ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-1769

Learning the Repair Urgency for a Decision Support System for Tunnel Maintenance

Y. Gatsoulis[1] and **M. O. Mehmood**[1] and **V. G. Dimitrova**[1] and **D. R. Magee**[1] and **B. Sage-Vallier**[2] and **P. Thiaudiere**[2] and **J. Valdes**[2] and **A. G. Cohn**[1]

Abstract. The transport network in many countries relies on extended portions which run underground in tunnels. As tunnels age, repairs are required to prevent dangerous collapses. However repairs are expensive and will affect the operational efficiency of the tunnel. We present a decision support system (DSS) based on supervised machine learning methods that learns to predict the risk factor and the resulting repair urgency in the tunnel maintenance planning of a European national rail operator. The data on which the prototype has been built consists of 47 tunnels of varying lengths. For each tunnel, periodic survey inspection data is available for multiple years, as well as other data such as the method of construction of the tunnel. Expert annotations are also available for each 10m tunnel segment for each survey as to the degree of repair urgency which are used for both training and model evaluation. We show that good predictive power can be obtained and discuss the relative merits of a number of learning methods.

1 INTRODUCTION

Complex decision making in domains with high impact, such as infrastructure management, is a challenging task that requires the consideration of a large number of parameters and their dependencies. For example, in the case of tunnel management, accurate pathology diagnosis and early risk assessment are critical for making cost effective maintenance plans. The common practice is that such decisions are made by a small number of domain experts, who follow their intuitions and apply tacit knowledge gained over many years of experience. This results in unsustainable subjective decision models, where it is common for experts to have no clear consensus on how the factors influence the outcomes and knowledge can be lost when experts leave.

We tackle these issues in the context of tunnel management within the EU project NETTUN[3]. Initial work [13] was focussed on following ontology engineering methodologies, where we engaged domain experts with extensive experience in tunnel management in a knowledge elicitation process to identify the concepts they consider and the rules they apply when diagnosing the pathologies of a tunnel based on its characteristics and inspection data.

Diagnosis of the pathologies is not sufficient, the experts are also required to assess the urgency of repairs that sections of tunnels require: the available maintenance budget must be prioritised, in particular taking account considerations relating to public safety. Compared to diagnosing pathologies the decision processes for prioritisation are more complex and difficult for the experts to articulate, and the ontology based methods for pathology diagnosis are not so appropriate since deciding priorities are not so much conceptual distinctions but a process of ranking. Furthermore, ontological models can have some limitations. Firstly, they may not be able to capture the true complexity of the decision process. Secondly, the process of validating these models is important, but laborious and slow. It is hard to identify missing or inaccurate rules, and some rules are "more reliable" than others, but experts typically cannot articulate this information. Furthermore, there are aspects of the decision process, such as risk assessment and potential for further degradation of tunnel portions, which take into account a number of parameters which experts find hard to specify declaratively.

To address these challenges, we adopt supervised machine learning models, taking advantage of the existence of provenance data with past observations and expert decisions. To the best of our knowledge this is the first time that machine learning methods are used in this domain for this task. A similar domain where such issues have been investigated is that of diagnosing the condition of water pipes, where most recent work focusses on Bayesian approaches [15, 8]. In our study we preferred to investigate other state of the art machine learning models (Section 3) that require little input from the experts since this is problematic as noted above.The methods we employ can also learn from the data with minimal pre-processing. Another closely related case study employed a Gaussian process model to classify surrounding rocks in tunnels, as this knowledge is important for their design and construction [16]. Although Gaussian processes are also able to provide probabilistic estimates, their performance tends to degrade in high dimensional problems when the number of features is a few dozens or more, such as the case here.

There appears to be very little work on DSS for tunnel maintenance. We have already mentioned [13] above which is concerned with pathology diagnosis, which is also the topic of [11] which also focusses on the uses of sensors to obtain a score per segment (somewhat similarly to the *cotation* score described below). There are a number of DSS to support other aspects of tunnel management, in particular, construction (e.g. [9]). There are also a variety of investigations into DSS for other kinds of transport infrastructure, e.g. highway maintenance [10], bridges [5], pavement maintenance [4], overpasses [17].

