Using Expert Knowledge to Generate Data for Broadband Line Prognostics Under Limited Failure Data Availability ⋆

Gishan Don Ranasinghe * **David Yearling** ** **Mark Girolami** *** **Ajith Kumar Parlikad** *

* *University of Cambridge, Department of Engineering, Institute for Manufacturing, 17 Charles Babbage Road, Cambridge, CB3 0FS, UK (e-mail: gd416, aknp2@ eng.cam.ac.uk)*
** *BT Applied Research, Adastral Park, Barrack Square, Martlesham, Ipswich, IP5 3RF, UK (e-mail: david.yearling@ bt.com)*
*** *University of Cambridge, Department of Engineering, Civil Engineering Building, 7A JJ Thomson Avenue, Cambridge, CB3 0FA, UK and The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK (e-mail: mag92@ eng.cam.ac.uk)*

Abstract: Due to exposure to the driving rain, water ingress can cause faults in electrical joints, junctions and distribution points in broadband lines. Over time, faulting behaviour may grow in magnitude eroding the electrical capability of these lines causing degradation of broadband service. Developing effective data-driven models for broadband line prognostics remains a challenge due to the limited failure data availability in the telecommunications industry. In order to address this problem, we present a technique for generating failure data that realistically reflect the behaviour of degrading broadband lines. To this end, we use the conditional generative adversarial network and more importantly, we control and direct the failure data generation process using expert knowledge on the water ingress failure cause. The proposed technique is evaluated using a real-world case study involving the time-to-failure prediction of two types of broadband lines in a south-west city in England. The prognostics performance is measured using the Kappa statistic and F-score. Benchmark performance is obtained using Random Oversampling, Synthetic Minority Oversampling and Adaptive Synthesis which can be used to oversample failure data by duplicating existing failure data or randomly generating data. Among these techniques, Random Oversampling achieved the best prognostics performance. It is shown that the proposed technique outperforms Random Oversampling technique by a large margin. More specifically, it increased the prognostics performance by 33% (Kappa statistic) and 25% (F-score) for Asymmetric Digital Subscriber Lines, and 17% (Kappa statistic) and 13% (F-score) for Very High Bitrate Digital Subscriber Lines compared to the Random Oversampling technique.

Keywords: Broadband line prognostics, Expert knowledge, Failure causes, Generative modelling, Limited failure data, Telecommunications equipment prognostics, Water ingress

1. INTRODUCTION

Broadband lines provide a signalling method for transporting multiple signals through coaxial cables, twisted pair and optical fibre transmission mediums. One of their main applications in the telecommunications industry is high-speed internet (Sundaresan et al., 2011). Although the number of consumers who adopt broadband internet delivered completely over optical fibre is increasing, many broadband lines continue to be served in part by metallic paths (i.e. paired wires). Paired wires serving each consumer pass through a variety of underground and overhead electrical junctions and joints, and typically end at distribution points (DPs) located at the top of telegraph poles. DPs subsequently connect to the consumer premises via a drop wire to provide broadband service.

Water ingress is a dominant cause of a variety of faults (e.g. corrosion and electrical shorts) in overhead electrical junctions, joints and DPs in broadband lines due to their exposure to the driving rain (Tencer and Moss, 2002). As this can be a gradual process, faulting behaviour may grow in magnitude eroding the electrical capability of a broadband line causing degradation of broadband service (Tencer and Moss, 2002). This degradation may result in the consumer experiencing dropping connection, poor speed or the complete failure of broadband service (i.e. broadband line failure).

⋆ This research was funded by the Engineering and Physical Sciences Research Council through a Doctoral Training Partnership grant (EP/M508007/1). This research was partly supported by the Next Generation Digital Infrastructure project (EP/R004935/1) funded by the Engineering and Physical Sciences Research Council and British Telecom (BT).

