Prediction of knee osteoarthritis progression from 2 year observational patient records

Somak Sanyal
Email:s.sanyal2@newcastle.ac.uk
School of Computing, Newcastle University, UK

August 2023

Abstract

Knee osteoarthritis (KOA) is a common chronic disease with no assured cure till date in which patients suffer from pain and irreversible progressive structural damage of cartilage and joints. In order to charter an effective clinical path and also for the development of disease-modifying osteoarthritis drugs (DMOADs), it is essential that physicians and researchers need to ascertain the nature and extent of future progression of the disease with reasonable accuracy. For the purpose, formulation of some robust techniques for such prediction emerges expedient. In this direction, the present study approaches the prediction of disease progression in KOA patients as a multi-class classification problem based on nearly 300 KOA patients' baseline feature datasets along with their progression measures (measured for a period of 2 years) as class labels. These datasets are taken from the APPROACH (Applied Public-Private Research enabling OsteoArthritis Clinical Headway) study, internally joined in combinations and are used to multiple machine learning (ML) algorithms (Random Forest, k Nearest Neighbors, Artificial Neural Networks, Decision Trees) to identify the best combinations of features and corresponding ML model that would give the most accurate predictions measured by F1 score macro and accuracy. In this study, the best model in terms of accuracy (66.6%) and macro F1 score (36.1%) was obtained from joining the Tomography and Questionnaires dataset and training on Random Forest classifier. The most important features contributing to this result are identified using Permutation Feature Importance. The individual contributions of each of these top features to the prediction are also visually represented with summary plots using SHAP.

1 Introduction

Knee Osteoarthritis (KOA) is a chronic degenerative disease of the knee joint with high incidence as well as prevalence rate and characterised by deterioration of the cartilage and changes in the underlying bones that leads to pain, mobility restrictions and resulting in worsening of quality of life in millions of people all over the world [1]. Although research for the assured cure of KOA is still underway, the disease generally lasts for long and progresses over time, following a definite pattern or trend. In most of the cases, treatment of this chronic disease is centred upon symptom management. Thus, ascertaining the current and probable future symptoms, based on the predictive nature of their progression

is of utmost importance to the physicians for planning the long-term course of treatment [2]. For credible prediction of the nature and extent of disease progression, an attempt has been made in this current study to develop a predictive model using supervised machine learning (ML). For this purpose, an observational set of longitudinal data of patients for 2 (two) years taken from the **A**pplied **P**ublic-**P**rivate **R**esearch enabling **O**steo**A**rthritis **C**linical **H**eadway (APPROACH) cohort database, containing different features of X-ray, MRI, Biomarker, Tomography, Clinical and Questionnaire Survey has been used with their disease progression measures for training the ML models.

It is important to note that the progression of the disease in patients being used in this study was noted during an observational period of only 2 years. In the context of the trend of progression of KOA which is a long lasting disease, it can be inferred that any patient showing signs of progression within just 2 years has a high progression rate and needs treatment on a priority basis. Patients having no or slow progression can be kept under observation for a longer period of time before planning for further treatment.

The predictions are made using the above-mentioned datasets as training sets, which are pre-processed, removed of any imbalance and trained into the ML models. Four patient class labels formed by combining 2 binary patient class labels based on different clinically relevant measures of progression of the disease, have been considered for training and testing the ML models which covers cases of non-progression of the disease and progression of either structural damage or general pain or both. This makes the current study a case of multi-class classification problem based on supervised machine learning. Selection of features has been made to identify which combination of features, when trained into ML models, would deliver the best result. During the course of the current study, different algorithms and learning configurations have been experimented to derive multiple predictive models. Nested cross validation is performed on each of these models to evaluate different hyperparameter settings on different subsets of the data, to identify the best set of parameters that increases the accuracy and F1 score of the predictions. Permutation Feature Importance technique is used to assess the contributions of individual features to the model's predictions. Finally, the contributions of the top features to the best predictions are analysed visually using SHAP and the model outputs are interpreted to estimate the best-performing combination(s) of features and ML model [3].

1.1 Motivation

As mentioned above, KOA follows a definite pattern or trend in patients based on their symptoms and medical parameters. Identifying these patterns or trends will help to predict the progression of the disease in patients. Using this predictive study, the nature and extent of progression of the disease could be projected for any particular patient and accordingly the time of intervention required by the physician and course of treatment can be decided. Also, such projections would facilitate effective development of Disease Modifying Osteoarthritis Drug (DMOAD) [4]. It can be determined which patient might expect a worsening in their condition, develop more severe symptoms and will need intervention of physicians first. There might be no definite cure for the disease till date, but this will give doctors an edge over the disease, to be prepared beforehand to slow down the progression. If in certain scenarios it is seen that there is no progression of the disease in a patient for a significant amount of time, it can be concluded that the patient is fine and would require no further treatment. Credible prediction of future progression of diseases like KOA is important since this would help in early intervention, and treat-

ment both in long and short-term, reducing unnecessary interventions, customised and personalised patient care, judicious planning for disease mitigation, resource allocation and also to carry out clinical trials for further research in this arena. Due to such high prevalence and increasing incidence of the disease, it can also well be estimated that direct and indirect costs of such chronic disease across the globe are also substantial [5]. Since detection of the disease is generally made at quite an advanced stage and accurate prediction of progression is difficult, many patients are finally subjected to total knee replacement (TKR) as an ultimate resort, which is highly expensive, somewhat painful and also compromises quality of life. Scientists and researchers world-wide are frantically working in quest of better understanding and more effective medical intervention for this global health issue. The current study to identify best possible combination of disease features and the best-fit ML models that deliver the most accurate prediction is an effort in this direction.

1.2 Aim

Predicting the progression of the disease is essential for patient stratification to be able to design tailored treatments and drug development [3,6]. Such prediction is envisaged to be carried out using observational data of patients and identifying the best possible combination of features and ML models that deliver the most accurate prediction for new patients.

1.3 Objectives

1. Understanding the APPROACH project and its cohort data.
2. Creating a data pre-processing pipeline to handle the missing values of mixed data types, scaling the data.
3. Oversampling the data to remove class imbalance and encoding the categorical values in the data.
4. Feature selection for dimensionality reduction and removing highly correlated features.
5. Creating the ML pipeline to implement the various ML algorithms and check the accuracy on the models.
6. Nested cross validation with hyper-parameter tuning for unbiased evaluation of the learning process.
7. Identify the most important 5 features which contribute to best results.
8. Interpreting the results to identify the combination of the best features and ML model and checking the contribution of the top features to the prediction.

2 Background Research

Extensive study of relevant scholarly papers and online resources have been carried out to get sufficient insights of related and allied works in the arena of application of ML for assessment of disease progression and substantiate the relevance of the present research. Salient features of this background study are presented herein below.

2.1 Knee osteoarthritis

Although research for complete understanding of the disease and its nature of progression are still underway, it is known that several biological, medical and environmental risk-factors do contribute to the development of the disease [7].Factors like age, obesity, diabetes, synovitis, stress, joint injuries, innate immunity, lower limb alignment and so on, along with the genetics of the patients also determine the degree of severity of the disease and its progression [1]. Medical practitioners agree that KOA in a patient is generally diagnosed quite late in the disease process, and thus disease modifying drugs may not be always effective. Moreover, the disease does not affect everyone in the same way, and the mechanisms that lead to the disease in different groups of patients are poorly understood. As such, diagnosing the disease at an early stage and classifying the patients based on their prognosis are essential to ascertain the disease progression and efficacy of the course of treatment [5]. KOA generally lasts for long and progresses over time, following a definite pattern or trend. In most of the cases, treatment of this chronic disease is centered upon symptom management. Thus, ascertaining the current and probable future symptoms, based on the predictive nature of their progression is of utmost importance to the physicians for planning of the long-term course of treatment of the patient [2].

2.2 APPROACH Consortium Data

APPROACH Consortium (a partnership of over 20 European clinical centres, research institutes, small enterprises and pharmaceutical companies) has developed a database of KOA patients to help analyse the disease progression through a 2-year observational study in 5 clinical centres from 4 European countries [5]. The cohort dataset containing multiple features of information collected from each of X-Ray reports, MRI Reports, Clinical Reports, Biomarkers, Tomography and Questionnaire Survey - the major sources for diagnosis and clinical assessment of KOA, have been used in the present study. The relevance and importance of such data sources are briefly discussed hereinbelow.

1. **X-Ray Data**: X-Ray Data of KOA patients are an important sources of valuable visual and quantitative information about the extent of joint degeneration and structural changes. This information can be effectively used for having a clear understanding of disease development and its management while helping development of predictive models so as to identify patients at higher risk of disease progression and contribute to meaningful guidance for treatment decisions. Many similar studies, like the ones mentioned in the background research section of this paper, use X-ray images with application of Machine Learning for studies related to KOA.

2. **MRI Data:** MRI (Magnetic Resonance Imaging) data provide with detailed information about soft tissues, cartilage, and other structural components of the knee joint. Comprehensive insights into the structural and tissue-level changes associated with KOA are obtained from MRI data that can be used to develop predictive models, monitor disease progression over time, deciding treatment modalities etc, leading ultimately to improving patient outcomes. Stratification of patients have been attempted in many studies by classifying MRI data images.

3. **Biomarkers:** Valuable information about molecular and biochemical information that reflects the underlying pathophysiology are obtained from Biomarker data of KOA patients. These are measurable indicators which are used for ascertaining

presence of disease, its severity and progression. They also provide insights into the molecular and biochemical changes associated with the disease and thus facilitate better understanding of the underlying mechanism of the disease, predicting its progression through analyses and interpretation of Biomarker data.

4. **Clinical:** Wide range of patient specific information are provided by clinical data of KOA patients viz. individual's health status, lifestyle and overall disease trajectory. This gives a holistic view of the patient's health and lifestyle as well as symptoms and functional status. Physicians and healthcare professionals combine the clinical data with other types of information (viz. biomarker, imaging and radiography etc.) so as to develop more accurate predictive models for KOA progression that finally leads to informed choice of treatment method and improved patient outcomes.

5. **Tomography:** Detailed three-dimensional images of the joint and surrounding structures are obtained from tomography data, such as CT (Computed Tomography) and MRI (Magnetic Resonance Imaging) of KOA patients thus allowing more accurate assessment of joint health and the changes over time .

6. **Questionnaire:** Questionnaire survey directly from the KOA patients are useful in securing first-hand information and renders valuable insights into their symptoms, functional limitations, and overall well-being. Such patient-reported information captures real-time, generally authentic and reliable inputs as well as trends of disease progression over time.

2.3 Related Works

1. The conventional diagnostic tools for OA severity assessments from radiographs etc involve a lot of subjectivity and can be enhanced through computer assisted methods. There have been several attempts for automation of such diagnosis and prediction of KOA progression with specific machine learning approaches viz. Deep Learning where raw radiographic image (XRay) data are directly used instead of depending solely on parameters defined by radiologists. Such effort reportedly yields substantially better prediction outcomes over the conventional reference methods. This also establishes that even the X-ray image alone can credibly contribute to the prediction of the disease progression [8].

2. In the study by Almajalid et.al [2], biomedical insights from magnetic resonance (MR) images of knee joints are used to predict osteoarthritis (OA). This research involved calculating the cartilage damage index (CDI) data from 36 specific locations within the tibiofemoral cartilage using 3D MR imaging. The feature dataset was then subjected to principal component analysis (PCA) for feature selection, whereas in the present study different techniques are used for feature selection. Then four different machine learning techniques (artificial neural network, support vector machine, random forest and naïve Bayes) to forecast the disease progression. This progression was assessed by quantifying changes in the KL grade, Joint Space Narrowing grade on the Medial Lateral compartment grade.

3. Schiratti et.al successfully attempted application of deep learning method using data on Joint Space Narrowing after 12 months of image acquisition to develop a predictive model for assessment of progression of KOA [9]. Although this research is based on Image Processing and developing a predictive model from image datasets

taken from a different source, this research does share similar procedure for measuring the structural progression (calculated using Xray images for minimum joint space width in the knee) and progression of pain (calculated using WOMAC score) as done in the Approach study whose datasets are used for this study [8,9].
Conventional osteoarthritis clinical trials are not very effective due to challenging selection of patients who may or may not show any disease progression during a trial period. In order to improve upon the efficacy of the clinical trials, selection criteria are required to be made more predictive of the disease progression. Widera P et. al have formulated such problem of patient selection as a classification task with multiple classes based on different measures of progression of the disease[3]. They have used longitudinal data sourced from different studies including the Approach Consortium Data, to test different algorithms and learning process configurations so as to identify the optimally performing machine learning models in terms of prediction errors and the impact of used features, for establishing their clinical relevance. Such approach based on application of machine learning reports much improved selection of relevant patients for trials, thus significantly reducing the number of patients who show no progression, thus leading to meaningful clinical trials. Although, the problem statement of this project is different from the current study, the working methodology such as some techniques of data pre-processing, performance measure, feature selection are very similar in respect of predicting the progression of the disease [3].

2.4 COMPUTATIONAL METHODS

Machine Learning (ML) is the study of how computer algorithms (i.e., machines) can "learn" complex relationships or patterns from empirical data and hence, produce (mathematical) models linking an even large number of covariates to some target variable of interest [10]. The importance of applying ML techniques to KOA has been well documented by Jamshidi et al. [11] and Kluzek and Mattei [12] in 2019.
In Machine Learning, a patient represented by various features in terms of their characteristics, risk factors, clinical data, medical image attributes and so on is considered as a sample. Such features are typically linked to form a multidimensional feature vector so as to promote the learning process [13].
Training or Learning can be of two types – either supervised or unsupervised. The data samples are assigned with a pair consisting of an input (typically a multi-dimensional feature vector) and a desired output value (e.g. a label having real-world meaning such as Cartilage Damage Index, Joint Space Narrowing, Kellgren Lawrence grades etc. in case of (KOA) in supervised learning. A function is generated to map every input (feature) to its associated output in the training phase. Such generated inferred function, in the testing phase, is used to map unknown inputs. On the other hand, in unsupervised training, ML techniques handle unlabelled dataset with an aim to find out patterns or trends in the dataset [14].
ML systems would have two distinct phases – a) the Training (Learning) phase and b) the Testing phase. In both the phases, the datasets or the samples are pre-processed in terms of noise elimination, removal of inconsistent data and discretisation and normalisation of data wherever required. Subsequently, the most relevant features are selected / extracted to identify the most important / informative feature subset used in learning / training phase would be applied for the testing phase. This phase is also known as **feature engineering** [13].

The feedback from the learning would further adjust pre-processing and feature selection process iteratively so as to refine the final learning or the trained model. Such trained models would now be tested for new dataset / samples in the testing phase. Based on features in the new samples the models are expected to make appropriate decision classifications.

Background study of some of the ML techniques, algorithms and evaluation metrics are noted as below:

SMOTE : With class imbalanced dataset, binary classifications become highly difficult. In cases of highly skewed class distribution ML methods would generate classifiers that consider only the majority class in a biassed manner with minority classes being ignored which makes the decision making process poor in spite of having fairly high accuracy level, in general. SMOTE oversampling is a commonly applied technique particularly when sample size is small like in the case of present study. When the sample size is small, undersampling runs the risk of significant loss of information and in such cases SMOTE oversampling may give better performance with high computational efficiency, yet simple. SMOTE, through interpolation of pre-existing samples tends to generate new samples from the minority class artificially. Such selection of pre-existing samples is done by identifying the nearest neighbours for each minority sample and thus forming a neighbourhood. Amongst many variants of SMOTE, SMOTENC is supposed to be an effective oversampling technique when the datasets contain both nominal and continuous features like in the case of present study.[15]

k-Nearest-Neighbours classifier : kNN classifier is to classify unlabeled observations by assigning them to the class of the most similar label. The k-Nearest-Neighbours (kNN) is a simple non-parametric classification method that works well in many problems. For the purpose of classifying any data record "A", its k nearest neighbours are retrieved to form a neighbourhood of "A". Choosing an appropriate value of k is however crucial since success of classification largely depends on this. A simplistic way to choose the k value is by running the algorithm iteratively many times with different k values and selecting the one that gives the best result. This is simple and fast to implement method [16,17].

