinicians with both patient education and signposting. This model took into account the patient's activity, symptoms and demographic information to determine if KOA was present at that point.

Validation of OAI model on OACTIVE Data

As the model was built using data from the OAI dataset, it includes variables that are not present in the OActive data. The common variables are shown in bold in Table .

***Table* 8.** *Variables in the OAI risk model for Propensity of Presenting.*

Variable	*Definition*
knee_stiff_day_limit	In the last 30 days, how many days has the subject limited activity due to knee pain/aching/stiffness? 0 days/ 1-7 days/ 7-14 days/ 14-21 days/ 21 or more days
Diff_upstr	Does the subject have difficulty getting upstairs? Yes/ No
Knee_swell	In the past 7 days, has the subject had knee swelling? Yes/ No
P01KPACT30	In the past 30 days, has the subject limited their activity due to knee pain? Yes/ No
Gender	What is the subject's gender? Male/ Female
B.LINE_SYMP	Does the subject present with knee pain today? Yes/ No
BMI overweight	What is the subject's BMI? Less than 25/ 25 or more
AGE	How old is the subject at the assessment? 45-50/ 50-55/ 55-60/ 60-65/ 65+

The variables missing were marginalised. To do this, each subject in the OActive data set was categorised into one of the 8 combinations of the binary variables in the model (Knee_swell, Gender and BMI overweight) and 5 Age bands, 40 possibilities in all, and the predictions of the OAI model were averaged over the training data filtered into the same combination.

The mean model predictions for the specific values of the four variables present in each row in the data were then used to calculate the AUROC for the OActive data (n=188).

The OActive data cohort is created as described in Figure .

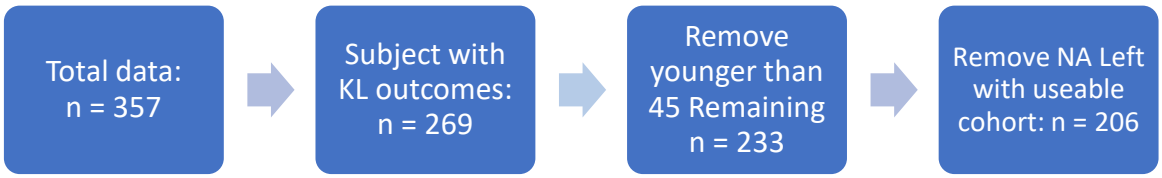


***Figure* 20.** *Visualisation of data cohort creation.*

OAI Test Data results

The reference result for the predictions on OActive data using the model developed with OIA data is the ROC curve for test data in OAI. This is shown in fig. 1. The ROC curve for the OActive data is shown in Figure 11.

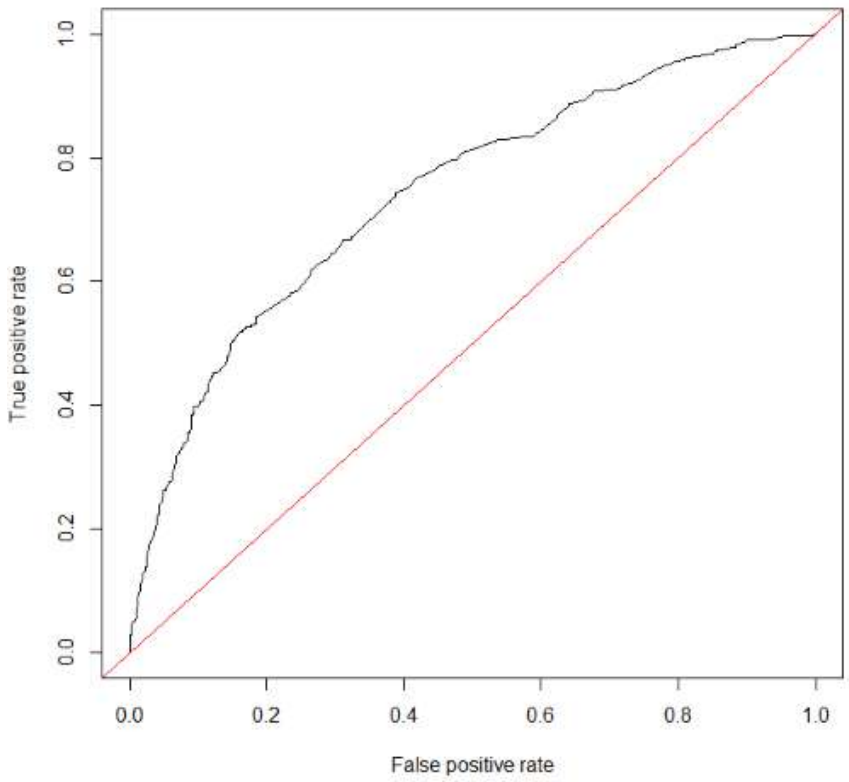


Figure 11. *ROC curve from the LogR model with the test OAI data. The AUROC for this curve is 0.7421.*

Table 9. *Confusion matrix.*

	Actual	
	0	1

Predictions	0	694	261
	1	127	272

Table 10. *Performance metrics for the OAI test data.*

Measure	*Value*
Accuracy	0.7134
Sensitivity	0.5103
Specificity	0.8453
PPV	0.7674
NPV	0.6332
AUROC [CI]	0.7421 [0.7152, 0.769]

OActive Data Validation results

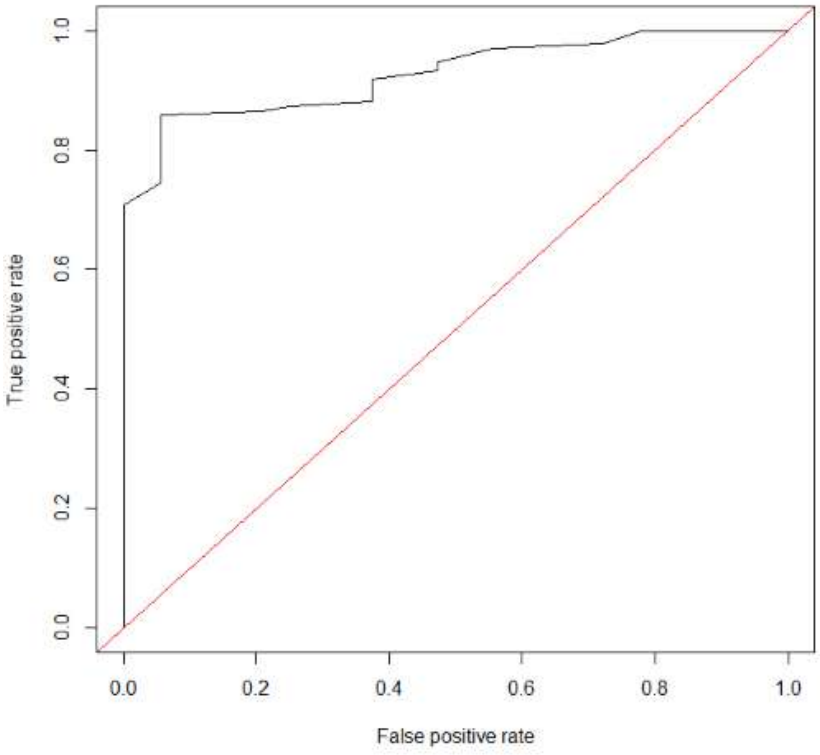


Figure 22. *ROC curve from the validation using the OACTIVE data, with the predictions calculated using the mean value of the corresponding predictions from the OAI training set. The AUROC is 0.9273.*

Table 11. *Confusion matrix.*

		Actual	
		0	1
Predictions	0	72	68
	1	0	66

Table 12. *Performance metrics for the OActive data.*

Measure	*Value*
Accuracy	0.6699

Sensitivity	0. 4925
Specificity	1.0000
PPV	1.0000
NPV	0.6634
AUROC [CI]	0.9273 [0.8938, 0.9608]

Misclassified Cases

In the OActive data, there is a high number of subjects who have been misclassified, as there are 68 cases when a subject with KOA is predicted to not have the disease out of the sample of 206 subjects. There is a level of misclassification in the OAI data also, with 398 subjects of the 1354 receiving a prediction that does not match with the diagnosis. Of the 398 misclassified cases, 261 are deemed to have no KOA when they do, in fact, have KOA.

OAI Test data

Looking closer at the test data gives an insight at the way the model predicts, and using the test data, it is possible to consider how all covariates, even those not present in the OACTIVE data contribute to the final prediction.

The following table considers only the cohort that have been misclassified as no KOA when there is evidence via x-ray that the subject has KOA.

Table 13. *Misclassification summary, no KOA when KOA is present.*

Cohort	*Insight*
143/261	No issue getting upstairs
212/261	No reported knee swelling in last 7 days
257/261	No pain/swelling/stiffness on the day of the baseline assessment (baseline symptoms)
222/261	Made no changes to their activity in the last 30 days as a result of knee pain
222/261	Not had to limit their daily activity due to knee stiffness
192/261	Have a BMI 25+
140/261	Are women.

Based on this finding, it is possible that the discrepancy between diagnosis and prediction falls in the group of symptomatic vs radiographic knee osteoarthritis.

Symptomatic OA is when a person experiences symptom, such as joint pain, aching and stiffness. Radiographic OA is found by observing features on an x-ray that suggest OA development. It is possible to have symptomatic OA without radiographic OA and vice versa. Up to 60% of people with radiographic KOA may not complain of symptoms. Sometimes, the lack of symptoms is backed up with less severe radiographic OA.

OACTIVE Validation

As the OACTIVE data does not contain all of the same variables as the OAI data, several assumptions relating to the predictions are made. By only considering gender, BMI, age and the presence of knee swelling, other symptoms can be left out entirely from the prediction model.

The OACTIVE data is collected across three centres; however, after removing missing values and selecting the variables required only data from two centres remain. The following tables show how the data is made up.

Table 14. *Data classification, actual and predicted by recruitment centre in the OActive study.*

Actual OA Classes

	0	**1**
ANIMUS	1	1
HULAFE	71	5
UNIC	0	128

Predicted OA Classes

	0	**1**
ANIMUS	2	0
HULAFE	76	0
UNIC	62	66

In the cases for the centre UNIC, 62 subjects should be predicted to have KOA, and the model has predicted no KOA. Similarly, HULAFE had five cases that should be KOA and the model has predicted no KOA.

Those subjects who were misclassified are likely to not suffer with symptoms, based on the small amount of data available, and therefore would not have symptomatic KOA, only radiographic KOA.

The only symptomatic variable present in the subset of the OActive data is *knee_swell*. This is self-reported and asks if the subject has experienced knee swelling in the last 7 days. Of the 68 subjects that have been misclassified from the OActive data, 63 subjects presented with no knee swelling in the previous 7 days. This may be an influential factor in the model determining between KOA and no KOA. It is worth noting that the variables that are not present in the OActive subset might make the difference in changing cases that were misclassified to class 1, as they hold predictive information about the subjects.

Because of this, one limitation using the OActive data is that the model can identify symptomatic KOA, but not radiographic KOA.

Validation of OACTIVE model on OAI Data

A logistic regression model using only the common variables of Age, Sex, knee swelling and BMI has been trained using OACTIVE data and tested with the OAI dataset.

Among the 206 subjects in the OACTIVE data 72 have no KOA at baseline, and the remaining 134 people have clinical KOA at the baseline assessment. The data was split into training and test sets, with the training set containing 104 subjects and the test set 102 subjects. Results are recorded for the test set only.

