Thrombosis prediction

AI approach

Lucio Honório Jr  
 Belo Horizonte Brazil  
 luciohcj@gmail.com

Luiz Augusto de Carvalho  
 Ontario Canada  
 ziulcarvalho@gmail.com

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

ABSTRACT

Total knee arthroplasty (TKA) is one of the most realized orthopedic procedures worldwide [1,2]. More than 600.000 of these surgeries are annually realized in USA [1]. Different complications were described and one of the most feared were venous thrombus embolism (VTE): Deep vein thrombosis (DVT) and pulmonary embolism (PE) 2.

Without prophylaxis DVT, symptomatic or not, could be detected in 41 to 84% of the image exams [3]. PE is not that common with variations between 1,5 and 10% but with mortality rates between 0,1 and 1,7% [4]. Many risk factors were described: age over 60 years, obesity, oral contraceptives, hormonal reposition therapy, bowel inflammatory disease, personal or familiar history of DVT or PE and long tourniquet time [5].

Basically there is mechanical and pharmacological methods to prevent PE and DVT. Up to 2015, many of them were used for prevention after TKA in \*\*\* name removed to first assessment. After that was initiated an institutional prevention protocol to standardize its prevention. We hypothesize that after protocol implantation smaller numbers of DVT and PE may be found. The objective of this study is verify if this new standardize protocol interfered in VTE incidence after primary TKA..

CCS

Computing methodologies, Artificial intelligence,

KEYWORDS

Mask RCNN, EfficientDet, Object detection, Data Augmentation.

INTRODUCTION

Venous thromboembolism (VTE) is known as a possible complication after total knee arthroplasty. In an attempt to standardize prevention measures against this phenomenon was implemented in a tertiary institution prevention protocol in 2015. The objective of this study is to evaluate the impact of this protocol for the prevention of venous thromboembolism in 2005 patients submitted to primary total knee arthroplasty.

The dataset provided has a quite limited number of files, 2005 this amount is very small for computer g models we are going to use. As we know the amount of input data plays a significant role in building high accuracy models.

ANALYTICAL SCHEMA

This work has as objective to create an AI model that based on the patients data predict the occurrence of thrombosis. Therefore our objective is to create a tool capable of based on ten patient parameters (eleven if we count the IMC which is obtained based on two of the ten) predict the occurrence of thrombosis. Note that although we have two types of thrombosis (Identified as TEP and TVP in the database) we are going to consider just a binary outbut: thrombosis or no thrombosis. Therefore we have a situation where regardless the patient had one or the two types the model predict Thrombosis. IN other words there is no discrimination about the type of the thrombosis. The spreadsheet bellow illustrate the logic:



We are going to build a basic model, which we are going to call “Baseline” and try to build upon it working, several parameters. This work is basically divided between the general description and evaluation of this baseline model and the description of the changes made on it and the evaluation of the results.

The first issue identified in the database provided was the unbalance between the positive and negative events:



This unbalance adds difficult to model the problem, we not only have few samples to work with but even fewer positives ones.

To circumvent this limitation we are going to follow two strategies:

We worked the changes in the baseline model following three types of strategies:

Data augmentation (Oversample): Using data augmentation techniques we evaluate the impact of the increase of the number of samples in our training base in the quality of our predictions.

Hyper parameters: Changing the hyper parameters we evaluate how the parameters impact the quality of our predictions.

Of course, due time constrains, it is not going to be possible to test every possible combination of data augmentation and Hyper parameters. We are presenting here the combinations which seemed more promising and we know that with enough time and data processing capability, best results can be achieved.

