Arbon, Edward; Smet, Peter; Gunn, Lachlan; McDonnell, Mark

Anomaly Detection in Satellite Communications Systems using LSTM Networks

Published in:
2018 Military Communications and Information Systems Conference, MilCIS 2018 - Proceedings

DOI:
10.1109/MilCIS.2018.8574109

Published: 12/12/2018

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Anomaly Detection in Satellite Communications Systems using LSTM Networks

Lachlan Gunn*, Peter Smet†, Edward Arbon†and Mark D. McDonnell‡

*Department of Computer Science, Aalto University, Finland, Email: Lachlan.Gunn@aalto.fi
†Defence Science and Technology Group, Adelaide, Australia, Email: Edward.Arbon@dst.defence.gov.au
‡Computational Learning Systems Laboratory, School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes SA 5095, Australia, Email: Mark.McDonnell@unisa.edu.au

***Abstract*—Most satellite communications monitoring tools use simple thresholding of univariate measurements to alert the operator to unusual events [1] [2]. This approach suffers from frequent false alarms, and is moreover unable to detect sequence or multivariate anomalies [3]. Here we consider the problem of detecting outliers in high-dimensional time-series data, such as transponder frequency spectra. Long Short Term Memory (LSTM) networks are able to form sophisticated representations of such multivariate temporal data, and can be used to predict future sequences when presented with sufficient context. We report here on the utility of LSTM prediction error as a de-facto measure for detecting outliers. We show that this approach significantly improves on simple threshold models, as well as on moving average and static predictors. The latter simply assume the next trace will be equal to the previous trace. The advantages of using an LSTM network for anomaly detection are twofold. Firstly, the training data do not need to be labelled. This alleviates the need to provide the model with specific examples of anomalies. Secondly, the trained model is able to detect previously unseen anomalies. Such anomalies have a degree of unpredictability that makes them stand out. LSTM networks are further able to potentially detect more nuanced sequence and multivariate anomalies. These occur when all values are within normal tolerances, but the sequence or combinations of values are themselves unusual. The technique we describe could be used in practice for alerting satellite network operators to unusual conditions requiring their attention.**

I. Introduction

A. Background

Satellite communications systems are subject to a wide range of anomalous behaviour; changes in the transmitter characteristics, the physical channel, and the receiver all manifest themselves in the received signal. The detection of anomalies in such systems is a complex problem, and one that is made more difficult by the unique nature of many anomalies [4] [5]. This precludes the supervised training of a classifier using representative examples of normal and anomalous signals. The approach preferred by commercial developers of satellite network monitoring and control systems uses simple univariate threshold-based detectors to flag anomalies to operators [1] [2]. In the case of frequency spectra, a representative template of a carrier is set as a baseline. Should any frequency differ from the corresponding template frequency by more than a given amount, an error is signalled. Such systems are notorious for flagging too many false positives, however, leading users to either ignore or switch off the thresholds [1].

Since it is not possible to detect novel or 'zero day' anomalies by their signatures, we [3] and others [6]–[8] have taken a different approach. Rather than using labelled anomalies, we train a model only with normal, unlabelled data, and flag any deviations from this model of normality. Recently, there has been interest in using recurrent neural networks, and in particular Long Short Term Memory (LSTM) networks, for the detection of anomalies [6]–[8]. Such networks can form sophisticated representations of high-dimensional time-series data. The trained network can then be used as a prediction model on new data, and its prediction error will reflect the degree to which the data is anomalous. The advantages of this approach are several. Since this is an unsupervised learning task, all training data can be both unlabelled, and nominally normal. This solves the problem of finding and tagging representative examples of anomalies, which are often difficult to obtain. In addition, the trained network is able to detect novel anomalies not in the training set. Such anomalies will stand out as a result of their unpredictability.

B. Proposed Anomaly Detection Framework

Here we present an LSTM-based anomaly detection system. The system is used to detect anomalous behaviour in high dimensional multivariate time series spectrum data from a satellite transponder. The network is presented with a spectral time series, typically between 24-64 consecutive time points, and trained to predict the spectrum several time-steps into the future. Following training, the magnitude of the prediction error vector can be used, either directly, or as a likelihood estimate from a mutivariate normal distribution [6], to flag anomalies. This model shows a substantial improvement over both threshold and moving average-based detectors, as well as a static baseline detector that predicts no change from the previous spectrum.

