A Comparative Study of Data Mining Techniques Applied to Renal-Cell Carcinomas

Ana Duarte[1], Hugo Peixoto[2(✉)], and José Machado[2]

[1] University of Minho, Campus Gualtar, Braga, Portugal
[2] Centro Algoritmi, University of Minho, Campus Gualtar, Braga, Portugal
hpeixoto@di.uminho.pt

Abstract. Despite being one of the deadliest diseases and the enormous evolution in fighting it, the best methods to predict kidney cancer, namely Renal-Cell Carcinomas (RCC), are not well-known. One of the solutions to accelerate the current knowledge about RCC is through the use of Data Mining techniques based on patients' personal and clinical data. Therefore, it is crucial to understand which techniques are the most suitable to extract knowledge about this disease. In this paper, we followed the CRISP-DM methodology to simulate different techniques to determine the ones with the best predictive performance. For this purpose, we used a dataset of 821 records of RCC patients, obtained from The Cancer Genome Atlas. The present work tests different Data Mining techniques, that can be used to predict the 5-year life expectancy of patients with renal cancer and to predict the number of days to death for patients who have a life expectancy of less than 5 years. The results obtained demonstrated that the best algorithm for estimating the vital status at 5 years was Random Forest. This algorithm presented an accuracy of 87.65% and an AUROC of 0.931. For the prediction of days to death, the best performance was obtained with the k-Nearest Neighbors algorithm with a root mean square error of 354.6 days. The work suggested that Data Mining techniques can help to understand the influence of various risk factors on the life expectancy of patients with RCC.

Keywords: Renal-Cell Carcinoma · Data Mining · Survival · Life expectancy · RapidMiner

1 Introduction

On a global scale, cancer is one of the major concerns of public health authorities. One of the most common type is kidney cancer, which causes approximately 430,000 new cases per year and corresponds to the 15th deadliest cancer worldwide [1]. The most representative form of kidney cancer is Renal-Cell Carcinoma (RCC), which accounts for about 90% of total cases, and refers to any malignant tumour that originates from the renal epithelium. This type of disease is divided into several histologic subtypes that have different specific characteristics. Clear-cell RCC (ccRCC), papillary RCC (pRCC) and chromophobe RCC (chRCC) are the most typical forms, and the other subtypes

S. Spinsante et al. (Eds.): HealthyIoT 2021, LNICST 432, pp. 53–62, 2022.
https://doi.org/10.1007/978-3-030-99197-5_5

represent only a residual proportion of the total incidence. The predominant subtype is ccRCC (≈75% of the total), followed by pRCC (≈10%) and chRCC (≈5%) [2–4].

The Tumour Node Metastasis (TNM) staging system is the most widely used tool for classifying malignant tumours, including RCC. According to the system's terminology, T refers to the size and extent of the primary tumour and indicates whether it has grown into nearby areas; N whether the tumour has spread to nearby lymph nodes; and M refers to metastasis, that is, if the cancer has spread to other parts of the body [5, 6]. In order to go into more detail about RCC, each of these letters can be divided into different groups, as summarized in Table 1.

Table 1. RCC classification according to the TNM staging system [5].

TNM	Meaning	Types
T1	Tumour is only in the kidney with 7 cm or less	T1a → greatest dimension <4 cm T1b → greatest dimension 4–7 cm
T2	Tumour is only in the kidney with more than 7 cm	T2a → greatest dimension 7–10 cm T2b → greatest dimension >10 cm
T3	Tumour has spread into major veins or perinephric tissues	T3a → has grown into the renal vein or its branches, or has invaded perirenal and/or renal sinus fat but not beyond Gerota fascia T3b → has grown into inferior vena cava T3c → has grown into vena cava above the diaphragm or has invaded the wall of the vena cava
T4	Tumour has proliferated to areas beyond Gerota fascia	
N	Tumour has spread to regional lymph nodes	NX → regional lymph nodes cannot be evaluated N0 → no regional lymph node metastasis N1 → metastasis in regional lymph node(s)
M	Presence or absence of distant metastasis	M0 → no distant metastasis M1 → the cancer has spread to distant organs

Medical knowledge about RCC has made great strides, and the main risk factors for developing the disease are already well-established. Smoking, physical inactivity, diabetes, hypertension, obesity, family history, age, gender, and ethnicity are some of the factors that have an impact on increasing the likelihood of developing RCC [7]. The steady increase in knowledge in this field has led to more sophisticated diagnostic techniques, resulting in a significant improvement in the 5-year overall survival rate, which has risen from 57% to 74% in recent years [3].

