ZGT Work Report - data imbalance problem in 30-days mortality prediction of elderly hip fraction patients

Changqing Lu

November 2020

1 Problem Statement

Hip fracture has been observed an increasing risk to the human beings all over the world. According to [1], a large number ($\approx$ 1.3 million) of people suffered from hip fractures in 1990, and the number is expected to be 3∼4 times until the Mid of the 21th century. Among these people, the elderly was reported a higher risk of mortality due to their poorer physical conditions and comorbidities. In this project, we proposed to develop a multi-modal well-performing model using machine learning techniques to predict the 30-days mortality of the elderly people who suffered a hip fracture based on the structured medical variables and X-ray images.

One of the important problems in the 30-days mortality prediction model development is to deal with the existing data imbalance, that most elderly patients who received the surgery after hip fracture can survive resulting in a small proportion of the positive samples. It was reported in [2] that, the 30-days mortality rates of the elderly hip fracture patients can reach around 10%. To our knowledge, machine learning techniques usually perform badly when they are trained on a heavily imbalanced dataset and even worse especially when the dataset is very small. In the dataset of our study, there are only 1056 patients in total, and among whom approximately 8% suffered the mortality within 30 days after receiving the surgery. However, the ability of the prediction to detect the risk of the mortality is significant, as it can provide the doctors suggestions and guidance for further medication of the patients. Therefore, a proper method employed to handle the data imbalance deserves to be investigated.

2 Related Work

Many previous work [3, 4, 5, 6] have been conducted on the 30-days mortality prediction of elderly hip fracture patients, utilizing various machine learning classification techniques. Generally, the commonly used variables include age, gender, pred-fracture mobility, fracture type etc, which is much less than that in our project. To have a comparable evaluation of the prediction models, the AUC (i.e. Area under the ROC curve of the prediction model) is reported. An acceptable prediction model should achieve an AUC of at least 0.7.

To deal with the data imbalance problems, overall, there are two categories of the most widely used methods: the outlier detection and oversampling methods. The former regards majority class as the normal data while the minority class as the outliers, and attempts to figure out the margins between them. The latter aims to give equal attention on both majority and minority classes in the training procedure of model fitting, and avoids the possible bias. Various approaches on both categories have been proposed and proved helpful and efficient in dealing with data imbalance issues. In this project, we intend to apply them on our 30-days mortality prediction and investigate their performance to provide further suggestions on the improvement of the prediction model.

3 Data Preparation

3.1 Dataset

A structured dataset consisting of 1026 patients with various hip fraction type recorded in the Ziekenhuis Groep Twente (ZGT, Hospital Group Twente) was provided. Each patient recording contains the variables extracted from different data sources of the hospital, e.g. non-private personal information, fracture types of the patients, lab test results from the laboratory and medications. Totally, 87 variables are available for one patient. Some of them are raw features while others are hand-engineered features that are extracted by medical technicians. In general, these variables can be grouped into separate categories, e.g. Demographics, Activities of Daily Living, Medications, Blood, Cardiology, Mobility, Radiology and Lab Tests. However, an overlapping issue has to be noted that, for specific lab test results, both the original measurement of a test and its evaluation result (i.e. normal/abnormal) are included, which might cause the redundancy of the feature information.

3.2 Pre-processing

To simplify the feature representations and prepare a complete and structured dataset for further 30-days mortality prediction, we performed a three-step pre-processing procedure on the original dataset. Firstly, we excluded the patient recordings which contains missing data and checked that this operation didn't affect the ratio of the majority and minority class in this data imbalance problem. Secondly, we categorize the labels of a feature into numerical factors, e.g., the evaluation result – *normal* and *abnormal* are categorized into 0 and 1 respectively. Thirdly, to handle with the overlapping issue that a lab test contains both measurement and evaluation result, we only keep the evaluation result, as generally an expert will analyze the mortality risk based on the evaluation results of different lab tests while only look into very few measurement results when necessary. We also used the Katz index of independence to represent the daily living activity capacity of a patient instead of including many sub-indexes, e.g. capacity to take a shower on his/her own. After the pre-processing procedure, 756 patient recordings are finally involved in the dataset, where 690 out of them survive while 66 out of them die after the hip fracture. Each recording consists of 58 features. We list their English translations (originally in Dutch) and feature grouping categories in Table 5 in the Appendix.