Random Forest classifier: In Random Forest Classifier, random vectors are sampled individually from the input vectors that generate multiple decision trees which in turn combine to form a cluster like a forest. An input vector is classified in the most popular class by the unit vote calculated by each of these decision trees. Random Forest adapts an Index (Gini Index) that measures the impurity of an attribute according to the classes. In Random Forest, The features are randomly selected in each decision split and for the purpose of improving the efficiency of prediction by random selection of features to reduce the correlation between each of the decision trees. There are many advantages of Random Forest viz. it eliminates the problem of over fitting, training data are less sensitive to outlier data, parameters can be set easily and thus the need for pruning the trees is eliminated, variable importance and accuracy is generated automatically etc.

Random Forest is capable of handling missing values, continuous, categorical and binary data and thus it serves the purpose of high dimensional data modeling. The need for pruning the decision trees is eliminated since the problems of over-fitting is taken care of by Random Forest is reinforced by bootstrapping and similar schemes. Random Forest is an efficient, interpretable and non-parametric classifier suitable for different types of datasets while ensuring high prediction accuracy. Of the popular machine learning methods, Random Forest definitely provides unique advantage of interpretation of model, accuracy of prediction and better generalizations which is achieved through random sampling and utilization of ensemble strategies [18]. Random forest algorithms are fast to

train and are powerful to work with problems of higher complexity. However, they are less interpretable and a bit slower than other algorithm [19].
Decision Tree classifier: Decision Tree is a Supervised learning technique widely used for classification but also used for regression problems. It is called a decision tree since like in a tree, it has the root node, which expands on further branches and leaves and gives a tree-like structure where internal nodes, branches and leaf nodes represent dataset features, decision rules and outcomes respectively. It is basically a schematic representation for getting all the possible solutions to a problem / decision based on given conditions. The decisions or the test are performed on the basis of features of the given dataset. The internal nodes compare attributes and decides on the downwards nodes based on the attributes and finally outcomes are derived as boolean decision values from the down most leaf nodes based on the idea of possible consequences. However, such decision trees are more suitable for application in cases of problems with low complexity [20,21].
Artificial Neural Network resembles functioning of a human brain and the computers are programmed so to behave like inter-connected brain cells. Generally ANN would have three layers of neurons : input that receives information; the hidden layer that is responsible for extracting patterns and performs the internal processing and output which generates and presents final network outputs. ANN is capable of parallel processing of multiple tasks. However, there is no particular guideline for deciding on the structure of ANN and it is accomplished through experience, trial and error and it is also dependent on appropriate hardware (processor). The major disadvantage of ANN is, it does not interpret or provide any insight about underlying logic of the testing solution it produces. Health-care providing organizations are utilizing artificial neural networks (ANN) to improve delivery of care at a reduced cost. [22]
Performance metrics: F1 Score is used to measure the classifiers in this study. The F1 score represents a balanced combination of two of the vital metrics parameters - precision and recall. In the context of medical literature, precision denotes to a positive predictive value, and recall corresponds with sensitivity. [3]
Performance metrics:Accuracy is another performance metrics used in this study. It is measured by calculating the number of correct predictions made out of the total predictions.

3 Materials and Methods

This project is a comparative study between different combinations of features of KOA patient's dataset and machine learning algorithms to see which combination provides the best prediction. This is based on a multi-class predictive classification model and the model performed is measured by accuracy and F1 score. In the entire study several ML algorithms and learning process configurations are tried to build the best performing model.

3.1 Dataset

The Approach consortium collected the following data features from nearly 300 KOA patients - Xray, MRI, Biomarker, Clinical, Tomography and Questionnaires. This data was collected through a 2-year observational study in 5 clinical centres from 4 European countries to help analyse knee OA progression. Such a secondary longitudinal dataset, a cohort for about 300 KOA patients has been made available and the baseline data

of these patients along with their progression measures are used for accomplishing the current project. Details of the datasets are listed below:

Dataset Name	Patient records	Features	Missing val
X-Ray	297	24	1.3%
MRI	297	133	3.6%
Biomarker	295	16	0.0%
Clinical	297	38	1.1%
Tomography	297	168	22.6%
Questionnaire	297	185	1.7%

3.2 Progression categories

In the data of APPROACH consortium, the patient categories are defined based on a 2 years observation time. Patients are split into 4 following categories:

1. a non-progressive category (N),
2. a progressive category (P) which involves inflammation, pain in general etc.,
3. a progressive category (S) which involves structural pain and damage and
4. a progressive category (P+S) which involves a combination of P and S.

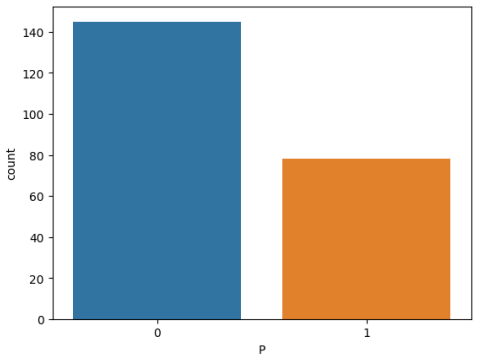
Categories have been defined based on different measures viz. symptoms of pain and structural damage at the beginning and at the end of the observation period. Pain has been measured using the pain subscale from the WOMAC (Western Ontario and McMaster Osteoarthritis Index) self-report questionnaire, that considers perceived level of pain during 5 different postures or activities [23] Structural damage has been measured in terms of joint space width (JSW) derived from radiographic readings [3].

3.3 Class Labels

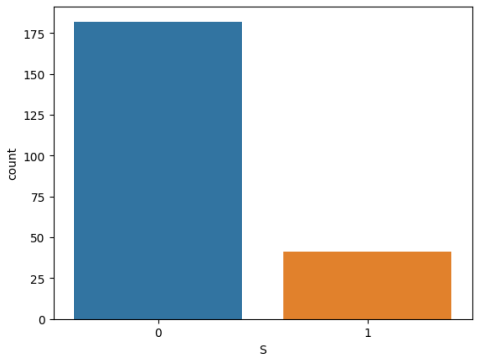
The dataset used in this study had two binary class labels 'P' (progression of pain) and 'S' (structural pain) for each patient. Due to patient withdrawals from the study over time, the labels were measured for 223 patients and the labels for the rest of the patients could not be measured. The binary class labels are then processed in accordance with the patient categorization done in the Approach study and changed to a multi-labelled class as shown below :

'P'	'S'	Class Label
0	0	0
0	1	1
1	0	2
1	1	3

In the original class labels, if a patient has 'P' value as 0 and 'S' value as 0, then the new label for the patient is shown as 0 ('N' in Approach study); if a patient had 'P' value as 0 and 'S' value as 1, then the new label assigned is 1 ('S' in Approach study); if a patient had 'P' value as 1 and 'S' value as 0, then the new label is shown as 2 ('P' in Approach study). Finally, if the patient had both 'P' value as 1 and 'S' value as 1, then the new label is 3 ('P+S' in Approach study). The label distributions in the new class are represented in Figure 2.



(a) Labels distribution for class 'P'



(b) Labels distribution for class 'S'

Figure 1: Class distribution of two binary classes 'P' and 'S' in original data

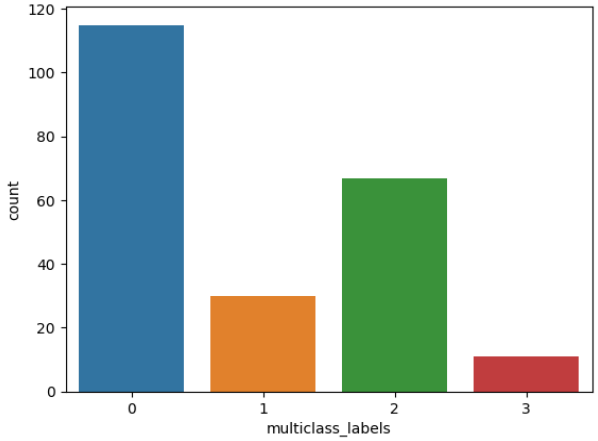


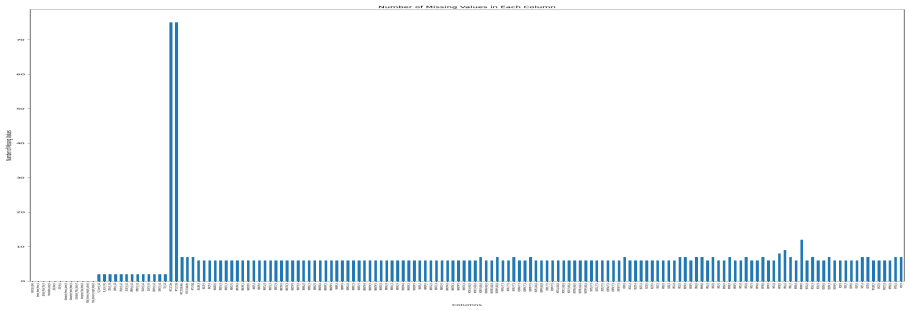
Figure 2: Label distribution after changing to a mutli-labelled class

3.4 Exploratory Data Analysis

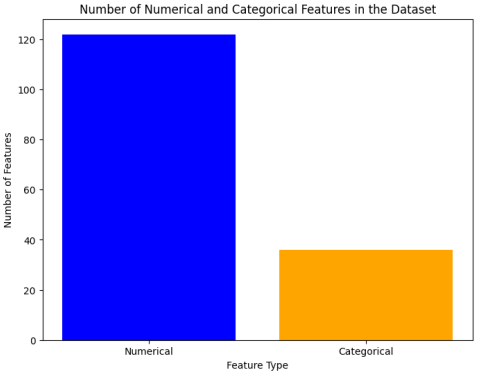
The XRAY, MRI, clinical, biomarkers, tomography and questionnaires datasets contain the feature data of the KOA patients. The combination in which the feature datasets are joined can impact the prediction criteria when trained on ML models. In this experiment, the feature datasets are joined in multiple combinations to train the ML models. Six of such joining combinations have been taken up in this work:

1. XRAY and MRI datasets
2. XRAY and clinical datasets
3. Biomarkers and clinical datasets
4. Tomography and Questionnaires datasets
5. XRAY, MRI, Biomarkers and clinical datasets
6. XRAY, MRI, Biomarkers, clinical, Tomography and Questionnaires datasets (all datasets combined together)

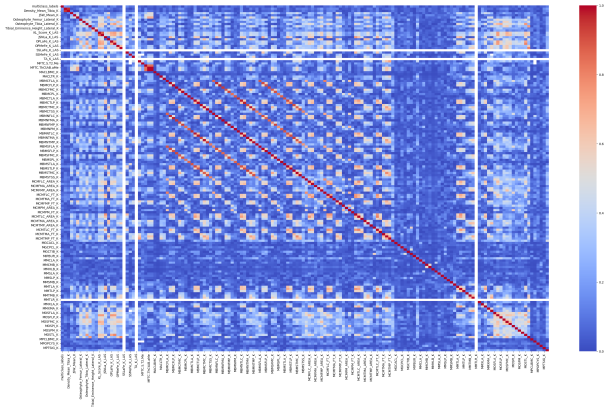
Each of these joined datasets are merged with the class labels. Each of the feature datasets had record of 295-297 patients. But the class labels were recorded for 223 patients. So, on merging the feature datasets with the class labels, the size of the datasets reduced to 221-223 patients.



(a) Missing values plot for each feature in Xray & MRI combined dataset

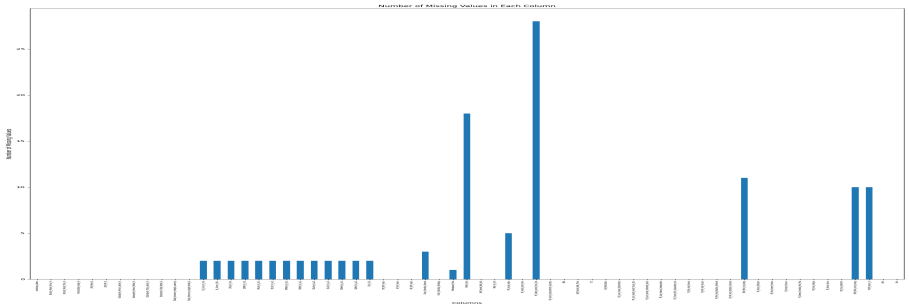


(b) Numerical and categorical features distribution

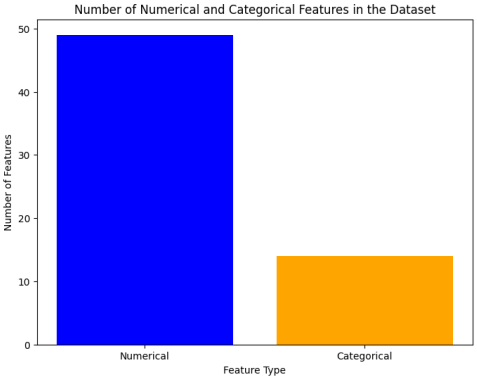


(c) Correlation heatmap

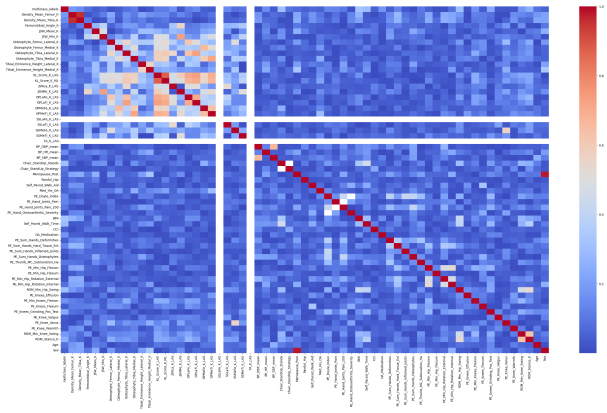
Figure 3: Missing values plot, numerical & categorical features distribution and correlation heatmap of Xray & MRI combined dataset



(a) Missing values plot for each feature in Xray & Clinical combined dataset

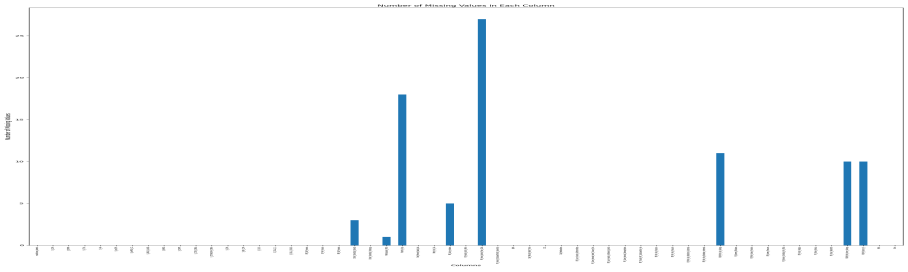


(b) Numerical and categorical features distribution

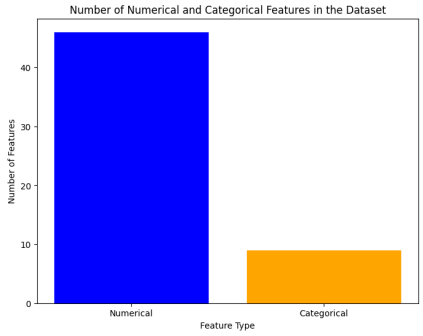


(c) Correlation heatmap

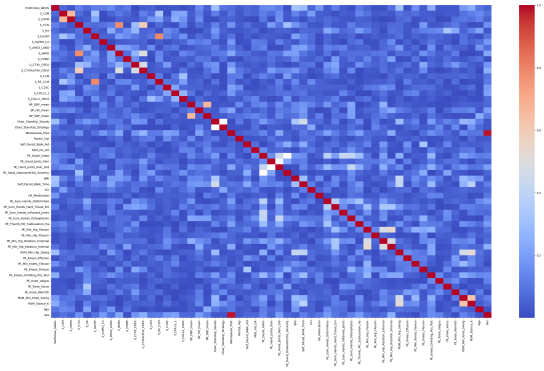
Figure 4: Missing values plot, numerical & categorical features distribution and correlation heatmap of Xray & Clinical combined dataset