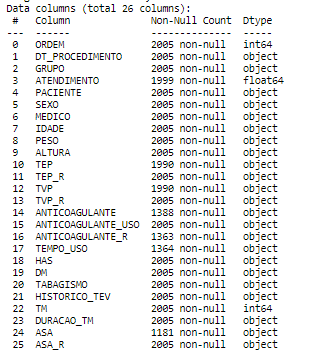
DATA EXPLORATION

Data from primary TKA patients realized between January 2011 and December 2017 were retrospectively collected from \*\*\* name removed to first assessment database, totalizing 2005 patients. This study was approved by the ethical committee of the Institution. An informed assigned consent was obtained before the study. No financial incentive was offered to any participant. Every primary TKA patients were included. Exclusion criteria was another surgical procedure together with TKA and infectious disease associated.

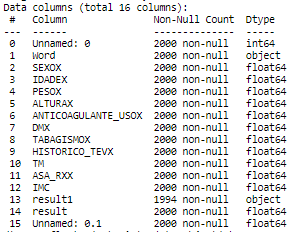
Every patient was operated on by the same surgical team with pneumatic tourniquet through anterior approach and medial parapatelar arthrotomy. Every TKA was with posterior cruciate sacrifice and both components were cementite at the same time. Patellar component was not used. Up to December 2014 there was not a padron regarding VET prophylaxis, pharmacological or mechanical, inside or outside the Institution. It was executed according each surgeon orientation to the patient. Starting in January, 2015 Madre Teresa Hospital standardized a clinical protocol to deal with VET prevention (Figure 1). In this protocol every patient received low molecular weight heparin (LMWH) subcutaneously 40 mg starting 6 hours after the end of the surgery and continuously every 24 hours up to the discharge to home. In this occasion was initiated oral anticoagulants (factor Xa inhibitors) up to 14 days in the post op. (PO). The same dose of LMWH was used every 12 hours if the patient had body mass index (BMI) over 30. Among mechanical measures, full weight bearing with a walker was initiated in the first day PO after peripheral nerve recovery or after intensive care unit release. The average length of stay in the hospital was 54 hours. Rehabilitation was prescribed to home for every patient and to ambulatory physiotherapy facilities as well.

Patient data analyzed were sex, age, weight, height, classification according American Anesthesiology Association (ASA), presence of diabetes, high blood pressure, smoking habits, time to the first walking, anticoagulant use after discharge, EP or DVT history and its occurrence up to 6 months PO. Only VE symptomatic cases were analyzed as well as those who needs any form of treatment. Asymptomatic cases were not studied.

The original CSV file layout is as follows:



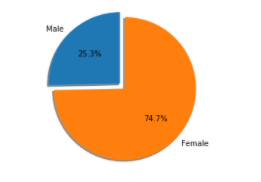
The treated database got the following layout:



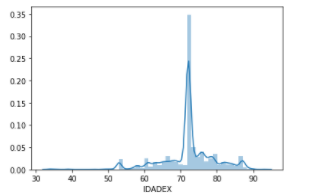
Note that the columns 0,1,13 and 15 have no relevance in the analytical model being there only for documentation and indexing propose. The parameters used as reference for prediction are the ones between 2 and 12 and the output is identified in columns 14 (result).

Analyzing the database regards the sex of the patients we have the following:



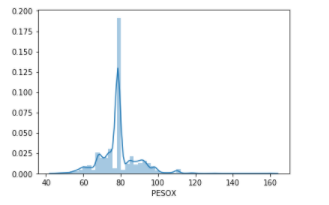


Analyzing the data regard the age of the patients we have the following:



Note the high concentration between 70 and 75 years old. The average age of the patients is 72 years.

When we look into the weight of the patients an interesting image emerges:

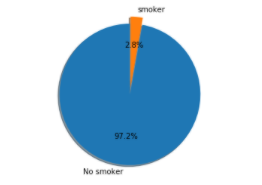


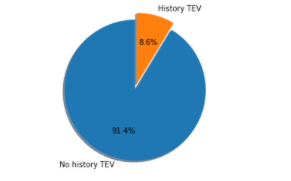
The average weight is 78.44Kg, a pretty high value. When correlating the high with the weight we have the IMC with the following average pattern:

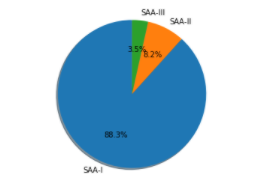


If we take the reference the recommended limits for woman and men (42 and 40 respectively) we can clearly see that our patients are basically old and overweigh.

