# AI prediction model for Knee Arthroplasty

## Computer Vision and Machine Learning

Source Code: github

Kaggle: mortenrosenquist

Morten Lyng Rosenquist
*Faculty of Technical Sciences*
*Aarhus University*
Aarhus, Denmark
201706031

April 15, 2022

***Abstract*—**

***Index Terms*—**

## I. Introduction

This thesis will analyze the development of an AI prediction model in a clinical setting regarding knee arthroplasty. Based on a patients demographics, lifestyle and clinical metrics, the model is to determine a patient's risk group. Knowing a patient is at risk can improve the chance of the knee arthroplasty surviving by tailoring an individual treatment plan. The thesis is done as a project in the course Computer Vision and Machine Learning at Aarhus University. The dataset is given together with a explanation of the features and description of the problem.

### *Support Vector Machine*

The model that will be used is a Support Vector Machine(SVM). It produces nonlinear boundaries by creating a linear boundary in a tranformed version of the feature space. With this functionality SVMs can be used for classification, regression and outlier detection. We will use it for classification, this is also called Support Vector Classification(SVC). In its simplest form the SVC creates an linear optimal seperating hyperplane between two seperated classes. Where classes can not be seperated by a linear boundary other kernels(polynomial, Radial Basis Function) can be utilized. There are hyper-parameters to be tuned of the SVC. The most important is the regularization parameter that determines the amount of punishment for misclassifications. Utilizing a polynomial kernel the degree of the kernel function is also to be tuned.

### *Imbalanced Data*

Building a model with imbalanced data can lead to trouble. As the model primarily sees the majority class it might not learn enough from the minority class. This can lead to a model predicting all new observations as the majority class. This will lead to a high accuracy and recall for the majority class but a recall of 0 on the minority class. This is basically an useless model, that wont predict any of the minority class observations. In the clinical setting this means that we wont place any patients in the risk group. There are several ways to mitigate the issues with imbalanced data. The majority class can be downsampled or the minority class can be oversampled. With regards to SVC we can add class weights to penalize or reward the classification of a class.

### *Feature Selection*

The dataset contains a range of features regarding the patients. Having data with a high dimensionality can lead to problems such as long training time, a complex model and overfitting. Therefore, it will prefered if we can reduce the dimensionality while still having good performance of the model. This can be done by selecting the features that contribute the most to the target variable. This is typically supervised and will keep features intact. Principal Component Analysis can also be used to reduce the amount of features. This will however transform the features based on variance. In regards to a clinical settings it would be interesting to see which features are the most valuable in terms of determining wether an patient is in the risk group.

### *Metrics*

## II. Methods

### *Preprocesssing*

The dataset contains two types of features - boolean (yes/no) and continous values (numeric). The dataset is not complete, meaning some observations are missing values in certain features. To overcome this the mean value is used for the numeric features and the most popular for the boolean features. As the features are in different units it is important to scale

the data. Without scaling certain features might dominate the other features. This is specially the case for SVC, because it tries to maximize the distance between the support vectors and the hyperplane. *sklearn* have different options for scaling; StandardScaler, MinMaxScaler and RobusScaler. The MinMaxScaler with default settings will suffice with a scaling of each feature individually between 0 and 1 .

cite

*Model training*

Feature selection

Pipeline

GridSearch

Write Results

Best Model

AUC

f1

Confusion Matrix

Write Discussion

TP, TF, FP, FF

Not optimal model, not detecting the FP

Write Conclusion

Other model

conclude

## III. Results

## IV. Discussion

## V. Conclusion