(a) Missing values plot for each feature in Biomarkers & Clinical combined dataset

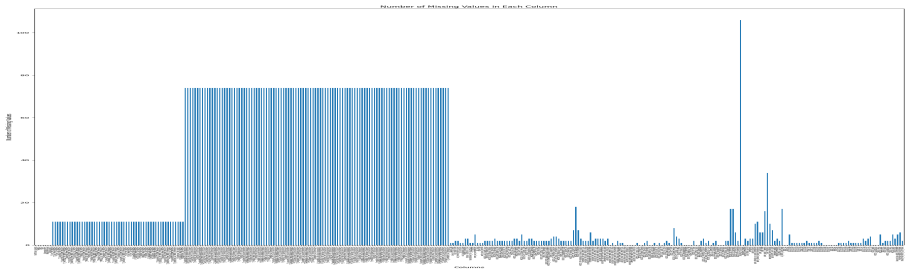


(b) Numerical and categorical features distribution

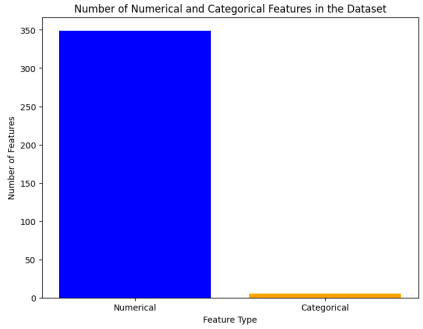


(c) Correlation heatmap

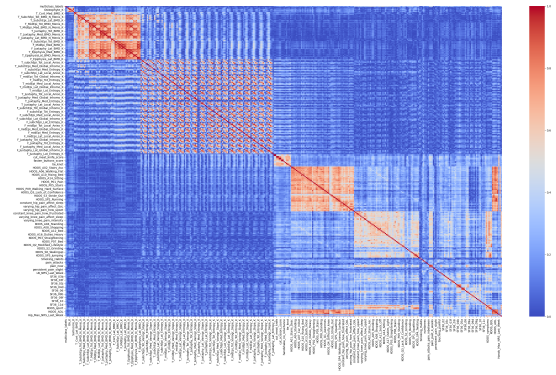
Figure 5: Missing values plot, numerical & categorical features distribution and correlation heatmap of Biomarkers & Clinical combined dataset



(a) Missing values plot for each feature in Tomography & Questionnaire combined dataset

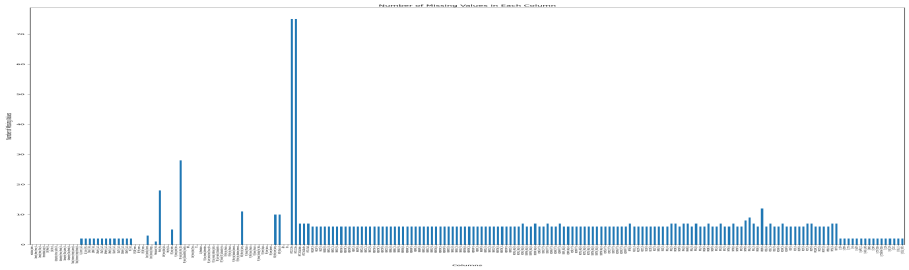


(b) Numerical and categorical features distribution

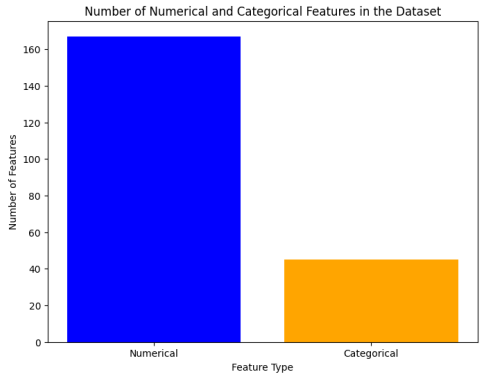


(c) Correlation heatmap

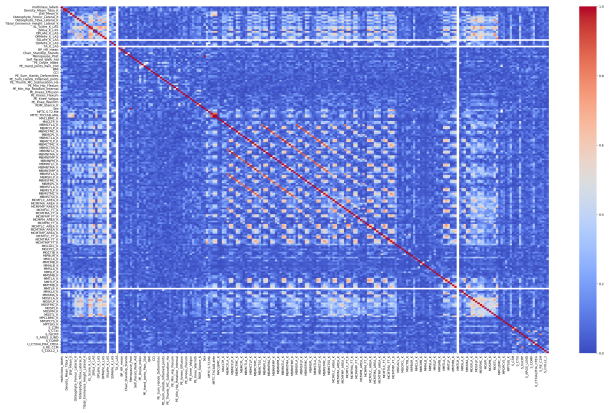
Figure 6: Missing values plot, numerical & categorical features distribution and correlation heatmap of Tomography & Questionnaire combined dataset



(a) Missing values plot for each feature in X-Ray, MRI, Clinical and Biomarkers combined dataset

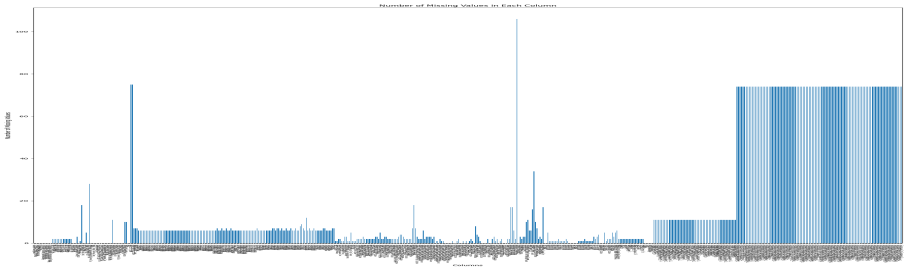


(b) Numerical and categorical features distribution

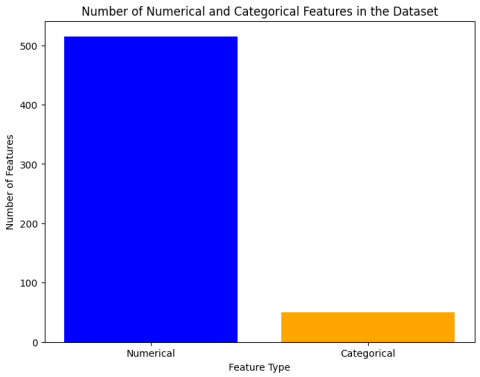


(c) Correlation heatmap

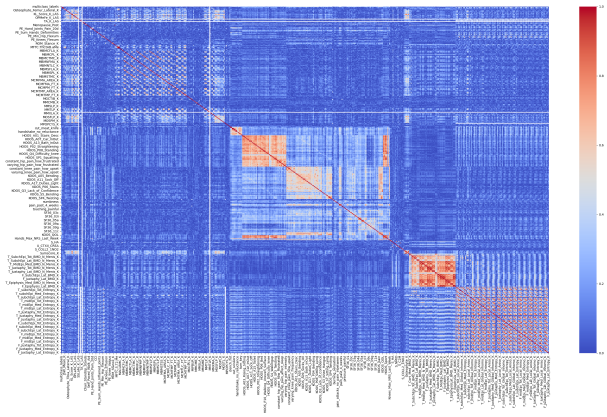
Figure 7: Missing values plot, numerical & categorical features distribution and correlation heatmap of X-Ray, MRI, Clinical and Biomarkers combined dataset



(a) Missing values plot for each feature in X-Ray, MRI, Clinical, Biomarkers, Tomography and Questionnaire combined dataset



(b) Numerical and categorical features distribution



(c) Correlation heatmap

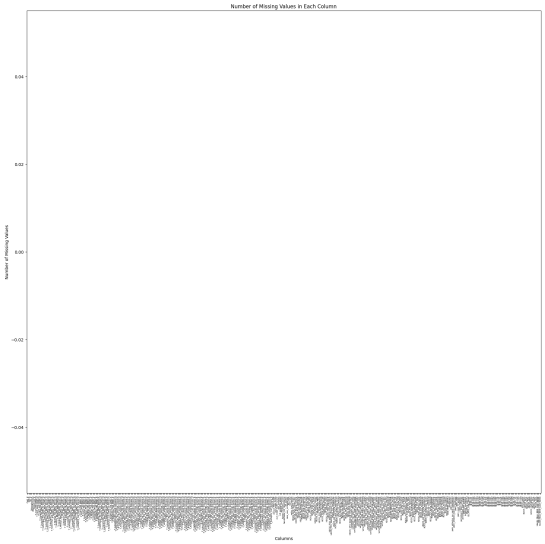
Figure 8: Missing values plot, numerical & categorical features distribution and correlation heatmap of X-Ray, MRI, Clinical, Biomarkers, Tomography and Questionnaire combined dataset

Exploratory Data Analysis is performed to have a better understanding of the datasets. All the datasets are checked for missing values in each of the features, the distribution of categorical and numerical features and the correlation between the features in the datasets. The plots are shown in Figures 3 - 8.

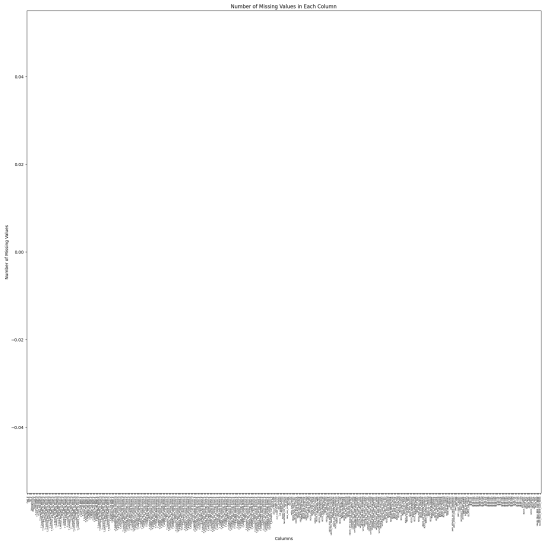
3.5 Data Pre-processing

All features with more than 60% missing values and all the features which had no-variance are dropped from the data. The dataset is then divided into training and testing data by keeping 20% of the total dataset for testing and using the rest for training the model. A data pre-processing pipeline is designed for handling the missing values and scaling the features in the dataset. The datasets contain both numerical and categorical values. The missing numerical values are imputed using the mean of its k nearest neighbors with the value of k set to be 5. The missing categorical attributes are imputed with the mode (most frequent) values of that feature. Feature scaling is performed on the numerical attributes. Since real-world data is used for this study, the features may have various magnitudes, ranges, and units. Thus, in order to standardize the data and have them within a common scale, feature scaling is performed. MinMaxScaler(), a scaling technique inside sklearn is used to scale each feature within the range (0,1). To avoid data leak, the data pre-processing pipeline is fit only on the training data and then used to transform both the training and testing datasets. One hot encoding is also performed on the categorical attributes. Even though some machine learning algorithms can work with categorical inputs, for most machine learning algorithms, the categorical data need to be

changed to numerical data before it is used to train the ML model. One hot encoding creates dummy binary columns for every unique value in the categorical data columns. The value is set as 1 in the new binary column if the binary column was the actual value in the feature of the original dataset, else the binary column value will be set as 0. However, since SMOTENC is used for oversampling the dataset after the pre-processing, it was not working when performed after one hot encoding since the categorical values were already encoded to numerical values. Hence, one hot encoding was not performed as a part of the data pre-processing pipeline and was performed only after oversampling was done.



(a) Missing values removed through imputation in X_train

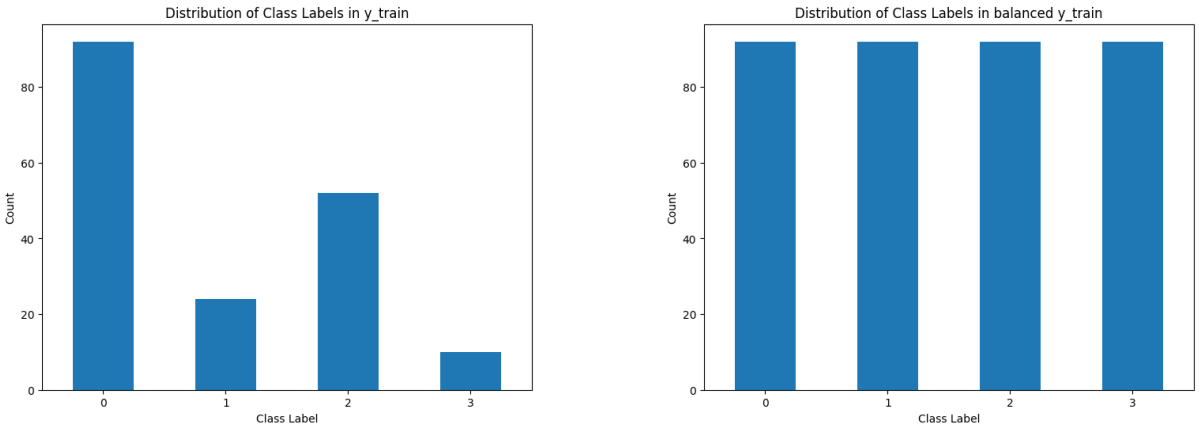


(b) Missing values removed through imputation in X_test

Figure 9

3.6 Oversampling

Oversampling was performed to remove the class imbalance in the dataset. Class imbalance in datasets cause bias which leads to improper model training where the models are biased in favour of the majority class. It is essential to balance the dataset before the model training so that equal weightage can be given to all the classes and any bias can be avoided. Oversampling of the minority class is performed to remove the class imbalance and obtain a balanced dataset for the classifier to yield better prediction [24]. Since the data that is being considered has both categorical and numerical data, SMOTENC is used [25]. SMOTENC was chosen over SMOTE for the oversampling as SMOTENC works well with both categorical and numerical data. Initially, SMOTE was used after one hot encoding the data. Even, though SMOTE works perfectly with numerical data, even after one encoding, it can't quite understand the categorical nature of the data.