The rest of the paper is organised as follows. Section 2 describes the tunnel data and the pre-processing steps undertaken to make them suitable for the machine learning methods used in this project, which

[1] School of Computing, University of Leeds, UK, emails: {y.gatsoulis, o.m.mehmood, v.g.dimitrova, d.r.magee, a.g.cohn}@leeds.ac.uk
[2] Société Nationale des Chemins de Fers Français (SNCF), France, emails: {bastien.sage-vallier, patrick.thiaudiere, joaquin.valdes}@reseau.sncf.fr
[3] EU FP7 "New Technologies for Tunnelling and Underground Works", http://www.nettun.org

are described in Section 3. Section 4 evaluates the chosen methods, while Section 5 discusses further work and concludes the paper.

2 DATA

We have data for 47 tunnels from a national rail company. For each tunnel, survey data was collected during periodical inspections (typically every four years), and in the dataset there are multiple inspections for most of the tunnels (between one and four for each tunnel) resulting in a total of 137 inspection surveys in the dataset. For the purposes of recording surveys, each tunnel is decomposed into 10m segments. This results in a total of 8283 segments or data points. The tunnels vary in length and the average number of segments is $\mu_l = 62$ with standard deviation of $\sigma_l = 70$. There are also characteristics describing the properties of the tunnel, which can be regarded as static (i.e. set at the time of tunnel construction, and not varying over time subsequently). These variables along with the urgency repair scale are explained in more detail in the following section.

2.1 Description

The static characteristics data of each tunnel segment are called *influencing factors*, of which there are 32. Examples of influencing factors are the climate of the area, the ground type the tunnel was built on, the lining type (see Figure 1a), etc. The influencing factors are nominal data and the possible values for each of them are different. For example, the climate influencing factor can be 'favourable conditions' or 'medium conditions' or 'hard conditions'; the ground type influencing factor can be one of the following values, 'altered rock', 'compact rock', 'soil' or 'mix ground'.

The tunnels are periodically inspected for potential problems, which are called *disorders*. In this dataset 24 disorders are present. Examples of disorders are moisture (Figure 1b), displacement of a lining element (Figure 1c), etc. The disorders are binary variables, representing their presence or absence, and unlike influence factors, they can change over time.

The experts in rail company have developed a model that aggregates the observed disorders in each 10m segment into a single numeric value, called the *cotation* value (0-100) which provides a summary of the degree of disorders in that segment (see Figure 2). The higher the value the worse the condition of the tunnel is. But it does not take account of influencing factors; so a lower cotation value may be more urgent to repair if its influencing factors are particularly egregious.

Lastly, we collected the experts' recommendations about the urgency of repairs for these 47 tunnels, and these values are treated as the ground truth, and as the target variable that a model learns. This was given in a scale from 1 to 5, which denote the following recommendations: 1: "the segment is good and no repairs are needed", 2: "pay attention to this segment but no repairs need to be planned at the moment", 3: "repairs need to be planned within u_3 time", 4: "repairs need to be planned within u_2 time", and, 5: "repairs need to be planned within u_1 time"; with $u_1 < u_2 < u_3$, i.e. 5 denotes that a tunnel segment requires the most urgent repairs (see example annotation in Figure 2).

2.2 Preparation

The dataset is biased towards the first two scales denoting 'no repair': category 1 accounts for 84% of the data, while category 2 accounts for a further 5% of it. Combined, they yield a ratio of 8:1 of 'no repair' versus 'repair'. Moreover, 11 of the tunnels and 44 of the periodic inspections are classified along their full length with the category good (1), hence, offering little further information about this category that is not covered from the rest of the data. These 11 "good tunnels" are filtered out before further processing, resulting in a remaining total of 37 tunnels, 93 inspections and 6211 segments/data points with a bias ratio between the $\{1, 2\}$ versus $\{3, 4, 5\}$ categories, i.e. 'no repair' versus 'repair', of approximately 6:1 (85% of the data), which is a small improvement from before.