4 Research Phase 1

In the first phase, we mainly considered two categories of methods to address the data imbalance problems in our study. One is the outlier detection methods, which can solve the data imbalance problem by regarding the minority class data as the outliers when the majority-minority ratio is high. The other is the oversampling techniques, which can solve the data imbalance problem by oversampling the minority class data to guarantee both classes receiving the equal attention from the classification model.

4.1 Outlier Detection Methods

Four of the most widely used outlier detection methods were used to distinguish the majority and minority classes: One-class SVM, Isolation Forest, Local Outlier Factor and DB SCAN.

One-class SVM One-class Support Vector Machine [7] originates from basic Support Vector Machine (SVM). A SVM constructs a hyperplane or a set of hyperplanes in a higher dimensional space to classify the instances via their attributes. Intuitively, a good separation of SVM is the hyperplane which builds the largest margin between two classes, as the larger the margin, the lower the generalization error of the classifier. One-class SVM followed the same directions to identify the outliers. However, instead of finding an optimal hyperplane that can separate two classes, One-class SVM is trying to find a hypersphere to encompass all normal instances. So that, data points which are not included in the hypersphere will be identified as the outliers.

Isolation Forest Isolation Forest [8] is an unsupervised learning algorithm for outlier detection mainly based on the principle that outliers are usually very rare in the dataset. Isolation Forest consists of many Isolation Trees which are generated from different sampling sets of the dataset. The construction of an isolation tree starts from randomly select an attribute and a value of this attribute as the guard condition of the root. Then, the root node classifies each recording according to its value of this attribute, where recordings with smaller values on this attribute will be separated into the left branch, and otherwise into the right branch. After that, such operation will be executed recursively until one of the two conditions are satisfied: 1) the remaining dataset at a node has only one recording or multiple identical recordings, or 2) the height of the tree has reached the limitation. Finally, outliers can be identified, if it can be classified into a leaf in a short path of the tree, as rarely appeared data will usually be separated as a shallow leaf in the training process.

Local Outlier Factor Local Outlier Factor [9] can be used to identify the outliers based on the concept of local density. At first, the k nearest neighbours theory is considered to produce the locality over the data point distribution. Further, the distance measured in the locality is employed to estimate the local density of it. Then, Local Outlier Factor identify the outliers via comparing the local density of a part to its neighbourhood, where points in regions of similar density compared to their neighbours are considered as normal points while points in regions which have a substantially lower density compared to their neighbours are regarded as outliers.

DB SCAN DB SCAN [10] is a very famous density-based data clustering algorithm to find the outliers as well. However, it works slightly differently from Local Outlier Factor. There are three steps in the DB SCAN algorithm, which are namely to identify the core points, reachable points and outliers respectively. Firstly, a distance ϵ and a number n are selected to determine the core points of a dataset, where points which have at least n neighbouring data points in the distance smaller than ϵ are regarded as core points. Secondly, directly reachable points whose distances from a core points is smaller than ϵ and reachable points that there exists a path to connect them to the core points via directly reachable points are identified. Finally, the remaining data points will be considered as the outliers.

4.2 Oversampling Techniques

We only considered the oversampling techniques instead of the undersampling techniques in our study, as our final dataset consists of only 756 patient recordings without missing data. Any abandonment of data can result in an imbalanced model fitting, due to the information loss, e.g. the majority-minority ratio changes.

4.2.1 Random Oversampling

Random Oversampling is a basic oversampling technique, where the existing examples of the minority class are randomly selected and appended to the dataset to guarantee the number majority and minority class examples used for model training are finally equal.

4.2.2 Synthetic Minority Oversampling Techniques

Another couple of the widely used oversampling techniques is the Synthetic Minority Oversampling Techniques [11] (SMOTE). While, different from the basic oversampling which duplicate minority classes, SMOTE generates synthetic data artificially to achieve a balanced class distribution according to certain rules.

SMOTE Basic SMOTE [11] aims to generate synthetic data on the connection line of two examples in minority classes which can indicate the commonality of that class. The generating process for one unit is as follow. Firstly, a minority class example is selected (denoted as X_i). After that, the distance between other minority class examples and that example are calculated and K nearest minority class examples (K is defined at the beginning) are collected. Then, one of the K nearest examples is randomly selected (denoted as X_j). Finally, the synthetic data is generated in the form of

$$X_{syn} = X_i + (X_i - X_j) \times \alpha \tag{1}$$

where alpha is a random number between [0,1]. This process is performed on each minority class example in iterations until the desired ratio between majority and minority classes is reached.