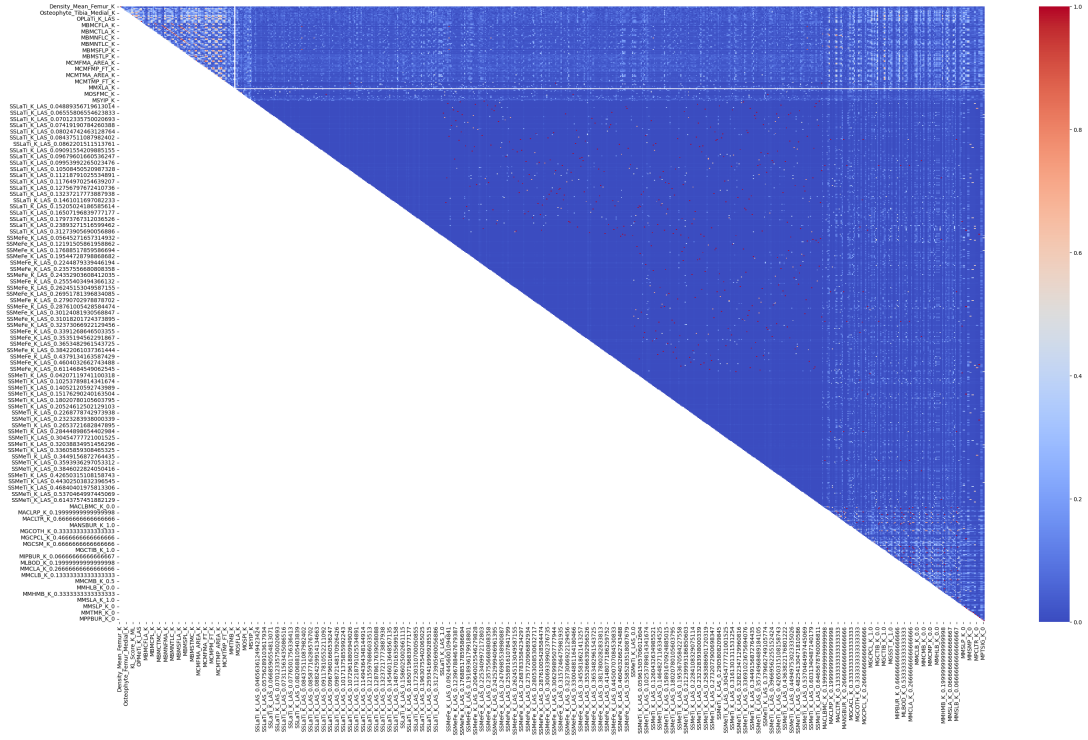
(a) Before oversampling

(b) After oversampling

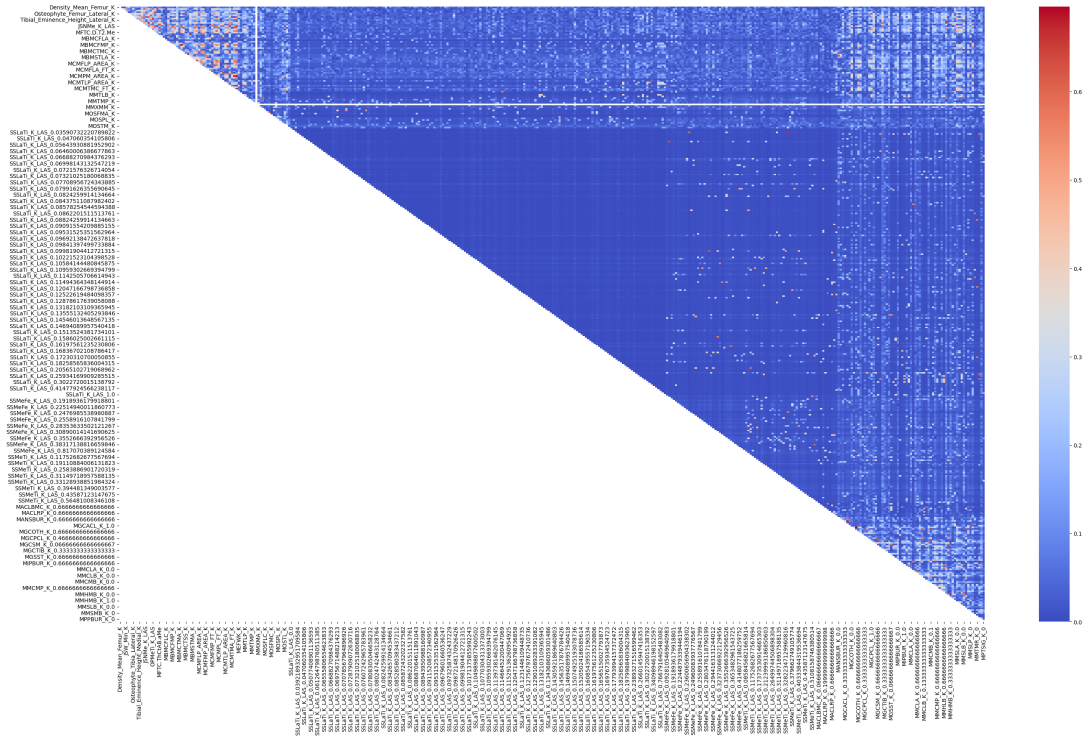
Figure 10: Oversampling the data

3.7 Feature selection using correlation matrix

The correlation matrix of the dataset is computed from which the correlations between each of the features could be seen. The lower triangle of the correlation matrix is masked and only the upper triangle is considered since these work like mirror images. As for the correlation coefficients, only their absolute values are taken into consideration. A threshold is set for 0.7 and one of the features having a correlation more than the threshold is dropped from the dataset. The action is performed on the training dataset and the features are selected by removing their highly correlated co-features. The same list of selected features are then kept for the test dataset. A heatmap plot of the correlation matrix is generated which shows the correlation between the features where lower correlations are identified by the color blue and the higher correlations are denoted by red. Another heatmap is generated after the correlated features are removed as shown in Figures 11 - 16.