As the most critical decision for the tunnel owner is whether a tunnel segment requires repairs or not we have collapsed these five categories into two by merging 1-2 together representing 'no repair', and 3-5 into 'repair'. Essentially, this makes the independent variable a binary one denoting whether a tunnel segment requires repairs or not. A further stage of model building can be used to distinguish between the urgency of repair in the former case (3,4,5).

In summary, the dataset used for model building consists of 37 tunnels comprising 93 period inspections and 6211 segments. The target variable is binary: 'repair' (3,4,5) vs 'no repair' (1,2) with a 1:6 ratio. Each instance is represented by an attribute vector consisting of 24 disorders, 32 influencing factors, and a cotation value.

3 METHODS

We conducted an initial investigation to verify the non-linearity nature of the problem using a logistic regression model for different sensitivity threshold values. The results of Figure 3 confirmed this hypothesis, as it can be seen that it performs inadequately regardless of the value of sensitivity. Factor analysis, dimensionality reduction and appropriate transformation could possibly help to improve its performance, but on the other hand there are other state of the art methods that are better suited in such cases and able to learn the non-linear relations of such complex data.

As a result, we decided to investigate the effectiveness of three popular state of the art models of machine learning: decision trees, random forests and support vector machines. The reasons for choosing these and a brief explanation of the models is given in this section.

3.1 Decision Trees

One of the desired requirements was for the tunnel diagnosis experts to be able to understand the reasons for the classification produced by the machine learning method; the ultimate decision as to the urgency of repair remains with a human, and the DSS is aimed to support their decision and recommendation. For this reason one of the models we investigated was a decision tree, as it offers excellent and fast explanations on the underlying reasoning while still performing sufficiently in most cases: the expert is able to inspect the decision tree and see why a particular categorisation has been made.

Decision trees are non-parametric machine learning methods that partition the state space using decision rules, so that training instances x_i of the same category y_i are grouped together. They are most commonly represented as a decision tree structure. The key question during training is which dependent variable to use for the split. In general, this is achieved by computing the impurity (H) of the dataset before and after the split for a given variable. So at each node m denote its data as Q. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets:

$$Q_{left}(\theta) = \{(x, y)|x_j \leq t_m\} \quad (1)$$

$$Q_{right}(\theta) = \{(x, y)|x_j > t_m\} = Q \setminus Q_{left}(\theta) \quad (2)$$

(a) Example of shotcrete lining type influence factor

(b) Example of moisture disorder

(c) Example of displacement of lining element disorder

Figure 1: Pictorial examples of particular values of some influencing factors and disorders of the tunnels

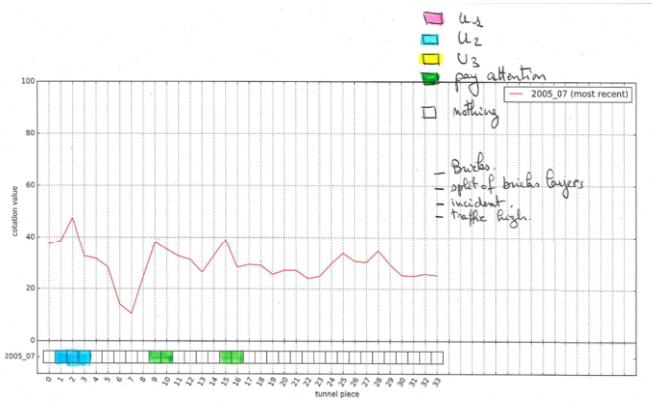

Figure 2: A graph of cotation values for an example tunnel. Also shown are the expert annotations for the ground truth for the urgency of repair: here "nothing" corresponds to repair urgency 1; "pay attention" to 2; U3 to 3; U2 to 4; and U1 to 5.