Table 15. *Results of logistic regression model.*

VARIABLE	**BETA**	**P-VALUE**
INTERCEPT	-2.80438	0.0712
AGE2 (50-55)	-0.08459	0.9550
AGE3 (55-60)	0.89955	0.5236
AGE4 (60-65)	2.84957	0.0575
AGE5 (65+)	23.00011	0.9950
BMI_BINS1 (BMI 25+)	0.44579	0.7198
SEX1 (FEMALE)	-0.10666	0.9228
KNEE_SWELL1 (YES)	20.62507	0.9959

Training Data **Test Data**

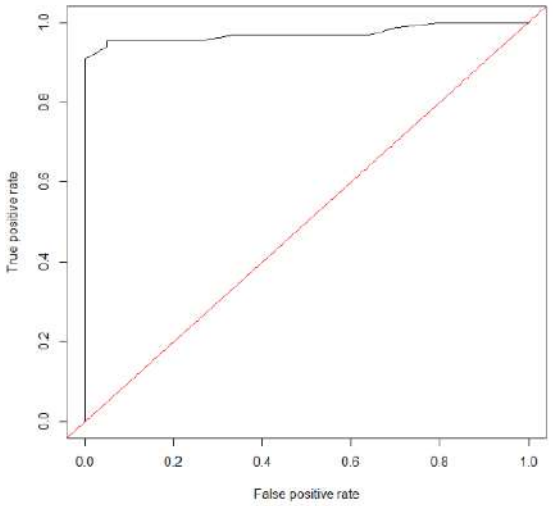


Figure 23. *The ROC curve from using the OACTIVE training data to test the model developed using the OACTIVE data. The AUC on the test data is 0.9722.*

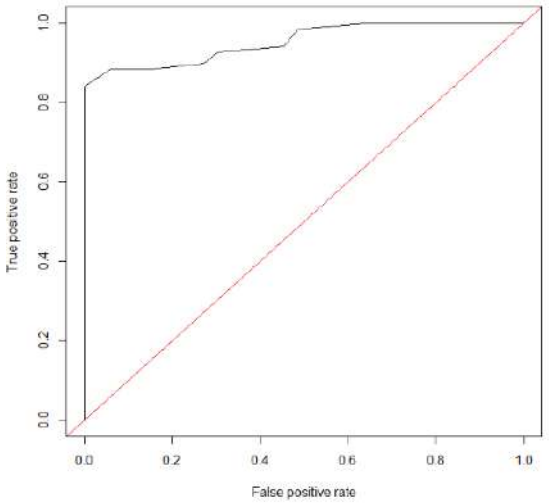


Figure 24. *The ROC curve from using the OACTIVE test data to test the model developed using the OACTIVE data. The AUC on the test data is 0.9532.*

Table 16. *Confusion matrix for the OActive model, tested on the OActive training data.*

		Actual	
		0	1
Predictions	0	37	4
	1	2	61

Table 17. *Confusion matrix for the OActive model, tested on the OActive test data.*

		Actual	
		0	1
Predictions	0	31	8
	1	2	61

Table 18. *Performance metrics for the OActive training data.*

Measure	Value
Accuracy	0.9423
Sensitivity	0.9385
Specificity	0.9487
PPV	0.9482
NPV	0.9391
AUROC	0.9722 [0.9406 – 1]

Table 19. *Performance metrics for the OActive test data.*

Measure	Value
Accuracy	0.902
Sensitivity	0.8841
Specificity	0.9394
PPV	0.9358
NPV	0.8901
AUROC	0.9532 [0.9176 – 0.9889]

The model was applied to the OAI dataset (n=2707) of whom 1627 do not present with clinical KOA at baseline and the remaining 1080 people do. This produced the curve in Figure 22, which is naturally less predictive than the full OIA model with 7 variables applied to the test data.

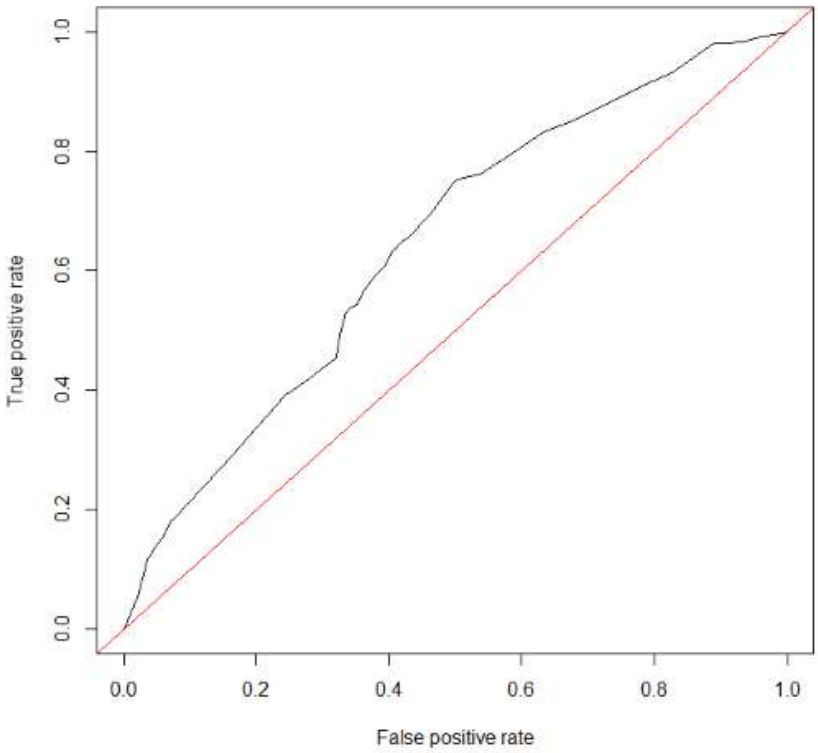


Figure 22. *The ROC curve from using the OAI data to validate the model developed using the OActive data, showing an AUC of 0.6567.*

Table 20. *Confusion matrix for the OActive model, validated on the OAI data.*

		Actual	
		0	1
Predictions	0	808	267
	1	819	813

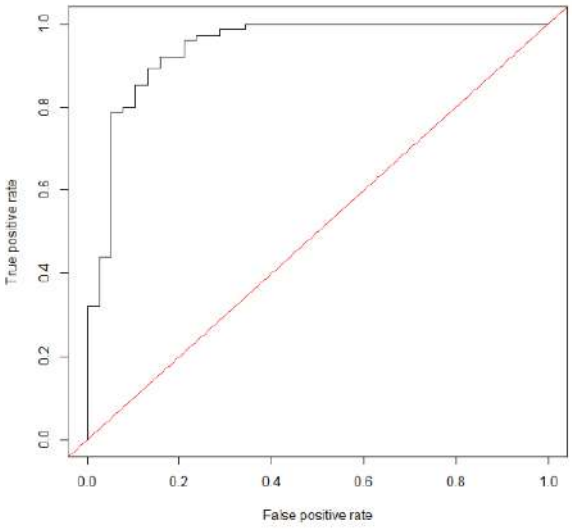
Table 21. *Performance metrics for the OAI validation data*

Measure	Value
Accuracy	0.5988
Sensitivity	0.7528
Specificity	0.4966
PPV	0.5993
NPV	0.6676
AUROC	0.6457 [0.6249 – 0.6665]

Biochemical variable model

Variables in model: COMP, HA, PIICP

Training Data
Training size: 113



The training ROC curve from using the biochemical model.

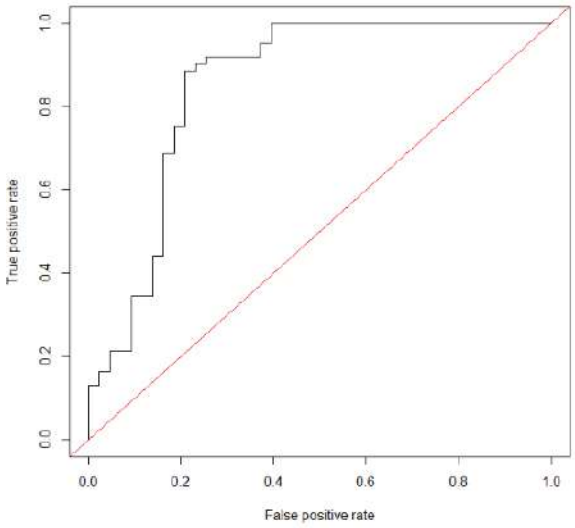
Confusion matrix for the training data.

		Actual	
		0	1
Predictions	0	31	267
	1	7	69

Performance metrics on training data

Measure	*Value*
Accuracy	0.885
Sensitivity	0.920
Specificity	0.8158
PPV	0.8332
NPV	0.9107
AUROC	0.9425 [0.8934 – 0.9915]

Test Data
Test size: 104



The test ROC curve from using the biochemical model.

Confusion matrix for the test data.

		Actual	
		0	1
Predictions	0	30	5
	1	13	56

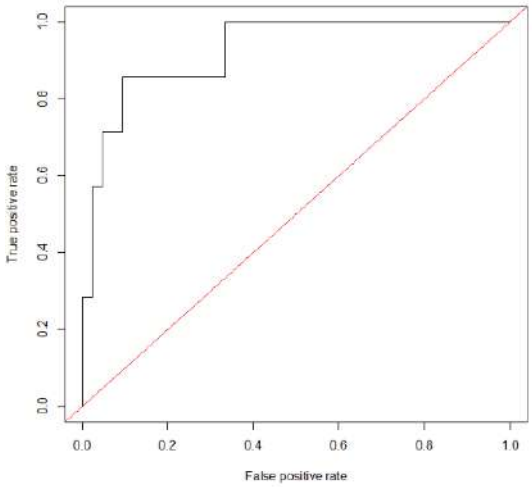
Performance metrics on test data

Measure	*Value*
Accuracy	0.8269
Sensitivity	0.9180
Specificity	0.6977
PPV	0.7523
NPV	0.8949
AUROC	0.8517 [0.7663 – 0.9371]

Biomechanical variable model

Variables in model: KCF1_Step, KCF2_Step, KCF1_Walk, KAM1_Walk, KAM1_Step

Training Data
Training size: 49



The training ROC curve from using the biomechanical model.

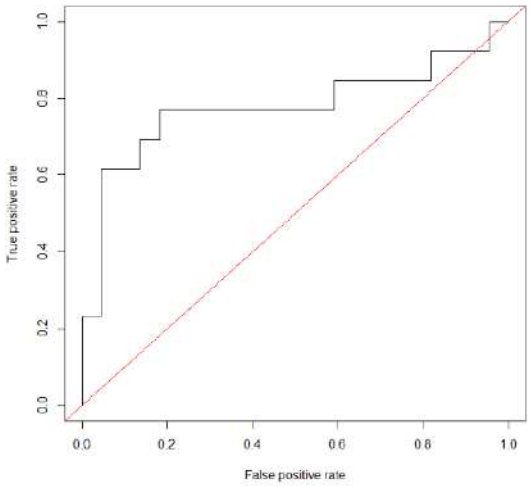
Confusion matrix for the training data.

		Actual	
		0	1
Predictions	0	41	3
	1	1	4

Performance metrics on training data

Measure	Value
Accuracy	0.9184
Sensitivity	0.57143
Specificity	0.97619
PPV	0.96
NPV	0.69492
AUROC	0.9252 [0.8274 – 1]

Test Data
Test size: 35



The test ROC curve from using the biomechanical model.

Confusion matrix for the test data.