Borderline SMOTE Borderline SMOTE [12] is another SMOTE oversampling technique, where more attention are paid to the minority class examples located near the borderline of the majority and minority classes. The motivation behind is that, the examples located near the borderline are usually more difficult to classify, which is even more serious for the minority class of an imbalanced dataset. The generating process works in the same way as the basic SMOTE, and the only difference is that the expected synthetic examples generation is based on the minority class examples located on the borderline.

Adaptive Synthetic Sampling Adaptive Synthetic Sampling [13] (ADASYN) is a SMOTE oversampling technique as well, where more attention are paid to the minority class examples located in the area with low distribution density. Similar to Borderline SMOTE, ADASYN intends to generate the synthetic data for the examples which are more difficult to classify due to their rare appearance. The generating process starts with a density distribution analysis of the minority class, reporting the classifying difficulty of each subset of that class and determining the desired number of synthetic data examples generated for each minority class example. After that, similar generating operations are applied as well.

5 Research Phase 2

According to the results of research phase, we find that outlier detection methods seems unsuitable for the data imbalance problem in the 30-days mortality prediction while oversampling techniques can improve the prediction performance slightly. However, the performance is still far away from the requirement by the experts that an AUC score larger than 0.7 is expected. Additionally, in the literature survey, we find that previous studies usually used a few important features (e.g. Katz index of independence and ASA score) to predict the 30-days mortality, while we used a structured dataset with many more features. Obviously, not all features are helpful in the prediction. On the contrary, some of them might bring much more confusion to the prediction. Therefore, we proposed a feature correlation/importance analysis to identify the most important ones.

5.1 Feature Correlation/Importance Analysis

To have a comprehensive correlation analysis on all variables, we selected four widely used methods from different perspectives to estimate the feature significance. Firstly, Pearson Correlation Coefficient and Distance Correlation Coefficient are employed to test the linear and non-linear correlation between two variables from a statistical point of view. Secondly, Maximal Correlation Coefficient is used to estimate how much information a variable contains about the other according to information theory. Thirdly, considering the dependence between covariates, we applied a classifier-based correlation analysis – Random Forest to investigate the most important features contributing to the 30-days mortality prediction.

Pearson Correlation Coefficient Pearson correlation coefficient [14] is a statistic that measures linear correlation between two variables. It is defined as the covariance of the two variables divided by the product of their standard deviations. When applied to a sample of two variables $X = [x_1, ..., x_n]$ and $Y = [y_1, ..., y_n])$, the calculating formula can be represented as the follow

$$C_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the mean value of the two variables X and Y. In the measurement, correlations are normalized to $[-1, 1]$, where 1 denotes total positive linear correlation, -1 denotes total negative linear correlation while 0 denotes no linear correlation.

Distance Correlation Coefficient Distance Correlation Coefficient [15] is a statistics measures the dependence between two paired arbitrary variables without necessarily equal dimensions. Compared to Pearson Correlation Coefficient, it considers the non-linear relationship between the two variables as well. It calculates the correlation of a sample of two variables X and Y using the following formula

$$dC_{XY} = \frac{dCov(X, Y)}{\sqrt{dVar(X)dVar(Y)}} \tag{3}$$

where the only difference from Pearson computation is that the distance covariance and distance variance are used instead of the covariance and variance. In the measurement, the distance correlation coefficient is 0 if and only if the random vectors are independent while other numbers indicates corresponding positive and negative correlations.

Maximal Information Coefficient Maximal information coefficient [16] measures the linear or non-linear association between two variables X and Y as well based on the mutual information. It uses binning to apply mutual information on continuous random variables. Compared to other mutual information methods, Maximal information coefficient additionally selects the number of bins and picks the maximum from possible choices. Finally, the strength of the correlation between two variables can be obtained.

Random Forest We applied the classifier-based method – Random Forest [17] to identify the feature importance, which can take the variable dependence into account. In this method, a random forest is built at first as a classification model to classify the two classes. To obtain the feature importance, one of the features will be randomly permutated at a time and corresponding classification error caused by the random shuffling will be measured, where a larger error means, the corresponding feature is more important in the classification. The reason behind is that, important features will have more impactful effects than others, and the information loss of them will accordingly decrease the performance more.