The impurity at m is computed using an impurity function $H()$:

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \quad (3)$$

and in this paper we used the Gini impurity:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \quad (4)$$

The final step is to select the variable for this node and the parameters that minimize the impurity, i.e. in this case the variable with the highest Gini gain:

$$\theta^* = \text{argmin}_\theta \, G(Q,\theta) \quad (5)$$

3.2 Random Forests

Another machine learning model that we investigated was random forests [1, 2], because they typically offer good classification performance at the trade-off of being harder to explain the underlying reasoning processes.

Random forests fall in the category of ensemble methods, which build estimators based on multiple weak classifiers, in this case short length decision trees, and specifically in this study these were CART models. Each tree is build on a random sample of the training dataset, and this process is known as bootstrap aggregating or bagging for short. Furthermore, each tree selects randomly a subset of the features which is used to train the decision tree on the random training subsample. The advantage of data and feature bagging is that the resulted meta-classifier has reduced variance and overfitting.

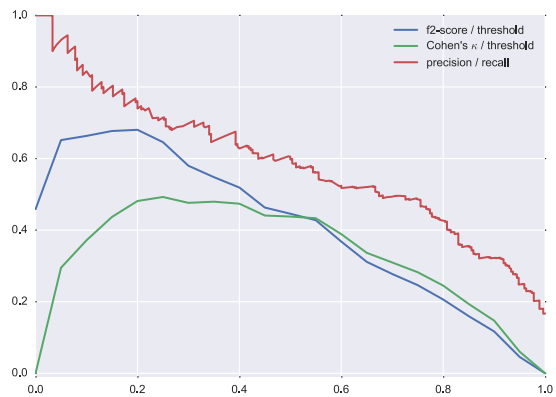

Figure 3: Logit model fitted on the dataset (best visualized in colour)

3.3 Support Vector Machines

Lastly we chose to investigate support vector machines since they are typically effective in high dimensional spaces, and from our past experience we have achieved better than state of the art results in similar problems [12].

A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space, which can be used for classification. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. The separating lines can be linear functions as well as non-linear ones. In this investigation a radial-basis function is used as the kernel for the separation.

4 RESULTS & DISCUSSION

Using the formatted data described in Section 2 we performed a cross-validation study with the methods presented in Section 3. The details of the experiments and the performance results are presented and discussed in this section.

4.1 Performance metrics

The performance metrics report are precision (p), recall (r) and f2-score ($f2$) for class 'repair', as well as Cohen's kappa coefficient (κ)

which is a more robust measurement of accuracy, particularly when there is class imbalance. When averages are reported these are the micro-averages, unless stated otherwise. We chose $f2$ rather than $f1$ since we wish to emphasise recall over precision – a false negative is potentially much more serious than a false positive.

4.2 Cross-validation

In order to test the performance of the models we conducted a leave-one-tunnel-out cross-validation (the whole tunnel, with all of its surveys). This type of cross-validation is more suitable in this case, as the typical procedure of performing k-fold cross-validation, i.e. by randomly populating the folds from the data seems to overestimate the performance of the classifiers, as shown by the results in Table 1 and Table 3 (presented in Section 4). Since adjacent tunnel segments are not truly independent instances (there are likely to be similar disorders in adjacent segments), by having the possibility one segment in the training data and its neighbour in the test data does not represent truly random sampling method. This is analogous to research in activity recognition from video data in which a preferred methodology is to leave-one-person-out [12].

Table 1: Micro performance metric averages showing that typical cross-validation (results shown for 5-folds) overestimates the true performance of the classifiers in comparison to the results shown in Table 3 where a leave-one-tunnel-out cross-validation methodology was used.

	precision (p)	recall (r)	f2-score ($f2$)	kappa (κ)
DT	.86	.87	.89	.87
RF	.91	.89	.89	.89
SVM	.92	.85	.86	.88

Leaving one tunnel out is more representative of the use case when a new tunnel comes in.