		Actual	
		0	1
Predictions	0	22	10
	1	0	3

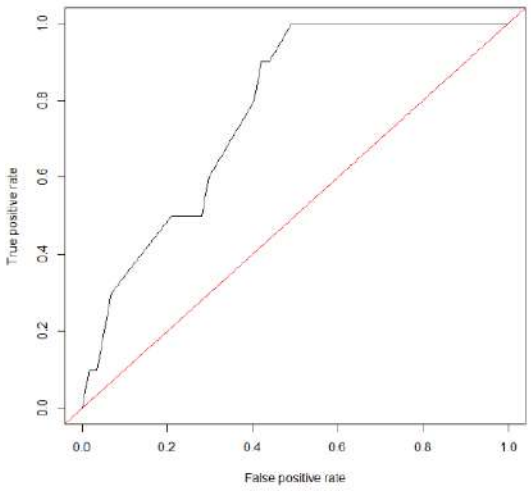
Performance metrics on test data

Measure	Value
Accuracy	0.7143
Sensitivity	0.23077
Specificity	1
PPV	1
NPV	0.56522
AUROC	0.7762 [0.5835 – 0.9689]

Socioeconomic variable model

Variables in model: household income, residency, level of education, parent's level of education

Training Data
Training size: 67



The training ROC curve from using the socioeconomic model.

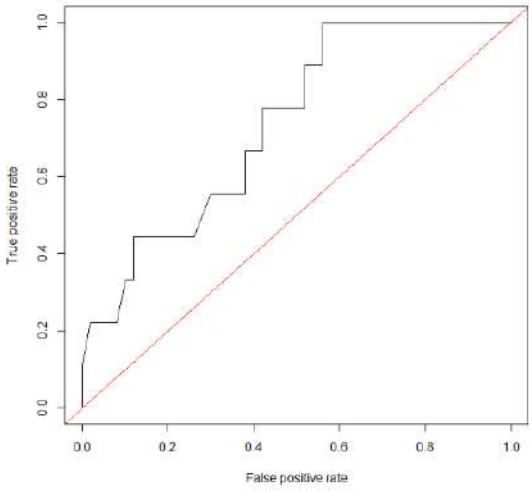
Confusion matrix for the training data.

		Actual	
		0	1
Predictions	0	56	9
	1	1	1

Performance metrics on training data

Measure	Value
Accuracy	0.8507
Sensitivity	0.1
Specificity	0.98246
PPV	0.85075
NPV	0.52190
AUROC	0.7737 [0.6454 – 0.902]

Test Data
Test size: 59



The test ROC curve from using the socioeconomic model.

Confusion matrix for the test data.

		Actual	
		0	1
Predictions	0	45	6
	1	5	3

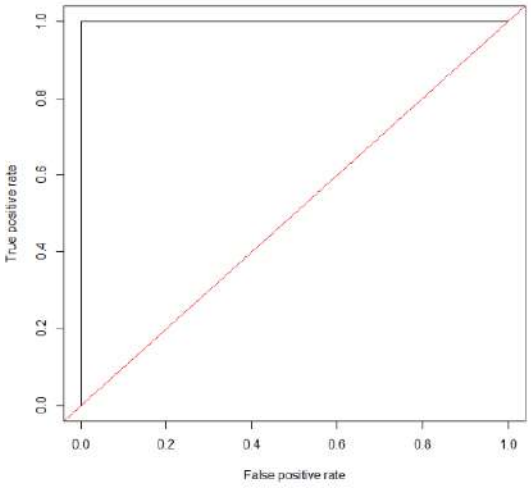
Performance metrics on training data

Measure	Value
Accuracy	0.8136
Sensitivity	0.3333
Specificity	0.9
PPV	0.76923
NPV	0.57447
AUROC	0.7356 [0.5736 – 0.8975]

Hybrid model 2 – Demographic and biochemical

Variables in model: Knee swell, sex, age, BMI, COMP, PIICP, HA

Training Data
Training size: 98



The training ROC curve from using the hybrid model.

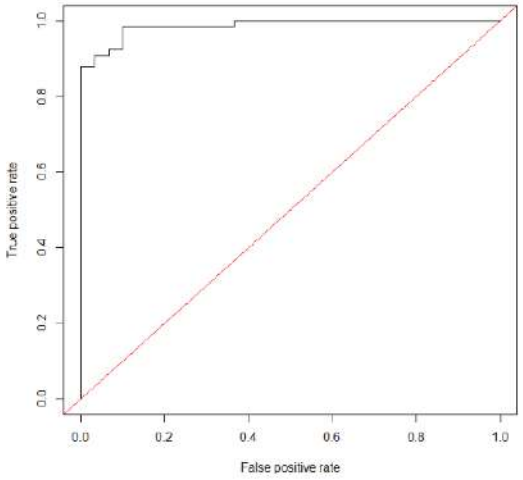
Confusion matrix for the training data.

		Actual	
		0	1
Predictions	0	34	0
	1	0	64

Performance metrics on training data

Measure	Value
Accuracy	1
Sensitivity	1
Specificity	1
PPV	1
NPV	1
AUROC	1

Test Data
Test size: 96



The test ROC curve from using the hybrid model.

Confusion matrix for the test data.

		Actual	
		0	1
Predictions	0	28	5
	1	2	61

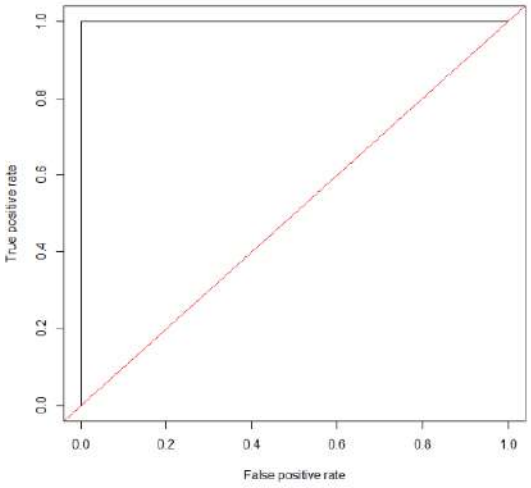
Performance metrics on training data

Measure	Value
Accuracy	0.9271
Sensitivity	0.9242
Specificity	0.9333
PPV	0.9327
NPV	0.9249
AUROC	0.9487 [0.9026 – 0.9949]

Hybrid model 3 – Demographic and biomechanical

Variables in model: Sex, Age, KCF1_Step, KAM2_Walk, KAM1_Step, KF_Walk

Training Data
Training size: 23



The training ROC curve from using the alternative hybrid model.

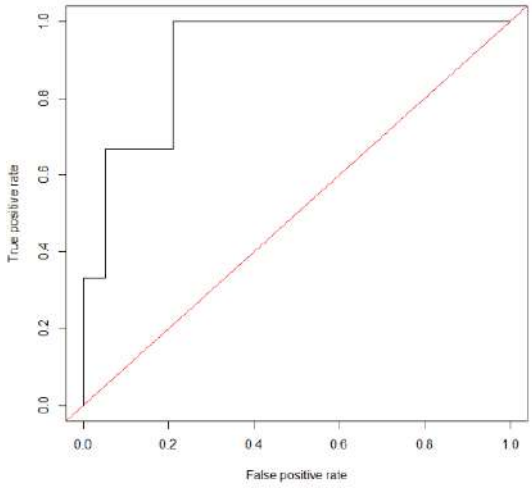
Confusion matrix for the training data.

		Actual	
		0	1
Predictions	0	18	0
	1	0	5

Performance metrics on training data

Measure	Value
Accuracy	1
Sensitivity	1
Specificity	1
PPV	1
NPV	1
AUROC	1

Test Data
Test size: 22



The test ROC curve from using the alternative hybrid model.

Confusion matrix for the test data.

		Actual	
		0	1
Predictions	0	18	1
	1	1	2

Performance metrics on training data

Measure	Value
Accuracy	0.9091
Sensitivity	0.66667
Specificity	0.94737
PPV	0.92683
NPV	0.73973
AUROC	0.9035 [0.7543 – 1]

The socioeconomic model has good performance on the test data, but any patterns in this could be covered by clinical features that are impacted by socioeconomic factors, such as education status leading to a more manual job that causes knee pain.

The models also perform well due to highly skewed data to the negative (no-KOA) class.

6. Validation of extracted MRI features

Data Presentation

The datasets given to be analysed are two. The first dataset (**diagnosisData**) contains the data of a regular diagnosis in OACTIVE's MRI collection dataset. The second (**earlyDiagData**) contains the data of a regular early diagnosis in OACTIVE's MRI collection dataset.

Both data sets consist of 2002 variables and one response variable. Note that all the variables in the dataset are numeric. Two variables of the data were removed as data indexes. The rest of the variables are separated into two major categories, femoral and tibial consisting of 1000 variables each. In each of those categories of the variables the variables further separate in two categories, one category for the mean values and one category for the standard deviation.

The response variable, for both datasets, is a two classes categorical variable. The first class, represented with "0", is the class of a person without osteoarthritis which will not present osteoarthritis in the near future. The second class, represented with "1", is the class with a person which in the near future will present osteoarthritis in one or both of his knees. From now on the second class represented with "1" will be considered as the positive class.

The first step of our analysis is a correlation analysis of the variables, for both datasets. To produce corelation analysis the Spearman correlation coefficient coefficient. In order to apply the above process, it was ensured that the correlation of the variables with themselves is excluded.

The "diagnosisData" dataset variables' correlation analysis, for all the features category, shown that there is a small number of highly correlated variables inside the dataset. In Figure 36 we can see how many variable pairs are correlated (positively or negatively) and in what degree.

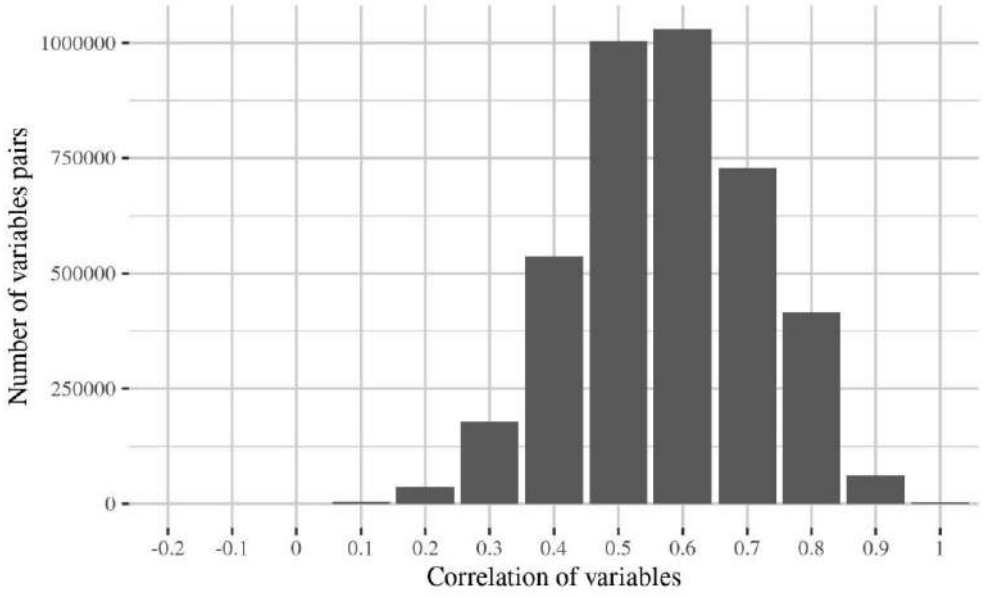


Figure 36. *diagnosisData correlation analysis.*

The "earlyDiagData" dataset variables' correlation analysis, for the femoral features category, similarly, shown that there is a small number of highly correlated variables inside the dataset. In Figure 37 we can see how many variable pairs are correlated (positively or negatively) and in what degree.