The objective of this paper is to present a technique for predicting the time-to-failure (TTF) of telecommunications broadband lines under the conditions of limited failure data availability. Predicting the TTF of broadband lines with minimal uncertainty would enable telecommunications service providers to identify line degradation and failure before consumers experience them. Consequently, the appropriate proactive interventions can be undertaken to prevent unplanned downtime of broadband service, and hence reduce consumer dissatisfaction whilst reducing maintenance costs.

Data-driven prognostics have become popular for electronic equipment prognostics since they can estimate prognostics model parameters from degradation patterns contained within condition monitoring and/or event data relating to past failures (Wang et al., 2020). However, the long-lasting problem with data-driven prognostics is that they rely on large amounts of historical failure data to estimate model parameters effectively (Wang et al., 2020). Nevertheless, historical failure data are limited due to two major reasons: (i) over-protective maintenance and replacement regimes; (ii) highly reliable equipment (Wang et al., 2020). This causes failures to be rare, and leads to the problem of limited failure data availability for data-driven prognostics of broadband lines which causes prognostics predictions to be associated with high uncertainty (Louzada et al., 2019). Thus, telecommunications service providers are affected by unsatisfied consumers due to unplanned downtime of broadband service and additional costs due to under maintenance, over maintenance and false alarms.

In Ranasinghe et al. (2019), we presented a methodology for generating failure data that realistically reflect the behaviour of degrading equipment (i.e. real-valued failure data) for prognostics under the conditions of limited failure data availability. It allows training datasets used for data-driven prognostics to be augmented so that an increased number of failure data samples is available for prognostics modelling. The methodology generates real-valued failure data by controlling and directing the failure data generation process using auxiliary information pertaining to the failure mode that needs predicting. More specifically, the noise being added to the newly generated failure data samples is conditioned on auxiliary information to prevent different modes of data being generated. Auxiliary information is additional information that adds value to the understanding of failure dynamics of the equipment of interest (e.g. equipment similarity information, expert knowledge on failure causes and failure modes and quality of equipment use). However, the current version of the methodology only provides a way to utilise equipment similarity information as auxiliary information. Hence, the use of other kinds of auxiliary information to generate real-valued failure data remains to be exploited.

In this paper, we present a technique for predicting the TTF of telecommunications broadband lines under the conditions of limited failure data availability. To this end, we extend the aforementioned methodology so that expert knowledge on broadband line failure causes (e.g. water ingress into electrical junctions, joints and DPs) can be used to generate real-valued broadband line failure data. Whilst we discuss all the aspects of the proposed technique (i.e. broadband line data preprocessing, approach to the TTF prediction of broadband lines and real-valued broadband line failure data generation), the key contributions of this paper are as follows: (i) empirical results obtained using existing oversampling techniques for broadband line prognostics under the conditions of limited failure data availability; (ii) extension to the real-valued failure data generation methodology which allows utilising expert knowledge on broadband line failure causes to control and direct the failure data generation process; (iii) empirical results which show that the proposed technique increases prognostics performance by a large margin compared to existing oversampling techniques.

Following the problem formulation presented in our previous paper (see Ranasinghe et al. (2019)), this paper commences by introducing historical datasets used in this work for prognostics modelling and the process followed for data preprocessing (Sec. 2). The approach to the TTF prediction of broadband lines and benchmark prognostics performance obtained using existing oversampling techniques are presented in Sec. 3. The extension to the real-valued failure data generation methodology is presented in Sec. 4. The results which show the improved prognostics performance are discussed in Sec. 5. The paper is concluded in Sec. 6.

2. DATA PREPROCESSING

Broadband lines provide internet using two line modes: Asymmetric Digital Subscriber Line (ADSL) and Very High Bitrate Digital Subscriber Line (VDSL). Their key difference is the download and upload speeds of internet service (Sundaresan et al., 2011). ADSL provides a maximum of 8 and 1 Megabits per second (Mbps) download and upload speeds respectively. VDSL is an improved version of ADSL and it provides 52 and 16 Mbps download and upload speeds respectively. We used time series data sampled from ADSLs and VDSLs that had a broadband connection failure due to faults in electrical junctions, joints and DPs. These data are sampled from real-world consumer broadband lines in a south-west city in England. Historically, broadband connection failures occurred in this area are due to faulting behaviour that is strongly correlated with extreme driving rain.