The tunnels vary in number of segments with the smallest having 3 segments (30m) and the longest 377 segments (3.7Km), with an average number of segments of $\mu_l = 62$, and standard deviation $\sigma_l = 70$. Very small tunnels, like the ones with 3 segments, can potentially bias the results in a positive or negative way. For this reason we decided to define a minimum number of segments that a tunnel must have in order to be considered as a cross-validation test case. The question then is whether choosing a minimum length of the tunnels to be cross-validated can also have any biased effect on the performance metrics of the classifiers. As such, we conducted an analysis by cross-validating for various lengths of the tunnels for a range of 10 minimum segments up to 100 minimum segments with a step of 5. Figure 4 shows the results a random forest classifier using 500 CART estimators.

It can be seen from the plot that choosing any particular minimum length does not influence significantly the results in some biased manner. Similar graphs were obtained for the rest of the models. As such, it is safe to choose a minimum length of 10 segments to have the most exhaustive cross-validation analysis. Only one tunnel had a length of less than 10 segments, resulting in a 36-fold cross validation (the 37th tunnel was still included in the training data in each fold).

4.3 Class imbalance

As described earlier the dataset even after the initial filtering still remains biased towards the 'no repair' class with a ratio of about

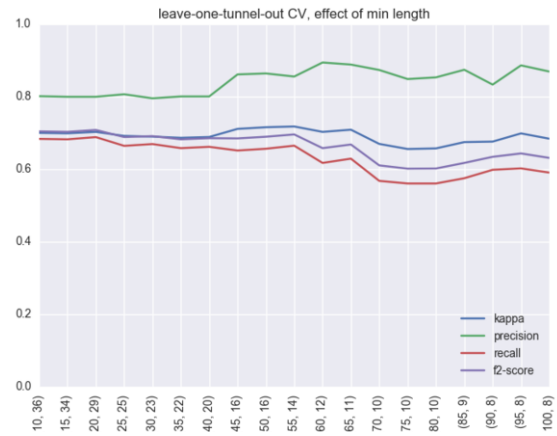

Figure 4: Investigation of the effect of number of folds in leave-one-tunnel-out cross-validation using a random forest classifier, the x-axis has i, j values where i is the minimum number of segments for a tunnel to be considered in the fold, and j is the number of tunnels with more segments than the minimum length (best visualized in colour)

6:1. Learning from imbalanced data is a common issue and a number of solutions have been proposed [7], mainly based on under-sampling the most popular class(es) or over-sampling the smaller one(s). We initially performed a k-fold cross-validation with a number of under/over-sampling methods and found that the most balanced results in terms of precision and recall were achieved with Tomek links under-sampling [14] and synthetic minority over-sampling technique (SMOTE [3]), which are both well-established and widely used methods that deal with the issue of class imbalance. We used both of these in the performance analysis of the classifiers in full leave-one-tunnel-out cross-validation, as well as tested the models with no sampling leaving the original training data unaltered.

4.4 Baseline

Table 2 presents the results from a weighted random guess according to the classes ratio in the training set on a leave-one-tunnel-out cross-validation.

Table 2: Class-portion-weighted random guess baselines

	precision (p)	recall (r)	f2-score ($f2$)	kappa (κ)
No sampling	.14	.14	.14	.00
Tomek links	.13	.13	.13	.00
SMOTE	.87	.53	.33	.00

4.5 Classifiers results

As described in Section 3 the classifier models we investigate are a CART tree, a random forest (RF) and a support vector machine (SVM). Table 3 shows the performance of the three models using no sampling, Tomek links under-sampling and SMOTE over-sampling (Section 4.3) in a leave-one-tunnel-out cross-validation (Section 4.2).