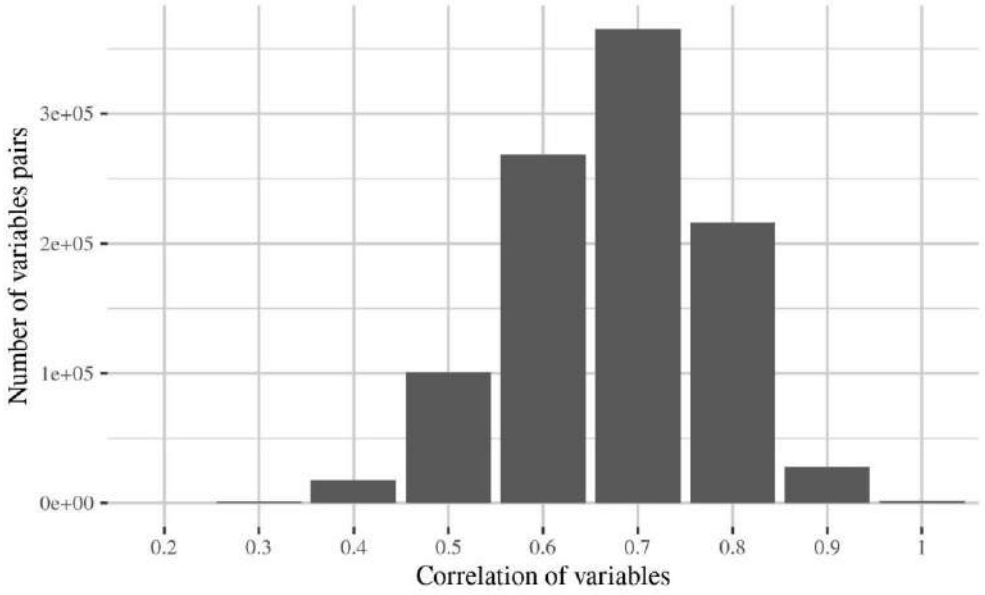


Figure 37. *earlyDiagData correlation analysis.*

A corelation analysis of the dataset was applied for all the possible combinations of the above data categories. For brevity, only the corelation, of the variables' categories combination which provide the best results will be provided. The correlation analysis of the selected variables saw that there are highly correlated variables inside the data set in the figure below we can see how many variable pairs are correlated (positively or negatively) and in what degree.

Highly correlated variables in the dataset are considered variables with corelation values over 0.7. The number of which will be presented in the table below for all the possible combinations of the OAI MRI variables' categories Table is for the "diagnosisData" dataset and Table is for the "earlyDiagData" dataset.

Table 30. *"diagnosisData" correlation table.*

CATEGORY COMBINATION NAME	NUMBER OF CORELATED VARIABLES
ALL	1999 / 2000
FEMORAL	1000 / 1000
FEMORAL MEAN	500 / 500
FEMORAL STD	500 / 500
MEAN	993 / 1000
STD	1000 / 1000
TIBIAL	999 / 1000
TIBIAL MEAN	492 / 500
TIBIAL STD	500 / 500

Table 31. *"earlyDiagData" correlation table.*

CATEGORY COMBINATION NAME	NUMBER OF CORELATED VARIABLES

ALL	1997 / 2000
FEMORAL	1000 / 1000
FEMORAL MEAN	500 / 500
FEMORAL STD	500 / 500
MEAN	992 / 1000
STD	1000 / 1000
TIBIAL	997 / 1000
TIBIAL MEAN	490 / 500
TIBIAL STD	500 / 500

In the above tables we can observe that there is a big number of highly correlated variables in all the variable categories, and in some cases the entire dataset is highly corelated. The smallest proportion of highly corelated variables can be found in the tibial mean variables' category for both datasets of the OACTIVE's MRI database.

Heatmap analysis

One important step in the data comprehension for both datasets is the visualization of the data in a heatmap. The goal behind this analysis is the comprehension of the unique characteristics of each variable with respect to the response variable.

The heatmap of the "diagnosisData" dataset is in Figure 38. The heatmap was created in the case of all the features category, as it gives the best results in the classification of the data. The lowest values are represented with green and the highest value is represented with the colour red. The horizontal white line separates the response variable's two classes. Above is represented the class "0" and below is represented the class "1". We can observe that for both cases the response isn't at all separable in any obvious way.

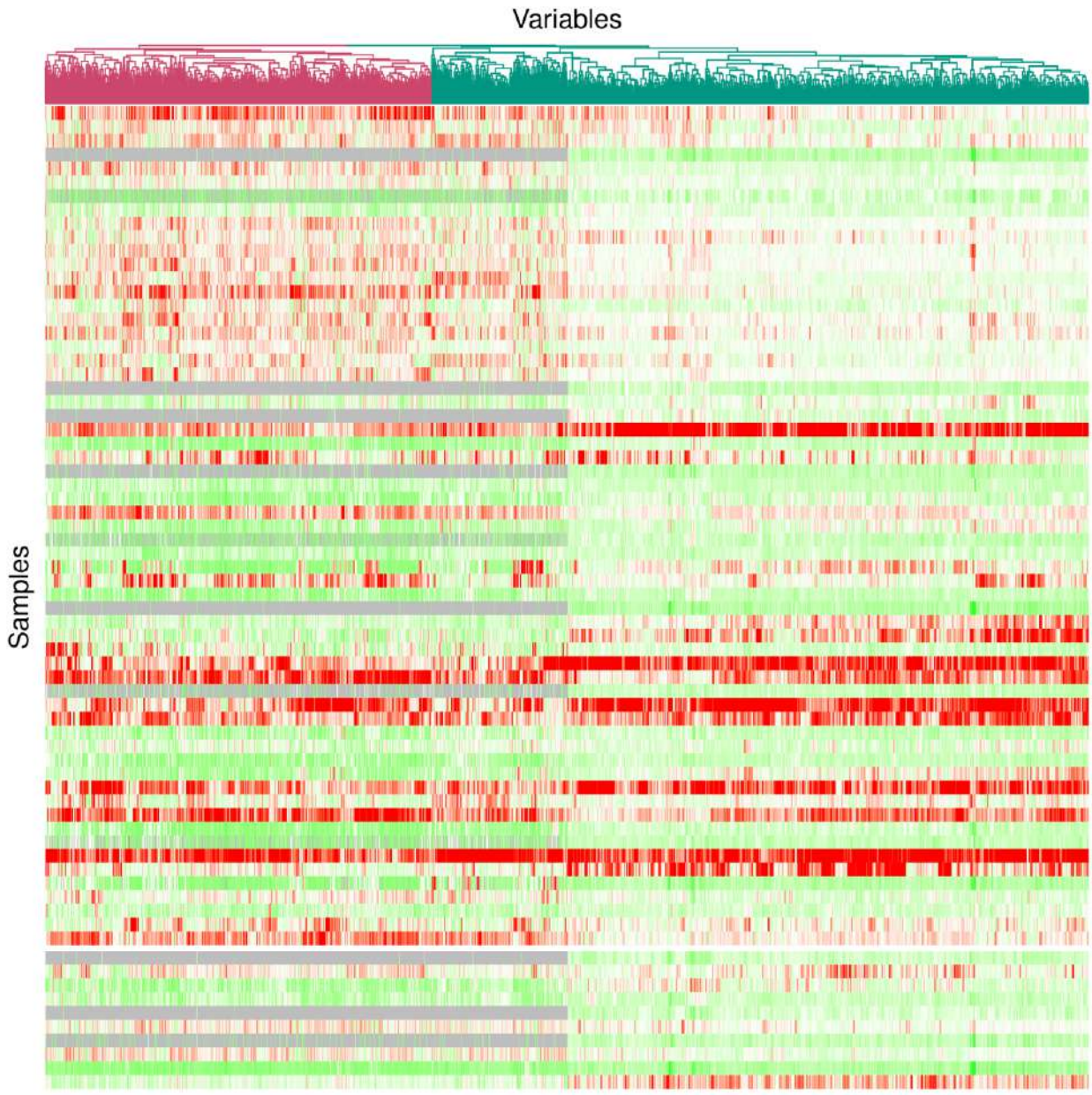


Figure 38. *diagnosisData heatmap.*

The heatmap of the "earlyDiagData" dataset is in Figure 39. The heatmap was created in the case of femoral features category, as it gives the best results in the classification of the data. The lowest values are represented with green and the highest value is represented with the colour red. The horizontal line in the middle of the heatmap, also in this case, separates the response variable's two classes. Above is represented the class "0" and below is represented the class "1". Similarly, we can observe that for both cases the response isn't at all separable in any obvious way.

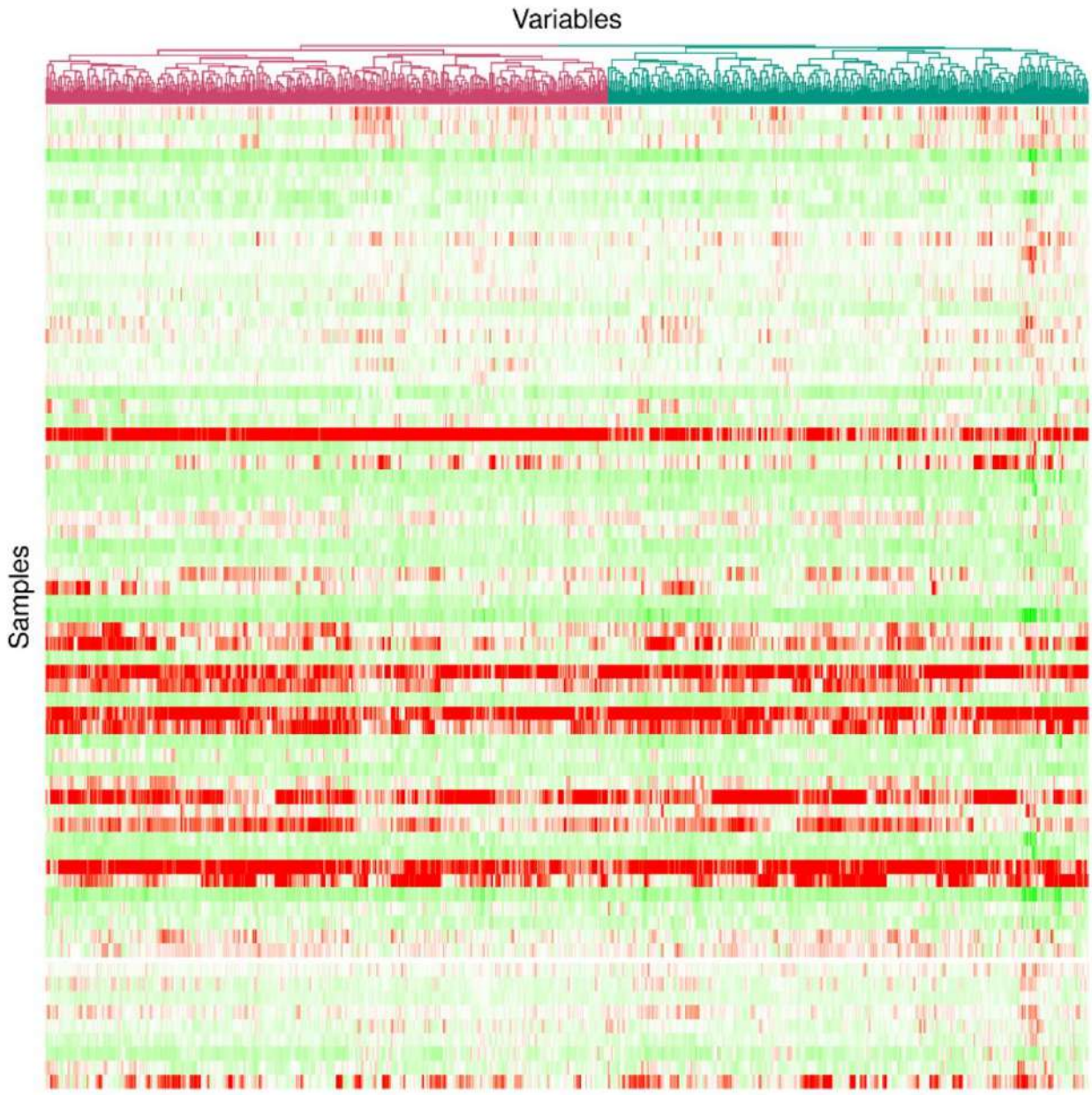


Figure 39. *earlyDiagData heatmap.*

Data pre-process

Data normalization

For the normalization of the data the technique of min-max scaling or min-max normalization. This technique is rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - (x)}{(x) - (x)},$$

where x is an original value, x ' is the normalized value. After the scaling of the data the ranges of the 100 first variables are sawn in the figure bellow.

