Detection of Knee Osteoarthritis Severity using a Fusion of Machine and Deep Learning models

MSc Research Project
Data Analytics

Karen Hernandez Abasolo
Student ID: X20118210

School of Computing
National College of Ireland

Supervisor: Dr. Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Karen Hernandez Abasolo		
Student ID:	X20118210		
Programme:	Data Analytics	**Year:**	2020-2021
Module:	MSc Research Project		
Supervisor:	Dr. Hicham Rifai		
Submission Due Date:	16/08/2021		
Project Title:	Detection of Knee Osteoarthritis Severity using a Fusion of Machine and Deep Learning Models		
Word Count:	9129	**Page Count:**	29

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Karen Hernandez Abasolo

Date: 16/08/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detection of Knee Osteoarthritis Severity using a Fusion of Machine and Deep Learning models

Karen Hernandez Abasolo
X20118210

Abstract

Knee Osteoarthritis accounts for more than 80% cases of arthritis impacting life quality of individuals. It is an irreversible disease that the only cure is the replacement of the knee, being important to diagnose it at early stages to prevent its progression. This study aims to improve the detection of Knee Osteoarthritis at all stages based on Kellgren-Lawrence scale using machine learning models such as Random Forest, Gradient Boosting and Xtreme Gradient Boosting trained with patient's information and deep learning models including DenseNet201 and InceptionResNetV2 trained with knee x-ray images., Their individual predictive capabilities are combined using late fusion strategy to select the final class. Machine learning models showed similar overall prediction performance between them although Deep learning models had higher efficiency showed in ROC curves compared to them, however, both altogether achieved better performance evaluated through Precision, Recall and F1-score. Moreover, using patient's data in machine learning models were identified the main features that influence the disease.

1 Introduction

Knee Osteoarthritis (KOA) is one of the most common degenerative diseases affecting elderly people in the world, it can limit the mobility of a person affecting daily life activities and even causing early retirement (Lespasio, M., 2017). Lim, K. Lau, C. S. (2011) predicts that this type of degenerative joint disease disorder will affect at least 130 million people across the world by 2050, of whom 40 million will be severely disabled by this condition. Moreover, when the disease is at the last stage the only treatment is a total knee replacement. So, it is recommended to identify Knee Osteoarthritis at first stages to avoid knee this medical procedure.

Knee Osteoarthritis (KOA) diagnosis depends on several criteria, for instance, Luyten, F. P. *et al.* (2012) proposed the criteria for early diagnosis of KOA as follows:

I. Pain in the knee
II. Radiographic grading based on Kellgren-Lawrence grade <2
III. Structural findings proven by arthroscopy and/or MRI

With the increase of medical imaging and electronic health records that are stored in repositories and available at any time such as Osteoarthritis Initiative (OAI) or Multicenter Osteoarthritis Study (MOST) with thousands of records including clinical, patient's information, anthropometrics, biomarkers, and imaging data. Researchers have been interested in contributing to this domain through novel machine and deep learning algorithms that help practitioners to diagnose, prognose, and take decisions more accurately, however, two approaches have been followed in this domain, studies that employed clinical information and those that have used imaging data to detect KOA.

Imaging applications are often considered as a “black box” that release an outcome without explanation, while diagnosis based on clinical data is not accurate. For that reason, automated detection, and classification of severity of KOA based on both clinical and imaging data might increase the performance of this task.

1.1 Research Question

To what extent fusion machine learning algorithms based on clinical data and deep learning methods based on imaging data can improve the detection accuracy of severity of Knee Osteoarthritis?

The fusion method incorporates machine learning models based on clinical data and deep learning based on x-ray images models to classify the severity of KOA.

1.2 Research Objectives

To carry out the current work, research objectives were identified, and they are outlined as follows.

1. Investigate the state-of-the-art related to prediction of Knee Osteoarthritis based on clinical and imaging data.
2. Implement and evaluate machine learning models based on clinical data.
3. Implement and evaluate deep learning models based on imaging data.
4. Merge the two methods and evaluate the performance.

1.3 Contribution

The major contribution of this project is the creation of a novel classification model that combines a deep learning approach that uses raw radiographic images and machine learning methods that use patient’s data.

The rest of the paper is sectioned as follows: Section 2 integrates the main related work completed in this domain. Section 3 describes the methodology used in this research followed by section 4 that explains the design overview of the proposed architecture while section 5 presents the implementation of the models. Evaluation of the results of each model is detailed in section 6. Finally, the paper is concluded with a discussion on findings, conclusion, and future work in section 7.

2 Related Work

Knee Osteoarthritis (KOA) diagnosis is a critical and time-consuming task for clinical practitioners. The recent advance in computer vision opened the door to the use of computer vision to diagnose several types of medical condition such as KOA. Previous work has either predicted the development of medical condition or classify it according to its severity. In general, studies predicting KOA are based on three main resources: imaging dataset which includes Magnetic Resonance Imaging (MRI) or X-Ray, and clinical information obtained from questionaries and biomedical data obtained from patient' consultation. The present literature review summarises the main findings in using computer vision and machine learning algorithms to either prognose or diagnose KOA condition.

2.1 Studies of Knee Osteoarthritis using clinical data

Kokkotis, Christos et al. (2020) predicted the possible development of a KOA condition, the study considered dataset integrated by physical activities indexes, questionnaire data, self-reported data about symptoms, and results from physical exams. This work uses a system which ranks the best features employing a voting system across different algorithms such as Pearson Correlation, Chi-squared, recursive Feature Elimination, Logistic Regression classifier, Random Forest, and Light Gradient Boosting. The most significant variables are selected and, six different machine learning methods are used to predict KOA. The best performance is achieved with Logistic Regression using around 40 variables, this work noted that most of the selected features is obesity, regardless of age or if a person did a surgery. The outcome is based on the Kellgren and Lawrence grade system, considering existence of KOA if KL is greater or equal than 2 or absence of KOA if KL is smaller than 2.

There are other similar studies that employed the same machine algorithms, however, Alexos, A. et al. (2020) used a different metric called Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) which is a pain scoring system from 0 (no pain) –to 100 (the most severe pain). The data was categorized into three classes: class 1 -pain decline, class 2 -no significant pain change, and class 3- pain increase. In this study, the best performing method is Random Forest with high accuracy of 84.3% using the first 25 features, however, incorporating more variables, the model accuracy decreases. Ntakolia, C. et al. (2020) established the same methodology Joint Space Narrowing (JSN) as the dependent variable, they utilized a range of risk factors selected according to a process which uses a filter, wrapper, and embedded tools before using them in the model. The results showed that for the left leg, the best model is Logistic regression with an accuracy of 78.3% using 164 features and for the right leg Support Vector Machine (SVM) performs better with an accuracy of 77.7% with 88 features. Taking these studies, they found that a mix of heterogenous characteristics from different categories builds a better performance if the aim is to predict KOA.

The last reviewed work in this category is presented by Christodoulou, E. et al. (2019). The dataset was integrated by 141 risk factors, 68 of them are related to any type of symptoms and 64 features described pain metrics. Furthermore, medical risk factors such as age, gender, hormonal status body weight and family history of disease were used to create clusters according to them, exactly six groups are created. The classification target was class 1 defined as incidence, in this state, participants do not have symptoms, but they are at risk factor due to the age, class 2 is called progression which involves all participants with frequent knee symptoms such as pain, aching or stiffness around the knee and class 3, non-exposed control group integrated by all participants who do not have symptoms neither present any risk factor. The dataset is unbalanced due to the last class, to face this problem two efficient techniques for handling it were applied, k- nearest Mist and SMOTE-SV; Following a data mining methodology was applied a DNN and ANN. For comparison purposes, other methods are applied such as decision trees, SVM, KNN, Adaboost, Random Forest and Linear Discriminant Analysis (LDA), excellent methods for classification tasks according to the literature. The best performance overall is through DNN model with 1 hidden layer and 50 nodes per layer, the accuracy is 79.39% while the rest of the algorithms reach an accuracy slightly higher than 50%. According to the subgroups, specifically results from gender, there is no difference greater than 0.5% between male and female subgroups concluding that gender does not represent an important factor in models. The authors concluded that working with models for specific groups get higher accuracy.

2.2 Studies of Knee Osteoarthritis using imaging data

X-ray and MRI images were used by clinicians to diagnose KOA however, this procedure is time consuming and subject to errors. Many studies have tried to use semi or fully automated methods to optimise the diagnosis process.

Four studies have used x-ray imaging datasets. Wahyuningrum, R. T. et al. (2019) classified the severity of KOA according to KL scale; the images were resized to have 400x100 pixels focusing on the knee joint, the original and this cropped image is stacked together to create sequential data. Three different deep learning models were used that are ImageNet pre-trained Convolutional Neuronal Network architectures, Visual Geometry Group (VGG-16), Residual Network (ResNet), and Densely Connected Convolutional Network (DenseNet). It was concluded that from all CNN architectures used, the best performance is with VGG-16 and it discriminated effectively in the different levels of KL. The accuracy obtained is 75.28%.