As expected, decision trees tend to perform overall worse than the other two methods in all sampling cases. An interesting result is that with Tomek links under-sampling they have a higher recall value ($r = 0.73$) than the others, i.e. they are able to recognise more of the 'repair' class than the other two methods, but their precision is much

Table 3: Performance metrics of the investigated models in a leave-one-tunnel-out cross-validation.

	precision (p)	recall (r)	f2-score ($f2$)	kappa (κ)
no sampling				
DT	.61	.65	.64	.57
RF	.79	.68	.70	.70
SVM	.82	.69	.71	.71
Tomek links under-sampling				
DT	.58	.73	.70	.59
RF	.80	.68	.70	.70
SVM	.81	.69	.71	.71
SMOTE over-sampling				
DT	.60	.68	.66	.57
RF	.76	.70	.71	.69
SVM	.59	.82	.76	.62

lower ($p = 0.58$), which means that they might become "annoying" to the experts if many segments that require no repairs are highlighted as needing so. It seems that under-sampling of the data allows the decision tree to better generalize by over-fitting even less on the 'no repair' class, since many of its training instances, which might be carrying "noisy" information has been removed.

A similar outcome appears with the SVM when SMOTE over-sampling is used. Its recall value is the highest among all models and all sampling methods ($r = 0.82$), given that SMOTE over-sampling possibly has amplified the 'repair' class, resulting in more and stronger support vectors. However, partly due to the trade-off between precision and recall, its precision value is one of the lowest ($p = 0.59$).

The best balance between precision and recall is given by RF and SVM under Tomek links under-sampling with similar performance metrics. RF also performed similarly with SMOTE over-sampling given its tolerance to over-fitting, however for a marginally worse recall, but about 4-5% better precision, the RF and SVM models with Tomek links seem better suited.

Lastly, all models perform significantly better than the baseline. However, the best results using leave-one-tunnel-out cross-validation were in the range of 0.7-0.8, which are worse than when using standard cross-validation, which were around 0.9. Given the number of variables and the number of tunnels, it is likely that more data and further investigation and discussion with the experts is needed.

4.6 Finding the most important factors

In the results presented in the previous section the full sets of disorders and influencing factors were used, which together with the cotation value account for 57 independent variables in total. We investigated whether all factors are equally important and if the models, despite their tolerance and the measurements we had taken, have still over-fitted. Figure 5 shows the cross-tabulations of the disorders from their contingency tables[4].

The graph shows that some disorders are more discriminative than others for separating the two classes. For example, it seems that disorders 7, 8, 17, 19, etc. occur more frequently when a tunnel needs repairs, while for example 1, 3, 6, etc. offer very little variation between the two classes. Due to the large number of possible values that the influencing factors can take, a similar cross-tabulations plot for them is not visually informative.

Ideally, it would be beneficial to perform some form of dimensionality reduction by factor analysis, such as principal component analysis or linear discriminant analysis. Factor analysis methods work

[4] Due to intellectual property reasons we do not display the names of all of the disorders; instead, they are numerically denoted.

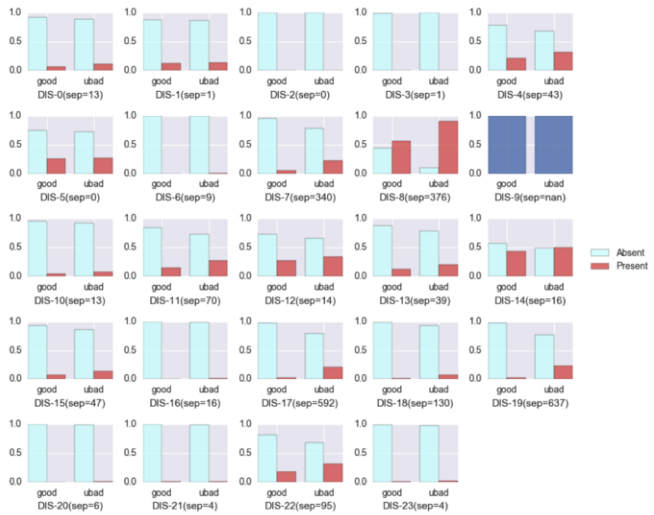

Figure 5: Cross tabulations of the disorders

on the assumption of continuous variables and thus not apply due to presence of categorical variables in our data. Instead, the models of the decision tree and of the random forest are able to provide an estimate on the importance of the features. For decision trees the importance of a feature is the Gini importance which is computed as the (normalized) total reduction of the criterion brought by that feature [2]; while for the random forest model the importance of the features is given by averaging the feature importance of each tree. The importance of each feature is a numerical value between 0-1, with higher values signifying more importance. It can be thought as the amount of variability that a feature explains in the dataset.