Dimensionality reduction

Because of the data's high dimensionality, it was dimmed necessary to implement dimensionality reduction techniques. The chosen examined techniques are the Principal Components Analysis (PCA) and the Random Projections (RP). The dimensionality reduction was applied for dataset cases of the OACTIVE's MRI dataset on the best results. As a result, for the "diagnosisData" dataset the dimensionality reduction was applied on all the features category and on the "earlyDiagData" it was applied on the femoral features category.

For the PCA technique, we calculate the principal components matrix and the projections of the data on the desired number of dimensions. The number of principal components was selected in two ways, first we arbitrary selected 2, 3, 4, and 5 principal components. The second way is the from the cumulative amount of variance explained by each principal component, from a literature review this number is 0.7. The number of principal components, that explain 0.7 of the total variance in the OACTIVE's MRI dataset, is 51 for the "diagnosisData" dataset and 51 for the "earlyDiagData" dataset.

For the RP technique, a projection matrix was created with use of Gaussian Distribution. The dimensions of the resulted vector space are selected in two ways, similar to the PCA technique, arbitrary and by a literature recommendation. Arbitrary we select 2, 3, and 4 dimensions. After a literature review, we implemented the Johnson-Lindenstrauss lemma with an error tolerance of 0.5. This methodology gave us a result of 364 dimensions for the "diagnosisData" dataset and a result of 332 dimensions for the "earlyDiagData" dataset.

Table 02. *"diagnosisData".*

	kNN			*LDA*			*mLR*			*SVM*		
	acc	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*
PCA 2dims	0.800	0.000	1.000	0.800	0.000	1.000	0.800	0.000	1.000	0.800	0.000	1.000
PCA 3dims	0.810	0.000	1.000	0.810	0.000	1.000	0.810	0.000	1.000	0.810	0.000	1.000
PCA 4dims	0.905	0.000	1.000	0.905	0.000	1.000	0.905	0.000	1.000	0.905	0.000	1.000
PCA 5dims	0.867	0.000	1.000	0.867	0.000	1.000	0.867	0.000	1.000	0.867	0.000	1.000
PCA 0.7var	0.905	0.000	1.000	0.905	0.000	1.000	0.286	1.000	0.211	0.857	0.500	0.895
RP 2dims	0.833	0.000	1.000	0.833	0.000	1.000	0.833	0.000	1.000	0.833	0.000	1.000
RP 3dims	0.889	0.000	1.000	0.889	0.000	1.000	0.815	0.000	0.917	0.889	0.000	1.000

RP 4dims	0.857	0.000	1.000	0.857	0.000	1.000	0.857	0.000	1.000	0.857	0.000	1.000
RP JL lemma	0.842	0.000	1.000	0.842	0.000	1.000	0.211	0.333	0.188	0.842	0.000	1.000

Table 03. *"earlyDiagData".*

	kNN			*LDA*			*mLR*			*SVM*		
	acc	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*
PCA 2dims	0.765	0.000	1.000	0.765	0.000	1.000	0.765	0.000	1.000	0.765	0.000	1.000
PCA 3dims	0.714	0.000	1.000	0.714	0.000	1.000	0.714	0.000	1.000	0.714	0.000	1.000
PCA 4dims	0.714	0.000	1.000	0.714	0.000	1.000	0.714	0.000	1.000	0.714	0.000	1.000
PCA 5dims	0.941	0.000	1.000	0.882	0.000	0.938	0.882	0.000	0.938	0.941	0.000	1.000
PCA 0.7var	0.842	0.000	1.000	0.842	0.000	1.000	0.211	1.000	0.063	0.158	1.000	0.000
RP 2dims	0.882	0.000	1.000	0.882	0.000	1.000	0.882	0.000	1.000	0.882	0.000	1.000
RP 3dims	0.810	0.000	1.000	0.810	0.000	1.000	0.810	0.000	1.000	0.810	0.000	1.000
RP 4dims	0.769	0.000	0.909	0.846	0.000	1.000	0.846	0.000	1.000	0.846	0.000	1.000
RP JL lemma	0.864	0.000	1.000	0.864	0.000	1.000	0.500	0.333	0.526	0.864	0.000	1.000

As it is consistent with the results of the classification without the application of dimensionality reduction, when we apply dimensionality reduction the results of the classification accuracy relations between the classification algorithms remain almost identical.

Visualization of the data

For the visualization of the data the Principal Components Analysis was utilized. The projection of the data on the first 2 principal components, for the variables categories all features, mean features, std features, femoral features and tibial features, are given in the figures bellow for both "diagnosisData" and "earlyDiagData" datasets.