Zhang, B. et al. (2020) kept the original image size containing right and left knees from X-ray images, ResNet-34 was used together with a Convolutional Block Attention Module (CBAM) which facilitated the knee joint localization, centre, medial and lateral parts of the knee that contributed to classify KL-0 and KL-1 considered by the literature a challenging task, the accuracy of the model is 74.81%.

Nasser, Y. et al. (2020) have used X-ray data to create a Discriminative Regularized Auto-Encoder (DRAE) to detect KOA using KL scale. DRAE improves classification tasks and as its name says, it discriminates properties forcing the network to capture them and maximizing the distance between classes. The main objectives of DRAE are minimising the intra-class and maximizing the inter-class. The process developed is semi-automated because anatomical segments are manually marked and detecting Regions of Interest (ROI) in advance. The model was compared to auto-encoder models: normal Auto-Encoder (AE) and Sparse Auto-Encoder (SAE), furthermore, other image classification methods were applied such as ResNet-101 and DenseNet-121 since the state-of-the art has used them. DRAE reached an accuracy of 82.53% over the other auto-encoder models, with 81.11% with SAE and 80.26% with AE model. The accuracy in traditional models is around 58-62% of accuracy, a big difference with DRAE.

Unlike the previous studies, Kwon, S. B. et al. (2020) used a dataset containing X-ray images plus gait analysis data which contains kinetic, kinematic, and spatial-temporal features from the knee, the hip, the ankle joint and spatiotemporal parameters detecting them when the knee is extended, in a knee abduction moment, knee rotational moment, and flexion of the knee, hip and ankle, another movement detected is cadence and stride length. This kind of data was not available as a public one and it was obtained in a Human Motion Analysis Laboratory. From these two different datasets are extracted features through the Inception-ResNetv2 which is a pre-trained CNN, once the feature selection are set, the classification according to KL grade is performed using a Support Vector Machine (SVM) getting an accuracy of 0.93, 0.82, 0.83, 0.88, and 0.97 for each grade in KL scale, concluding that basing the model on X-ray images and gait data can improve the classification, however, as it was mentioned before, Gait data is not easy to find available, these studies are possible when the research is in coordination with laboratories in this domain.

2.3 Studies of Knee Osteoarthritis using clinical and imaging data

Only few studies that take into consideration two sources of data and implement integral models for a full understanding of the disease.

A comparative model that integrates x-ray and patient's data is presented by Abedin, J. et al. (2019), they used Elastic Net (EN), Random Forest (RF) and a Linear Mixed Model (LMM) to predict the several levels of KOA using a set of variables from OAI dataset, then, a CNN regression is implemented to predict the same output but using plain x-ray images. The results showed that those approaches give a similar performance compared through RMSE, with 0.974 for machine learning algorithms and 0.993 for the implemented CNN. It is remarkable that using patient's information gives a clear overview of characteristics that influence the outcome. Like studies that have analysed this data, they identify a difference when patients have had a surgery, patient's sex, pain, and symptoms to the right or left knee, among other variables. The authors suggested that a combination of both data may represent a way to improve the prediction accuracy of KOA.

A multimodal study that inspired the present research based on plain radiographs and clinical data is proposed by Tiulpin, A. et al. (2019). They built several experiments considering for the first one: Age, Sex and BMI factors; the second adds Injury, Surgery and WOMAC variables, another one takes as input variable KL-grade and the last one incorporates plain radiographs. The scheme of prediction is based on their own scale y=0 means no progression, y=1 progression within 60 months and y=2 progression after 60 months. Performance of these models is assessed according to ROC and Average Precision (AP), their results are 0.79 and 0.68, respectively. The main contribution is the combination of a CNN trained with x-ray images and clinical variables mentioned before through a GBM-based fusion that reached the best prediction performance unlike the other experiments. However, clinical variables used in this study are limited to six of a large number contained in OAI dataset.

We provide a summary that reported relevant information before conducting the current work presented in Table 1.

Table 1. Comparison of the Existing Work

References	Data	Outcome variable	Techniques	Results / Findings
Lazzarini, N. (2017)	Clinical variables Food and pain questionnaires Biomechanical markers from middle-aged women	Incidence / No incidence of KOA	Ranked Guided Iterative Feature Elimination (RGIFE) Random Forest	Generation of sub models with different outcome. The best performing model was KL incidence OA outcome
Halilaj, E. (2018)	High risk subjects according to knee pain, aching, stiffness, knee replacement, family history of OA, BMI	Joint Space Narrowing (JSN) Pain Score (WOMAC)	Clustering methodology LASSO	High accuracy using radiographic progression as the outcome with 2 visits. Pain progression presents higher accuracy with only 1 visit.
Christodoulou, E. et al. (2019)	141 risk factors: 60 symptoms variables 68 pain variables	Class 1: Incidence Class 2: Progression Class 3: Non-exposed	DNN, ANN, Decision Trees, SVM, KNN, AdaBoost, Random Forest, Linear and Discriminant Analysis	Higher accuracy when subgroups with variables suh as gender, sex, age or weight are created
Kokkotis, Christos et al. (2020)	physical activities indexes questionnaire data self-reported symptoms physical exams	Class 1: KOA K>=2 Class 2: no KOA KL 0-1	XGBoost Random Forest Decision Trees Naïve Bayes Support Vector Machine (SVM) K-nearest Neighbor (KNN) Logistic Regression	It remarks the importance of correct features to predict KOA. Most of the selected features are related to symptoms, obesity, whether the person has faced a surgery and age
Alexos, A. et al. (2020)	25 several features from OAI study selected through a counting system with Feature Selection	WOMAC	Decision Trees K Nearest Neighbors Support Vector Machine Random Forest XGBoost Naïve Bayes	Random Forest performed the best score with a few numbers of variables
Ntakolia, C. et al. (2020)	Anthropometrics, Behavioral, Quality of Life, Medical history, medical imaging outcome, Nutrition, Physical Activity and Physical Exam	Joint Space Narrowing (JSN)	Gradient Boosting Model Multilayer Perceptron Logistic Regression Naïve Bayes Gaussian Random Forest Support Vector Machine	The study divides the outcome according to Left and Rigth Knee. A heterogeneous mix of features maximize the performance of the models
Lim, J., Kim, J. and Cheon, S. (2019)	Clinical variables demographic and personal features biomechanical markers	Binary outcome Osteoarthritis (Yes/No)	DNN with scaled-PCA	AUC of 76.8% including demographics variables
Moustakidis, S. *et al.* (2019)	Joint symptoms disability, function, and general health	Class 1: Incidence Class 2: Progression Class 3: Non-exposed	Linear Discriminant Analysis (LDA), Decision Trees, SVM Gaussian, KNN, AdaBoost, Random Forest and DNN	The superiority of DNN to diagnose KOA. Implementation of feature subgroups exploration
Bandyopadhyay, S. and Sharma, P. (2016)	X-ray images	Normal or Affected Knee X-ray image according to KL scale	Random Forest	Better accuracy using a feature set (Texture, Haralick, First Four Moments, Statistical + Region Properties)
Antony, J. *et al.* (2017)	X-ray images	KL scale	Fully Convolutional Network (FCN) Convolutional Neural Network (CNN)	Multi-class classification accuracy 60.3% Classification of KOA conditioned to KL scale with grade 0-2 is challenging due to the small variations
Tiulpin, A. *et al.* (2019)	X-ray images	KL scale	ResNet-34 Deep Siamese Convolutional Neural Network	Ability of models to learn important OA features transferable to a different dataset. Creation of attention maps that highlight features influencing the network decision.