Table 4: Five most important disorders and influencing factors according to the decision tree and random forest models

disorder	importance	influence factor	importance
rock faults	.07	climate	.04
rock deterioration	.05	lining type	.04
falling/missing of lining element	.05	discontinuities density	.03
leakage	.05	drainage system	.03
moisture	.05	water flow/load	.03

The variable with the highest importance value is that of cotation, with an average importance between the decision tree and the random forest of 0.66. This is expected as the experts use soft thresholds that directly give a classification, in many cases regardless of the other factors when this value is high enough. To better understand the importance of the disorders and the influence factors, we excluded the cotation value and tested the models with only the disorders and the influence factors.

Table 4 shows the five most important factors for each of the disorders and influencing factors. The 15 most important disorders and 8 most important influencing factors are able to explain 80% of the variability of the dataset. Further, Table 5 shows the association between influencing factors and disorders by listing the five most correlated ones according to Cramer's V [6] and with $\phi_c > 0.37$. Some of these relations are explained by the tunnel experts. For example, the presence of a waterproofing system, which protects from extrados hydraulic pressure, can influence diagonal cracks occurring in situations like landslides, lateral thrust, or differential settlement. For a tunnel with an unlined longitudinal profile, the joint rock patterns

and cracks may create polyhedron which are potentially unstable – an example of relationship between tunnel shape and rock elements. Further discussions with the experts has also revealed that our system is identifying patterns which they would over-fit to a set of tunnels, yet this would not benefit a holistic approach. For example, hydraulic overpressure may not be a common occurrence but must be factored in, if it occurs.

Table 5: Five most correlated disorders and influencing factors according to Cramer's V co-efficient

disorder	influencing factor	Cramer's V (ϕ_c)
diagonal cracks	waterproofing system	0.53
missing rock elements	tunnel shape	0.53
diagonal cracks	strain anisotropy	0.52
diagonal cracks	tunnel age	0.47
rock deterioration	tunnel shape	0.43

4.7 Classifier results when using the most important factors

Following the analysis from the previous section for the most important factors, we used the set of the 23 disorders and influence factors together with the cotation value to retest the models. The results are shown in Table 6.

Table 6: Performance metrics of the investigated models using the most discriminant disorders and influence factors in a leave-one-tunnel-out cross-validation.

	precision (p)	recall (r)	f2-score ($f2$)	kappa (κ)
no sampling				
DT	.62	.71	.69	.60
RF	.80	.69	.71	.71
SVM	.82	.68	.70	.70
Tomek links under-sampling				
DT	.61	.72	.69	.60
RF	.79	.70	.72	.71
SVM	.81	.69	.71	.71
SMOTE over-sampling				
DT	.52	.68	.65	.52
RF	.78	.70	.72	.70
SVM	.58	.81	.75	.61

It can be seen that the results are fairly similar to the ones before when using the complete set of variables. This means that 24 factors out of the total 57 are sufficient for classifying the data equally well, i.e. a dimensionality reduction of nearly 50%.

5 CONCLUSIONS

In this paper we have presented the results of a decision support system based on state of the art machine learning methods for the domain of tunnel maintenance by a European national rail operator. This is a critical application domain, as many businesses rely on reliable and safe transport infrastructure, while the high cost of the repairs (and disruption to journeys during their implementation) dictate careful financial and operation planning.

A specialised cross-validation procedure was employed to avoid misleading overestimated results compared to standard methods. The performance metrics have demonstrated a good level of effectiveness of the algorithms. To deal with the class imbalance, since as expected health portions are many more than unhealthy ones, we utilized popular and well-tested methods of under- and over-sampling. However, the results showed that there were no significant differences, which implied that the class bias in this case is not detrimental to the effectiveness of the algorithms. Also, it was shown that similar results can be obtained with half the features from the original set, which are able to explain the majority variability of the dataset and has the potential to reduce the possibility of overfitting to the training set.