5.2 Feature Selection: Voting Mechanism

An interesting finding is obtained from the correlation analysis that various correlation analysis methods showed similar results when estimating feature importance. Even though, there are still slightly differences between different methods. To determine the important features that will be used for further prediction, we designed a voting mechanism based on both feature importance results and literature suggestions. The voting mechanism is that, if a feature is suggested as important by one of the feature importance methods or the literature, it will be given a star symbol $\star$ accordingly. All features will be compared eventually, and those whose importance are more than two stars will be decided as the final feature representations used for the prediction. Here, we list the selected features with their significance in Table 1. Based on that, we performed the oversampling techniques again for the prediction mmodel using important features.

Feature	**Significance**	**Feature**	**Significance**
Feature type	$\star\star\star\star\star$	Care path and history	$\star\star\star\star$
Gender	$\star\star\star\star$	Pre fracture mobility	$\star\star\star\star$
Age	$\star\star\star\star$	Heart axis orientation EKG	$\star\star\star$
Katz index of independence	$\star\star\star\star$	Therapy type	$\star\star\star$
ASA score	$\star\star\star\star$	N05 (psycholeptics)	$\star\star\star$
Blood pressure systolic	$\star\star\star\star$	Memory problem	$\star\star$
Blood pressure diastolic	$\star\star\star\star$	UREU (renal)	$\star\star$
Width of QRS complex EKG	$\star\star\star\star$	CRP (infection)	$\star\star$

Table 1: Feature significance results

6 Results

6.1 Evaluation Metrics

In this study, the prediction problem is actually a binary classification problem. The positive class is the patients who have deceased in 30-days after a hip fracture surgery while the negative class is the patients who have survived 30-days. The 'True Positives', 'True Negatives', 'False Positives' and 'False Negatives' are defined as 'Correctly predicted samples which are originally positive', 'Correctly predicted samples which are originally negative', 'Wrongly predicted samples which are originally negative' and 'Wrongly predicted samples which are originally positive' respectively.

Based on that, the evaluation metrics can be defined as follows:

Specificity, which shows the ability to capture the patients who are going to survive, is

$$Specificity = \frac{TN}{FP + TN} \tag{4}$$

Precision, which shows the ability to be precise in predicting patients who are going to die, is

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall, which shows the ability to ability to capture the patients who are going to die, is

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

F1-score, which shows the balance between **Precision** and **Recall**, is

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

Accuracy, which shows the ability to correctly predict samples over all samples, is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Area Under the Receiver Operating Characteristic curve (AUC), that measures the area under the ROC curve where x-axis is the True Positive Rate and y-axis is the False Positive Rate, can show the ability of the model in this binary classification problem. According to [18], an AUC between 0.7 and 0.8 is considered as acceptable discrimination, an AUC between 0.8 and 0.9 is considered as excellent discrimination and an AUC above 0.9 is considered as outstanding discrimination.

6.2 Experimental Setup

To have a global evaluation on each method we used in our two research phases, we performed the 5-fold cross validation experiments. In the cross validation, the dataset were split into five parts. At each time, four out of them were used as the training set while the remaining one was used as testing set. The final result will be estimated on the combination of the five testing sets, which is actually the global performance on the dataset.

For outlier detection methods, we just followed their corresponding standard instructions in the Python package: sklearn. For oversampling methods, we chose the support vector machine (SVM) with the linear kernel function as the classifier, because it showed the best performance when compared to other classifiers, according to the pilot study from Berk Yenidogan et al..

6.3 Outlier Detection Methods

The results of the outlier detection methods are shown in Table 2. Obviously, outlier detection methods are not suitable in solving the data imbalance problem here, as they can only obtain an AUC between 0.5 and 0.6. Additionally, none of the four selected outlier detection methods has an outstanding effect when compared to another.