Wahyuningrum, R. T. et al. (2019)	Cropped X-ray images	KL scale	VGG-16 Residual Net (ResNet) DenseNet	Accuracy of 75.28 % with CNN-LSTM model
Bany Muhammad, M. *et al.* (2019)	X-ray images	KL scale	VGG, ResNet, Inception Customize Base Model Architecture	Refine accuracy of CNN with a customized ensemble model architecture
Zhang, B. et al. (2020)	X-ray images	KL scale	ResNet-34 Convolutional Block Attention Module (CBAM)	Future resolution and contrast enhancement in imaging data improve the diagnosis using CNN models
Nasser, Y. et al. (2020)	X-ray images	KL scale	Discriminative Regularized Auto-Encoder (DRAE) ResNet-101 and Dense-121.	DRAE obtained better performance than classical Autoencoder models and CNN
Kwon, S. B. et al. (2020)	X-ray images Gait analysis data	KL scale	Inception-Resnetv2 Support Vector Machine	Gait data and radiographic improve the accuracy of KOA classification
Abedin, J. *et al.* (2019)	Questionnaire data and X-ray images	KL scale	Elastic Net Random Forest Linear Mixed Model (LMM) Convolutional Neural Network (CNN)	Identification of useful explanatory features before x-ray imaging examination. Elastic Net and LMM presents higher prediction accuracy based on RMSE
Tiulpin, A. *et al.* (2019)	X-ray and clinical data	No knee OA progression Progression within the next 60 months Progression after 60 months	CNN Logistic Regression Gradient Boosting Machine (GBM)	AUC of 0.79 It yields better prediction performance that using a single type of data

2.4 Summary

State-of-the-art outlined the usage of machine and deep learning in KOA diagnosis and prediction, demonstrating improvements in them based on reference studies. In this domain, several data sources are considered as inputs where it is found medical images such as MRI and X-ray, biomarkers, clinical or patient data information related to pain, symptoms, physical activities that contribute to the development of the disease. However, what is observed is that most of these studies suggest as a future work the implementation of a process that associate both type of data such as it has been done by Abedin, J. et al. (2019), researchers believe that incorporating imaging and explanatory variable of the disease will enhance the performance of the suggested models. In a systematic review of multimodal fusion models conducted by Huang, S.-C. *et al.* (2020) is remarked that combining clinical and imaging data leads to a higher diagnostic accuracy and enable medical staff to interpret imaging results.

3 Research Methodology

3.1 Introduction

After assessing the requirements of the research, Knowledge Discovery in Database (KDD) methodology will be followed step by step in both modelling processes. The research focuses on predicting Knee Osteoarthritis disease through the late fusion of two models, machine learning models fed with clinical data and deep learning techniques fed by x-ray images. In the first stage, three machine learning algorithms are applied Random Forest, Gradient

Boosting and XGB while in the second stage, pre-trained models such as Dense201 and InceptionResNetV2 are applied to detect the level of severity through medical images.

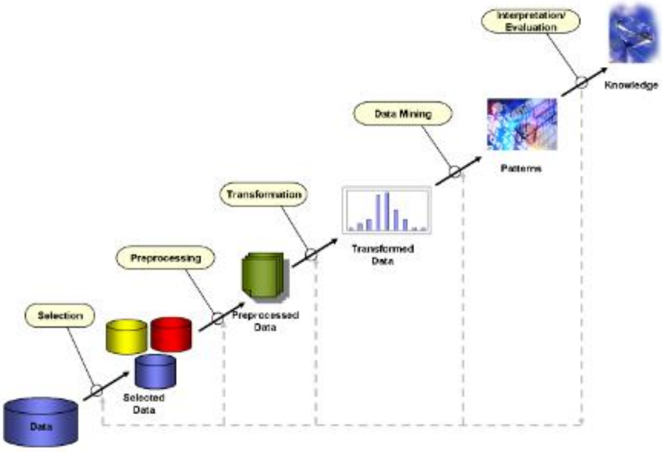


Figure 1. KDD Methodology (Guerra-Hernández, A. and Mondragon-Becerra, R., 2008)

3.2 Data Collection

The research was conducted using information from a dataset provided by Osteoarthritis Initiative (OAI) that is a longitudinal, prospective, and observational study of knee osteoarthritis applied in 4,796 participants. It contains knee images origin from x-ray and MRI mechanisms as well, clinical data of the patient obtained from personal questionaries. Although the study comprises more than 6 examinations through 14 years, our research utilizes only data from the baseline. As it was mentioned earlier, the classification is based on Lawrence-Kellgren metric that rank knee osteoarthritis severity from 0 to 4, this metric was provided by radiologist in the study.

3.3 Ethics Implications

OAI study realise all datasets and manuals to be used for research purpose. All clinical data are identified by an ID per patient, and all medical images are from both legs, but it is not possible to link them and track patient's personal information. With this action, confidentiality is well preserved.

3.4 Data Pre-processing

3.4.1 Clinical Data Pre-processing

For the implementation of machine learning models, clinical data such as anthropometrics (weight, height, BMI) and patient's answers to diverse questionaries that examinate pain, nutrition habits, physical activities, and symptoms are included. A complete list of them that we have chosen as possible predictors is provided in a supplementary file.

At this stage, it is vital to perform data cleaning, data wrangling and data preparation; variables containing a high rate of missing values should be evaluated to consider if they will be kept or dropped from our data frame. Also, exploratory data analysis is performed that gives us an overall sight of our data, it helps to understand the relation between outcome variable versus predictors and identify outliers.

One Hot Encoding is used on categorical variables and standardization is applied on ordinal and numerical variables to get a dataset that ranges between 0 and 1. Finally, due to the large number of variables, feature selection is performed to keep only meaningful variables.

3.4.2 Image Pre-processing

Deep learning requires good quality images. The area of interest for the left and right knee have been zoomed and converted into 254x254 pixels. In addition, data augmentation technique is used such as shifting, rotating, flipping and creates synthetic images from the originals (Minh, T. N. et al., 2018) to increase the data set.

3.5 Modelling

After pre-processing both data, three models of machine learning are fed with clinical data while x-ray images are used in pre-trained deep models at this stage. Both approaches have been implemented by diverse researchers. Only a few have done a complementation; what is novelty in this research is the fusion approach of them to improve the prediction of all different levels of severity in KOA.

3.5.1 Machine Learning Models

Random Forest

Random forest is an ensemble algorithm that includes many decisions trees, each of them learns from a random sample of data. The trees have used these data samples several times, because of this, the variance in each tree is high but the distribution of the random forest is generally low. The outcome of the algorithm is the average of the predictions for all trees in the random forest. Its strengths are better accuracy even when we are dealing with a large number of variables and its robustness to outliers and noise (Leo Breiman Statistics, L. B, 2001).

Gradient Boosting

Unlike random forest that ensemble trees by averaging, Gradient Boosting consist in a different ensemble formation. With boosting, new models are learning considering error from past iterations to provide a more accurate estimation. Compared to single machine learning algorithms, they can reach a better accuracy. One of the advantages of this ensemble method is the freedom to design the model choosing the most suitable loss function (Natekin, A. and Knoll, A., 2013).

XGBoost

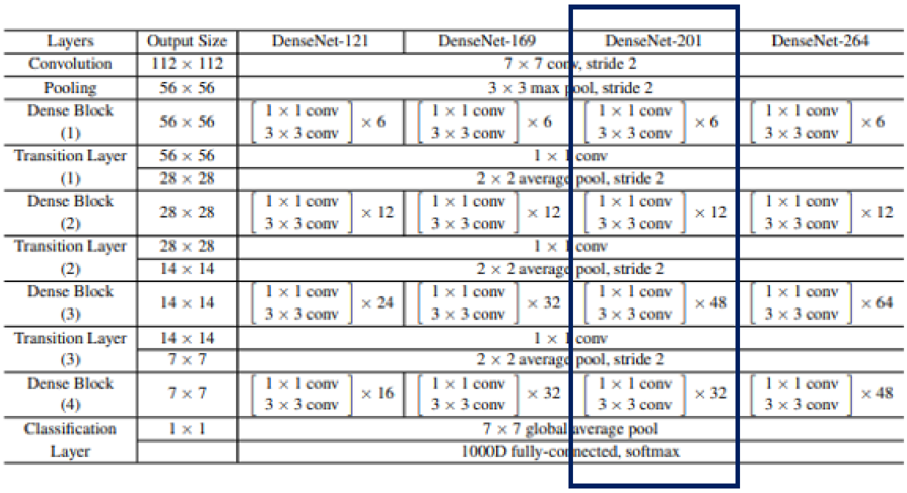
Extreme Gradient Boosting (XGB) is a scalable end-to-end boosting system that process data 10 times faster and more accurate than other popular solutions. It has been implemented in many data science problems. The algorithm builds trees and apply boosting technique, after that, a score that provides the significance of each variable in the data is calculated by the amount that each variable improves the performance in that node (Chen, T. and Guestrin, C., 2016).

3.5.2 Pre-trained models (Transfer Learning)

The literature review proved that deep neural networks have been very effective in terms of image recognition that is the main challenge predicting knee osteoarthritis severity (Chen, P. et al.; 2019), (Tiulpin, A. et al., 2018), (Antony, J. et al., 2016). Convolutional Neural Network (CNN) includes convolutional, pooling and fully connected layers and requires a large amount of labelled data. To overcome the large size data requirement, transfer learning can be used which utilizes pre-trained models. In medical diagnosis tasks the following architectures have been implemented.