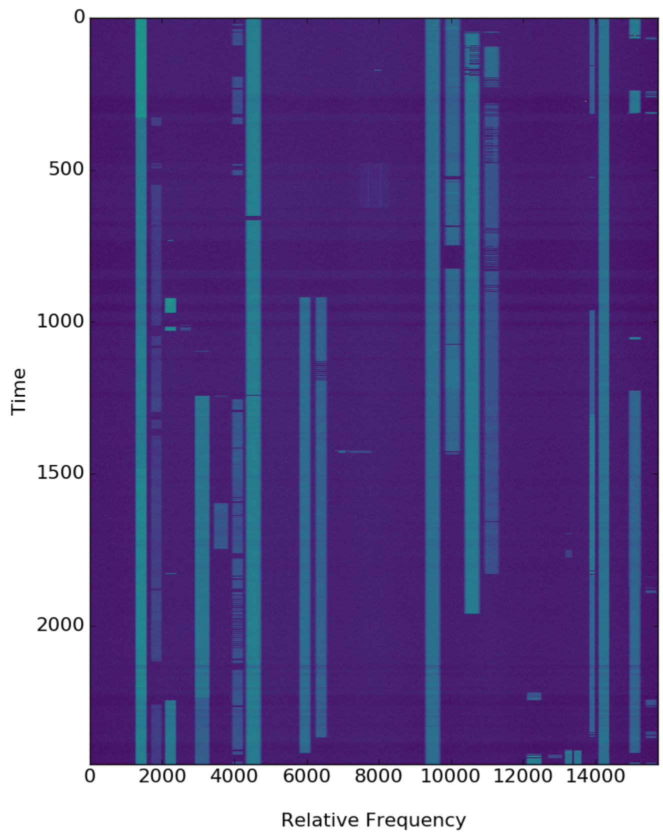

Fig. 1. Waterfall diagram of the spectral data set. Spectra are collected at 30-minute increments, with each spectrum including 15731 frequency bins spread across a bandwidth of $\approx$ 60 MHz – a bin width of 3904 Hz. The resolution bandwidth is 7.646 kHz.

II. Methods

A. Network Parameters

A three layer, 200 x 200 x 200 LSTM network was trained, using the RMSprop optimizer with a learning rate of 0.0001. This learning rate was selected after trying rates of 0.1, 0.01, and 0.001, as suggested by [9]. The input size was a vector of 24 contiguous historical time points, each consisting of a spectrum snapshot of 500 frequencies. These training samples represent the context from which predictions must be learned. The output size was 2500, representing the 500 predicted frequencies at 1-5 future time-steps. Dropout of 30% was used between layers to prevent overtraining [10] [11], and all the raw data were normalized to the range [0,1] [11]. There were a total of 2456 consecutive spectrum measures, taken at 30 minute intervals, which were split 70:30 into training and validation sets. The entire data set is shown as a waterfall diagram in Figure 1. For the experiments described herein, a subset of the transponder data, composed of 500 frequencies and representing a single carrier, was used. The LSTM network model was trained until there was no improvement in prediction error on the validation data. This was typically between 30-60 epochs.

TABLE I

Maximum prediction errors $\pm$ std for the LSTM and baseline predictors at 5 levels of lookahead. Advantage represents the % decrease in prediction error of the LSTM relative to the baseline model. Carrier data were used for these figures.

Lookahead	LSTM	Baseline	Advantage
1	0.348 $\pm$ 0.060	0.473 $\pm$ 0.079	35.8%
2	0.352 $\pm$ 0.066	0.474 $\pm$ 0.078	34.8%
3	0.356 $\pm$ 0.070	0.477 $\pm$ 0.080	34.0%
4	0.359 $\pm$ 0.074	0.486 $\pm$ 0.092	35.3%
5	0.362 $\pm$ 0.077	0.492 $\pm$ 0.092	36.0%

TABLE II

Mean prediction errors $\pm$ std for the LSTM and baseline predictors at 5 levels of lookahead. Carrier data were used for these figures.