Method	Specificity	Precision	Recall	Accuracy	F1-score	AUC
One-class SVM	0.916	0.134	0.136	0.848	0.135	0.531
Isolation Forest	0.904	**0.154**	0.182	0.841	**0.167**	**0.545**
Local Outlier Factor	**0.920**	**0.154**	0.152	**0.853**	0.153	0.538
DB SCAN	0.838	0.118	**0.227**	0.784	0.155	0.532

Table 2: Results on outlier detection methods

6.4 Oversampling Methods

The results of the oversampling methods where SVM is used as the classifier are shown in Table 3. Obviously, all oversampling methods improve the classification results more or less, as the AUC of them increase by 0.1 compared to the classification where no oversampling technique is used. Additionally, among the four oversampling methods, Borderline SMOTE obtains the best performance. The reason behind may be that many minority classes are located near the borderline which needs more attention during the training process. However, with the help of the oversampling techniques, classification performances are still unsatisfactory where the AUC scores can only reach between 0.6 and 0.7, while an AUC between 0.7 and 0.8 can be considered as acceptable. Therefore, further study to improve the classification performance is required.

Oversampling Technique	Specificity	Precision	Recall	Accuracy	F1-score	AUC
No Oversampling	**0.974**	**0.182**	0.061	**0.894**	0.091	0.521
Random Oversampling	0.757	0.161	**0.489**	0.733	0.242	0.626
SMOTE	0.774	0.167	0.471	0.748	0.247	0.625
Borderline SMOTE	0.796	0.175	0.453	0.766	**0.252**	**0.628**
ADASYN	0.766	0.161	0.470	0.740	0.240	0.619

Table 3: Results on oversampling methods

6.5 Oversampling Methods over Feature Selection

The results of the oversampling methods that are applied over selected features are shown in Table 4. Compared to Table 3, it indicates that feature selection greatly improves the classification performance a lot, as the AUC scores increase by 0.1 again. Among the four oversampling methods, Borderline SMOTE still obtains the best performance, with an accuracy of 0.745, a F1-score of 0.319 and an AUC score of 0.721, which outperforms a previous work in this field [5]. The performance increase obtained after the feature selection illustrates that, not all features are required in the 30-days mortality prediction. On the contrary, the inclusion of unnecessary features might bring too much confusion for the classification model, resulting in performance decrease accordingly.

Oversampling Technique	Specificity	Precision	Recall	Accuracy	F1-score	AUC
Random Oversampling	0.719	0.192	0.700	0.717	0.301	0.713
SMOTE	0.720	0.195	**0.708**	0.719	0.306	0.719
Borderline SMOTE	**0.751**	**0.208**	0.682	**0.745**	**0.319**	**0.721**
ADASYN	0.714	0.190	0.700	0.713	0.299	0.712

Table 4: Results on oversampling methods after feature selection

7 Conclusion and Future Work

In the study, we mainly investigate the possible solutions to deal with the data imbalance issue existing in the 30-days mortality prediction of elderly hip fracture. Based on the results we obtained in Section 6, we can conclude that outlier detection methods are not suitable for the data imbalance issue here, while oversampling methods can help improve the performance a bit. However, due to a large amount of variables we use for the prediction model, the prediction performance is affected by the redundant and confusing information. To find out the independent and important features out of all variables we have, we performed the feature selection stage. Results show that such feature selection is necessary in the 30-days mortality prediction of elderly hip fracture, as the prediction performance using importance features is improved a lot.

In the future work, it is suggested to have a deeper and more comprehensive feature selection for this project. Due to the lock down situation recently, the feature selection in our study is limited to mathematical ways. In addition, to remove the redundant information and simplify the prediction model, e.g., for the lab test results, we only preserved their evaluation result (i.e. normal or abnormal). In fact, it is necessary to discuss them with experienced experts and determine final significant features accordingly. Additionally, since our task is mainly to solve the data imbalance issue in this project, we didn't take a look at the performance of other classifiers here. Therefore, it is also interesting to test their capacity of predicting the 30-days mortality of elderly hip fracture and explore their effectiveness combined with the feature selection.

References

[1] B. Gullberg, O. Johnell, and J. Kanis, "World-wide projections for hip fracture," *Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA*, vol. 7, no. 5, p. 407—413, 1997.

[2] F. Hu, C. Jiang, J. Shen, P. Tang, and Y. Wang, "Preoperative predictors for mortality following hip fracture surgery: a systematic review and meta-analysis," *Injury*, vol. 43, no. 6, p. 676—685, June 2012.