The 5-year survival rate is a widely used indicator to measure the severity of cancer and represents the percentage of patients who are still alive after a 5-year period since the diagnosis of the disease. For kidney cancer specifically, the 5-year survival rate is

usually estimated by considering only the locations to which the cancer has spread, as shown in Fig. 1. The information illustrated in Fig. 1 is based on SEER database [8].

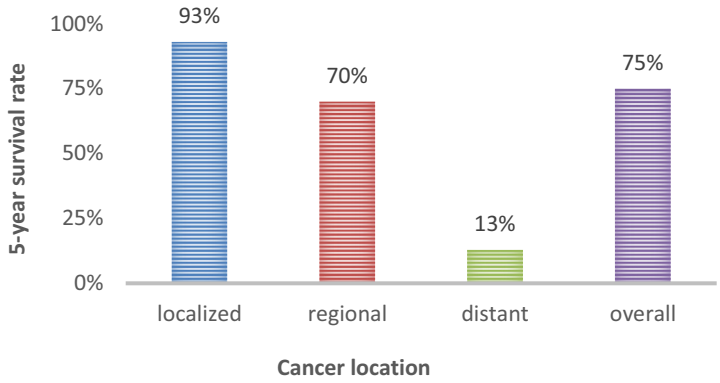

Fig. 1. 5-year survival rate according to the tumor's location.

However, the estimate of the 5-year survival rate could be improved by including more features beyond tumour location, and by considering the interactions that exist between them. In order to achieve a more accurate estimate of 5-year survival rate, we can use Data Mining (DM) techniques combining multiple features. The two main advantages of these techniques are the possibility of developing complex predictive models that are close to reality and the fact that they do not require long studies that consume many resources. DM techniques can be easily adapted to predict the survival estimate for different time periods with small adjustments to the models in a relatively simple way.

In this context, the present work aims to compare the predictive capacity of different DM algorithms to determine the best suited models to predict whether a given patient has a life expectancy of more or less than 5 years and, if less, to estimate the expected days of life.

The paper is organized into five main sections. This first section gives a brief overview of kidney cancer, RCC and DM techniques. The second section presents some related work, and the third section describes the materials and methods. Finally, Sect. 4 analyses and discusses the obtained results and Sect. 5 summarizes the main conclusions and some possible future work.

2 Related Work

In recent years, some approaches have been proposed in the literature to investigate the use of DM techniques in kidney diseases. At this level, in 2014, Zeenia Jagga and Dinesh Gupta used J48, Random Forest (RF), Sequential Minimal Optimization (SMO) and Naïve Bayes (NB) techniques to identify whether a kidney cancer is at an "early stage" (stage i or stage II) or at a "late stage" (stage iii or stage iv). The dataset used comprised a total of 62 genes and the results obtained demonstrated that the RF algorithm

presented the best predictive values, with a sensitivity of 89%, an accuracy of 77%, and an Area Under Receivers Operating Curve (AUROC) of 0.8 [9].

In its turn, in 2019, El-Houssainy A. Rady and Ayman S. Anwarb tested the Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Radial Base Function (RBF) algorithms for predicting the stage of chronic kidney disease. For this purpose, they used a dataset of 24 attributes and concluded that the algorithm with the best values was PNN, with an accuracy of 99.7%, a precision of 98.7% and an F-Measure of 99.4% [10].

More recently, in 2020, Aranuwa Felix Ola tested J48, Logistic Model Tree (LMT), M5P, REPTree, Hoeffding Tree (VFDT), Decision Stump (DS) and RF algorithms using Weka. The analysed data contained information about the patient's gender, age, lifestyle, gender and hereditary disorder, chemical and industrial exposure, and patient complaints. In this study, the algorithm with the best predictive ability was J48, with an accuracy of 74.7%, an F-Measure of 61.4%, a precision of 68.7% and a recall of 71.4% [11].