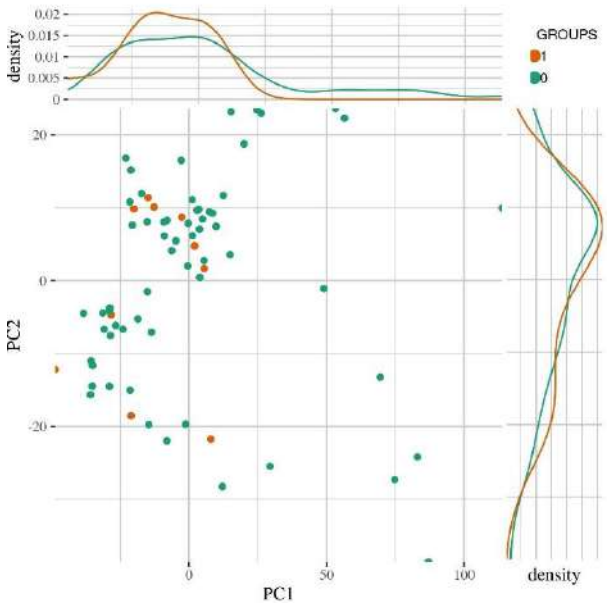


Figure 00. *"diagnosisData" all features.*

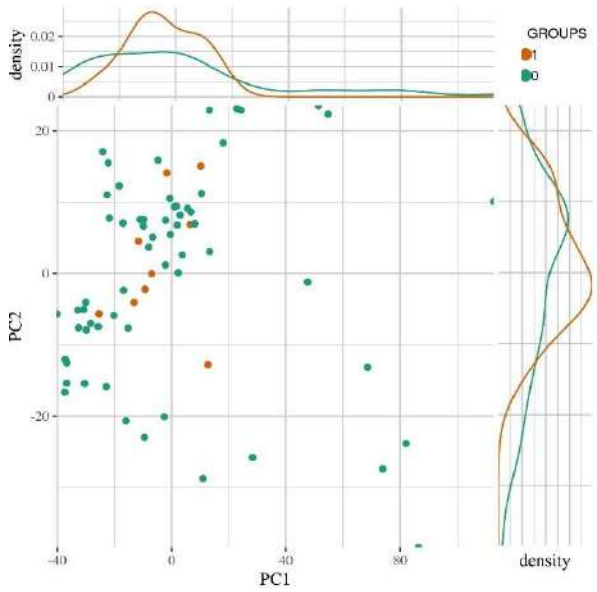


Figure 01. *"earlyDiagData" all features.*

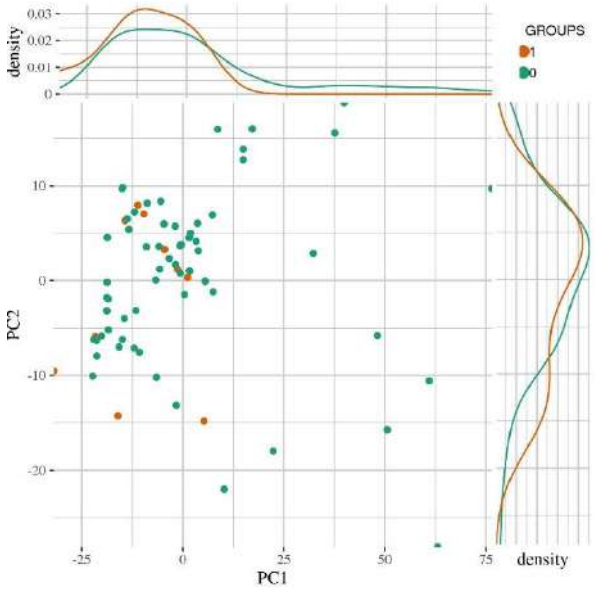


Figure 02. *"diagnosisData" mean features.*

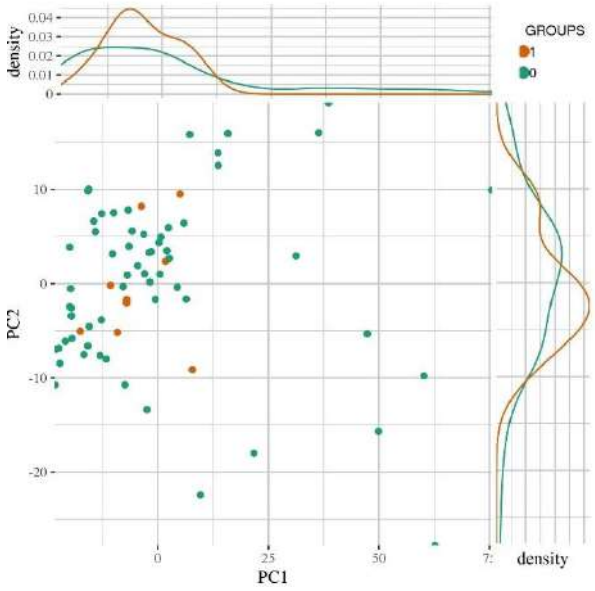


Figure 03. *"earlyDiagData" mean features.*

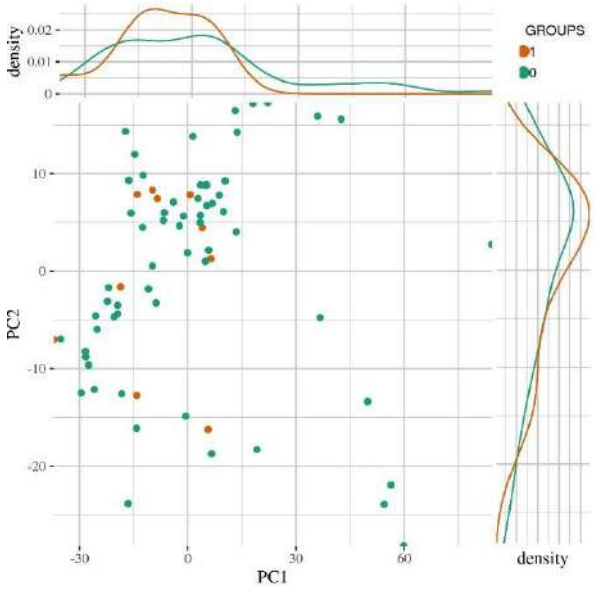


Figure 04. *"diagnosisData" std features.*

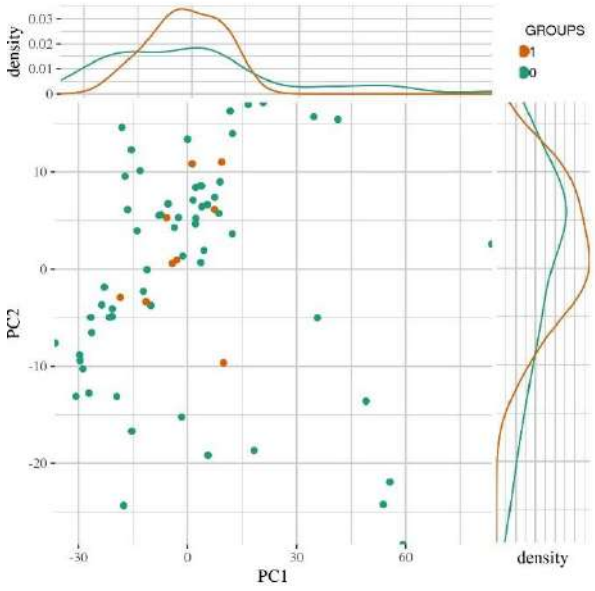


Figure 05. *"earlyDiagData" std features.*

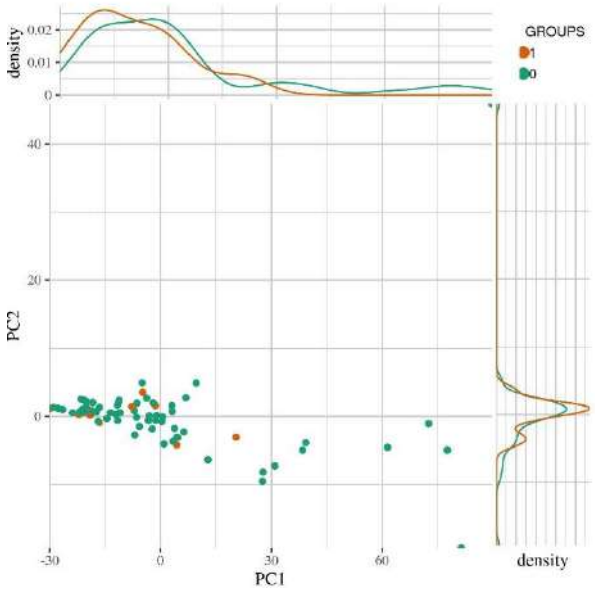


Figure 06. *"diagnosisData" femoral features.*

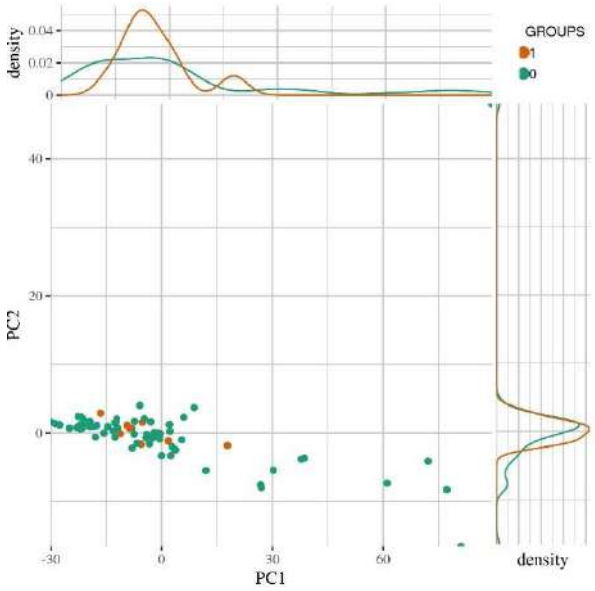


Figure 07. *"earlyDiagData" femoral features.*

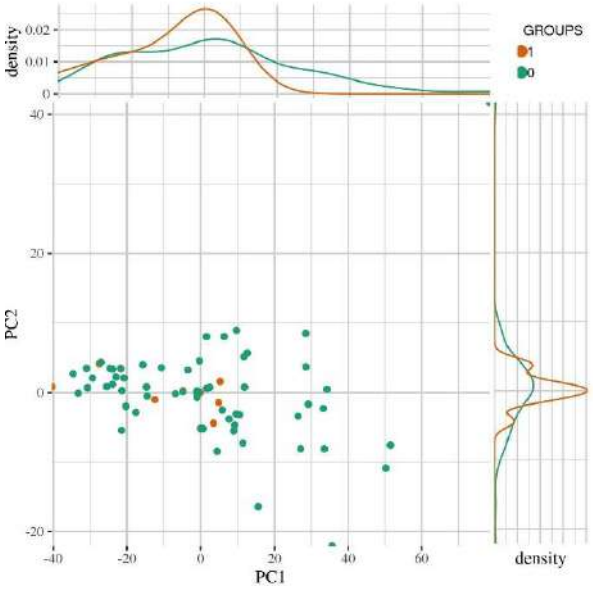


Figure 08. *"diagnosisData" tibial features.*

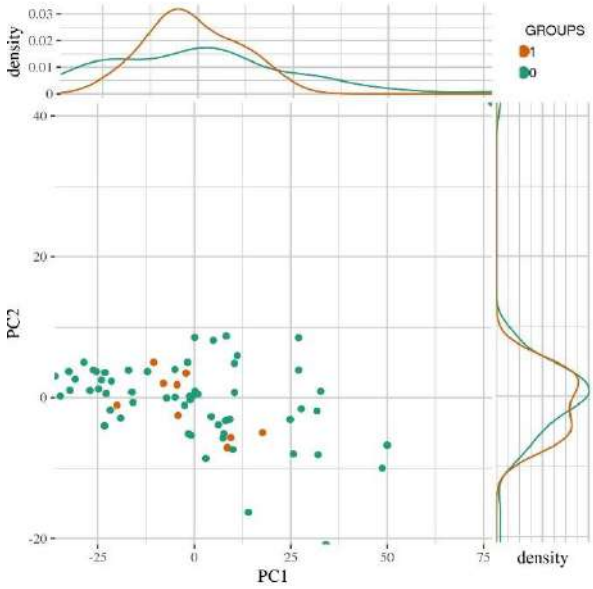


Figure 09. *"earlyDiagData" tibial features.*

As it is immediately apparent from the representation of the data, the Classes "0" and "1" aren't segregated at all, in all the cases. As the density plots of each principal component represent the classes have almost identical distribution in the space. From these visualizations an initial conclusion is that we have an almost impossible classification problem to tackle, in all the cases and in both the datasets.

Classification tasks

The classification algorithms used for this data set are the Multinomial Logistic Regression (MLR), the Linear Discriminant Analysis (LDA), the k-Nearest Neighbours (kNN) and finally the Support Vector Machine (SVM). The results of the classifications can be seen on the tables (Table 34 for "diagnosisData" and Table 35 for "earlyDiagData" datasets) for both datasets.

The aforementioned algorithms are applied on almost all the variable categories of the OAI MRI dataset. Specifically, on all the variables, on the mean variables, on the std variables, on the femoral variables, on the femoral mean variables, on the femoral std variables, on the tibial variables, on the tibial mean variables, and on the tibial std variables. The measures used are accuracy, sensitivity and specificity.

Table 34. *diagnosisData results table.*

	kNN			*LDA*			*mLR*			*SVM*		
	acc	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*
all	0.944	0.000	1.000	1.000	1.000	1.000	0.222	1.000	0.176	0.944	0.000	1.000
femoral	0.783	0.000	1.000	0.783	0.000	1.000	0.478	0.400	0.500	0.783	0.000	1.000
femoral mean	0.895	0.000	1.000	0.789	0.000	0.882	0.526	0.500	0.529	0.895	0.000	1.000
femoral std	0.864	0.000	1.000	0.864	0.000	1.000	0.818	0.333	0.895	0.864	0.000	1.000
mean	0.864	0.000	1.000	0.864	0.000	1.000	0.227	0.667	0.158	0.864	0.000	1.000
std	0.789	0.000	0.833	0.632	0.000	0.667	0.053	0.000	0.056	0.947	0.000	1.000
tibial	0.842	0.000	0.941	0.842	0.500	0.882	0.105	1.000	0.000	0.105	1.000	0.000

tibial mean	0.850	0.000	1.000	0.850	0.000	1.000	0.150	1.000	0.000	0.150	1.000	0.000
tibial std	0.778	0.000	1.000	0.778	0.000	1.000	0.444	0.750	0.357	0.222	1.000	0.000

Table 35. *earlyDiagData results table.*

	kNN			*LDA*			*mLR*			*SVM*		
	acc	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*	*acc*	*sens*	*spec*
all	0.857	0.000	1.000	0.857	0.000	1.000	0.286	1.000	0.167	0.857	0.000	1.000
femoral	0.842	0.000	1.000	0.789	0.333	0.875	0.579	0.667	0.563	0.842	0.000	1.000
femoral mean	0.941	0.000	1.000	0.882	0.000	0.938	0.588	1.000	0.563	0.941	0.000	1.000
femoral std	0.900	0.000	1.000	0.900	0.000	1.000	0.700	0.500	0.722	0.900	0.000	1.000
mean	0.923	0.000	1.000	0.923	0.000	1.000	0.269	0.500	0.250	0.308	1.000	0.250
std	0.875	0.000	0.933	0.938	0.000	1.000	0.250	0.000	0.267	0.938	0.000	1.000
tibial	0.875	0.000	1.000	0.625	0.333	0.667	0.333	1.000	0.238	0.875	0.000	1.000
tibial mean	0.897	0.000	1.000	0.517	0.000	0.577	0.379	0.333	0.385	0.897	0.000	1.000
tibial std	0.950	0.000	1.000	0.850	0.000	0.895	0.500	0.000	0.526	0.950	0.000	1.000

In the results we can observe that there is a diversity of the classification accuracy, sensitivity, and specificity based on the variable categories that are used. An expected diversity was observed between the algorithms results, because of the different natures of each algorithm. An unexpected diversity was observed between some of the categories contained in both "diagnosisData" and "earlyDiagData" datasets.

From the above results we can conclude that the results of the LDA algorithm had the best results, between the results of all the classification algorithms for the "diagnosisData" dataset and the mLR algorithm had the best results for the "early diagnosis" dataset. Except for the largest classification accuracy in combination with satisfactory sensitivity and specificity values, we can observe a slight imbalance between the prediction of the positive and negative class of the response variable.