When we look other aspects of the patients:

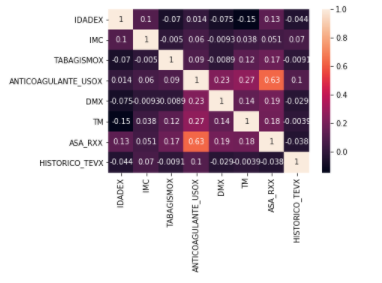






As can be seen, besides being predominately female, old and overweigh there is no defined characterization of the patients.

Trying to see if there is any direct correlation between the variables analyzed we found the following:



We can see that there some correlation between use of anti-coagulant and DM, TM and SAA (ASA) classification.

DATA PREPARATION

When analyzing the database we identified 809 out of 2005 patients where some information was missing. Given the limited number of samples we decided not to adopt a strategy where we simply eliminate the sample with some information missing. Our strategy was to fill the gaps with the average value of the column. This strategy is important because allows us to keep the samples and extract the useful information from the other columns.

Other important treatment adopted was due the unbalancing of the data. The original database had just 2,24% of positive outputs. To deal with this problem we used a technique called “Oversample”. It consists basically in randomically select a given number of samples of each class in a way to keep the number of outputs in each class equal. In our specific case we built a new database with 1.000 samples of each output. That means the positive output was repeated 20 times (we had only 45 positive outputs in the original file) and the negative ones 0.5 times.

BASELINE MODEL

The first step is to develop a baseline model. In our specific case we are going to use as our basic model random forest

* Preparing the model
* Training the model
* Evaluating the model
* Making predicitions.

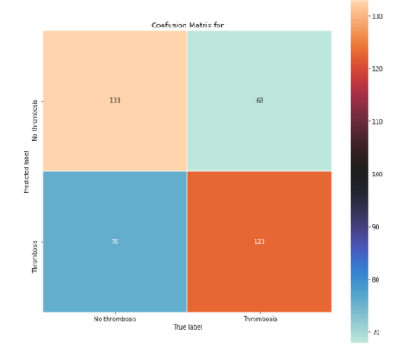
We are going to separate the oversampled database into training and test in a proportion of (80% training and 20% test). That means we are going to use 1.605 samples to train out model and apply this trained model to predict the output of 400 samples. Doing that we are going to check how many of these 400 the model predict correctly.

We are going to test several techniques:

* Linear Regression
* Decision tree
* Random Forest
* SVC

Once we identify which one of these models worked better we are going to adjust the hyper-parameters to improve the quality of the prediction even further.

Model 1 – Linear regression

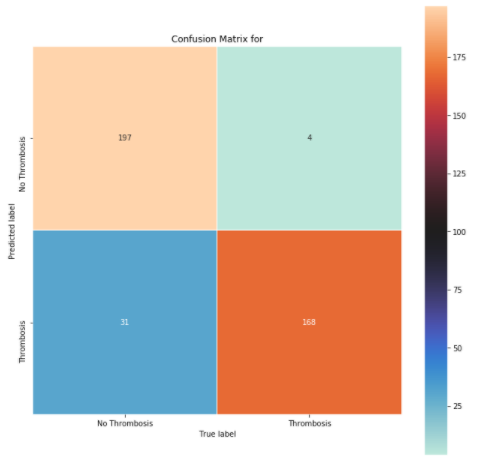


This model got it right 64% of the samples. Note that in this particular problem is important to get the false negative as smaller as possible. That means a good model may guesses the patient having thrombosis when in fact it hadn’t but what it cannot do is to predict that a patient is not going to have it and in the end he ends-up having.