A flowchart of the process followed for preprocessing ADSL and VDSL datasets is shown in Fig. 1. First, low variance features are removed due to their low predictive power. Then datasets converted into run-to-failure datasets by removing the parts of time series belong to the time before the start of equipment degradation and after the failure.

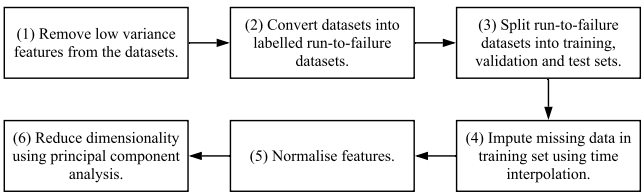


Fig. 1. Flowchart of the data preprocessing process.

The TTF of broadband lines is predicted using the fixed time window approach which requires labelling pre-failure

time series into segments (also see Fink et al. (2015)). In this study, run-to-failure data are segmented into 5 time windows and each of them has a fixed length of 1 day. Thus, the segments are *1 day before the failure*, *2 days before the failure*, *3 days before the failure*, *4 days before the failure* and *5 days before the failure*. Then all the data samples are labelled with the corresponding time window identity. That is, data samples belong to *1 day before the failure* segment is labelled with *1*, data samples belong to *2 days before the failure* segment is labelled with *2*, data samples belong to *3 days before the failure* segment is labelled with *3* and so on.

The labelled run-to-failure datasets are then split into training, validation and test sets containing 60%, 20% and 20% of the data samples contained in the original datasets respectively. The training set is used to train prognostics models, the validation set is used for hyperparameter tuning and the test set is used to evaluate prognostics models on previously unseen data. Missing data in the training ADSL and VDSL sets are imputed using time interpolation.

Prior to performing principal component analysis (PCA), the data are normalised in order to transform all the features into a comparable scale. PCA is then used to reduce the dimensionality of datasets from 25 features to 7 principal components (PCs). The reduction of feature space allows predictive models to improve their learning rates and reduce computation costs. The cumulative explained variance ratio obtained by the first 7 PCs for ADSL dataset is 66% and for VDSL dataset is 77%. These PCA transformed datasets are used to develop and evaluate prognostics models in the next section.

3. PROGNOSTICS MODELLING

In this section, the approach to the TTF prediction of broadband lines is presented first. Prognostics performance evaluation methods and benchmark prognostics performance used to evaluate the proposed technique are discussed next.

3.1 Time-to-failure Prediction of Broadband Lines

The TTF prediction of broadband lines is modelled as a multi-class classification problem as follows: given a data sample $x \in X$ and labels $y \in Y$ (i.e. 1 to 5 labels created for pre-failure time series segments in the previous section), calculate the conditional probability $Pr(y \mid x)$. The label with the highest $Pr(y \mid x)$ is the estimated label y' for the data sample x. Thus, the time series segment indicated by y' is the TTF failure of the broadband line. For example, if the segment indicated by the estimated label y' is 3, then the TTF is 3 days.

We developed multi-class classifiers using the following predictive algorithms: random forest (RF), k-nearest neighbour (kNN), decision tree (DT), support vector machine (SVM) with radial basis function kernel, adaptive boosting (Adaboost), multi-layer perceptron (MLP) and Naive Bayes (NB).