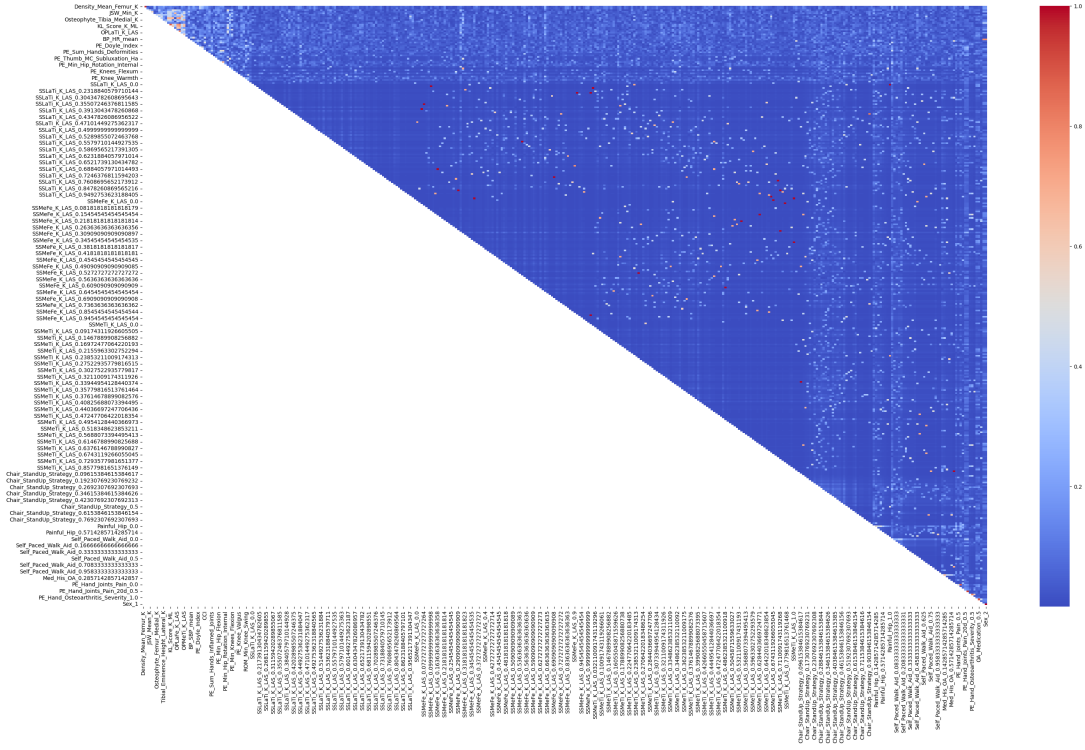


(a) Correlation heatmap before feature elimination

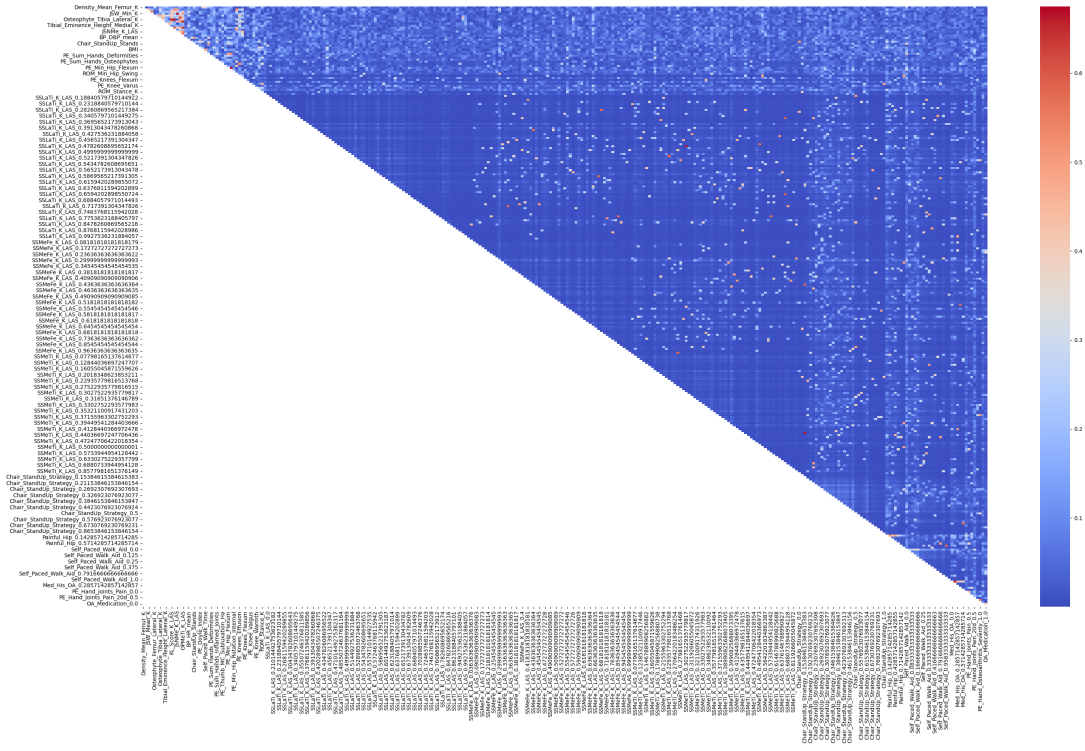


(b) Correlation heatmap after feature elimination

Figure 11: For XRAY and MRI combined dataset

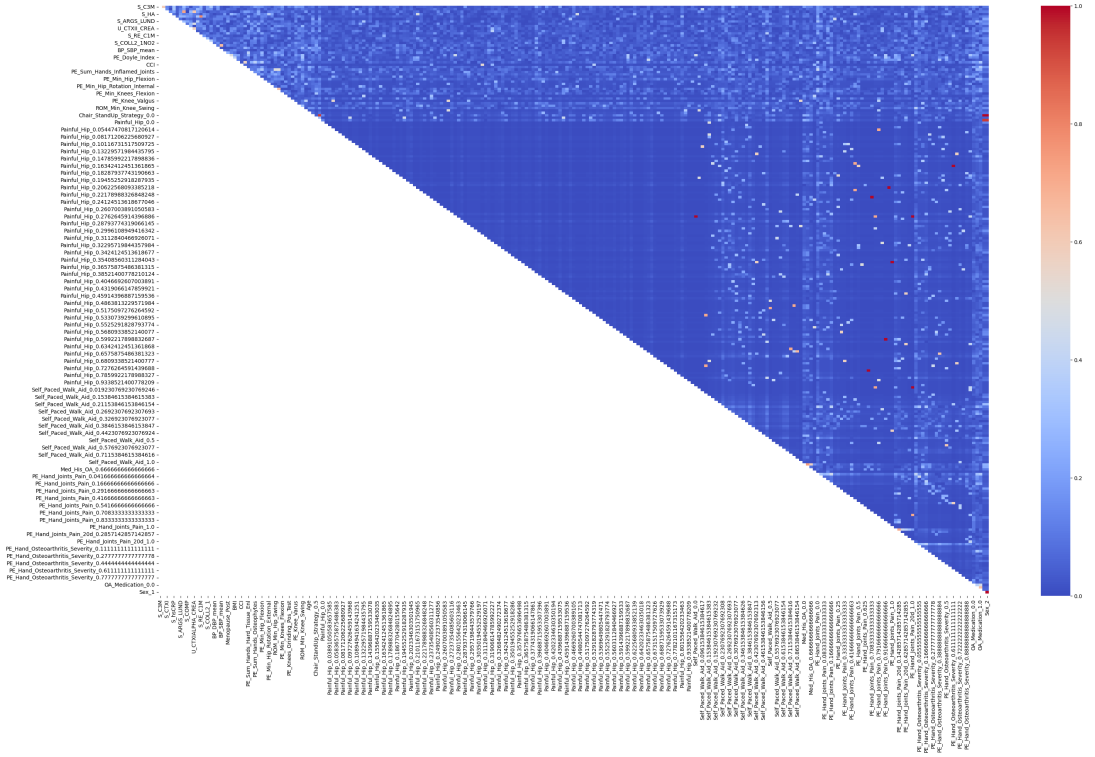


(a) Correlation heatmap before feature elimination

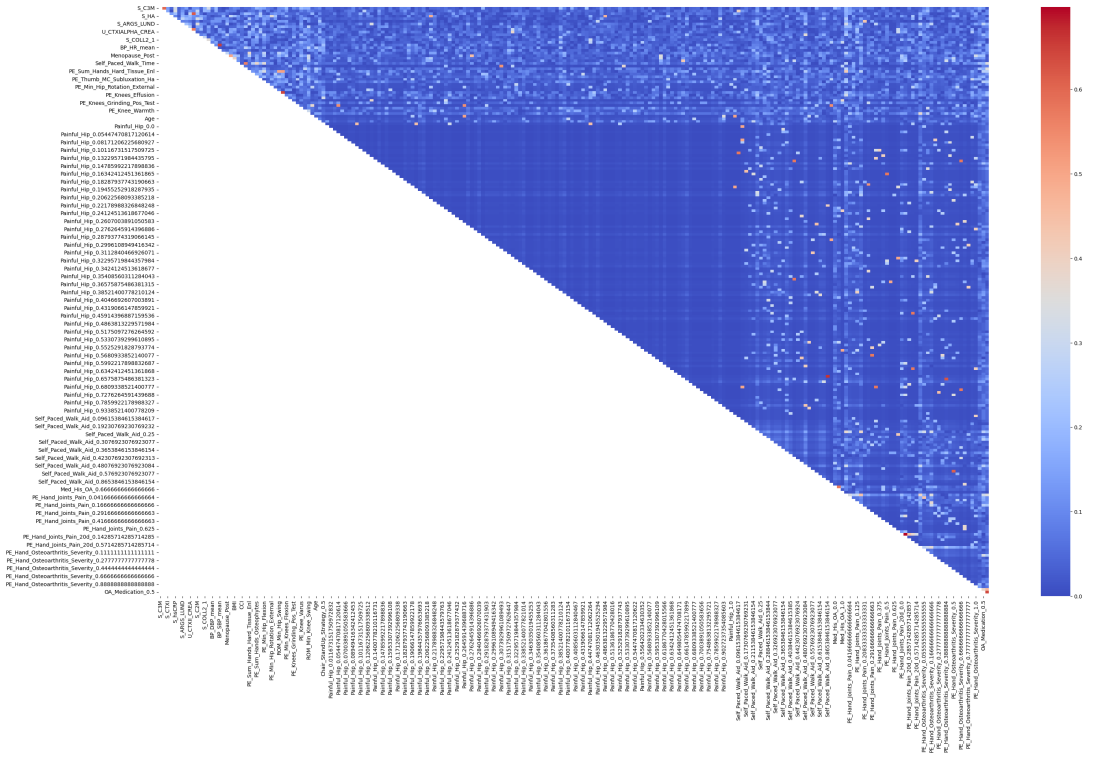


(b) Correlation heatmap after feature elimination

Figure 12: For XRAY and Clinical combined dataset

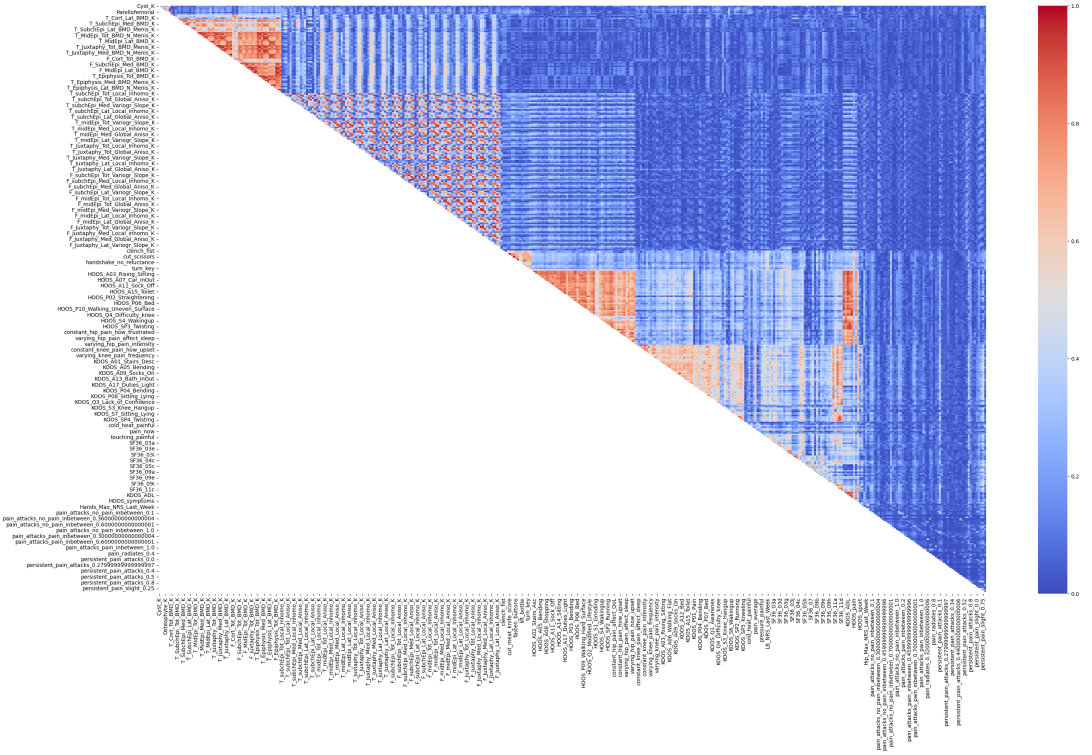


(a) Correlation heatmap before feature elimination

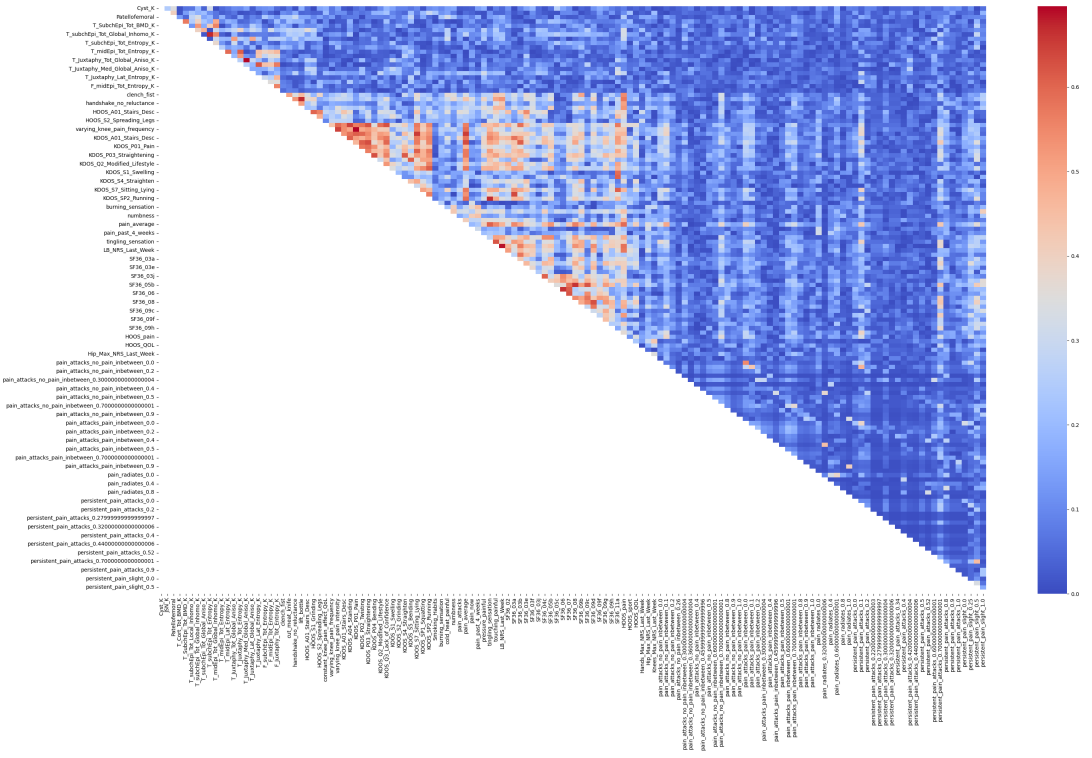


(b) Correlation heatmap after feature elimination

Figure 13: For biomarkers and clinical combined dataset

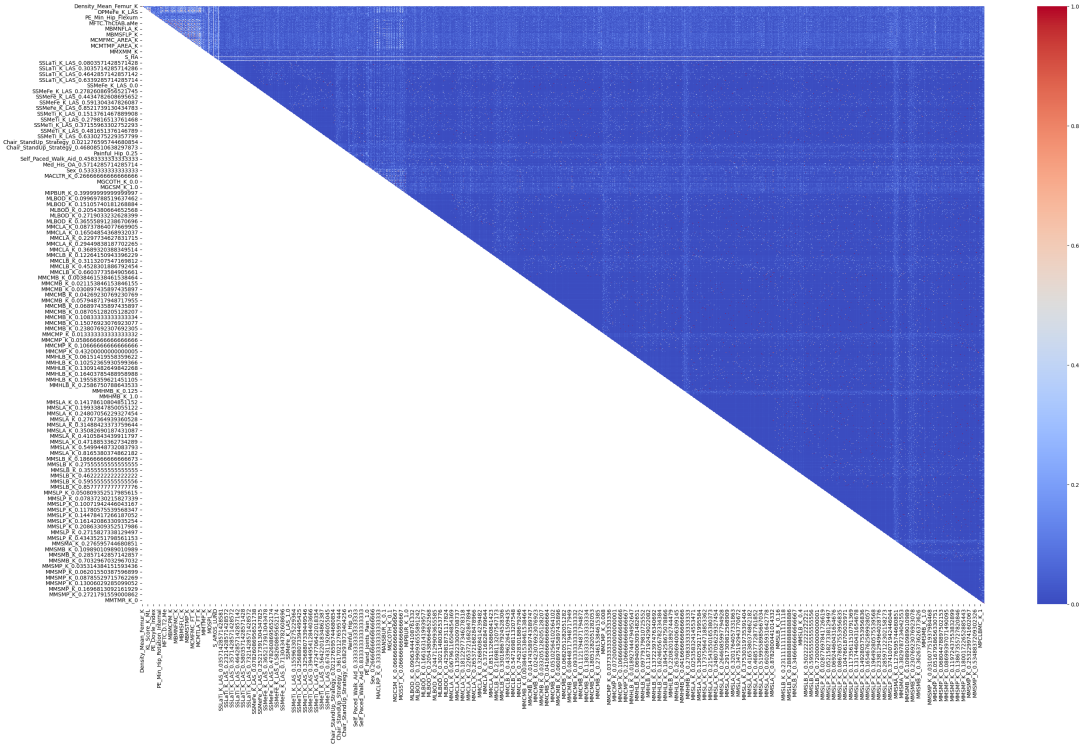


(a) Correlation heatmap before feature elimination

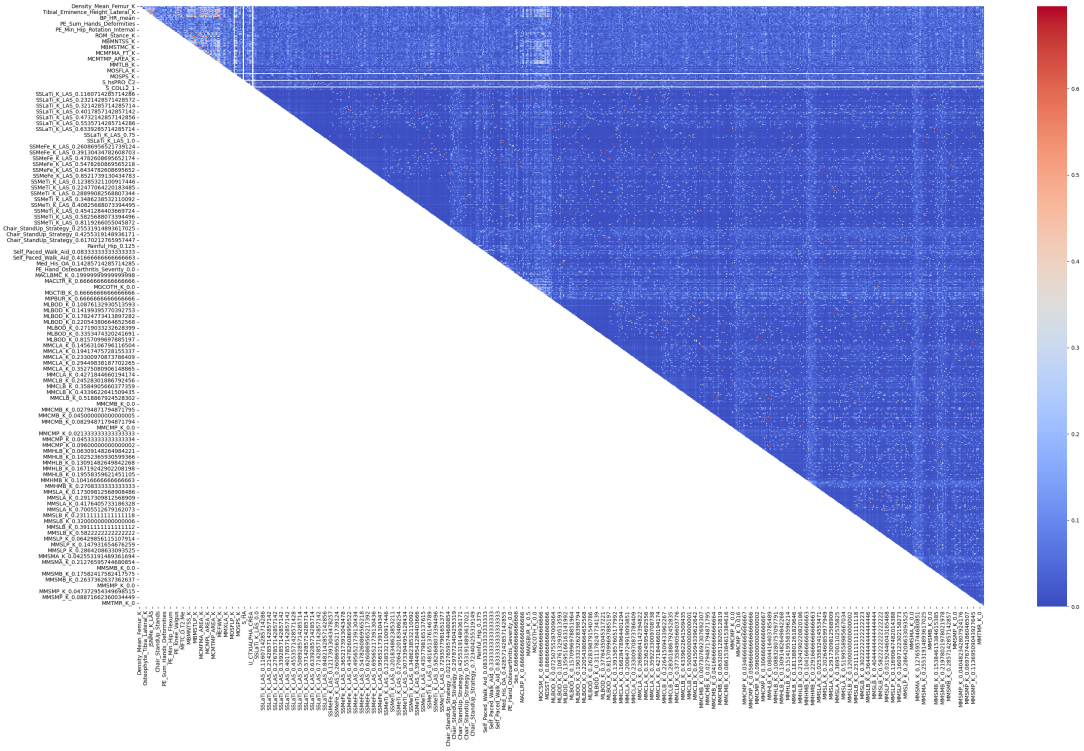


(b) Correlation heatmap after feature elimination

Figure 14: For Tomography and Questionnaires combined dataset

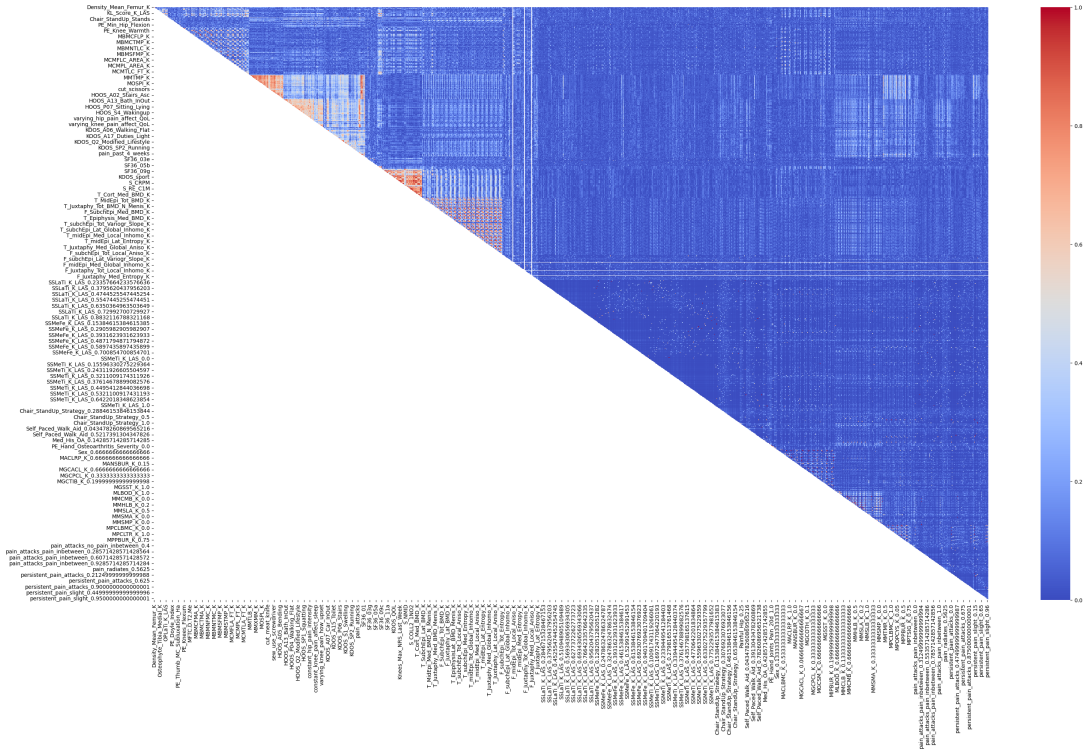


(a) Correlation heatmap before feature elimination

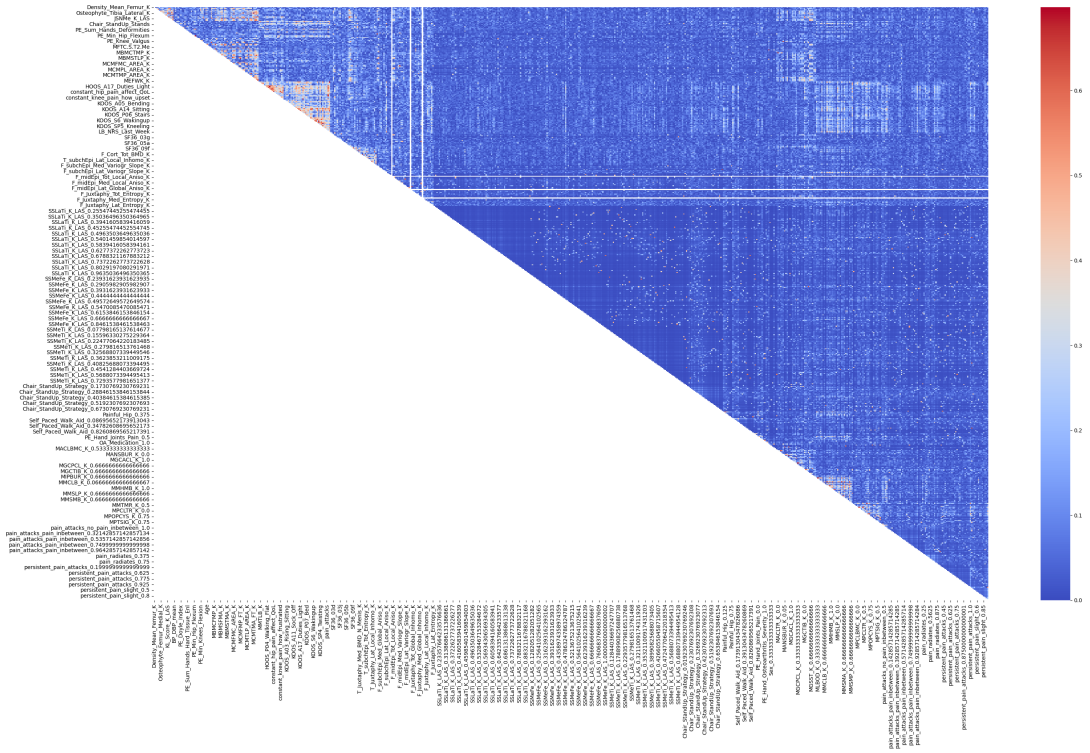


(b) Correlation heatmap after feature elimination

Figure 15: For XRAY, MRI, Clinical and Biomarkers dataset



(a) Correlation heatmap before feature elimination



(b) Correlation heatmap after feature elimination

Figure 16: For XRAY, MRI, Clinical, Biomarkers, Tomography and Questionnaires data

3.8 Model Training and Evaluation

Four machine learning algorithms are used in this study – RandomForest Classifier, k Nearest Neighbor classifier, Decision Tree classifier and Artificial Neural Networks.
For cross validation, Stratified K Fold CV is used with 3 splits for hyperparameter tuning. For K nearest neighbour model, a list of nearest neighbours is used as hyperparameters. For the Random Forest classifier, potential values of estimators and maximum tree depths is used. For Decision Tree classifier only values for maximum tree depths are used and for Artificial Neural Networks, a list of the hidden layer sizes are used. The classifer

Table 1: Models parameters with ranges

Classifier	Parameter	Range
KNN	n_neighbors	[3, 5, 7,11,15,20]
Random Forest	n_estimators	[100, 200, 300, 400, 500]
	max_depth	[None, 10, 20, 30, 40, 50]
ANN	hidden_layer_sizes	[(100,), (50, 50), (30, 30, 30)]
Decision Tree	max_depth	[None, 10, 20, 30, 40, 50]

algorithms, the parameters and the Stratified K Fold CV together is considered to be a model and Halving Grid Search cross validation is used on top of it to perform a grid search with halving after each step to find the best performing parameters. The best models for each of the classifiers after the hyperparameter tuning are used to predict the outcomes of the test datasets. The performance is measured by calculating the accuracy of the models and macro F1 score.
Permutation Feature importance is used to identify the top performing features for each of the models and then used to create new training and datasets. All the models are then trained on the top performing features datasets.
For k nearest neighbours and artificial neural networks models, a SHAP KernelExplainer and for random forest and decision tree models, a SHAP TreeExplainer is created using the newly trained models and the predict method. SHAP values for each of the models are computed using the top features test datasets for each of the models and a summary plot is generated to visualize the impact of the top features on the predictions.