Model 2 – Decision Tree

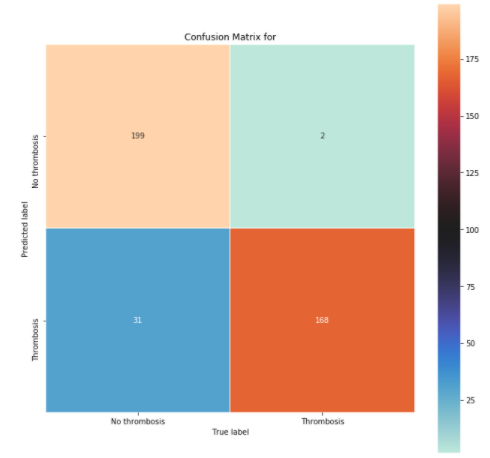
.



Note that this model seems to be much better than the previous one. It not only got a much higher percentage of right predictions but a very low percentage of false negatives.



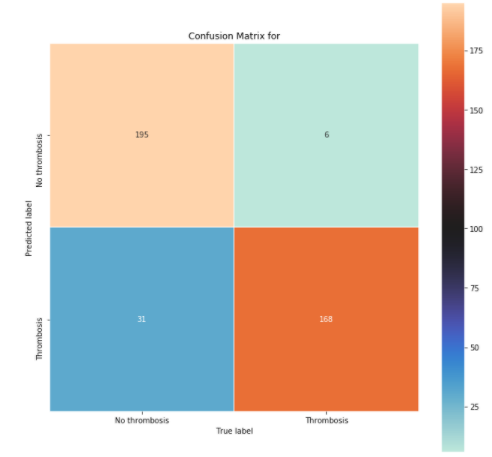
Model 3 – Random Forest



Note that we got a small but important improvement in the false-negative using this model (50% smaller). The number of false positive remained equal the previous model. Overall this model got right 0,50% more than the previous one but this improvement come from the fase-positives what in our context is very important.



Model 4 – SVC (Vector Classifier)



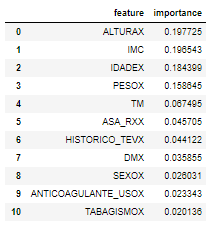
This model proved to be better than Linear regression but inferior than both Decision three and Random Forest.



Therefore becomes clear que the best model to predict the occurrence of thrombosis is Random Forest (Among the four models tested). It got 91.75% correct, 0.5% of false negatives 7.75& false positives. That means for each 200 patients the model will predict correctly 184 (having or not Thrombosis), predict incorrectly that 15 will have and predict incorrectly that 2 will not have.

THE IMPORTANCE OF THE PARAMETERS

Here we are going to evaluate the impact of each one of the parameters in the decision process of the model. This is an important evaluation because tells us what is really being important when taking the decision to classify a patient as “No thrombosis” or “Thrombosis”:

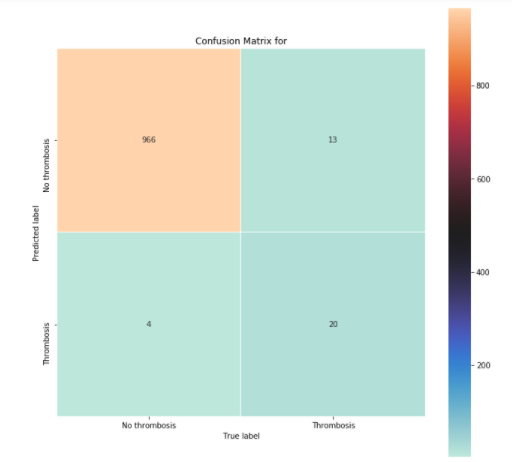


Here some very interesting things can be seen: IMC seems to be the most important factor in the definition (Considering that IMC is a composition of ALTURA and PESO (Height and Weight), followed by the age. A way down we have TM, SAA classification and History of TEV. Sex, use of anticoagulant and Tabagism seem to have almost no importance regards having or not Thrombosis.