DenseNet201

While traditional CNN has a single connection to each layer, Dense Convolutional Network (DenseNet) creates a network that connects all layers. There is no need to re-learn redundant feature map because each layer reads data obtained from the precedent one and preserve meaningful knowledge transferred to the next layer. Consequently, its layers are very narrow, however, the outcome is based on all features maps obtained in the whole network. As is seen in Figure 2, dense blocks are added that has a convolutional layer with a residual concept followed by a transition layer on DenseNet201 architecture (Huang, G. *et al.*, 2019).



Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 64$
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Figure 2. DenseNet-201 Architecture (Huang, G. *et al.*, 2017)

InceptionResNetv2

A combination between Inception-V4 architecture with residual connections trained on the ImageNet validation set, that has shown improvements in recognition performance. The

Inception architecture contains Inception blocks that are parallel layers (1x1 Conv, 3x3 Conv and 5x5 Conv) with their output filter banks followed by a single output vector that represent the input for the next layer, on the other hand, ResNet schema overcomes the degradation problem caused by deeper networks, creating layers that fit a residual mapping (Szegedy, C. et al., 2016).

3.5.3 Fusion of medical images and clinical data

The aim of data fusion from several modalities is to improve the model used to diagnose the diseases. In this work, late fusion was chosen as the modality to use (Huang, S.-C. *et al.*,2020).

Late Fusion

Late or decision-level fusion is the process of incorporating outcomes from multiple classifiers. Each classifier chooses a separate source of dataset to train the model. The rules to fusion different models employ averaging, majority voting, weighted voting or a metaclassifier. These aggregation functions are generally chosen empirically. Figure 3 illustrates how the process works.

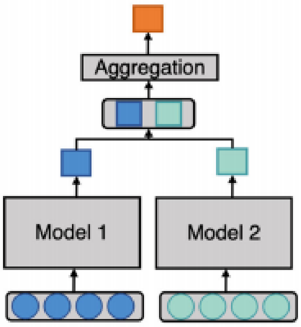


Figure 3. Late Fusion Strategy (Huang, S.-C. *et al.*, 2020)

The methodology in the current work for this stage is based on Qiu, S. *et al.* (2018) who performed a fusion of deep learning models using different voting techniques.
Firstly, the probability predictions from Random Forest, Gradient Boosting and XGboost are taken to perform mean voting as the final prediction, defined as:

$$P_i^{mean} = (P_i^{RF} + P_i^{GB} + P_i^{XGB})$$

Where P_i^{RF}, P_i^{GB} and, P_i^{XGB} are the probabilities from Random Forest, Gradient Boosting and XGBoost respectively, and P_i^{mean} is the mean of the i^{th} class of severity of KOA. Then, max voting takes the prediction with the highest probability as $max(P_1^{mean}, P_2^{mean}, P_3^{mean}, P_4^{mean}, P_5^{mean})$, and according to the value i^{th} is assigned the final prediction.

Finally, majority voting is used to compute the final prediction considering three independent labels, $\hat{y}_1$ is the prediction value from DenseNet201, $\hat{y}_2$ comes from InceptionResNetV2 model and, $\hat{y}_3$ is the final prediction from machine learning models computed before. Figure 4 illustrates the process in detailed.

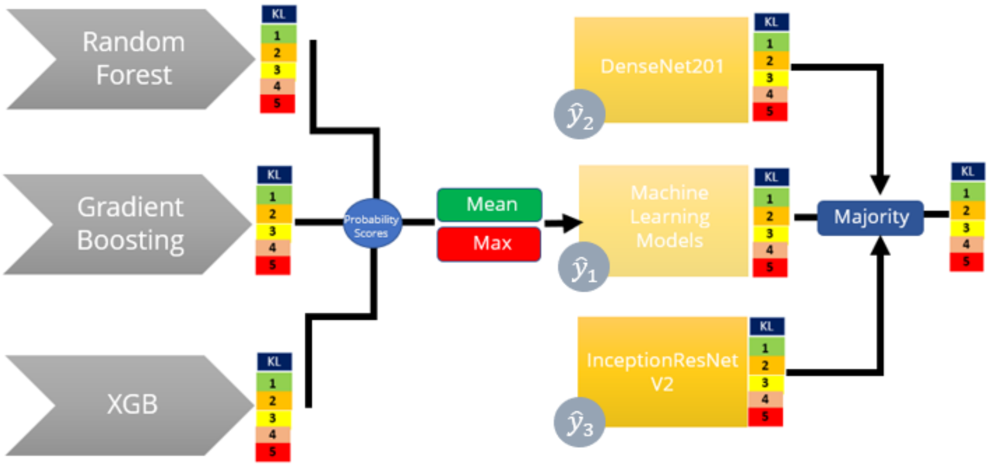


Figure 4. Fusion of Machine and Deep Learning Models

3.6 Evaluation

To evaluate the performance of the fusion models several metrics are used. First, confusion matrix for a multi-label problem shows the results obtained against real values. In addition, Sensitivity, Specificity, and Accuracy. Although accuracy is the most used metric, evaluating multi-class tasks with it may present a bias towards the majority classes. Other metrics such as F-measure that combine precision and recall is more effective to evaluate when an instance has been correctly classified according to the class (Branco, P., Torgo, L. and Ribeiro, R., 2015)

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Another well-known measure is Received Operating Characteristics (ROC) curve that plot True Positive Rate versus False Positive Rate at all classification thresholds will be our main metric to evaluate performance of the proposed models.

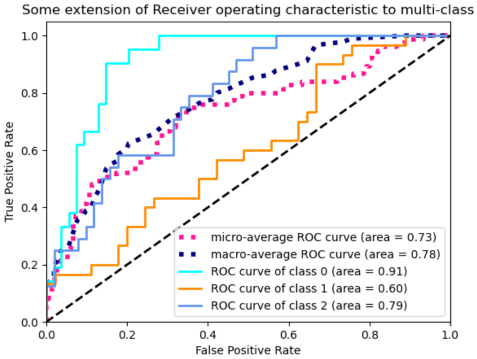


Figure 5. ROC curve plot[1]

[1] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

4 Design Specification

To implement an efficient system able to predict knee osteoarthritis on their respective severity, an architecture design (Figure 6) is developed.

For training deep learning is used TensorFlow, specifically Keras a free and open-source that has a repository with pre-trained models such as DenseNet201 and InceptionResNetv2 that allow to transfer learning to our image dataset. The images pre-processing stage, visualization of data, implementation, and evaluation are carried out in Google Colab environment due to the memory and time optimization running our code.

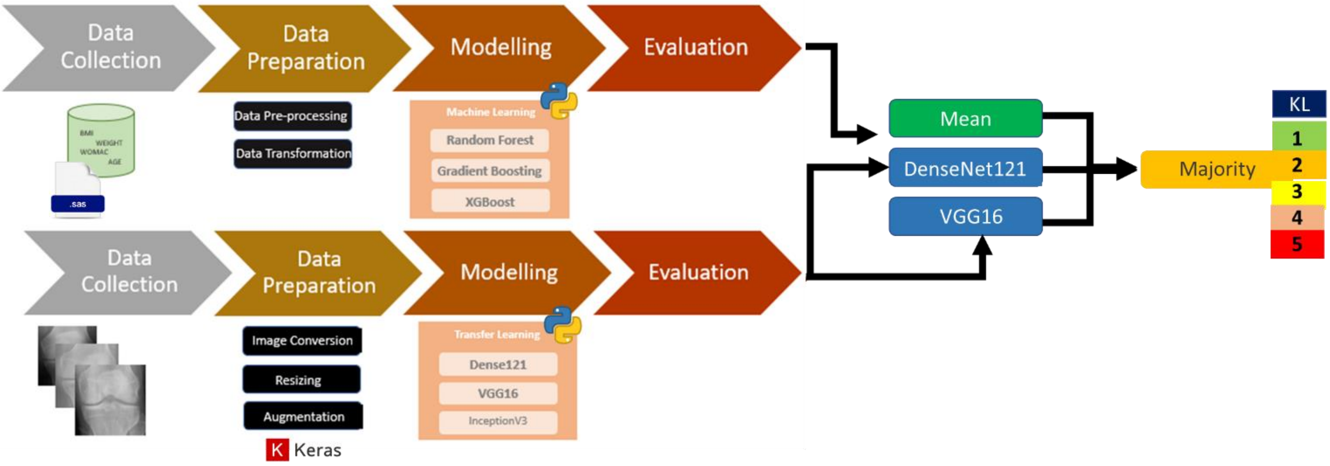


Figure 6. Design architecture

5 Implementation

In this section, the implementation of the proposed methodology to predict KOA severity using machine and deep learning methods is explained in detail, following the design flow presented above.