3.2 Evaluation Methods

The prognostics performance produced by classifier-based prognostics models is measured using the F-score and Cohen's Kappa statistic. F-score is the weighted harmonic mean of precision and recall normalised between 0 (i.e. worst value) and 1 (i.e. best value). However, F-score can be affected by statistical fluke (Powers, 2015). Hence, when measuring prognostics performance we also employ the Kappa statistic. It can be used as a statistical method for identifying whether a classifier simply guesses at random (Powers, 2015). Kappa statistic is always less than or equal to 1. Values of 0 or less indicate a poor classifier and conversely, 1 indicates a classifier that does not guess at random. A widely accepted schema for the Kappa statistic is shown in table 1 (Landis and Koch, 1977). The null hypothesis (H_0) used in this schema is: *the classifier performance is not due to random chance*. Thus, when measuring prognostics performance for each prognostics model, we first observe the value of Kappa statistic to identify whether the classifier performance is affected by statistical fluke. If the classifier performance is not affected by statistical fluke (i.e. Kappa statistic is in almost perfect agreement with H_0), we use the F-score of the classifier to quantify the prognostics performance.

Table 1. Schema for Cohen's Kappa statistic (Landis and Koch, 1977)

Kappa statistic range	Strength of agreement with H_0
Less than 0	Poor (i.e. due to random chance)
0 to 0.2	Slight
0.21 to 0.4	Fair
0.41 to 0.6	Moderate
0.61 to 0.8	Substantial
0.81 to 1	Almost perfect

3.3 Benchmark Prognostics Performance

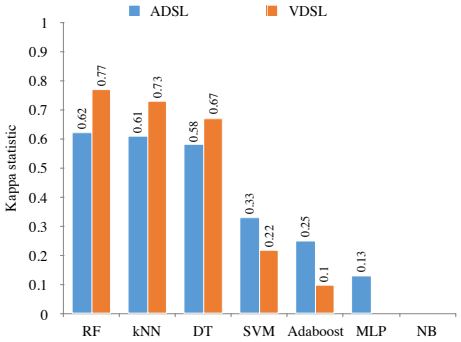
The proposed technique for predicting the TTF of broadband lines under the conditions of limited failure data availability is evaluated against the following benchmarks:

Benchmark 1: Performance obtained when prognostics models are trained on the original training dataset (i.e. the training dataset that is not augmented) and evaluated on the test dataset.

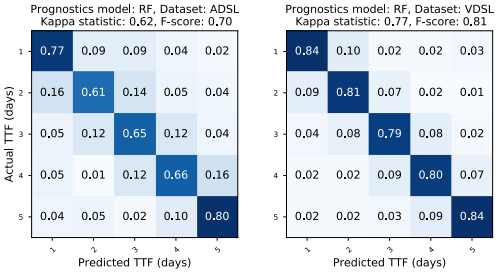
Benchmark 2: Performance obtained when prognostics models are trained on the training dataset that is augmented using existing oversampling techniques and evaluated on the test dataset.

Fig. 2 shows the Kappa statistic, confusion matrixes and F-scores obtained by prognostics models for Benchmark 1. It can be observed that the RF classifier-based prognostics model has obtained the best Kappa statistic value for ADSL (0.62) and VDSL (0.77) datasets. This means there is substantial agreement that the prognostics model performance (i.e. F-scores obtained by the classifier) is not due to random chance. The F-scores are 0.7 and 0.81 for ADSL and VDSL datasets respectively.

Fig. 3 shows the Kappa statistic, confusion matrixes and F-scores obtained by prognostics models for Benchmark 2. In contrast to Benchmark 1, training datasets are now augmented using the following oversampling techniques:



(a) Kappa statistic of all prognostics models for ADSL and VDSL.



(b) Confusion matrix, Kappa statistic and F-score of best prognostics model for ADSL (left) and VDSL (right).