Being old and over-weight seems to be the defining factors, having a pre-existing illness adds to your chance to having problems. Pretty common sense conclusion.

VALIDATION

We select a 50% subset of the original file (not balanced) and ask the trained model to classify it for us.



As we can see the results we a bit worse than the one obtained with the test set:

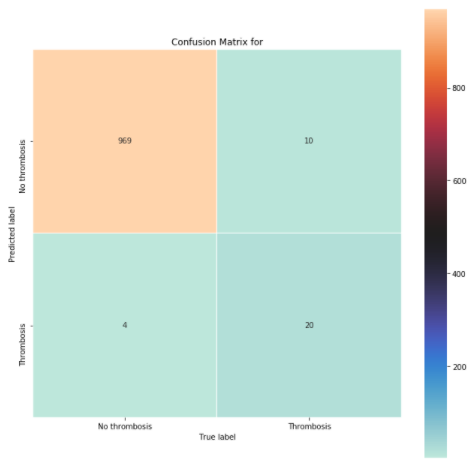


This is an indication that there is some overfitting in the training dataset, what is expected in a context where we had so few positive events to work with.

If we analyze the minority class (thrombosis) e can see that the model was capable of identifying 20 out of the 33 (60% of the events).

TRYING TO IMPROVE THE MODEL

One strategy to avoid overfitting is known “ensembling” it consists basically combining the models and by majority vote classify the item. Combining the four models we manage to archive a small improvement:





Small improvement in the total percentage of correct predictions but almost 30% reduction on the false negative.

If we think that the unbalanced class (thrombosis) is 2,24% of the samples that means the model gets it right 55% of the times.

REFERENCES

HCPUnet, Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality. Disponível em: http://hcupnet.ahrq.gov

Ethgen O, Bruière O, Richy F, Dardennes C, Reginster JY. Health related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. J Bone Joint Surg Am. v.86-A, n.5, p.963-974. 2004.

Khokhar A. et al. Venous thromboembolism and its prophylaxis in elective knee arthroplasty: An international perspective. The Knee. v. 20, p.170-176. 2013

Sivashankar Chandrasekaran, Siva Kumar Ariaretnam, Jason Tsung and David Dickison. Early mobilization after total knee replacement reduces the incidence of deep venous thrombosis. ANZ J Surg. v.79, p.526–529. 2009.

Anderson, M. Dunbar, J. Murnaghan, S.R. Kahn, P. Gross et al Aspirin or Rivaroxaban for VTE Prophylaxis after Hip or knee arthroplasty. N Engl J Med p.378:388. 2018

Andersen P,K, Kehlet H. Should deep venous thrombosis prophylaxis be used in fasttrack hip and knee replacement? Acta Orthopaedica. v.83, n. 2 p.105–106. 2012.

O’Reilly RF, Burgess I, Zicat BA. The prevalence of venous thromboembolism after hip and knee replacement surgery. Med J Aust. v.182, p.154-159. 2015.

Song K, Xu Z, Rong Z, Yang X, Yao Y, Shen Y, et al. The incidence of venous thromboembolism following total knee arthroplasty: a prospective study by using computed tomographic pulmonary angiography in combination with bilateral lower limb venography. Blood Coagul Fibrinolysis. v.27, n.3, p.266–269. 2016.

Quinlan DJ, Eikelboom J, Dahl O, et al. Association between asymptomatic deep vein thrombosis detected by venography and symptomatic venous thromboembolism in patients undergoing elective hip or knee surgery. J Thromb Haemost. v.5, p. 1438-1443. 2007.

Howie C, Hughes H, Watts AC. Venous thromboembolism associated with hip and knee replacement over a ten-year period: a population-based study. J Bone Joint Surg Br. v87-B p.1675–1680. 2005.

Zhang et al. Risk factors for venous thromboembolism of total hip arthroplasty and total knee arthroplasty: a systematic review of evidences in ten years. BMC Musculoskeletal Disorders. p.16-24. 2015.