Two approaches are engineered in parallel employing distinct type of data but belonging to the same patient. In the implementation of Random Forest, Gradient Boosting and XGBoost our source data comes in sas files, and they are converted to a dataframe that will be processed and modelled in Python, a general-purpose programming language, using available libraries such as pandas and NumPy. These methods, their validation and evaluation are provided by scikit-learn package. They are written, edited, and modified in Jupyter notebook hosted by Anaconda

5.1 Data Collection

We utilised data from OAI study at the baseline. For our research, clinical and x-ray images from participants was taken. Clinical variables were chosen considering possible risk factors that influence the progression of the disease, metrics obtained from MRI or biospecimen were excluded because it is considered complex data, unlike info derived from questionaries or simple anthropometrics obtained in a regular medical visit and X-ray images that is a quick and low-cost procedure to examinate the narrow space between bones.

5.2 Data Preparation

5.2.1 Clinical Data

Cleaning and Transformation

The first step was to merge records from two different files, one of them contains clinical variables while the other one has two important attributes “P0SEX” and “P02RACE”. At the beginning, we had 4,796 rows. Then, after analysing data was decided to drop columns with more than 50% of missing values such as “V00KOOSFX3”, “V00HOURWK” and “V00INCOME”. The same dataframe was duplicated to create a single row that contains a unique identifier to right or left knee due to KL metric is assigned to each side of the knee.

In this stage, a wide range of variables are identified as categorical, and they must be encoded. Numerical and ordinal variables are normalized to have all values between 0 and 1, adopting for this task, Min-Max normalization technique.

Feature Selection

There are 176 features after completing all pre-processing steps, however, as literature suggest, dimensionality reduction improves the performance of our model, reducing time and storage space, furthermore, with less variables, it allows us to interpret data easily. This task is performed using Recursive Feature Elimination (RFE) through Gradient Boost to select the most important features. After few iterations, 30 features are selected that achieved the best accuracy. These variables are listed in a supplementary table.

Exploratory Analysis

At the end of the pre-processing step, the final dataset contains 4,806 rows related 2,403 patients. The distribution of our data according to KL grade is illustrated in Figure 5. The figure shows that the dataset is imbalanced, most of the data belongs to class 0 and only a small portion belongs to class 4.

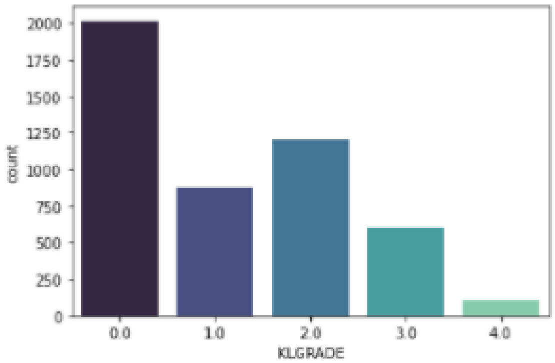
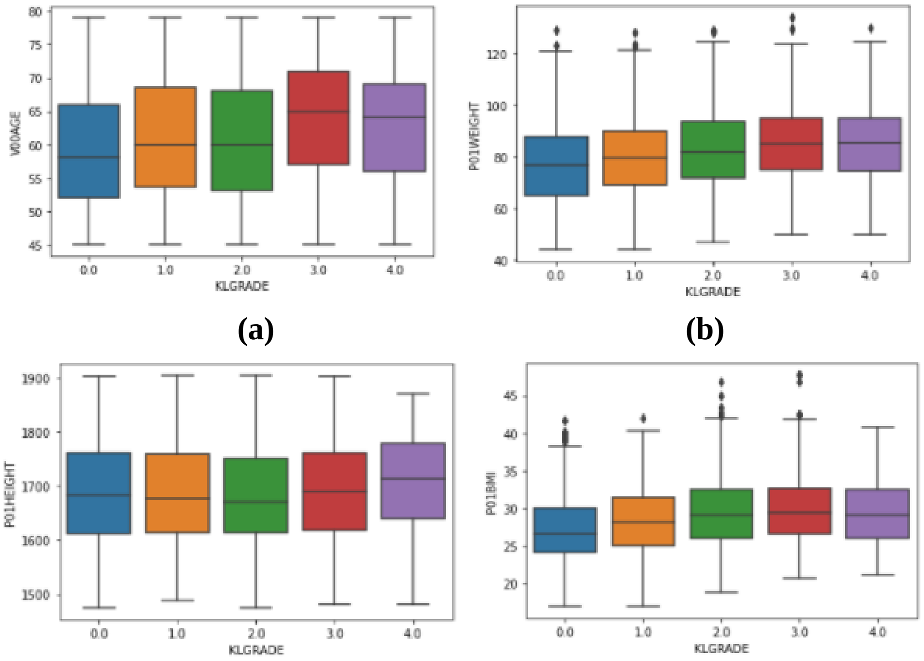


Figure 7. Distribution of clinical dataset according to target variable

The following blox-pots in Figure 6 display the distribution of different patient features against our target variable. As is observed, the severity of KOA increases in elderly people and the same trend is observed with height, weight, and Body Mass Index (BMI), the higher the value, the severity of the disease grow.



(a) (b)

(c) (d)

Figure 8. Boxplots of patient features by KL grade severity: (a) age (b) weight (c) height and (d) BMI

5.2.2 Imaging Data

The original x-ray images at the baseline period have already been cropped[2], obtaining the region of interest (ROI) which is the anterior knee joint view. The dimension of them is 224x224 pixels. Images are split into train, test, and validation data, however, for the purpose of our research, this split must match the ratio (70:30) in the other models that use clinical data according to the division of clinical data, for that reason, we re-arrange all images using a high-level file operation, shutil library in Python.

Firstly, a customise function, that apply gaussian blur technique to reduce noise in the images and equalise histograms is applied. Then, several augmentation techniques are used such as image rotation in a range of 15, zooming in a range of 0.1, shear mapping images randomly, and flipping half of the pictures horizontally. The images in figure 7 is a sample that show how x-ray images look, after applying augmentation strategy.

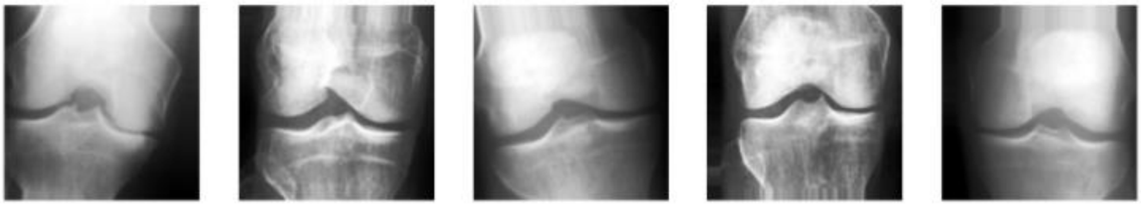


Figure 9. A sample of Knee x-ray images after pre-processing

[2] Chen, Pingjun (2018), "Knee Osteoarthritis Severity Grading Dataset", Mendeley Data, V1, doi: 10.17632/56rmx5bjcr.1

5.3 Machine Learning Models

After processing clinical data, we split it into two datasets, 70% for training and 30% for testing.

Random Forest

Initially, a Random Forest classification model was fitted with 20 trees in the forest and a value of 4 as the deep of each three. Then, an exhaustive search was implemented using GridSearchCV with 3-fold cross validation to fine-tune the model. The best parameters are shown in Table 2.

Table 2. Hyperparameters in Random Forest model

Parameter	Value
Bootstrap:	TRUE
max_depth:	90
max_features:	2
min_samples_leaf:	3
min_samples_split:	8
n_estimators:	300

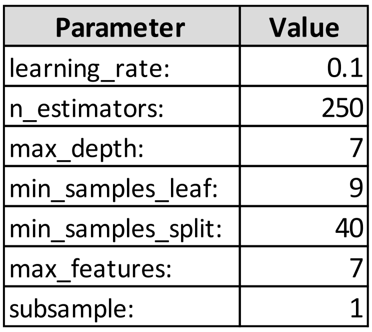
Two additional models are created to deal with imbalanced data. The first strategy consists in Synthetic Minority Oversampling Technique (SMOTE) that augment data through synthetic examples of the minority class, however, it does not relevant information (Chawla, N. V. *et al.*, 2002).The second strategy used is Class Weighting that is part of the RandomForestClassifier class, with “balanced” option, the weights are adjust inversely proportional to class frequencies in the input data[3].

Gradient Boosting

The second model is another machine learning algorithm that predicts class of the disease. Similar to Random Forest, hyper parametrisation is utilised for better performance. The initial training process consisted in 100 boosting stages to execute with maximum nodes in the tree with a learning rate of 0.1. After GridSearchCV with the same number of fold cross validation than RF, the best parameters were summarised in Table 3.

[3] *sklearn.ensemble.RandomForestClassifier — scikit-learn 0.24.2 documentation* (no date) *Scikit-learn.org*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Accessed: July 20 , 2021).