From the above results, in conclusion, the best classification accuracy can be provided from the tibial features category for the "diagnosisData" dataset and the femoral features category for the "earlyDiagData" dataset. Bellow, are presented the Ro Curves which illustrate the results, only, of the bast categories for the two datasets:

- MLR

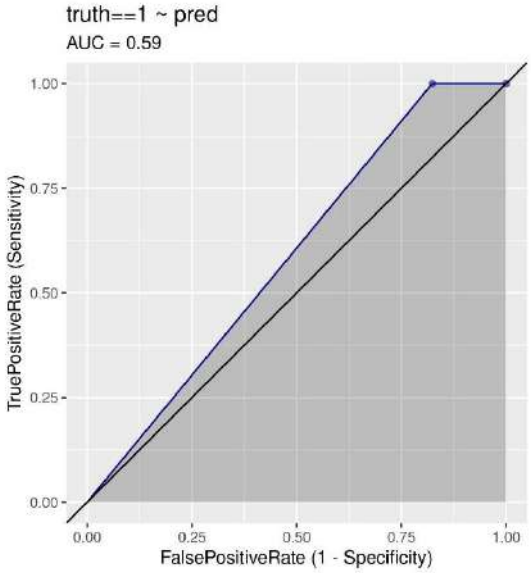


Figure 00. *"diagnosisData".*

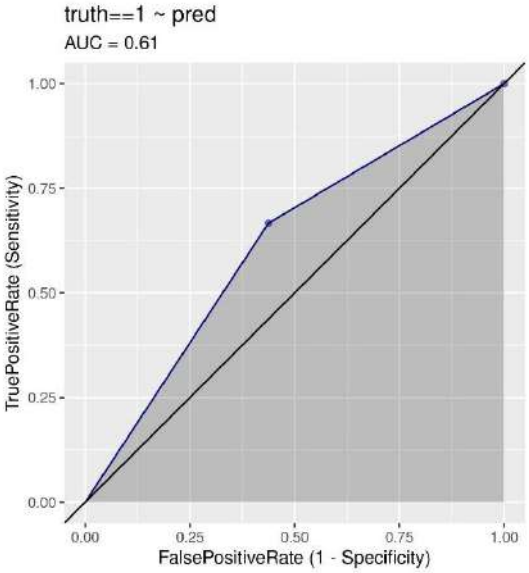


Figure 51. *"earlyDiagData".*

- LDA

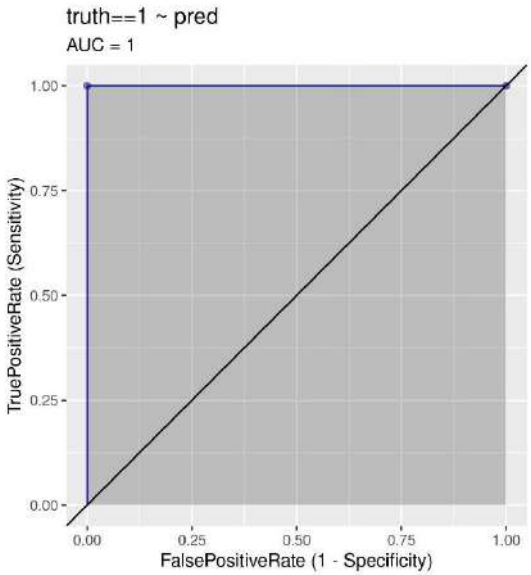


Figure 02. *"diagnosisData".*

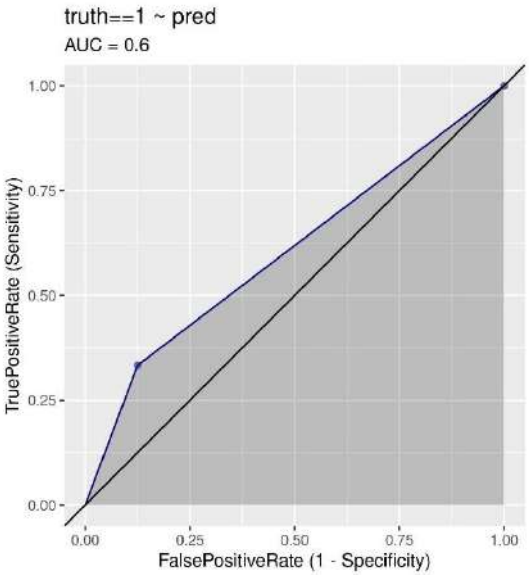


Figure 03. *"earlyDiagData".*

- kNN

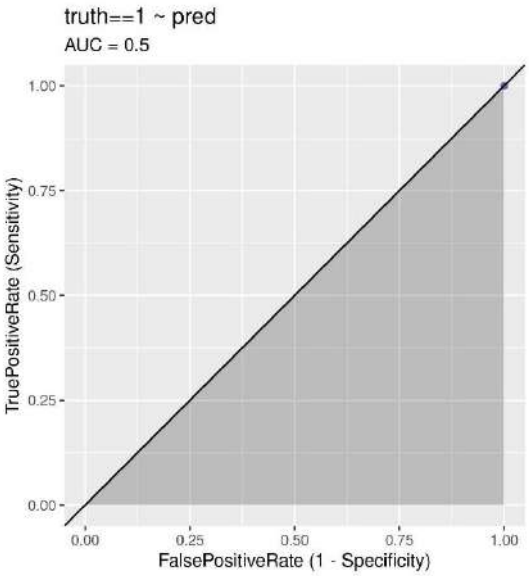


Figure 04. *"diagnosisData".*

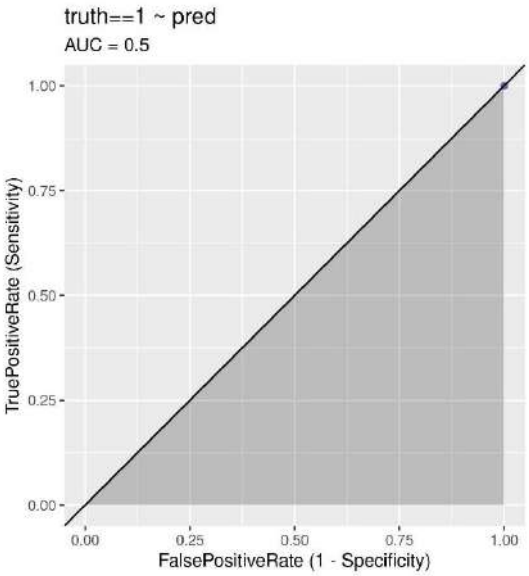


Figure 05. *"earlyDiagData".*

- SVM

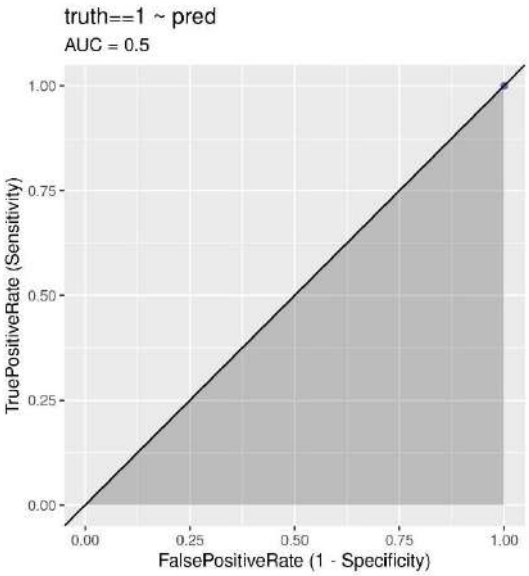


Figure 06. *"diagnosisData".*

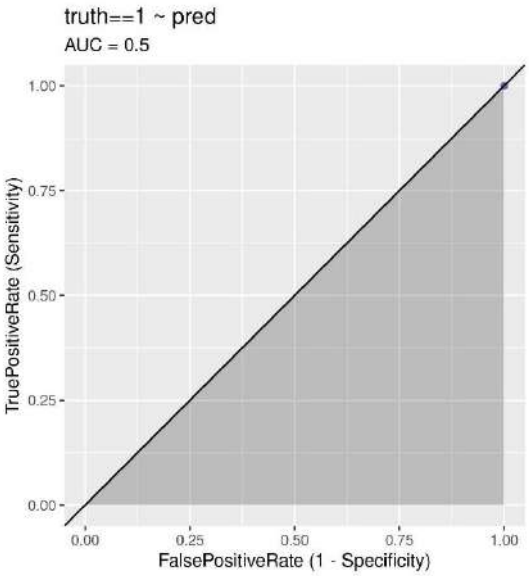


Figure 07. *"earlyDiagData".*

The extensive analysis presented in this report shows us the challenges presented in the OACTIVE's MRI data collection. The findings of this analysis gave us valuable insight into this data collection. We found that it is a dataset that contains a large number of correlated variables in both the case of "diagnosisData" and "earlyDiagData" datasets. We can also support our previous claim from the classification models application. These datasets are not so separable, as shown even in all classification tasks. Finally, form the dimensionality reduction analysis we found that we can reduce the dimensions on the dataset with a relatively small cost in the classification accuracy.

7. Personalised models based on OACTIVEs data

In this Section we worked for evaluation of Early Diagnosis and Diagnosis of KOA. Data were obtained from the OACTIVEs database.

A. Early Diagnosis

In task for early diagnosis of KOA, we proposed two approaches. Data were obtained from the OACTIVE's database. Specifically, the current study includes clinical data and MRIs features from baseline (PCA dimensions from Section 5). To cope with the certain task, we worked on two different datasets, which are described as follow:

- **Dataset A**: Clinical data from baseline
- **Dataset B**: Clinical data plus extracted MRIs features from baseline

Furthermore, our study was based on Kellgren and Lawrence (KL) grade as outcome for the classification.

Subsequently, the samples of the dataset A and dataset B were divided into two categories, as follows:

- Class 1: Early-KOA: This class comprises participants who have KL 1 at baseline. These participants had KL grades equal 1 in at least one of the two knees or in both.
- Class 2: Non-KOA: This class involves participants with KL 0 at baseline. In particular, these participants do not have KOA in any of their knees.

Methodology

The machine learning (ML) methodology for knee osteoarthritis (KOA) diagnosis proposed in this work includes four processing steps: data pre-processing of the collected clinical data (Dataset A and Dataset B), feature selection, learning process (Logistic Regression) and evaluation of the classification results. To evaluate the predictive capacity of the selected feature subset, a repeated 5-fold cross-validation process was adopted using the aforementioned classifier. More details about the proposed methodology are presented in Deliverables 6.3 and Deliverable 6.5.

Results

In this section, we present the most important risk factors as they have been selected by the proposed hybrid feature selection (FS) methodology. Furthermore, the overall performance of the models is presented in relation to the number of features and then reference is made to the models with the highest accuracies.

Dataset A: Clinical data from baseline

Table 36 gives the ranking of the first 35 selected features along with the associated votes that were assigned to each one.

Table 36. *Feature ranking after FS hybrid methodology*

	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	smoking	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
2	piicp_pg_ml_	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
3	familyoahistory	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
4	restingvas	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	5/6
5	personalhistoryofhandoa	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
6	painside	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	5/6
7	musclestrength_mrc_right[kneeextensors]	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	5/6
8	il_1E'β‰¤_pg_ml_	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
9	highbloodpressure	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
10	serumha_ng_ml_	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	4/6
11	mass_kg_	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	4/6
12	age_years_	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	4/6
13	personalhistoryofhipoa	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	3/6
14	bmi	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
15	walkingvas	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
16	timesincepainstart	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
17	serumcomp_ng_ml_	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
18	rightkneealignment	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
19	musclestrength_mrc_left[kneeflexors]	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
20	musclestrength_mrc_left[kneeextensors]	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
21	maritalstatus	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
22	levelofeducation[individual]	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
23	leftkneealignment	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
24	joint_effusion	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
25	height_m_	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
26	anycurrentmedication	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	2/6
27	alcohol	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
28	tnf_E'B±_pg_ml_	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
29	sex	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
30	regularsportleisureactivity	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
31	previouskneeinjuries[right]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
32	previouskneeinjuries[left]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
33	occupationalrisk	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
34	musclestrength_mrc_right[plantarflexors]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
35	musclestrength_mrc_right[kneeflexors]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6

Dataset B: Clinical data plus extracted MRIs features from baseline

Table 37 gives the ranking of the first 35 selected features along with the associated votes that were assigned to each one.