3 Materials and Methods

The dataset used for the present study was obtained from the Genomic Data Commons of the National Institutes of Health (NIH) of the USA, which provides clinical and genomic data from cancer research programs for research purposes [12]. The dataset includes data from The Cancer Genome Atlas (TCGA) projects – TCGA-KIRC, TCGA-KIRP and TCGA-KICH – and relates to patients with RCC. TCGA selects different cancers for study which have poor prognosis capacity and a high public health impact [13].

The data processing and the construction of the predictive models were performed using RapidMiner software, following the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology [14]. This methodology enumerates a series of steps that enable the implementation of data mining projects from situations in real context [15]. In the case of present work, we followed the CRISP-DM methodology to create the DM models for each of the proposed objectives: Determine whether a patient is likely to be alive 5 years after the diagnosis of RCC and predict life expectancy in the cases where the prognosis is worse.

3.1 Data Preprocessing

The original dataset consists of 821 records and 154 personal and clinical attributes, including *vital_status* and *days_to_death* attributes. Based on these two parameters, it is possible to apply different DM algorithms to respond to the outlined objectives.

At this stage, the raw data were treated in order to:

- Check their consistency by analysing redundant attributes – The original dataset contained some redundant attributes, such as age in years and age in days. In these cases, the inconsistent records were removed.
- Replace missing data with existing redundant attributes – In the case of existing redundant fields, the missing values were replaced by the values of the duplicated columns.

- Eliminate irrelevant columns for the DM process; columns without any values filled in; or fields that refer to duplicate/redundant data.

Since the dataset contained redundant pathological and clinical data, the latter columns were eliminated as pathological data are more reliable. In case of inconsistencies, the priority was given to pathological data. If there were missing values in the pathological data, those cells were replaced with the corresponding clinical values.

At the end of these steps, the attributes were reduced to the following 11: *age_at_index, ajcc_pathologic_m, ajcc_pathologic_n, ajcc_pathologic_t, ajcc_pathologic_stage, gender, primary_diagnosis, prior_malignancy, race, days_to_death and vital_status*.

Considering the classification of the TNM staging system, the consistency of the data was checked using the following rules [5]:

- If Stage I then M0, N0, T1
- If Stage II then M0, N0, T2
- If Stage III then (M0, N1, T1 or T2) or (M0, any N, T3)
- If Stage IV then (M0, any N, T4) or (M1, any N, any T)

Moreover, these rules also allowed the derivation of some missing values related to M, N, T, and stage. For example, for the records referring to stage I or II with missing values in *ajcc*_pathologic_*n*, we have replaced those missing values by N0. Afterwards, since *ajcc_pathologic_stage* contained some missing values that could not be inferred by the rules, the corresponding rows were removed. In addition, some records contained the non-specific histological term "renal cell carcinoma, NOS" in the *primary_diagnosis* column [16]. These values were replaced by the most common specific type of RCC ("clear cell adenocarcinoma, NOS").

The basic statistics of the numeric and nominal attributes after the data preprocessing are presented in Table 2 and Table 3, respectively.

Table 2. Summary statistics for numerical attributes

Numeric attribute	Min	Max	Average	Standard deviation	Missing values
age_at_index	17	90	60.341	12.393	0.37%
days_to_death	0	3615	910.129	731.614	74.97%

From the data analysis, it can be observed that the "alive" and "dead" values of the class label *vital_status* were not balanced. In order to obtain more balanced data, we have simulated the application of different methodologies, including SMOTE, but the best option was to perform a simple oversampling of duplication of records with respect to the "dead" value to improve the system's performance [17, 18]. After this procedure, the number of "dead" records was increased to 606.