Table 3. Hyperparameters in Gradient Boosting model



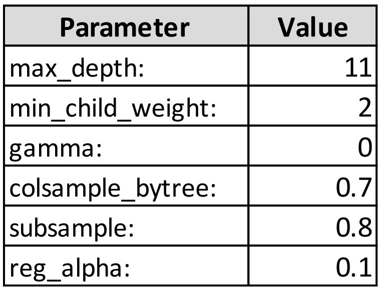
Parameter	Value
learning_rate:	0.1
n_estimators:	250
max_depth:	7
min_samples_leaf:	9
min_samples_split:	40
max_features:	7
subsample:	1

Same than RF, a new model with SMOTE technique is implemented to evaluate whether this strategy improves the classification.

XGBoost

The last model is trained following the principle of gradient boosting is XGBoost, in contrast to Gradient Boosting, XGBoost computes the second partial derivatives of the loss function which improves model generalization. XGBoost was set with initial values, and after tuning the model, the best parameters found are listed below where is notable that the number of trees is larger than RF and GB.

Table 3. Hyperparameters in XGBoost model



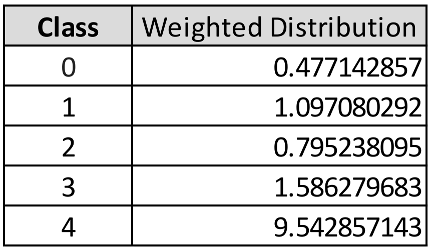
Parameter	Value
max_depth:	11
min_child_weight:	2
gamma:	0
colsample_bytree:	0.7
subsample:	0.8
reg_alpha:	0.1

A last XGBoost model employing SMOTE technique is created. Each trial maximizes accuracy through Stratified cross validation and all metrics to evaluate their performance are implemented, described in detail in Evaluation section.

5.4 Transfer Learning Models

As mentioned earlier, the advantage of using transfer learning is that requires a small amount of data for training the model. Due to the imbalance dataset, synthetic samples corresponding to the minorities classes are created in addition to reducing images corresponding to the majority class. Table 4 shows the resulting distribution according to the correspondent class.

Table 4. Weighted Distribution of training dataset



Class	Weighted Distribution
0	0.477142857
1	1.097080292
2	0.795238095
3	1.586279683
4	9.542857143

DenseNet201

DenseNet201 model is implemented using pre-trained weights from “ImageNet”. Through the first layer, input data is passed with a shape of (224x224x3) being the height and width of the image. After stacking Densenet201 scheme, a BatchNormalization, a Dense layer with 512 as filters and activation “relu”, and a final dropout layer with the rate of p=0.3 are placed with 5 as the number of filters which represents the number of classes of KL severity. The optimizer is Adam, and categorical crossentropy is set as the loss function.

To deal with what the model is training, a callback class is created that monitors the improvement of training accuracy epoch by epoch, an earlystopping that will be applied after 3 epochs with no changes, and a threshold of 0.9 that define the level of accuracy that we want to achieve. The first model is trained with 12 epochs, but following a suggested code implementation[4], a strategy to fine-tune the model is kept those first epochs frozen and then, adding 15 more.

InceptionResNetV2

InceptionResNetV2 is a pre-trained model available in Keras application[5]. The model uses pre-trained weights from ImageNet. We implemented with similar parameters used in DenseNet201. After Inception ResNetV2 architecture was attained a batch normalization, a fully connected, dropout and the final layers. Unlike DesNet201, this model contains more 55 million of trainable parameters.

5.5 Fusion of machine and deep learning models

After implementing individual machine learning models, their probability scores were averaged and then, the highest value was considered the first candidate to be the final prediction. The second candidate outcome came from DenseNet201 model, and the third result was taken from InceptionResNet201. Thus, the final prediction class is calculated through majority voting between these three results. However, as a rule of thumb, if the instance has three different results, the definitive class was taken from DenseNet201 which is the model with the best performance.

[4] https://www.kaggle.com/gpiosenka/notebookf03b3b1161

[5] https://keras.io/api/applications/inceptionresnetv2/

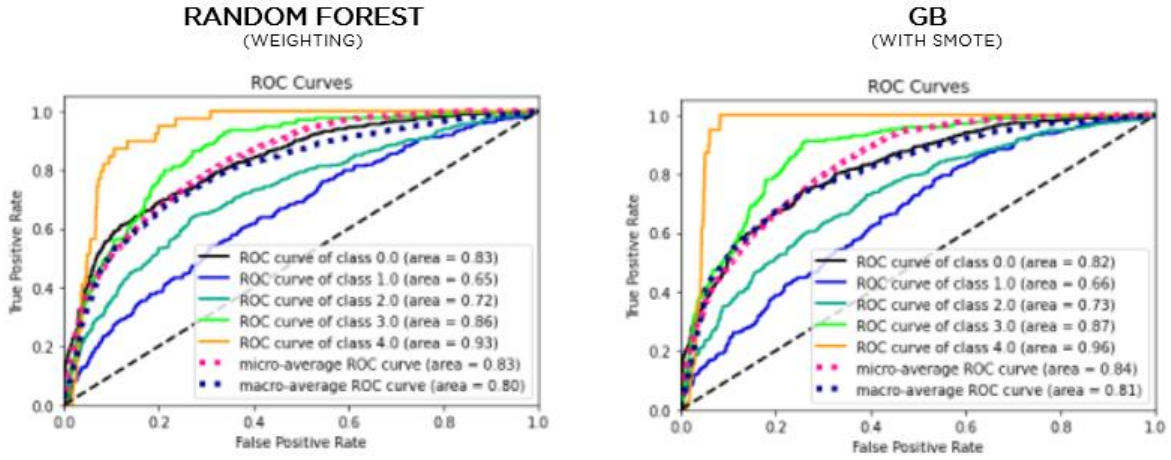
6 Evaluation

The last part of the proposed methodology consists of analysing the performance of all implemented models. However, it is important to keep in mind that our main objective is to prove if better predictions will be reached fusing machine and deep learning methods with both, clinical and imaging data.

6.1 Machine Learning Models

ROC curves are considered and micro average precision to evaluate each method. As was mentioned earlier, the ratio of data is 70:30 for training and test data. To prevent overfitting in the scores and bias due to the imbalanced sampling is implemented 10-Fold Stratified cross-validation. Three experiments were performed applying Random Forest (the hyperparameter model, with SMOTE technique and Weighting classes) while for GB and XGBoost two models were executed, hyperparameter model without and with SMOTE. Through the analysis of ROC curves and each classification report was found that implementing SMOTE technique improve the models slightly. According to weighted average precision and F1-score, we choose the best version in each model: Weighted Random Forest, Gradient Boosting with SMOTE and XGB without SMOTE.

Assessment of the models through ROC is presented in figure 8. Our three models have similar performance, the best performance is through Gradient Boosting with micro-average 84% AUC. It is shown that with different thresholds, class 4 has the highest true positive rate.



RANDOM FOREST
(WEIGHTING)

GB
(WITH SMOTE)

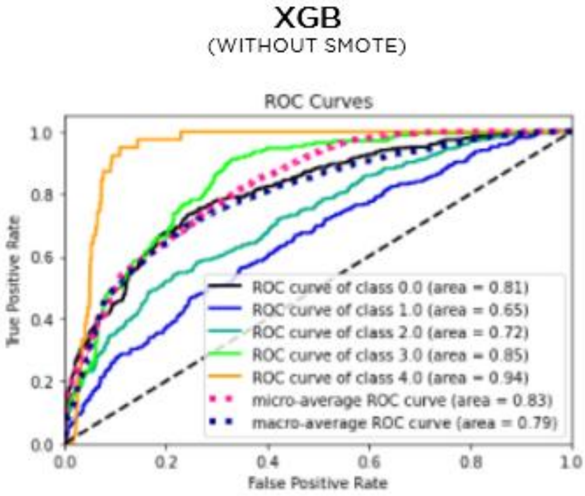


Figure 10. ROC curves for Random Forest, Gradient Boosting, and XGB models

6.1 Transfer Learning Models

To predict in the same sample was considered to split train and test datasets according to machine learning models, however, 10% from training data was taken to create a validation set. A complete evaluation of deep learning models by loss and accuracy of training and validation data is calculated by each epoch of the process.

DenseNet201

As was discussed in section 2, ROC curves are one of the best methods to evaluate a multi-label classification performance. Figure 12 shows the ROC of DenseNet201, in contrast to machine learning models there is an improvement of the area per class, machine learning models reported an area around 0.72 for class 2 while in DenseNet201 is 0.81, being 10% higher.