[3] G. Holt, R. Smith, K. Duncan, D. Finlayson, and A. Gregori, "Early mortality after surgical fixation of hip fractures in the elderly - an analysis of data from the scottish hip fracture audit," *The Journal of bone and joint surgery. British volume*, vol. 90, pp. 1357–63, 11 2008.

[4] L. Kirkland, D. Kashiwagi, M. Burton, S. Cha, and P. Varkey, "The charlson comorbidity index score as a predictor of 30-day mortality after hip fracture surgery," *American journal of medical quality : the official journal of the American College of Medical Quality*, vol. 26, pp. 461–7, 03 2011.

[5] W. Nijmeijer, E. Folbert, M. Vermeer, J. Slaets, and J. Hegeman, "Prediction of early mortality following hip fracture surgery in frail elderly: The almelo hip fracture score (ahfs)," *Injury*, vol. 47, 07 2016.

[6] J. Karres, N. Kieviet, J. Eerenberg, and B. Vrouenraets, "Predicting early mortality following hip fracture surgery: the hip fracture estimator of mortality amsterdam (hema)," *Journal of Orthopaedic Trauma*, vol. 32, p. 1, 09 2017.

[7] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*. The MIT Press, 1999, pp. 582–588.

[8] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422.

[9] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "Lof: Identifying density-based local outliers." vol. 29, 06 2000, pp. 93–104.

[10] A. Ram, J. Sunita, A. Jalal, and K. Manoj, "A density based algorithm for discovering density varied clusters in large spatial databases," *International Journal of Computer Applications*, vol. 3, 06 2010.

[11] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, 06 2002.

[12] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887.

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[14] W. contributors, “Pearson correlation coefficient — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=991140692, 2020.

[15] ——, “Distance correlation — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=Distance_correlation&oldid=978003107, 2020, [Online; accessed 30-November-2020].

[16] ——, “Maximal information coefficient — Wikipedia, the free encyclopedia,” https://en.wikipedia.org/w/index.php?title=Maximal_information_coefficient&oldid=974810958, 2020, [Online; accessed 30-November-2020].

[17] A. Cutler, D. Cutler, and J. Stevens, *Random Forests*, 01 2011, vol. 45, pp. 157–176.

[18] D. W. Hosmer and S. Lemeshow, *Applied logistic regression.* John Wiley and Sons, 2000.

A Appendix

No.	Feature Name	Feature Group
1	Gender	Demongraphics
2	Age	
3	Prone to delirim	Cognitive Problems
4	Memory problem	
5	Previous confusional state	
6	Katz index of independence	Activities of Daily Living
7	Prone to under nutrition	Assessment
8	Short nutritional assessment score	
9	Unintentionally lost weight	
10	Loss of appetite	
11	Drink or tube feeding	
12	American Society of Anesthesiologists (ASA) physical status classification	
13	Blood pressure systolic	Blood
14	Blood pressure diastolic	
15	Blood thinner medication	
16	Width of QRS complex EKG	Cardiology
17	Heart axis orientation EKG	
18	Risk of falling	Falling
19	A fall accident in past 6 months	
20	Fracture type	Fracture
21	Laterality	
22	Therapy type	
23	Care path and history	
24	Pre fracture mobility	Mobility
25	A02 (drugs for acid related disorders)	Medication
26	A10 (drugs used in diabetis)	
27	B01 (antithrombotic agents)	
28	B02 (antihemorrhagics)	
29	B03 (antianemic preparations)	
30	C01 (cardiac therapy)	
31	C03 (diuretics)	
32	C07 (beta blocking agents)	
33	C08 (calcium channel blockers)	
34	C09 (agents acting on the renin-angiotensin system)	
35	C10 (lipid modifying agents)	
36	L04 (immunosuppressants)	
37	M01 (anti-inflammatory and antirheumatic products)	
38	N05 (psycholeptics)	
39	R03 (drugs for obstructive airway diseases)	
40	HB	Lab Test
41	HT	
42	CRP	
43	LEUC	
44	THR	
45	ALKF	
46	GGT	
47	ASAT	
48	ALAT	
49	LDH1	
50	UREU	
51	KREA	
52	GFRM	
53	NA	
54	XKA	
55	GLUCGLUC	
56	X-ray finding 1	Radiology
57	X-ray finding 2	
58	X-ray finding 3	

Table 5: Pre-processed features and belonging groups