Table 3. Summary statistics for nominal attributes

Nominal attribute	Values	Number of records	Missing values
ajcc_pathologic_m	M0/M1	712/90	0.12%
ajcc_pathologic_n	N0/N1/N2	655/40/6	12.70%
ajcc_pathologic_stage	Stage I/II Stage III/IV	433/96 179/95	0.00%
ajcc_pathologic_t	T1/T1a/T1b T2/T2a/T2b T3/T3a/T3b/T3c T4	46/229/164 81/18/15 14/158/62/2 14	0.00%
Gender	Male/Female	536/267	0.00%
primary_diagnosis	CCA/PA/RCC	468/262/59	1.74%
prior_malignancy	Yes/No	113/690	0.00%
Race	White/Black nr/asian/native	653/116 18/14/2	0.00%
vital_status	Alive/Dead	601/202	0.00%

3.2 Data Cleansing

After inserting the preprocessed dataset into RapidMiner, we built a block called "Data Cleansing", which consists of several processes that perform the following additional data treatment operations:

- The records of patients who died that were associated with "0" in the *days_to_death* field were removed.
- The missing values in *days_to_death*, *primary_diagnosis* and *age_at_index* attributes were replaced by *zero*, *"Clear cell adenocarcinoma, NOS"* and by the mean age, respectively.
- The records were filtered to include only patients who died within 5 years of the diagnosis date (*days_to_death* $<= 1825$), and patients under 75 years (*age_at_index* $<= 75$).
- Records containing outliers were removed.

Depending on whether the purpose of the prediction is to classify the vital status as "alive" or "dead" or to estimate the number of days until death, the system then performs different operations. In the first case, the *days_to_death* column was removed and in the second case, the nominal attributes were converted to numeric values and the rows related to the vital status "alive" were removed.

3.3 Modeling and Validation

To model the problem, we followed 2 different methodologies, depending on the expected target. Target 1 aims to determine the vital status of a patient 5 years after the diagnosis

of the disease. Target 2 intends to estimate the life expectancy for patients who are not expected to survive longer than 5 years. For each of these objectives, we tested different DM techniques in order to obtain reliable predictive models. The techniques that led to the best results were:

- DM techniques for Target 1 = {Random Forest, Rule Induction, Generalized Linear Model, Logistic Regression, k-Nearest Neighbors}
- DM techniques for Target 2 = {Linear Regression, Generalized Linear Model, Neural Net, Deep Learning, k-Nearest Neighbors}

For the construction of the predictive models, for the training and validation phases, a 10-fold cross-validation approach was used for target 1, and a leave-one-out cross-validation was used for target 2. In this step, the parameters were optimized to obtain the highest values for accuracy, precision, recall and AUROC, in the case of target 1, and the lowest Root Mean Square Error (RMSE) in the case of target 2.

Accuracy represents the percentage of correct predictions, precision refers to the percentage of positive predictions (dead) that were correctly identified, and recall indicates the percentage of positive cases that were correctly identified. On the other hand, the AUROC indicator provides information about the quality and robustness of the model and the RMSE measures the deviations of the prediction errors.

In the specific case of medical predictions, it is more important to minimize false negatives than false positives. False negatives, in this context, correspond to the patients who died but which the algorithm predicts to be "alive" and false positives correspond to the patients who survived but that were predicted to be "dead". A situation in which a healthy person is falsely identified as sick can be corrected by complementary testing. However, if a sick person is falsely diagnosed as healthy, this false prognosis may delay the treatments needed to combat the disease. Thus, for target 1 is important to obtain high recall values. In the case of target 2, the performance of the models was analysed considering the indicator RMSE.

In addition to checking the performance of the models, it is also important to analyse whether they are overfitted or not. In order to test the models, the data were split into two subsets: one with 80% of the data for training and validation, and a second subset with 20% of the data for an independent test of the models.

4 Results and Discussion

To select the algorithm that best fits each target, we compared the values of the main performance indicators obtained during the train and test phases. For the classification of the patients' vital status, these values are summarized in Table 4.

As indicated in Table 4, the models have similar measurements in both training and testing phases. Therefore, it was concluded that the created models were not overfitted. Regarding the ability to predict the patients' vital status, the technique that presented the best results was RF, which achieved high values for accuracy (87.65%), precision (82.36%), recall (92.23%) and AUROC (0.931) and an acceptable runtime. Some of the considered parameters for this algorithm were 40 trees, "gini-index" criterion, and

Table 4. Performance of algorithms in classification process (alive/dead).