4 RESULTS

The results for each of the combinations with the different models are shown below:

Table 2: X-Ray and MRI Dataset

Classifier	Accuracy	Macro F1 Score	Best Parameter
KNN	33.3%	22.6%	'n_neighbors': 3
Random Forest	46.6%	32.2%	'max_depth': 20, 'n_estimators': 100
ANN	37.7%	19.9%	'hidden_layer_sizes': (50, 50)
Decision Tree	44.4%	33.4%	'max_depth': 50

Table 3: X-Ray and Clinical Dataset

Classifier	Accuracy	Macro F1 Score	Best Parameter
KNN	46.6%	35.7%	'n_neighbors': 3
Random Forest	44.4%	23.8%	'max_depth': 20, 'n_estimators': 500
ANN	42.2%	28.3%	'hidden_layer_sizes': (100,)
Decision Tree	26.6%	14.7%	'max_depth': None

Table 4: Biomarkers and Clinical Dataset

Classifier	Accuracy	Macro F1 Score	Best Parameter
KNN	35.5%	27.9%	'n_neighbors': 3
Random Forest	42.2%	22.7%	'max_depth': 50, 'n_estimators': 300
ANN	42.2%	28.5%	'hidden_layer_sizes': (30,30,30)
Decision Tree	37.7%	27%	'max_depth': 40

Table 5: Tomography and Questionnaires Dataset

Classifier	Accuracy	Macro F1 Score	Best Parameter
KNN	33.3%	27.4%	'n_neighbors': 3
Random Forest	66.6%	36.1%	'max_depth': 10, 'n_estimators': 100
ANN	33.3%	20.2%	'hidden_layer_sizes': (100,)
Decision Tree	46.6%	34.7%	'max_depth': 20

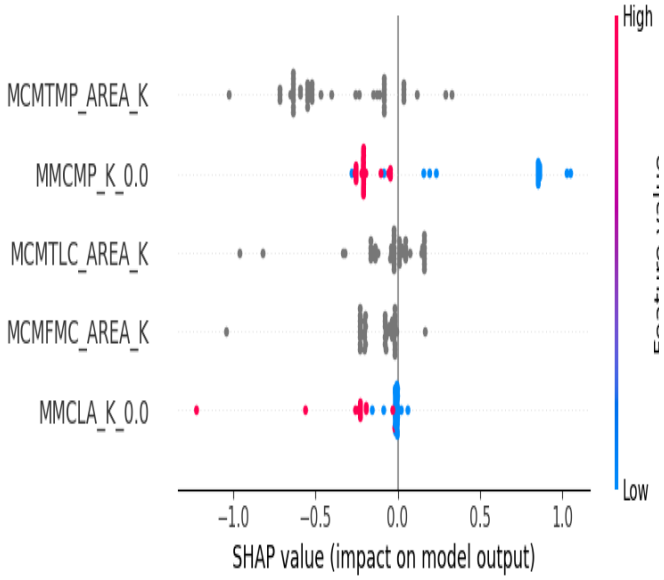
Table 6: X-Ray, MRI, CLinical and Biomarkers Dataset

Classifier	Accuracy	Macro F1 Score	Best Parameter
KNN	28.8%	11.8%	'n_neighbors': 3
Random Forest	53.3%	29.7%	'max_depth': 20, 'n_estimators': 500
ANN	42.2%	24.9%	'hidden_layer_sizes': (100,)
Decision Tree	24.4%	14.1%	'max_depth': None

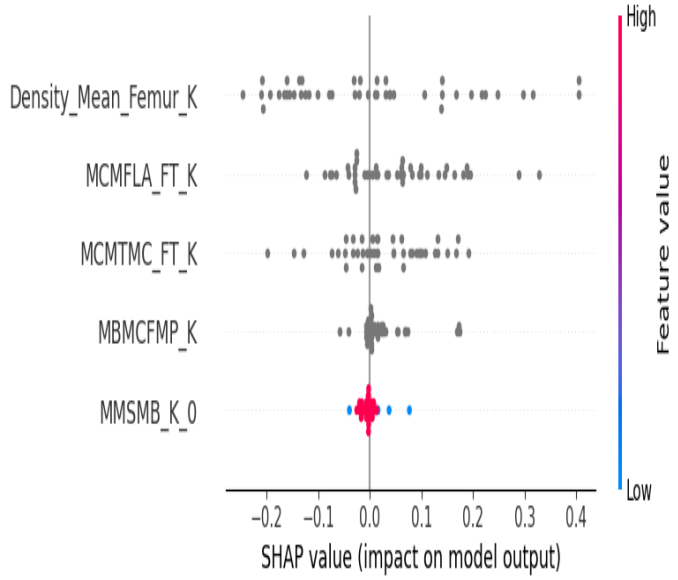
Table 7: X-Ray, MRI, CLinical, Biomarkers, Tomography and Questionnaires Dataset

Classifier	Accuracy	Macro F1 Score	Best Parameter
KNN	31.1%	25.2%	'n_neighbors': 3
Random Forest	62.2%	33.5%	'max_depth': 30, 'n_estimators': 500
ANN	55.5%	34.8%	'hidden_layer_sizes': (100,)
Decision Tree	42.2%	26.5%	'max_depth': 40

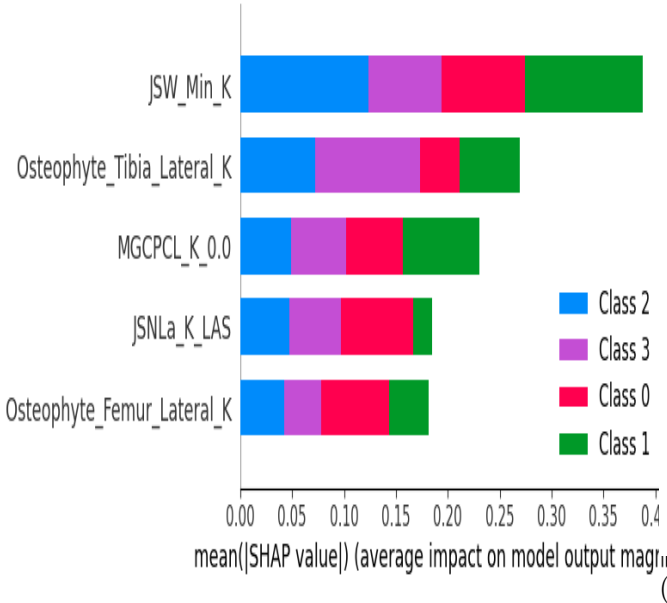
The shap plots (for XRAY and MRI combined datasets) plotted by training the models using the top features:
For KNN:'MMCLA_K_0.0', 'MCMTMP_AREA_K', 'MCMTLC_AREA_K', 'MCMFMC_AREA_K', 'MMCMP_K_0.0'
For Random Forest CLassifier: 'Density_Mean_Femur_K', 'MCMFLA_FT_K', 'MCMTMC_FT_K', 'MBMCFMP_K', 'MMSMB_K_0'
For ANN: 'MGCPCL_K_0.0', 'Osteophyte_Tibia_Lateral_K', 'JSW_Min_K', 'Osteophyte_Femur_Lateral_
'JSNLa_K_LAS' For Decision Tree Classifier: 'MMCLA_K_0.0', 'MBMSTMC_K', 'Osteophyte_Tibia_Medial_K', 'MMHMB_K_1.0', 'JSW_Mean_K'



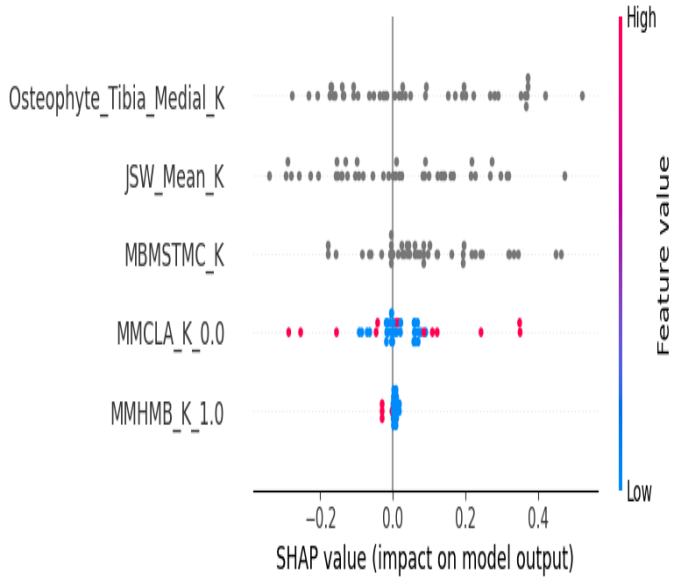
(a) SHAP Summary Plot for KNN model



(b) SHAP Summary Plot for Random Forest CLassifier model



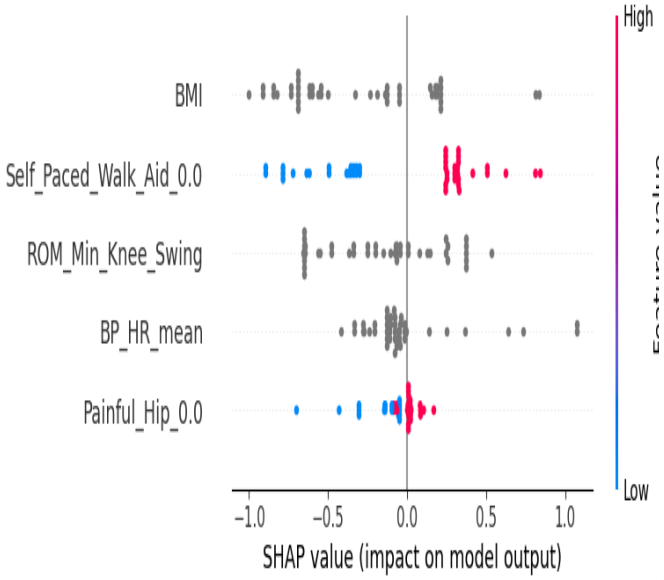
(c) SHAP Summary Plot for ANN model



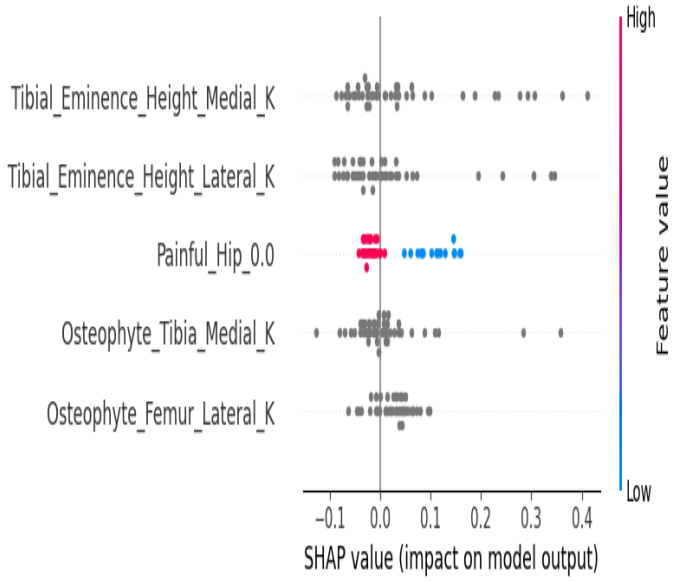
(d) SHAP Summary Plot for Decision Tree classifer model

The shap plots (for XRAY and Clinical combined datasets) plotted by training the models using the top features:
For KNN: 'Painful_Hip_0.0', 'BMI', 'BP_HR_mean', 'Self_Paced_Walk_Aid_0.0', 'ROM_Min_Knee_Swing'

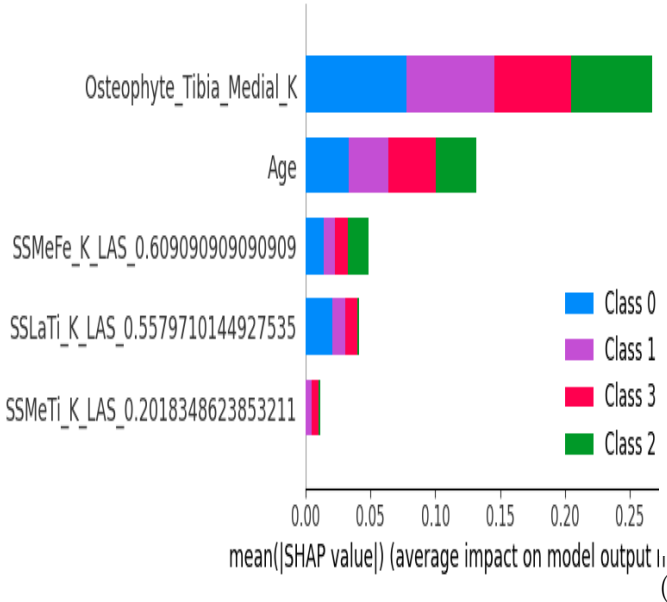
For Random Forest CLassifier: 'Osteophyte_Tibia_Medial_K', 'Osteophyte_Femur_Lateral_K','Tibial_Em
'Painful_Hip_0.0', 'Tibial_Eminence_Height_Lateral_K'
For ANN: 'SSMeFe_KLAS_0.609090909090909', Osteophyte_Tibia_Medial_K', 'SSMeTi_K_LAS_0.20183
'SSLaTi_K_LAS_0.5579710144927535 For Decision Tree Classifier: 'PE_Sum_Hands_Deformities',
'Osteophyte_Tibia_Lateral_K', 'OPMeFe_K_LAS', 'BMI', 'PE_Sum_Hands_Inflamed_Joints'



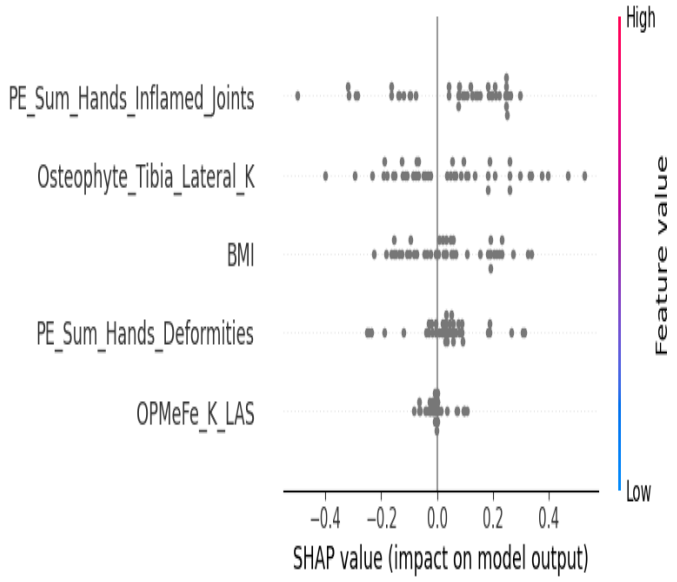
(a) SHAP Summary Plot for KNN model



(b) SHAP Summary Plot for Random Forest CLassifier model



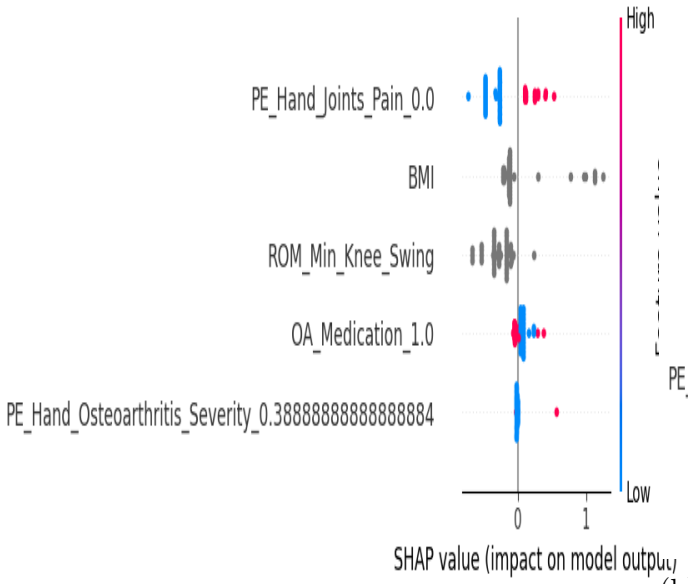
(c) SHAP Summary Plot for ANN model



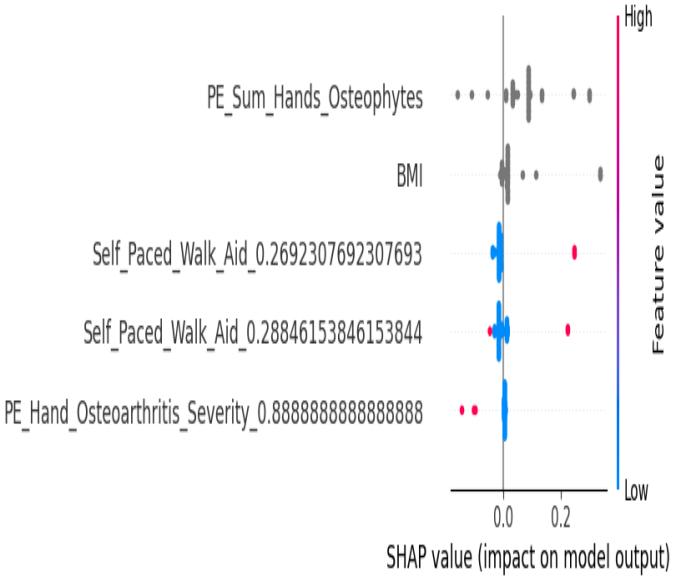
(d) SHAP Summary Plot for Decision Tree classifer model