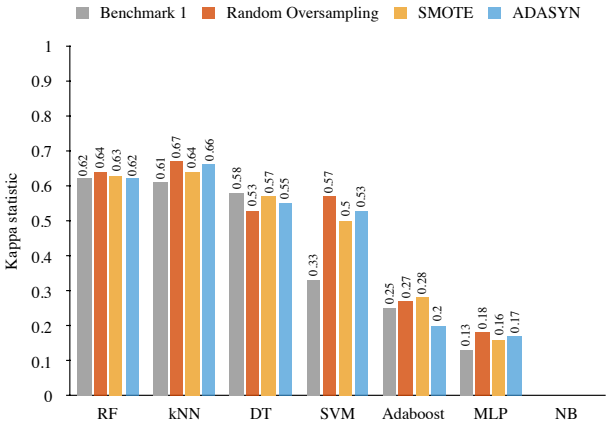
Fig. 2. Performance obtained for Benchmark 1.

Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthesis (ADASYN). The kNN classifier-based prognostics model and Random Oversampling technique have obtained the best Kappa statistic value for ADSL (0.67) and VDSL (0.78) datasets. However, this is a marginal increase in Kappa statistic compared to Benchmark 1 (i.e. 8% increase for ADSL and 1% increase for VDSL). Hence, the Kappa statistic is still in substantial agreement with the prognostics model performance. The F-scores are also only marginally improved compared to the Benchmark 1 (i.e. 4% increase for ADSL and 1% increase for VDSL). This marginal increase in prognostics performance is since Random Oversampling, SMOTE and ADASYN either duplicate existing failure data or randomly generate data (Weiss, 2004). Therefore, they do not introduce new and realistic failure data samples to augment training datasets (Weiss, 2004). Hence, the fundamental problem of limited failure data availability is not addressed sufficiently.

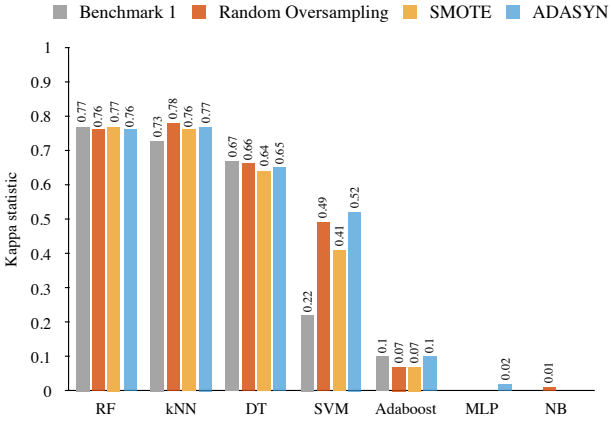
It can be concluded that Benchmark 1 and 2 failed to obtain almost perfect agreement for the Kappa statistic. Hence, there is low confidence in F-scores produced by prognostics models. In the following section, we show that the proposed technique enables increasing confidence in prognostics model performance by obtaining almost perfect agreement for the Kappa statistic. Moreover, it enables improving F-scores by a large margin compared to the benchmarks.

4. GENERATING REAL-VALUED BROADBAND LINE FAILURE DATA

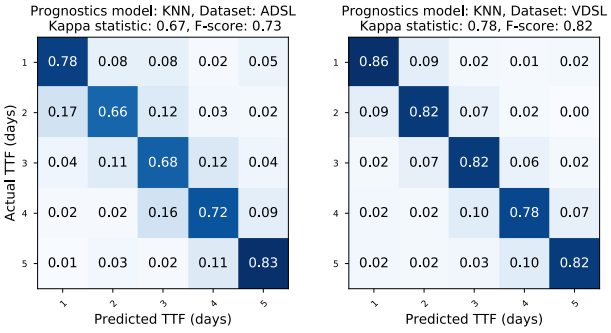
The methodology for generating real-valued failure data consists of three phases (see Fig. 4). A detailed description



(a) Kappa statistic of all prognostics models for ADSL.



(b) Kappa statistic of all prognostics models for VDSL.



(c) Confusion matrix, Kappa statistic and F-score of best prognostics model for ADSL (left) and VDSL (right).