Table 37. *Feature ranking after FS hybrid methodology*

	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	serumha_ng_ml_	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
2	piicp_pg_ml_	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
3	smoking	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
4	personalhistoryofhandoa	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
5	painside	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	5/6
6	il_1E'β‰¤_pg_ml_	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
7	familyoahistory	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	5/6
8	PC9	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	5/6
9	mass_kg_	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	4/6
10	age_years_	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	4/6
11	PC42	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	4/6
12	PC38	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	4/6
13	PC32	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	4/6
14	walkingvas	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
15	personalhistoryofhipoa	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	3/6
16	joint_effusion	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	3/6
17	bmi	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
18	PC8	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
19	PC50	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
20	PC47	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
21	PC35	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
22	PC33	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
23	PC28	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
24	PC15	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
25	PC11	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	3/6
26	serumcomp_ng_ml_	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
27	restingvas	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
28	regularsportleisureactivity	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
29	previouskneeinjuries[right]	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
30	musclestrength_mrc_right[kneeextensors]	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
31	maritalstatus	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
32	leftkneealignment	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
33	highbloodpressure	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	2/6
34	height_m_	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
35	PC7	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6

Comparative Analysis

Figure 58 depicts the testing performance (%) of the Logistic Regression model with respect to the number of selected features on Dataset A and Dataset B. In particular, the model with Dataset A failed in this task, recording low testing performances. In contrast, the other model which based on Dataset B had

an upward trend in the first 5-9 features, followed by a steady testing performance. Specifically, the Logistic Regression model on Dataset B with respect to selected features showed an upward trend in 5-9 features, with a maximum (94%) at 9 features. Then with the addition of new features, it showed a relatively stable testing performance. Best overall performance was achieved by Logistic Regression on Dataset B at 9 features whereas the inclusion of additional features led to a small reduction in the accuracies achieved.

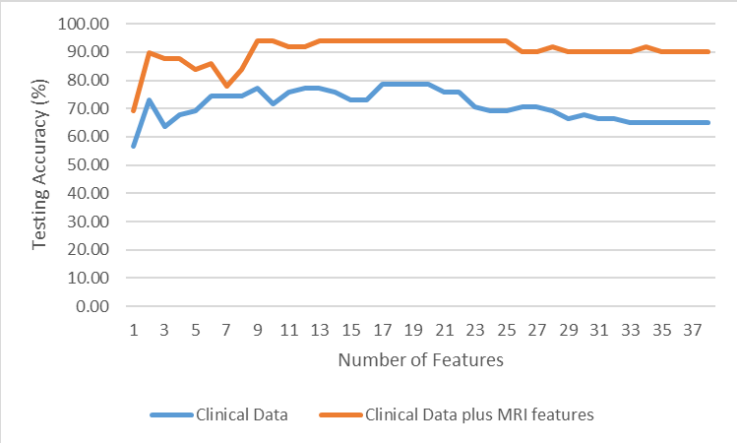


Figure 58. *Learning curves with testing accuracy scores for Logistic Regression model trained on feature subsets of increasing dimensionality.*

B. Diagnosis

In task of KOA diagnosis, we worked with clinical data from baseline from all individuals with or without Knee Osteoarthritis. Furthermore, our study was based on Kellgren and Lawrence (KL) grade as outcome for the classification.

Subsequently, the samples of the dataset C and dataset D were divided into two categories, as follows:

- Class 1: KOA: This class comprises participants who have KL >=2 at baseline. These participants had KL grades equal or higher than 2 in at least one of the two knees or in both.
- Class 2: Non-KOA: This class involves participants with KL0-1 at baseline. In particular, these participants do not have KOA in any of their knees.

Methodology

The machine learning (ML) methodology for knee osteoarthritis (KOA) diagnosis proposed in this work includes four processing steps: data pre-processing of the collected clinical data (Dataset C and Dataset D), feature selection, learning process (Logistic Regression) and evaluation of the classification results. To evaluate the predictive capacity of the selected feature subset, a repeated 5-fold cross-validation process was adopted using the aforementioned classifier. More details about the proposed methodology are presented in Deliverables 6.3 and Deliverable 6.5.

Results

In this section, we present the most important risk factors as they have been selected by the proposed hybrid feature selection (FS) methodology. Furthermore, the overall performance of the model is presented in relation to the number of features and then reference is made to the models with the highest accuracies.

Table 38 gives the ranking of the first 35 selected features along with the associated votes that were assigned to each one.

Table 38. *Feature ranking after FS hybrid methodology*

	Features	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	tnf_E'B±_pg_ml_	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
2	piicp_pg_ml_	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
3	musclestrength_mrc_right[kneeflexors]	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
4	joint_effusion	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
5	il_1E'β‰¤_pg_ml_	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
6	anycurrentmedication	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	6/6
7	musclestrength_mrc_left[kneeflexors]	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	5/6
8	musclestrength_mrc_right[plantarflexors]	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	5/6
9	kneepain	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	5/6
10	personalhistoryofhandoa	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	4/6
11	musclestrength_mrc_left[plantarflexors]	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	4/6
12	levelofeducation[individual]	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	4/6
13	musclestrength_mrc_right[kneeextensors]	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	3/6
14	smoking	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
15	serumcomp_ng_ml_	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
16	personalhistoryofhipoa	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	2/6
17	occupationalrisk	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
18	maritalstatus	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
19	knee_morphology	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
20	familyoahistory	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	2/6
21	bmi	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
22	alcohol	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	2/6
23	timesincepainstart	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
24	sex	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
25	serumha_ng_ml_	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
26	rightkneealignment	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
27	regularsportleisureactivity	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
28	previouskneeinjuries[right]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
29	previouskneeinjuries[left]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
30	painside	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
31	musclestrength_mrc_left[kneeextensors]	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
32	mass_kg_	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
33	leftkneealignment	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
34	il_6_pg_ml_	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6
35	height_m_	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	1/6

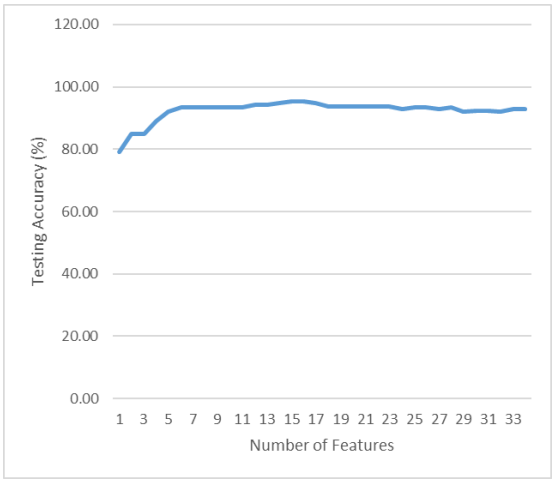


Figure 59. *Learning curve with testing accuracy scores for Logistic Regression model trained on feature subsets of increasing dimensionality.*

Figure 59 depicts the testing performance (%) of the Logistic Regression model with respect to the number of selected features. Logistic Regression model with respect to selected features showed an upward trend in 3-6 features, with a maximum (93.38%) at 6 features. Then with the addition of new features, it showed a relatively stable testing performance. Best overall performance was achieved by Logistic Regression at 6 features whereas the inclusion of additional features led to a small reduction in the accuracies achieved.

Overall, understanding the inner workings of ML algorithms is of utmost importance. The proposed methodology (Deliverable 6.3 and Deliverable 6.5) is based on a hybrid approach that combines a robust feature selection technique with a well-known classifier that enhances our understanding of the methodology applied in the diagnosis of KOA. In early diagnosis, we demonstrated that the additional MRI features (PCA dimensions from Section 5) contributed to this task. Specifically, with fewer features we achieve higher accuracies. Understanding the contribution of risk factors is a valuable tool for creating more powerful, reliable and non-invasive diagnostic tools in the hands of physicians.

8. Conclusions

This deliverable (Deliverable D9.2) describes the evaluation results and outcomes of the clinical studies in OACTIVE. Initially, an extensive analysis of OACTIVE's database was presented. We started our analysis by presenting the main characteristics of both datasets, in terms of correlation, with a mixed association technique, in terms of the response variable's separation, by creating a heatmap of the feature's vs the samples, and finally, in terms of variable importance analysis, by utilizing the Random Forests Variable Importance. Afterwards, we visualised the data into two dimensions by utilising the Principal component analysis (PCA) technique., and we colour each sample with the class it represents. Finally, various several classification models (e.g., Multinomial Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbours, Random Forests and XGBoost) were applied.

Classical statistical models were found to be effective to model the clinical, biomarker and socioeconomic data acquired by OACTIVE. Furthermore, the previously validated models from the OAI data set were successfully validated also on the OACTIVE data. Given the specificities of cohort definitions in each of

the three data acquisition studies, and also from the limited sample sizes, we report the statistical results as exploratory, with the aim of identifying key markers of interest for hypothesis generation in future studies.

The extensive analysis in Section 5 shows us the challenges presented in the OACTIVE MRI dataset. The findings of this analysis gave us valuable insight into this database. We found that both of the datasets contain a large number of correlated variables. This results in a not so separable dataset, as shown even in all classification tasks. Finally, form the dimensionality reduction analysis we found that we can reduce the dimensions on the dataset with a relatively small cost in the classification accuracy.

End, we worked in diagnosis of KOA for different stages. The first approach concerns early diagnosis and the second one diagnosis of KOA only with clinicals data. In these tasks a machine learning workflow for diagnosis of KOA is provided. In particular, we observed that in the work for early diagnosis geometric features contribute significantly. Overall, understanding the inner workings of ML algorithms is of utmost importance. The proposed methodology is based on a robust feature selection technique. Understanding the contribution of risk factors is a valuable tool for creating more powerful, reliable, and non-invasive diagnostic tools in the hands of physicians.

9. References

1. Emery CA, Whittaker JL, Mahmoudian A, Lohmander LS, Roos EM, Bennell KL, Toomey CM, Reimer RA, Thompson D, Ronsky JL, Kuntze G, Lloyd DG, Andriacchi T, Englund M, Kraus VB, Losina E, Bierma-Zeinstra S, Runhaar J, Peat G, Luyten FP, Snyder-Mackler L, Risberg MA, Mobasheri A, Guermazi A, Hunter DJ, Arden NK. Establishing outcome measures in early knee osteoarthritis. Nat Rev Rheumatol. 2019 Jul;15(7):438-448. doi: 10.1038/s41584-019-0237-3.

2. Loeser RF, Goldring SR, Scanzello CR, Goldring MB. Osteoarthritis: a disease of the joint as an organ. Arthritis Rheum. 2012 Jun;64(6):1697-707. doi: 10.1002/art.34453.

3. Favero M, Ramonda R, Goldring MB, Goldring SR, Punzi L. Early knee osteoarthritis. RMD Open. 2015 Aug 15;1(Suppl 1):e000062. doi: 10.1136/rmdopen-2015-000062.

4. Jamshidi A, Pelletier JP, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nat Rev Rheumatol. 2019 Jan;15(1):49-60. doi: 10.1038/s41584-018-0130-5.

5. Luyten FP, Bierma-Zeinstra S, Dell'Accio F, Kraus VB, Nakata K, Sekiya I, Arden NK, Lohmander LS. Toward classification criteria for early osteoarthritis of the knee. Semin Arthritis

Rheum. 2018 Feb;47(4):457-463. doi: 10.1016/j.semarthrit.2017.08.006.