The shap plots (for Biomarkers and Clinical combined datasets) plotted by training the models using the top features:
For KNN: 'BMI', 'PE_Hand_Joints_Pain_0.0', 'OA_Medication_1.0', 'PE_Hand_Osteoarthritis_Severity_0.
'ROM_Min_Knee_Swing'
For Random Forest CLassifier: 'Self_Paced_Walk_Aid_0.2692307692307693', 'BMI', 'PE_Sum_Hands_Ost
'Self_Paced_Walk_Aid_0.28846153846153844', 'PE_Hand_Osteoarthritis_Severity_0.8888888888888888'

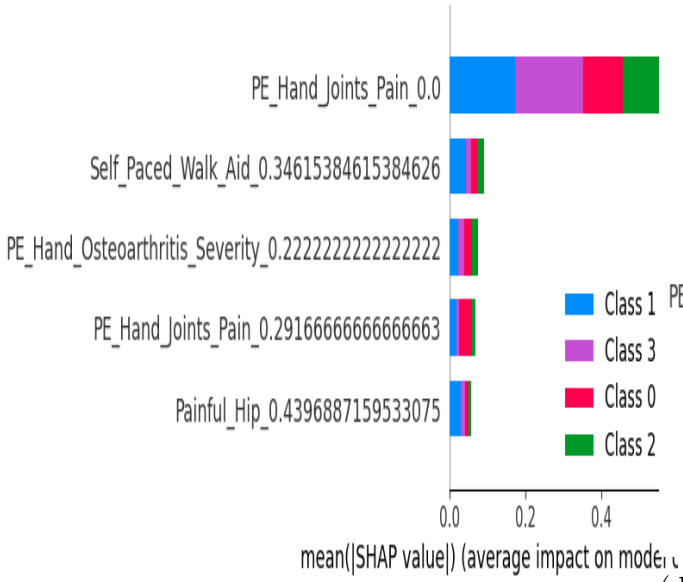
For ANN: 'PE_Hand_Joints_Pain_0.0', 'PE_Hand_Osteoarthritis_Severity_0.2222222222222222', 'PE_Hand_Joints_Pain_0.29166666666666663', 'Self_Paced_Walk_Aid_0.34615384615384626', 'Painful_Hip_0.4396887159533075' For Decision Tree Classifier: 'PE_Hand_Joints_Pain_0.0', 'S_HA', 'S_CTXI', 'PE_Min_Knees_Flexion', 'ROM_Min_Hip_Swing'



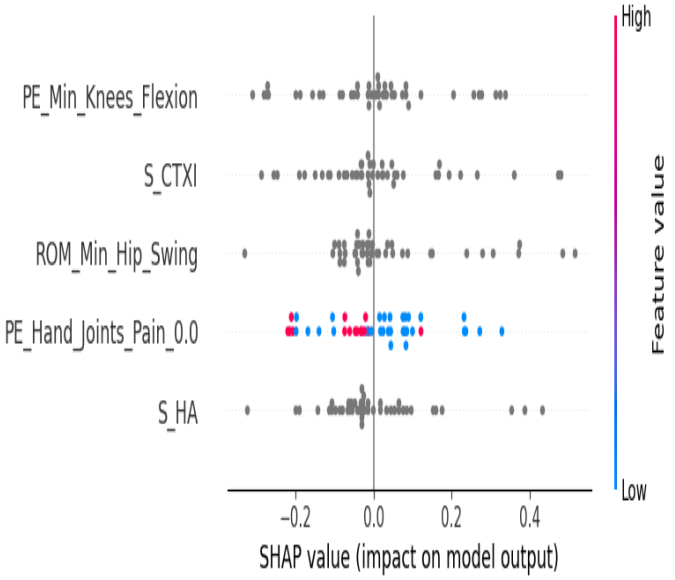
(a) SHAP Summary Plot for KNN model



(b) SHAP Summary Plot for Random Forest CLassifier model



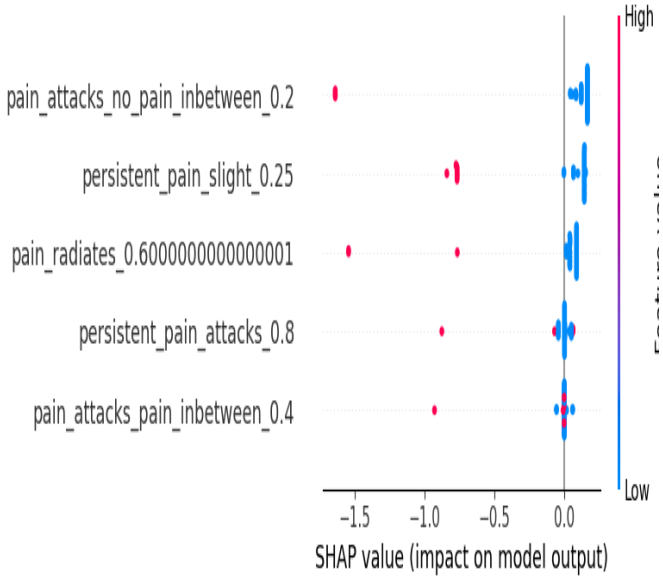
(c) SHAP Summary Plot for ANN model



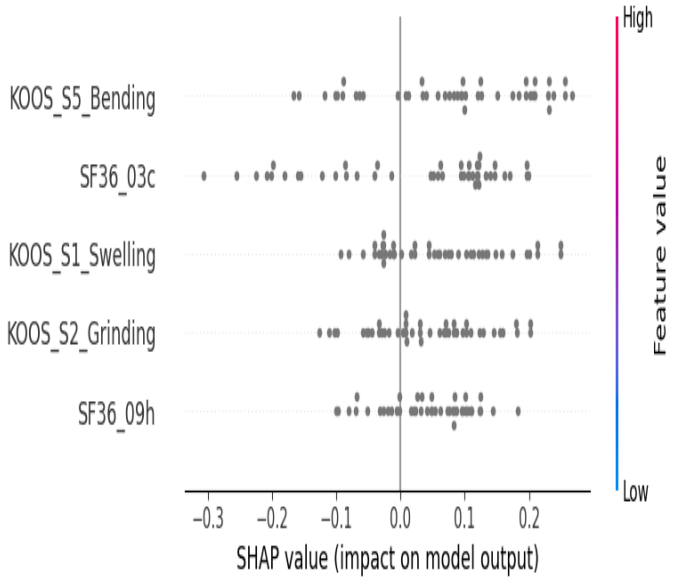
(d) SHAP Summary Plot for Decision Tree classifer model

The shap plots (For Tomography and Questionnaires combined datasets) plotted by training the models using the top features:
For KNN: 'persistent_pain_attacks_0.8','persistent_pain_slight_0.25', 'pain_radiates_0.6000000000000001' 'pain_attacks_pain_inbetween_0.4', 'pain_attacks_no_pain_inbetween_0.2'
For Random Forest CLassifier: 'KOOS_S2_Grinding', 'KOOS_S5_Bending', 'KOOS_S1_Swelling', 'SF36_03c', 'SF36_09h'
For ANN: 'persistent_pain_attacks_0.4', 'T_midEpi_Lat_Entropy_K', 'constant_knee_pain_affect_QoL', 'pain_attacks_pain_inbetween_0.4', 'pain_attacks_no_pain_inbetween_0.8' For Decision Tree

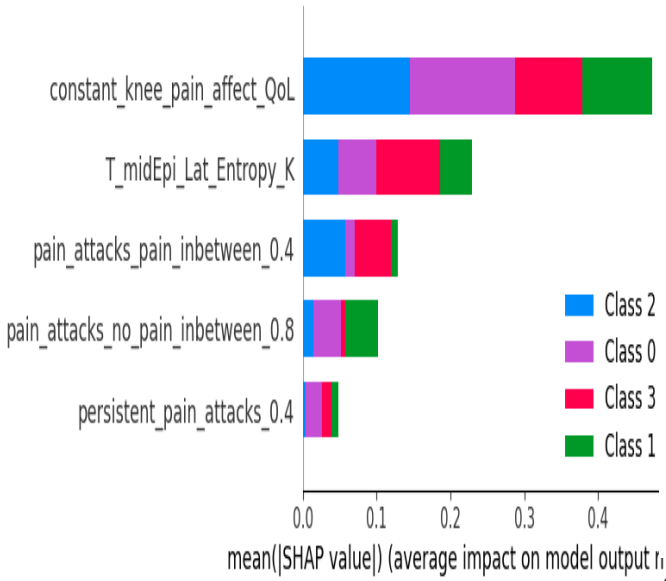
Classifier: 'SF36_03c', 'T_Juxtaphy_Tot_Global_Aniso_K', 'KOOS_S5_Bending', 'SF36_09g', 'T_Juxtaphy_Tot_Entropy_K'



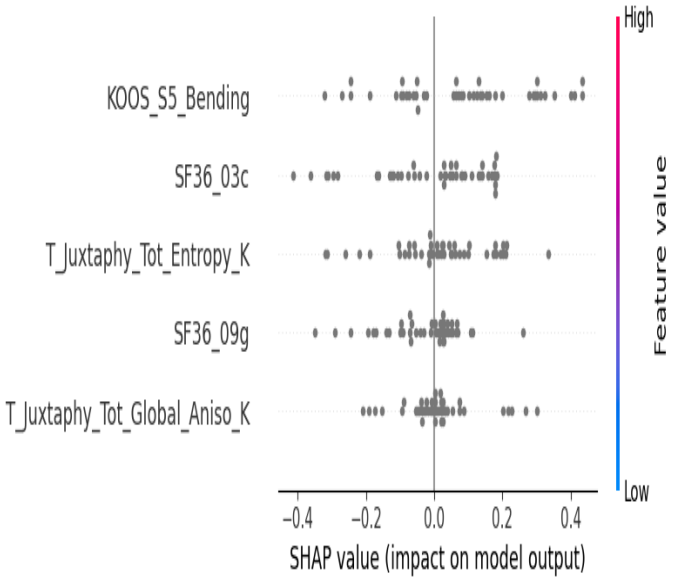
(a) SHAP Summary Plot for KNN model



(b) SHAP Summary Plot for Random Forest CLassifier model

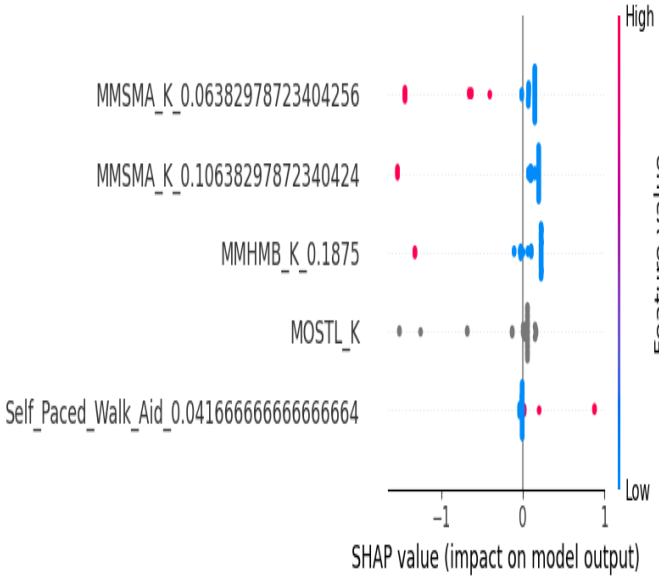


(c) SHAP Summary Plot for ANN model

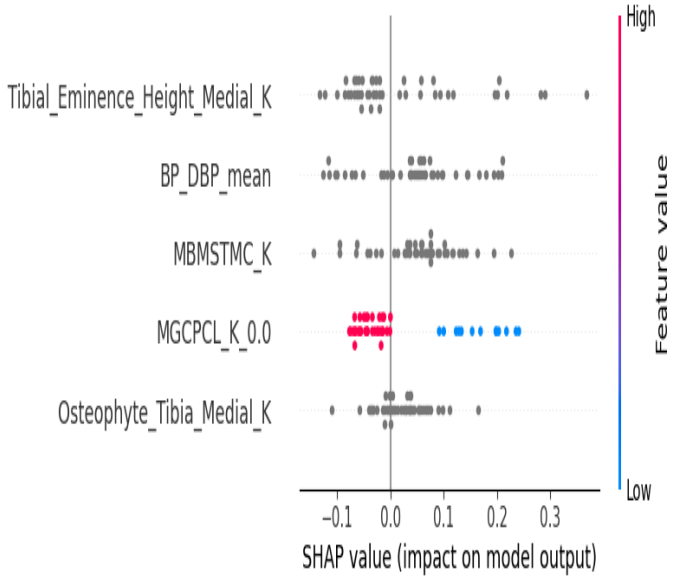


(d) SHAP Summary Plot for Decision Tree classifer model

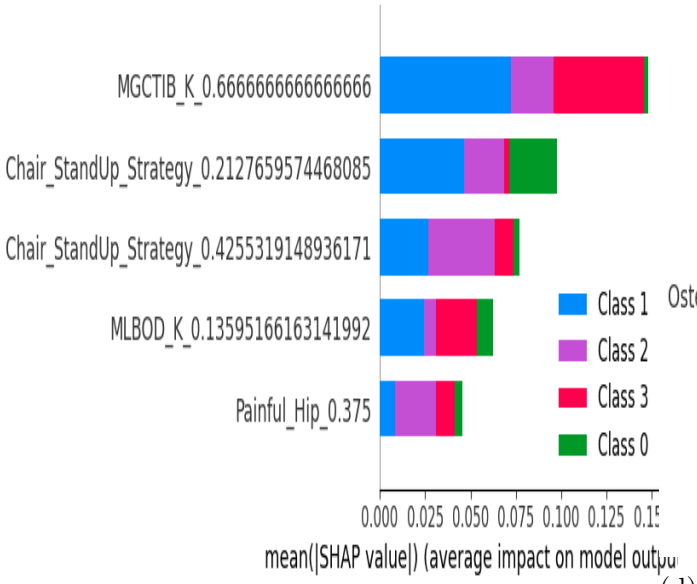
The shap plots (For XRAY, MRI, Biomarkers and clinical combined datasets) plotted by training the models using the top features:
For KNN: 'MMHMB_K_0.1875', 'MOSTL_K', 'Self_Paced_Walk_Aid_0.041666666666666664', 'MMSMA_K_0.06382978723404256', 'MMSMA_K_0.10638297872340424'
For Random Forest CLassifier: 'Osteophyte_Tibia_Medial_K', 'BP_DBP_mean', 'Tibial_Eminence_Height_Medial_K', 'MGCPCL_K_0.0', 'MBMSTMC_K'
For ANN: 'MLBOD_K_0.13595166163141992', 'MGCTIB_K_0.6666666666666666', 'Chair_StandUp_Stra 'Painful_Hip_0.375', 'Chair_StandUp_Strategy_0.2127659574468085' For Decision Tree Classifier: 'ROM_Min_Hip_Swing', 'Osteophyte_Femur_Medial_K', 'MMCMP_K_0.008', 'PE_Min_Hip_Flexur 'PE_Min_Knees_Flexion'