Fig. 3. Performance obtained for Benchmark 2.

of these phases is provided in Ranasinghe et al. (2019). The limitation of the current version of the methodology is that it only provides a way to utilise equipment similarity information as auxiliary information. In this section, we extend the methodology so that expert knowledge on broadband line failure causes can be used as auxiliary information, and hence generate real-valued failure data for predicting the TTF of broadband lines under the conditions of limited failure data availability.

The extension involves following the process outlined in Fig. 5 for Phase 1 (i.e. identification and conversion of auxiliary information). Phase 2 and 3 of the methodology remain unchanged.

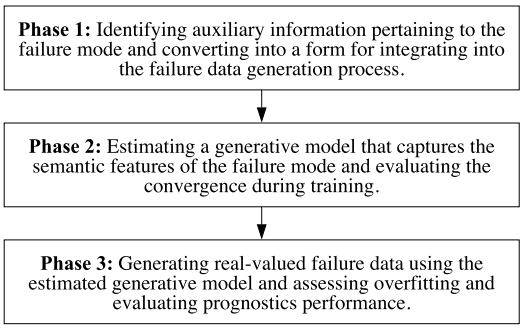


Fig. 4. Diagram outlining the three phases of the methodology for generating real-valued failure data.

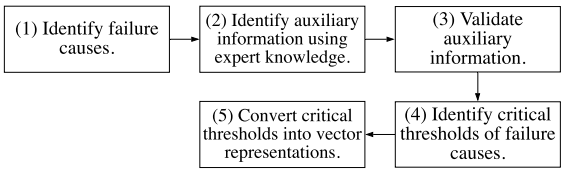


Fig. 5. Diagram outlining the steps for identifying and converting expert knowledge on failure causes.

(1) Identify failure causes: Fault tree analysis and historical maintenance records are used to identify failure causes of the failure modes that need predicting (i.e. corrosion and electrical shorts in electrical joints, junctions and DPs). We identified that water ingress is a dominant failure cause of broadband line connection failures.

(2) Identify auxiliary information using expert knowledge on failure causes: To reiterate, the data used in this study are sampled from broadband lines in a south-west city in England. Once water ingress is identified as a dominant failure cause, expert knowledge acquired from maintenance engineers is used to identify auxiliary information related to the failure cause. Maintenance engineers provided two pieces of auxiliary information based on their experience on historical broadband line failures occurred in the south-west city in England: (i) an increase in broadband line failures is expected when it is raining; (ii) an increase in broadband line failures is expected when it is raining and when prevailing winds are easterly. This is since anecdotally, engineering practice favoured placing overhead joints and DPs on the east side of telegraph poles as prevailing winds (and consequently driving rain) are typically westerly or south-westerly.

(3) Validate auxiliary information: When observing the number of failures occurred during different weather conditions (i.e. rain, drizzle, clouds and clear), it was identified that 84% of failures were occurring when it was raining for the majority of the week that the failure has occurred. In order to conduct a more robust experiment, auxiliary information identified from expert knowledge is validated using statistical hypothesis testing. To this end, we used historical maintenance records and weather reports (obtained from the OpenWeatherMap API). Whilst the former provides the date and time of failures, the latter provides rainfall levels and direction of wind when it was raining.

Two statistical hypothesis tests are developed using the following null hypotheses: (i) *there is an increase in broadband line failures when it is raining*; (ii) *there is an increase in broadband line failures when it is raining and when prevailing winds are easterly.* The objective of statistical tests is to identify whether the corresponding null hypothesis can be rejected. For the first test, probability values (p-value) of 0.92 (for ADSL) and 0.9 (for VDSLs) are obtained. This means, there is weak evidence against the null hypothesis, thus it is retained. For the second test, p-values of 0.03 (for ADSL) and 0.01 (for VDSL) are obtained. This means, there is strong evidence against the null hypothesis, thus it can be rejected.