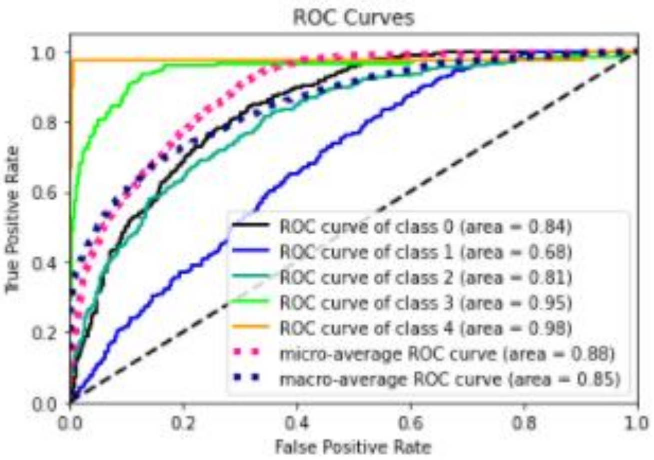


Figure 11. ROC curves for DenseNet201

InceptionResNetv2

The performance of InceptionResNetV2 through ROC curve in all classes of severity of KOA is shown in Figure 12. As DenseNet201, all classes performed better than machine learning models with a micro-average ROC curve of 0.88. Comparing values, DenseNet201 performed slightly better than InceptionResNetV2.

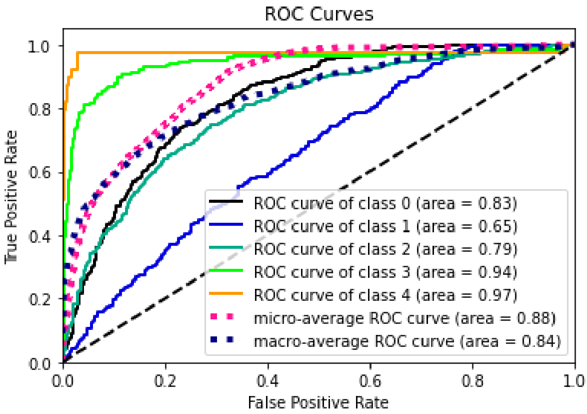


Figure 12. ROC curves for InceptionResNetV2

6.3 Evaluation of machine and transfer learning models in terms of primary metrics

In addition to ROC curves, we present a detailed comparison tables of all performed models in test data, categorized by severity grade of KOA. In terms of Precision, the measure dictates the patients that were correctly identified according to the KL grade out of all the patients having that level either predicted or true cases. KL grade 4 is the best class correctly predicted by DenseNet201 with 88%, KL grade 0 is the next class with the highest precision value in all performed models with around 70%, the worst result was in KL grade 1, having the best precision rate in Random Forest model.

Table 5. Precision from different models for each severity level of KOA

Precision	Machine Learning models			Deep Learning models	
Severity Level	RF	GB	XBG	DenseNet201	InceptionResNetV2
0	0.69	0.68	0.69	0.71	0.71
1	0.36	0.32	0.36	0.3	0.28
2	0.46	0.47	0.48	0.59	0.55
3	0.43	0.43	0.44	0.66	0.73
4	0.14	0.2	0.03	0.88	0.78

Recall is the rate of patients identified correctly by class out of all predicted values for that class. Similar performance than Precision, the best correctly classified grade of KOA is 4 with 90% of recall. DenseNet201 achieved a recall value of almost twice better than machine learning models in class 3. For class 0, Random Forest presents the same recall rate than deep learning models.

Table 6. Recall from different models for each severity level of KOA

Recall	Machine Learning models			Deep Learning models	
Severity Level	RF	GB	XBG	DenseNet201	InceptionResNetV2
0	0.71	0.68	0.74	0.71	0.7
1	0.27	0.26	0.29	0.37	0.31
2	0.47	0.47	0.5	0.44	0.5
3	0.53	0.51	0.46	0.8	0.78
4	0.15	0.31	0.03	0.9	0.82

In our research project, due to classifying correctly all classes is important, F1-score is computed as a trade-off between precision and recall. Table 7 shows the results of this metric performed in all models. The results are similar like those presented in Precision and Recall, however, InceptionResNetV2 presents slightly better value than DenseNet201 in class 2 and 3.

Table 7. F1-score from different models for each severity level of KOA

F1-score	Machine Learning models			Deep Learning models	
Severity Level	RF	GB	XBG	DenseNet201	InceptionResNetV2
0	0.7	0.68	0.71	0.71	0.7
1	0.31	0.29	0.32	0.33	0.29
2	0.46	0.47	0.49	0.5	0.52
3	0.48	0.47	0.45	0.73	0.75
4	0.15	0.24	0.03	0.89	0.8

In table 8 is displayed an overall of Precision, Recall, F1-score and Accuracy as a weighted average between classes. The performance between machine learning models is similar, around 53% in all metrics. On the other hand, deep learning models using x-ray images show a higher performance in all metrics, being DenseNet201 the model with the highest weighted average metrics. In consequence, in our fusion model if there is no value found through majority voting between $\hat{y}_1, \hat{y}_2$ *and* $\hat{y}_3$, the outcome will be considered as $\hat{y}_1$ being the predicted value from DenseNet201 model.

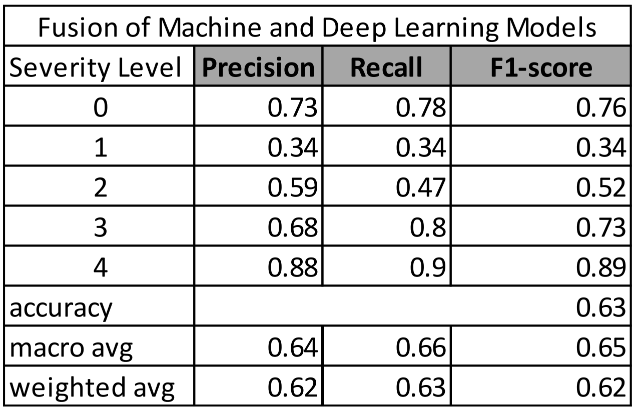
Table 8. Weighted metrics from Machine and Deep Learning Models

Metric	Machine Learning models			Deep Learning models	
	RF	GB	XBG	DenseNet201	InceptionV2
Weighted Avg Precision	0.53	0.52	0.53	0.6	0.59
Weighted Avg Recall	0.53	0.52	0.54	0.59	0.59
Weighted Avg F1-score	0.53	0.52	0.53	0.6	0.59
Accuracy	0.53	0.52	0.54	0.59	0.59

6.4 Evaluation of proposed Fusion Model

Given the outcomes from three machine learning (Random Forest, Gradient Boosting and XGBoost) and two deep learning models (DenseNet201 and InceptionResNetV2), a final prediction was calculated through the fusion of them explained in section 5. The predictive performance of this method is shown in table 9. The fusion model outperformed individual models, generating higher weighted precision, recall, F1-score and accuracy.

Figure 9. Classification Report for Fusion of Machine and Deep Learning Models



Fusion of Machine and Deep Learning Models			
Severity Level	**Precision**	**Recall**	**F1-score**
0	0.73	0.78	0.76
1	0.34	0.34	0.34
2	0.59	0.47	0.52
3	0.68	0.8	0.73
4	0.88	0.9	0.89
accuracy			0.63
macro avg	0.64	0.66	0.65
weighted avg	0.62	0.63	0.62

7 Discussion

Predicting severity of knee osteoarthritis disease involves a fully understanding of what factors influence this progression with the assessment of x-ray images at the same time. The current research project proposed a fusion of models that incorporate both data.

Firstly, machine learning algorithms yielded ROC of 0.81-0.82, 0.65-0.66, 0.72-0.73, 0.85-0.87 and 0.93-0.96 for 0 to 4 level of KOA, respectively being capable to detect the disease with patient's data. In these three models was observed that KL grade 1 and 2 presented the worst performance in ROC curve and primary metrics such as Precision, Recall and F1-score due to the similarity in the predictor variables at these stages. As was seen in the plots in explanatory section, average age is around 60 years old, average height of the patients is 1700 mm, and average weight is about 80 kg for these two classes.

Through the implementation of these models, the most important variables that contributed to predict KOA were identified that confirm state-of-the-art findings. At the beginning, we had more than 170 features, however, after computing feature selection, we took only 30. Moreover, we computed the importance of feature in each model and the findings were that P01OAGRDL, P010AGRDRL, P01BMI, V00ABCIRC, P01LXRKOA, P01HEIGHT, and P01KSURGR are the variables that most influence the progression. P01OAGRDL, P010AGRDRL, and P01LXRKOA are calculated from x-ray images while others such as V00ABCIRC, P01BMI are derived from the weight of the patient, P01HEIGHT reflects that taller a person is, highest the chance to suffer a higher severity of the disease, and P01KSURGR that explore if the patient has had a surgery or arthroscopy before. With this knowledge, it may represent a first suspicion when a patient is examined.