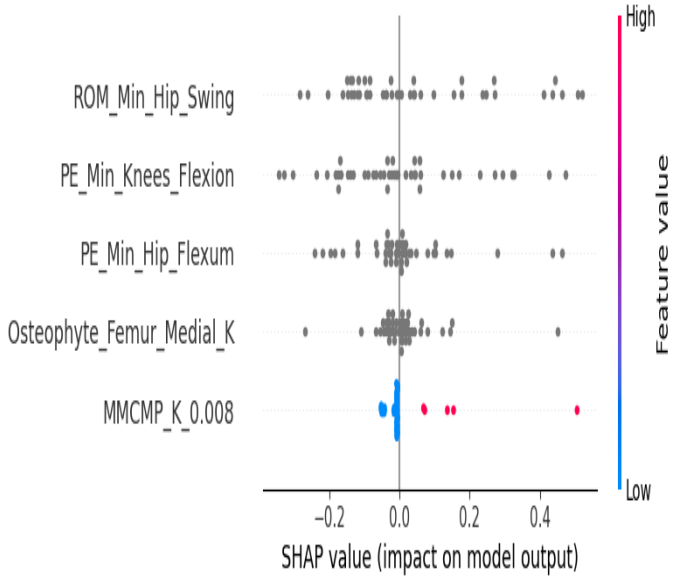
(a) SHAP Summary Plot for KNN model



(b) SHAP Summary Plot for Random Forest CLassifier model

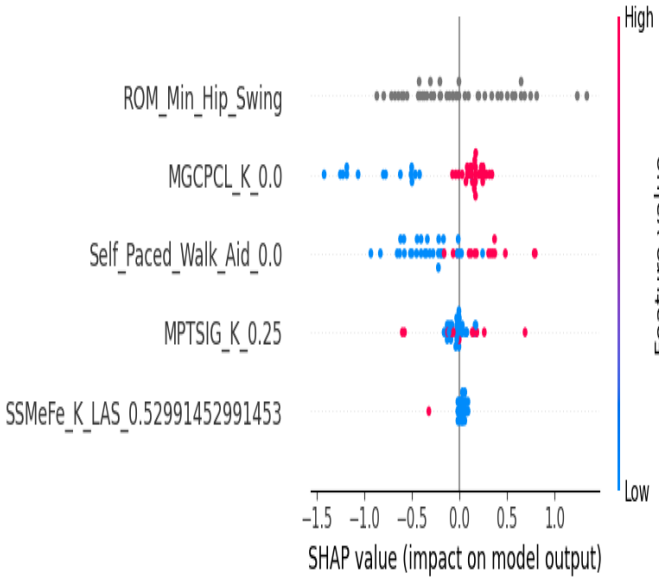


(c) SHAP Summary Plot for ANN model

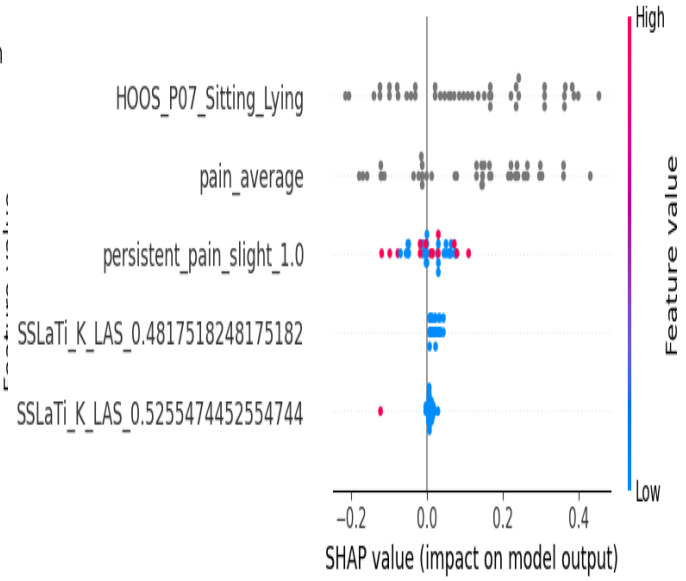


(d) SHAP Summary Plot for Decision Tree classifer model

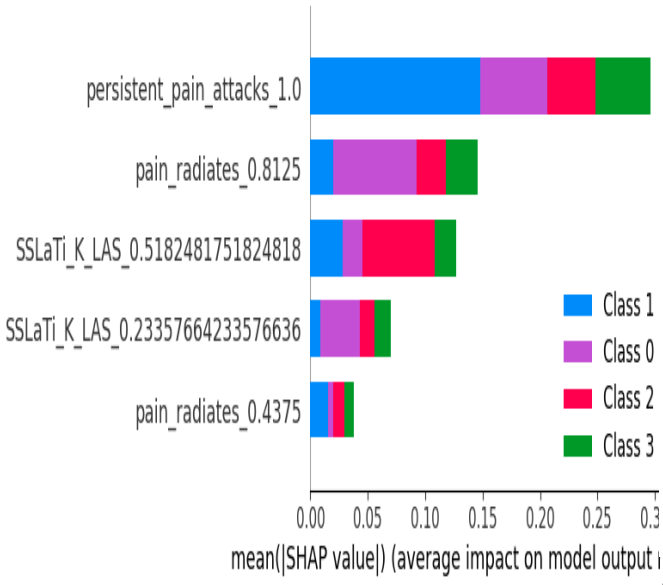
The shap plots (For XRAY, MRI, Biomarkers, clinical, tomography and Questionnaires combined datasets) plotted by training the models using the top features:
For KNN: 'MPTSIG_K_0.25', 'MGCPCL_K_0.0', 'Self_Paced_Walk_Aid_0.0', 'ROM_Min_Hip_Swing', 'SSMeFe_K_LAS_0.52991452991453'
For Random Forest CLassifier: ['HOOS_P07_Sitting_Lying', 'pain_average', 'persistent_pain_slight_1.0', 'SSLaTi_K_LAS_0.4817518248175182', 'SSLaTi_K_LAS_0.5255474452554744']
For ANN: 'persistent_pain_attacks_1.0', 'pain_radiates_0.8125', 'SSLaTi_K_LAS_0.23357664233576636', 'SSLaTi_K_LAS_0.5182481751824818', 'pain_radiates_0.4375' For Decision Tree Classifier: 'constant_knee_pain_affect_QoL','MPCLTR_K_0.0', 'MBMSFLP_K', 'varying_hip_pain_intensity', 'KOOS_A09_Socks_On'



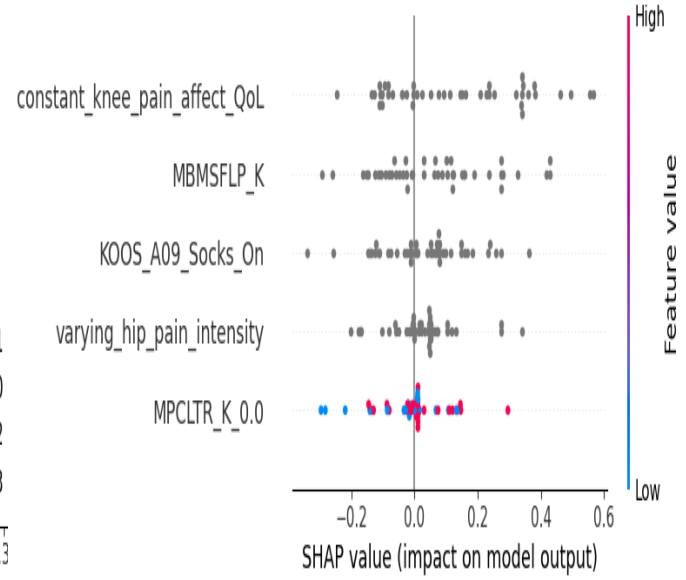
(a) SHAP Summary Plot for KNN model



(b) SHAP Summary Plot for Random Forest CLassifier model



(c) SHAP Summary Plot for ANN model



(d) SHAP Summary Plot for Decision Tree classifer model

5 CONCLUSION

In conclusion, this is a comparative study between different combinations of features and ML models for predicting the progression of Knee Osteoarthritis (KOA) in patients. Disease features of nearly 300 patients were taken from the the APPROACH study, pre-processed, oversampled, removed of correlated features and then used to train several machine learning algorithms. The best model was found to be a Random Forest classifier trained on a combination of Tomography and Questionnaires data, with an accuracy of 66.6% and a macro F-1 score of 36.1%. The best parameters were found to be max depth': 10, 'n estimators': 100. The best features were found to be 'KOOS_S2_Grinding', 'KOOS_S5_Bending', 'KOOS_S1_Swelling', 'SF36_03c', 'SF36_09h'. Shap was used to check the contribution of each of the features to the prediction. The 'KOOS_S5_Bending' feature was found to be the highest contributor. In general, Random Forest classifier performed well with almost all the combinations of datasets used.

By using a range of data sources, including Biomarker, Clinical, Tomography, and Questionnaire data, the study is able to develop a comprehensive predictive model that takes into account multiple factors that may influence disease progression. The results of this study have important implications for the future management of KOA, and may help to improve patient outcomes by enabling more accurate prediction of disease progression. The limitations of this study as well as the future scope of work in this area have also been discussed in the relevant sections for further guidance and continuation of research in this important arena involving human well-being.

6 LIMITATIONS OF THIS STUDY

1. This study being a multiclass calssification model, the progression of the disease can only be understood through the multiple class labels which are associated with feature sets. The degree of progression can be assumed by considering the time period for which the progression is observed and associating it with the progression label.

2. The dataset used for this study had feature details for about 300 patients, which when combined with their progression labels the dataset size reduced to about 220. Usually ML learning algorithms need to train on much larger datasets for learning to achieve good accuracy and F1 score.

3. In this study, one hot encoding was performed to encode the categorical values. One hot encoding increases the dimensionality of the datasets which stands out to be a problem for some ML algorithms.

4. For selecting the top feature, permutation feature importance was used. Application of more powerful feature selection tools like Recursive Feature elimination could yield better results.

7 FUTURE SCOPE

1. A regression approach to the same disease progression problem can be conducted to have a continuous progress prediction which can be more effective for tailoring treatments.

2. Testing model with larger datasets can help understand the performance of the model used in this study. If the performance is not satisfactory, attempt can be made to first train this model with larger datasets and then conduct the test.

3. Even though, RandomForest classifiers showed the best result in this study, RandomForest classifiers can work directly with categorical values. In fact, performing one hot encoding results in increased dimensionality of the datasets which causes more problems for RandomForest classifier and reduces its performance. Attempt can be made to use RandomForest classifier directly with the categorical values for better results.

4. Parameter tuning can be done with more number of nearest neighbours (for KNN), depth of trees and estimators (for Random Forest and Decision Trees) and hidden layer sizes (for ANN) to see if the performance improves.. More folds of cross validation can also be performed.

5. Other learning techniques of feature scaling, feature imputation, feature selection, etc. can be used to check for better results.

8 REFERENCES

1. Lespasio MJ, Piuzzi NS, Husni ME, Muschler GF, Guarino AJ, Mont MA. Knee osteoarthritis: a primer. The Permanente Journal. 2017;21.

2. Du Y, Almajalid R, Shan J, Zhang M. A novel method to predict knee osteoarthritis progression on MRI using machine learning methods. IEEE transactions on nanobioscience. 2018 May 24;17(3):228-36.

3. C. Widera P, Welsing PM, Ladel C, Loughlin J, Lafeber FP, Petit Dop F, Larkin J, Weinans H, Mobasheri A, Bacardit J. Multi-classifier prediction of knee osteoarthritis progression from incomplete imbalanced longitudinal data. Scientific Reports. 2020 May 21;10(1):8427.

4. Qvist P, Bay-Jensen AC, Christiansen C, Dam EB, Pastoureau P, Karsdal MA. The disease modifying osteoarthritis drug (DMOAD): is it in the horizon?. Pharmacological research. 2008 Jul 1;58(1):1-7.

5. About approach [Internet]. [cited 2023 Aug 28]. Available from: https://www.approachproject.eu/approach

6. Angelini F, Widera P, Mobasheri A, Blair J, Struglics A, Uebelhoer M, Henrotin Y, Marijnissen AC, Kloppenburg M, Blanco FJ, Haugen IK. Osteoarthritis endotype discovery via clustering of biochemical marker data. Annals of the Rheumatic Diseases. 2022 May 1;81(5):666-75.

7. Courties A, Sellam J, Berenbaum F. Metabolic syndrome-associated osteoarthritis. Current opinion in rheumatology. 2017 Mar 1;29(2):214-22.

8. Tiulpin A, Klein S, Bierma-Zeinstra SM, Thevenot J, Rahtu E, Meurs JV, Oei EH, Saarakkala S. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Scientific reports. 2019 Dec 27;9(1):20038.

9. Schiratti JB, Dubois R, Herent P, Cahané D, Dachary J, Clozel T, Wainrib G, Keime-Guibert F, Lalande A, Pueyo M, Guillier R. A deep learning method for predicting knee osteoarthritis radiographic progression from MRI. Arthritis Research & Therapy. 2021 Dec;23:1-0.

10. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Frontiers in bioengineering and biotechnology. 2018 Jun 27;6:75.

11. Jamshidi A, Pelletier JP, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. Nature Reviews Rheumatology. 2019 Jan;15(1):49-60.

12. Kluzek S, Mattei TA. Machine-learning for osteoarthritis research. Osteoarthritis and cartilage. 2019 Jul 1;27(7):977-8.

13. Zheng A, Casari A. Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc."; 2018 Mar 23.

14. Dayan P, Sahani M, Deback G. Unsupervised learning. The MIT encyclopedia of the cognitive sciences. 1999 Oct:857-9.

15. Maldonado S, López J, Vairetti C. An alternative SMOTE oversampling strategy for high-dimensional datasets. Applied Soft Computing. 2019 Mar 1;76:380-9.

16. Kokkotis C, Moustakidis S, Papageorgiou E, Giakas G, Tsaopoulos DE. Machine learning in knee osteoarthritis: A review. Osteoarthritis and Cartilage Open. 2020 Sep 1;2(3):100069.

17. Zhang Z. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine. 2016 Jun;4(11).

18. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI). 2012 Sep 1;9(5):272.

19. Tan H. Machine learning algorithm for classification. InJournal of Physics: Conference Series 2021 Aug 1 (Vol. 1994, No. 1, p. 012016). IOP Publishing.

20. Singh J, Joshi G, Tiwari H, Tiwari I. Applied Machine Learning for and Analysis of RR Lyrae Variable Stars. In2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on 2018 Aug 30 (pp. 247-252). IEEE.

21. Decision tree algorithm in Machine Learning - Javatpoint [Internet]. [cited 2023 Aug 29]. Available from: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

22. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PloS one. 2019 Feb 19;14(2):e0212356.

23. Bellamy N. WOMAC: a 20-year experiential review of a patient-centered self-reported health status questionnaire. The Journal of rheumatology. 2002 Dec 1;29(12):2473-6.

24. Kumar P, Bhatnagar R, Gaur K, Bhatnagar A. Classification of Imbalanced Data:review of methods and applications. IOP Conference Series: Materials Science and Engineering. 2021;1099(1):012077. doi:10.1088/1757-899x/1099/1/012077

25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321–57. doi:10.1613/jair.953