To conclude, the increase in broadband line failures when it is raining is identified as a valid piece of auxiliary information. However, there is no strong evidence to support the increase in broadband line failures when it is raining and when prevailing winds are easterly.

(4) Identify critical thresholds: In this step, we identify what thresholds of rainfall impact broadband line failures the most. First, weather data are used to categorise rainfall into the following thresholds: light rain, moderate rain, shower rain and heavy rain. Then each failure is tagged based on what threshold of rainfall occurred for the majority of the week that the failure has occurred. Shower and heavy rainfall thresholds produce the highest number of failures per unit (i.e. per day) compared to light and moderate rainfall thresholds. Thus, shower and heavy rainfall thresholds are identified as critical thresholds of rainfall causing water ingress into electrical joints, junctions and DPs in broadband lines. These critical thresholds are then integrated as auxiliary information into the failure data generation process.

(5) Convert critical thresholds into vector representations: In order to integrate auxiliary information into the failure data generation process, we first convert it into an abstract form. This allows broadband line-specific information to be generalised to all the broadband lines that have failed under the failure modes that need predicting. For instance, if the rainfall in a particular location where a broadband line is located at (during the degradation period of electrical joints, junctions and DPs) is recorded as *the rainfall at the location where broadband line A, B and C located at increased from moderate to shower rain*, once converted into the abstract form this information becomes *some variable X increases*. Thus, specific terms such as broadband line A, B and C, rainfall and numerical thresholds are ignored. Then the abstracted information is converted into the statistical form by representing it as some continuous variable C. The continuous variable C can be converted into a distribution between some values y_0 and y_1. Finally, this distribution can be represented as a vector Y containing some values $\{y \in Y \mid y_0 < y < y_1, \text{and } y \text{ increases}\}$.

As mentioned at the beginning of this section, Phase 2 and 3 of the methodology remain unchanged and directly used to generate real-valued broadband line failure data using the converted auxiliary information. We generated 10,000 ADSL and 10,000 VDSL real-valued failure data samples and then augmented the original ADSL and VDSL training datasets.

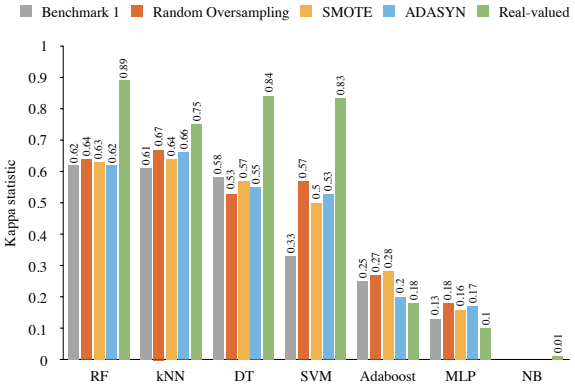
5. RESULTS AND DISCUSSION

The prognostics performance obtained when prognostics models are trained on the augmented training datasets and evaluated on the test datasets is shown in Fig. 6. The RF-based prognostics model has obtained the best value for the Kappa statistic for ADSL and VDSL. It can be observed that the Kappa statistic for ADSL is increased by 33% compared to the previous best performance (i.e. kNN and Random Oversampling). The Kappa statistic for VDSL is increased by 17% compared to the previous best performance (i.e. kNN and Random Oversampling). This means the proposed technique achieved the almost perfect agreement for the Kappa statistic by outperforming Benchmark 2 by a large margin, and hence improved the confidence in prognostics model performance.