Current future work is focusing on integration of the system to the end-users' site, as well as further discussions with the experts and extensive testing with further data. We also plan to hierarchically refine the 1-5 classification (in particular to split the "repair" case (3-5) into those which are most urgent and those which are not).

Acknowledgements

This work is part of the NeTTUN project (http://nettun.org), funded by the EC 7th Framework under Grant Agreement 280712.

REFERENCES

[1] L. Breiman, 'Random forests', *Machine Learning*, **45**(1), 5–32, (2001).
[2] L. Breiman and A. Cutler, *Random Forests*, Online (http://www.stat.berkeley.edu/~breiman/RandomForests), 2004.
[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, **16**, 321–357, (2002).
[4] Jui-Sheng Chou, 'Web-based CBR system applied to early cost budgeting for pavement maintenance project', *Expert Systems with Applications*, **36**(2, Part 2), 2947–2960, (March 2009).
[5] Michael P. Enright and Dan M. Frangopol, 'Maintenance planning for deteriorating concrete bridges', *Journal of Structural Engineering*, **125**(12), 1407–1414, (1999).
[6] Cramr Harald, *Mathematical Methods of Statistics*, p282, Princeton University Press, 1946.
[7] H. He and E.A. Garcia, 'Learning from Imbalanced Data', *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263–1284, (2009).
[8] Z. Li, B. Zhang, Y. Wang, F. Chen, R. Taib, V. Whiffin, and Yi Wang, 'Water pipe condition assessment: a hierarchical beta process approach for sparse incident data', *Machine Learning*, **95**(1), 11–26, (jun 2014).
[9] V. Likhitruangslip and P. G. Ioannou, 'RISK-SENSITIVE DECISION SUPPORT SYSTEM FOR TUNNEL CONSTRUCTION', in *ASCE Geotechnical Special Publication*, (2004).
[10] Bhoj Raj Pantha, Ryuichi Yatabe, and Netra Prakash Bhandary, 'Gis-based highway maintenance prioritization model: an integrated approach for highway maintenance in nepal mountains', *Journal of Transport Geography*, **18**(3), 426 – 433, (2010). Tourism and climate change.
[11] Noureddine Rhayma, Aurlie Talon, Pierre Breul, and Patrick Goirand, 'Mechanical investigation of tunnels: risk analysis and notation system', *Structure and Infrastructure Engineering*, **12**(3), 381–393, (2016).
[12] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. G. Cohn, and D. C. Hogg, 'Qualitative and Quantitative Spatio-temporal Relations in Daily Living Activity Recognition', in *Proc. of 12th Asian Conference on Computer Vision*, pp. 115–130, Singapore, (2014).
[13] D. Thakker, V. Dimitrova, A.G. Cohn, and J. Valdes, 'PADTUN - Using Semantic Technologies in Tunnel Diagnosis and Maintenance Domain', in *ESCW2015*, (2015).
[14] I. Tomek, 'Two Modifications of CNN', *IEEE Transactions on Systems Man and Communications*, **6**, 769–772, (1976).
[15] C. Wang, Z. Niu, H. Jia, and H. Zhang, 'An assessment model of water pipe condition using Bayesian inference', *Journal of Zhejiang University SCIENCE A*, **11**(7), 495–504, (July 2010).
[16] Yan Zhang, Guoshao Su, and Liubin Yan, 'Classification of surrounding rocks in tunnel based on Gaussian process machine learning', in *2011 International Conference on Electric Technology and Civil Engineering (ICETCE)*, pp. 3971–3974. IEEE, (April 2011).
[17] Jana elih, Anej Kne, Aleksander Srdi, and Marjan ura, 'Multiplecriteria decision support system in highway infrastructure management', *Transport*, **23**(4), 299–305, (2008).