Algorithm		Accuracy (%)	Precision (%)	Recall (%)	AUROC	Runtime* (ms)
Random Forest	Train	87.65	82.36	92.23	0.931	345
	Test	86.60	80.61	91.86	0.949	
Rule Induction	Train	80.82	78.18	78.90	0.838	1139
	Test	77.32	75.00	73.26	0.822	
Generalized Linear Model	Train	78.88	79.43	64.39	0.800	198
	Test	76.92	69.81	64.91	0.794	
Logistic Regression	Train	76.70	78.83	65.83	0.821	149
	Test	72.68	73.85	57.14	0.816	
k-Nearest Neighbors	Train	79.29	81.03	70.79	0.888	115
	Test	74.74	73.42	67.44	0.842	

* Total runtime for training and testing phases calculated with Jackhammer extension.

a "maximal depth" of 15. The obtained performance metrics enable to consider that RF is suitable for determining life expectancy at 5 years and can be used as a complement to medical diagnosis.

For the analysis of the life expectancy estimation, the values obtained in terms of RMSE are illustrated in Fig. 2.

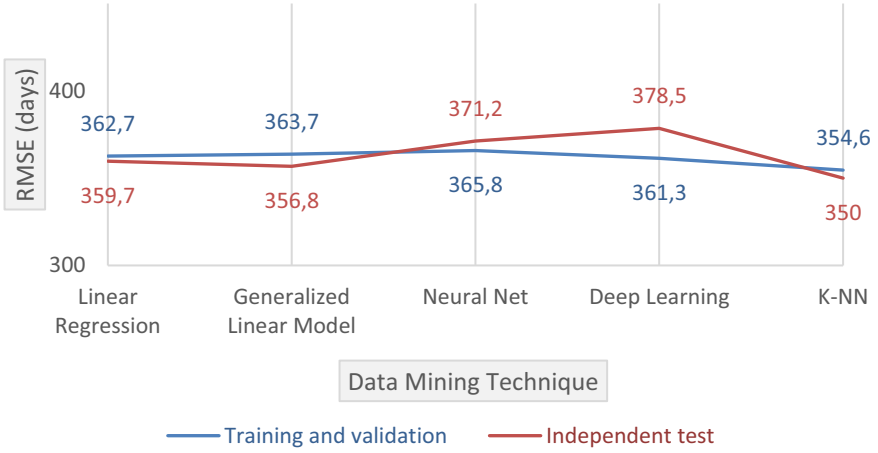

Fig. 2. Comparison of the RMSE between the training and testing phases for each DM technique.

Despite the high margin of error found for each algorithm (~1 year), it is possible to get a general idea of the number of years of life that are expected. From this point of view, the best performance values were obtained using the k-Nearest Neighbors (k-NN) technique, which presented a RMSE of 354.6 days, and the Deep Learning (DL) technique, which presented a RMSE of 361.3 days. In turn, the Linear Regression (LR) technique presented a RMSE of 362.7 days. This technique has the advantage of providing an easy-to-understand and easy-to-use prediction mechanism that can be used for an initial assessment of the severity of the disease. The equation obtained by the LR algorithm was:

$$\begin{aligned} days\_to\_death = {} & 86.317 \times ajcc\_patho \log ic\_n \\ & - 129.988 \times ajcc\_patho \log ic\_stage + 775.348 \end{aligned}$$

This equation was obtained considering only the *ajcc_pathologic_n* and *ajcc_pathologic_stage* attributes, since they were the only ones that had a p-value of less than 0.05, which makes them the most important parameters for predicting life expectancy. The p-values obtained were 0.079 for *age_at_index*, 0.048 for *ajcc_pathologic_n*, 0.001 for *ajcc_pathologic_stage*, 0.124 for *ajcc_pathologic_t*, 0.512 for *gender*, 0.991 for *primary_diagnosis*, and 0.619 for *prior_malignancy*.

In terms of execution times, the k-NN and LR algorithms were the fastest with 19 and 63 ms, respectively. On the other hand, DL and NN were the slowest algorithms with 3027 and 844 ms, respectively.