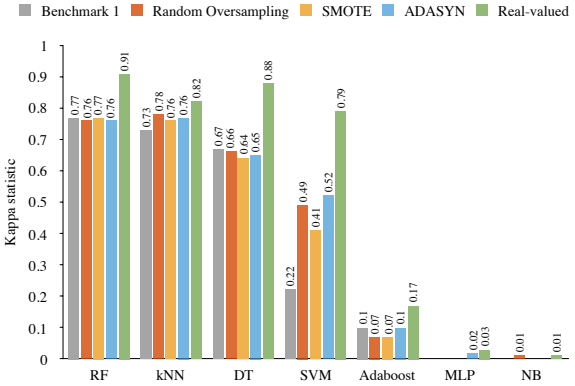
As shown in Fig. 6, the confusion matrixes and F-scores are also significantly improved. More specifically, F-score of ADSL is increased by 25% compared to the previous best performance (i.e. kNN and Random Oversampling) and VDSL is increased by 13% compared to the previous best performance (i.e. kNN and Random Oversampling).

6. CONCLUSION

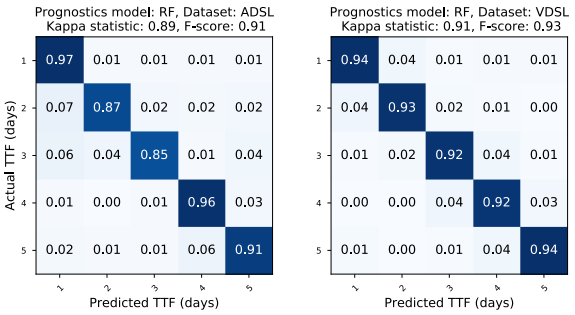
In this paper, a technique for predicting the time-to-failure of telecommunications broadband lines under the conditions of limited failure data availability is presented. This technique extends the methodology presented in our previous paper (Ranasinghe et al., 2019) so that real-valued broadband line failure data can be generated using expert knowledge on the water ingress failure cause. The impact of the research presented in this paper is that the proposed technique allows predicting real-world broadband line failures with minimal uncertainty when real broadband line failure data are limited. This enables telecommunications service providers to proactively undertake the appropriate interventions to prevent unplanned downtime of broadband service, and hence reduce consumer dissatisfaction whilst preventing costs due to over maintenance and false alarms.



(a) Kappa statistic of all prognostics models for ADSL.



(b) Kappa statistic of all prognostics models for VDSL.



(c) Confusion matrix, Kappa statistic and F-score of best prognostics model for ADSL (left) and VDSL (right).

Fig. 6. Performance obtained for real-valued failure data generation.

REFERENCES

Fink, O., Zio, E., and Weidmann, U. (2015). A classification framework for predicting components' remaining useful life based on discrete-event diagnostic data. *IEEE Transactions on Reliability*, 64(3), 1049–1056.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Louzada, F., Cuminato, J.A., Rodriguez, O.M., Tomazella, V.L., Ferreira, P.H., Ramos, P.L., Niaki, S.R., Gonzatto, O.A., Perissini, I.C., Alegría, L.F., et al. (2019). A repairable system subjected to hierarchical competing risks: Modeling and applications. *IEEE Access*, 7, 171707–171723.

Powers, D.M. (2015). What the f-measure doesn't measure: Features, flaws, fallacies and fixes. *arXiv preprint arXiv:1503.06410*.

Ranasinghe, G.D., Lindgren, T., Girolami, M., and Parlikad, A.K. (2019). A methodology for prognostics under the conditions of limited failure data availability. *IEEE Access*, 7, 183996–184007.

Sundaresan, S., De Donato, W., Feamster, N., Teixeira, R., Crawford, S., and Pescapè, A. (2011). Broadband internet performance: a view from the gateway. *ACM SIGCOMM computer communication review*, 41(4), 134–145.

Tencer, M. and Moss, J.S. (2002). Humidity management of outdoor electronic equipment: methods, pitfalls, and recommendations. *IEEE Transactions on Components and Packaging Technologies*, 25(1), 66–72.

Wang, J., Liang, Y., Zheng, Y., Gao, R.X., and Zhang, F. (2020). An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. *Renewable Energy*, 145, 642–650.

Weiss, G.M. (2004). Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1), 7–19.