Additionally, deep learning models had higher performance results as we analysed in evaluation section, with ROC curve values of 0.83-0.84, 0.65-0.68, 0.79-0.81, 0.94-0.95, and 0.97-0.98 for 0 to 4 level of KOA, respectively. However, it is remarkable that occurs the same trend in KL grade 1 and 2 like machine learning models, this is because, even with knee x-ray images is difficult to distinguish the structural differences in the knee joints.

In the last stage of the work, a fusion process was carried out that proved the effectiveness of integrating these two forms of independent data, achieving better performance in weighted average precision, recall and F1-score. The fusion method was built using predicted and real values of testing data, thereby, ROC curve was not calculated because the lack of probability scores for these outcomes at different thresholds. Figure 16 is a confusion matrix that classify the predictions made by the fusion method against real values. KL grade 4 is the class with the smallest number of instances where most of the cases were correctly predicted, in medical terms that means that automated systems can assist accurately in classifying the worst stage of the disease. However, as it was mentioned earlier is important to increase the reliability of the system in diagnosing earlier stages such as 1 and 2 to apply the correspondent treatment to stop the progression of the disease.

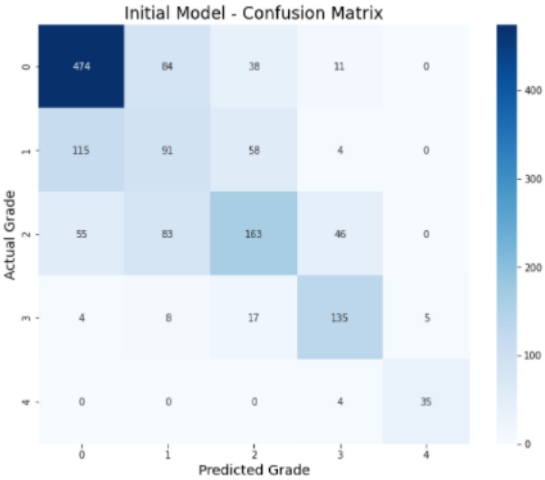


Figure 16. Classification matrix for Fusion model

8 Conclusion and Future Work

In this research project, we present a fusion system based on machine and deep learning methods to predict the severity of Knee Osteoarthritis. Ensemble models such as Random Forest, Gradient Boosting and Xtreme Gradient Boosting were trained with patient's data to predict each level of the disease according to Kelgren-Lawrence scale. DenseNet201 and InceptionResNet were the transfer learning models applied to achieve the same task utilising knee x-ray images. The main difference between this work and what state-of-the-art has developed in this domain is the usage of these two types of data.

The aim of the current work was to prove if there is a better performance using the proposed fusion system than employing a single model either machine or deep learning, which was demonstrated by the results obtained. Deep learning fed with x-ray images were more powerful in predicting the severity of KOA across all levels. However, incorporating patient's data allows to identify factors that determine the progression of the disease which is important for medical staff to deliver an integral diagnosis.

Although the proposed methodology contributes to better prediction of the disease, there are some limitations to consider. The models have been executed in a single dataset where all records were obtained from American people involved in OAI study. Besides, data used in this work comes only from the baseline of the study.

According to these limitations, in future work more data should be incorporated to train the models and enhance their performance, it is also recommended to use a different dataset to test them in order to analyse how they work in a different set of conditions and create a fusion model capable of being applicable at any patient cohort.

9 Acknowledgement

I would like to thank Dr. Hicham Rifai for the excellent guidance and valuable feedback throughout the supervision sessions for this research project. I would also like to thank my mother, father, and brothers for always trusting in me to achieve one my dreams and, my friends in Ireland for their endless motivation.

References

Abedin, J. *et al.* (2019) "Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images," *Scientific reports*, 9(1), p. 5761.

Alexos, A. et al. (2020) "Prediction of pain in knee osteoarthritis patients using machine learning: Data from Osteoarthritis Initiative," in 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA. IEEE, pp. 1–7.

Antony, J. *et al.* (2016) "Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks," *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1609.02469 (Accessed: August 10, 2021).

Antony, J. *et al.* (2017) "Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks," *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1703.09856 (Accessed: August 3, 2021).

Bandyopadhyay, S. and Sharma, P. (2016) "Detection of Osteoarthritis using Knee X-Ray Image Analyses: A Machine Vision based Approach." Available at: https://www.semanticscholar.org/paper/b5c55c6c40c389cb203bb654c78b2a3a306ffe4e (Accessed: August 4, 2021).

Bany Muhammad, M. *et al.* (2019) "Deep ensemble network for quantification and severity assessment of knee osteoarthritis," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, pp. 951–957.

Branco, P., Torgo, L. and Ribeiro, R. (2015) "A survey of predictive modelling under imbalanced distributions," *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1505.01658 (Accessed: August 1, 2021).

Chawla, N. V. *et al.* (2002) "SMOTE: Synthetic minority over-sampling technique," *The journal of artificial intelligence research*, 16, pp. 321–357.

Chen, P. *et al.* (2019) “Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss,” *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, 75, pp. 84–92.

Chen, T. and Guestrin, C. (2016) “XGBoost: A Scalable Tree Boosting System,” *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1603.02754 (Accessed: August 10, 2021).

Guerra-Hernández, A. and Mondragón-Becerra, R. (2008) “Explorations of the BDI Multi-agent support for the knowledge Discovery in Databases process.” Available at: https://www.semanticscholar.org/paper/690a3945f3bfd1445c68ef5697367066898b1b11 (Accessed: August 4, 2021).

Halilaj, E. *et al.* (2018) “Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative,” *Osteoarthritis and cartilage*, 26(12), pp. 1643–1650.

Huang, G. *et al.* (2017) “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Huang, G. *et al.* (2019) “Convolutional networks with dense connectivity,” *IEEE transactions on pattern analysis and machine intelligence*, pp. 1–1.

Huang, S.-C. *et al.* (2020) “Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines,” *npj digital medicine*, 3, p. 136.

Kokkotis, Christos et al. (2020) “A Machine Learning workflow for Diagnosis of Knee Osteoarthritis with a focus on post-hoc explainability”, en 2020 11th International Conference on Information, Intelligence, Systems and Applications IISA. IEEE.

Kwon, S. B. et al. (2020) “Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images”, IEEE access: practical innovations, open solutions, 8, pp. 120597–120603.

Lazzarini, N. *et al.* (2017) “A machine learning approach for the identification of new biomarkers for knee osteoarthritis development in overweight and obese women,” *Osteoarthritis and cartilage*, 25(12), pp. 2014–2021.

Leo Breiman Statistics, L. B. (2001) “Random Forests,” in *Machine Learning*.

Lespasio, M. (2017) “Knee Osteoarthritis: A Primer,” *The Permanente journal*, 21(4), pp. 16–183.

Luyten, F. P. *et al.* (2012) “Definition and classification of early osteoarthritis of the knee,” *Knee surgery, sports traumatology, arthroscopy: official journal of the ESSKA*, 20(3), pp. 401–406.

Minh, T. N. *et al.* (2018) “Automated image data preprocessing with deep reinforcement learning,” *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1806.05886 (Accessed: August 4, 2021).

Moustakidis, S. *et al.* (2019) “Application of machine intelligence for osteoarthritis classification: a classical implementation and a quantum perspective,” *Quantum Machine Intelligence*, 1(3–4), pp. 73–86.

Natekin, A. and Knoll, A. (2013) “Gradient boosting machines, a tutorial,” *Frontiers in neurorobotics*, 7. doi: 10.3389/fnbot.2013.00021.

Nasser, Y. et al. (2020) “Discriminative Regularized Auto-Encoder for early detection of knee OsteoArthritis: Data from the OsteoArthritis Initiative”, IEEE transactions on medical imaging, 39(9), pp. 2976–2984.

Ntakolia, C. et al. (2020) “A machine learning pipeline for predicting joint space narrowing in knee osteoarthritis patients,” in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, pp. 934–941.

Qiu, S. *et al.* (2018) “Fusion of deep learning models of MRI scans, Mini-Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment,” *Alzheimer’s & dementia (Amsterdam, Netherlands)*, 10(1), pp. 737–749.

Szegedy, C. *et al.* (2016) “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1602.07261 (Accessed: August 1, 2021).

Tiulpin, A. *et al.* (2018) “Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach,” *Scientific reports*, 8(1), p. 1727.

Tiulpin, A. *et al.* (2019) “Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data,” *Scientific reports*, 9(1), p. 20038.

Zhang, B. *et al.* (2020) “Attention-based CNN for KL grade classification: Data from the osteoarthritis initiative,” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 731–735.