5 Conclusions and Future Work

This paper aimed to identify the most suitable DM techniques and their parameters for predictions related to RCC. The DM techniques used allow the construction of complex prediction models that take into account the influence of multiple attributes simultaneously. The constructed models allow predicting whether a given patient has a life expectancy of at least 5 years and, in the worst scenarios, they allow calculating the patients' life expectancy. These models can be used as a valuable tool to complement medical diagnosis.

By simulating different techniques and optimizing their parameters, we have verified that the RF algorithm has a high efficiency to characterize the vital status of RCC patients at five years. On the other hand, the LR algorithm provides a simple and easy to understand mechanism to calculate the life expectancy of RCC patients, with a margin of error of about 1 year.

Although the present study is based on real data, it would be interesting for future work to replicate the experience with a larger dataset, also including more columns corresponding to other important risk factors, such as genes, workplaces exposures, or body mass indexes. A larger dataset could even contribute to improve the predictive models.

Acknowledgements. This work is funded by "FCT—Fundação para a Ciência e Tecnologia" within the R&D Units Project Scope: UIDB/00319/2020.

References

1. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. Cancer J. Clin. **71**, 209–249 (2021). https://doi.org/10.3322/caac.21660
2. Hsieh, J.J., et al.: Renal cell carcinoma. Nat. Rev. Dis. Prim. **3**, 1–19 (2017). https://doi.org/10.1038/nrdp.2017.9
3. Choueiri, T.K., Motzer, R.J.: Systemic therapy for metastatic renal-cell carcinoma. N. Engl. J. Med. **376**, 354–366 (2017)
4. Dizman, N., Philip, E.J., Pal, S.K.: Genomic profiling in renal cell carcinoma. Nat. Rev. Nephrol. **16**, 435–451 (2020). https://doi.org/10.1038/s41581-020-0301-x
5. Brierley, J.D., Gospodarowicz, M.K., Wittekind, C. (eds.): TNM Classification of Malignant Tumours. Wiley Blackwell (2017)
6. National Cancer Institute: Cancer Staging. https://www.cancer.gov/about-cancer/diagnosis-staging/staging. Accessed 08 June 2021
7. Scelo, G., Larose, T.L.: Epidemiology and risk factors for kidney cancer. J. Clin. Oncol. **36**, 3574–3581 (2018). https://doi.org/10.1200/JCO.2018.79.1905
8. American Cancer Society: Survival Rates for Kidney Cancer. https://www.cancer.org/cancer/kidney-cancer/detection-diagnosis-staging/survival-rates.html. Accessed 08 June 2021
9. Jagga, Z., Gupta, D.: Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. BMC Proc. **8**, 1–7 (2014). https://doi.org/10.1186/1753-6561-8-S6-S2
10. Rady, E.-H.A., Anwar, A.S.: Prediction of kidney disease stages using data mining algorithms. Inf. Med. Unlocked. **15**, 100178 (2019). https://doi.org/10.1016/j.imu.2019.100178
11. Ola, A.F.: A model for prediction of kidney cancer using data analytics technique. Am. J. Data Min. Knowl. Discov. **5**, 27–36 (2020). https://doi.org/10.11648/j.ajdmkd.20200502.12
12. Grossman, R.L., et al.: Toward a shared vision for cancer genomic data. N. Engl. J. Med. **375**, 1109–1112 (2016). https://doi.org/10.1056/nejmp1607591
13. National Cancer Institute: TCGA Cancers Selected for Study. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers. Accessed 17 June 2021
14. RapidMiner. https://rapidminer.com/. Accessed 07 May 2021
15. Morais, A., Peixoto, H., Coimbra, C., Abelha, A., Machado, J.: Predicting the need of neonatal resuscitation using data mining. In: Procedia Computer Science, pp. 571–576. Elsevier B.V. (2017). https://doi.org/10.1016/j.procs.2017.08.287
16. Dickie, L., Johnson, C., Adams, S., Negoita, S.: Solid Tumor Rules. National Cancer Institute, Rockville, MD (2020)
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
18. Peixoto, C., Peixoto, H., Machado, J., Abelha, A., Santos, M.F.: Iron value classification in patients undergoing continuous ambulatory peritoneal dialysis using data mining. In: Proceedings of the 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE), pp. 285–290. SCITEPRESS (2018). https://doi.org/10.5220/0006820802850290