Lookahead	LSTM	Baseline	Advantage
1	0.079 $\pm$ 0.066	0.102 $\pm$ 0.081	29.2%
2	0.082 $\pm$ 0.071	0.107 $\pm$ 0.086	30.1%
3	0.085 $\pm$ 0.075	0.111 $\pm$ 0.091	30.6%
4	0.088 $\pm$ 0.079	0.115 $\pm$ 0.095	30.8%
5	0.090 $\pm$ 0.082	0.118 $\pm$ 0.099	31.0%

B. Model Training and Statistics

Each spectrum snapshot consisted of 500 frequencies, and the model received as input random slices consisting of 24 such consecutive snapshots. The network was then trained to predict the next five spectrum traces, using the ‘regression’ model, with mean square error as the loss function. In this model, a prediction is represented as a single expected value for each of the 500 frequencies. For evaluation purposes, the model’s prediction error was taken as the absolute difference between the actual and predicted next frequencies. The effectiveness of the LSTM was compared to a baseline model. The static baseline model always predicts that the next trace will be identical to the last trace. In the absence of any other information, this represents a ‘best effort’ predictor. The relative advantage of the LSTM to the baseline model was calculated as the difference of their prediction accuracy scores, and expressed as a percentage of the LSTM score. For completeness, the LSTM was also compared to predictions based on the average of the training data, and the moving average of the previous ten data points. The prediction errors of the models were compared using an independent samples t-test. The test compared 702 LSTM to 702 baseline prediction errors (df = 1402), using the mean absolute prediction error computed across the 500 frequencies for each data point. The models were significantly different across all lookahead levels, but only level one results, which showed the smallest differences between mean absolute prediction errors, are reported here.

TABLE III
MEAN PREDICTION ERRORS $\pm$ STD FOR THE LSTM AND BASELINE PREDICTORS AT 5 LEVELS OF LOOKAHEAD ASSESSED ON DATA WHERE NO CARRIER WAS PRESENT.

Lookahead	**LSTM**	**Baseline**	**Advantage**
1	0.087 $\pm$ 0.073	0.119 $\pm$ 0.101	37.7%
2	0.087 $\pm$ 0.073	0.120 $\pm$ 0.101	37.5%
3	0.088 $\pm$ 0.074	0.121 $\pm$ 0.102	37.1%
4	0.089 $\pm$ 0.074	0.121 $\pm$ 0.102	36.8%
5	0.089 $\pm$ 0.075	0.122 $\pm$ 0.102	36.6%

III. Results

For both the LSTM and baseline predictors, the errors at all five lookahead levels were dominated by the high degree of noise in the data. Maximum absolute prediction errors for spectra from a carrier ranged from 35% to 49% (Table I), while the mean errors were between 8% and 12% (Table II). The mean prediction error increased to between 9% and 12% when predictions related to frequency spectra in the absence of a carrier (Table III). Figures 2 and 3 are heat maps representing the absolute prediction errors for predictions at 1-5 levels of lookahead for both the LSTM and baseline models. The predictions are for the validation data only. As expected, the error for both models increases when asked to predict the spectrum of a carrier further out in time (increasing lookahead, Tables I and II, Figure 2). This trend is less apparent when the LSTM model is trained on frequency data in the absence of a carrier (Table III and Figure 3). Interestingly, the margin separating the mean prediction error of the two models also increases with increasing lookahead (from 29% to 31%, Table II). This indicates that the LSTM has formed a prediction model that is able to predict trends over multiple time steps in unseen spectrum data. This more sophisticated model allows the LSTM to increasingly outperform static predictions as the prediction horizon increases.

It is clear that the mean prediction errors of the LSTM are consistently lower than that of the baseline model. This holds true in both the presence (two-sample $t(1402) = 19.5$, $p < 0.0001$) and absence of a carrier signal (two-sample $t(1402) = 99$, $p < 0.0001$). For carrier-based predictions, the maximum error for the LSTM model was around 35.5%, compared to the 48% baseline (Table I). This represents an improvement of 12.5% in absolute terms. The maximum prediction error is closely aligned to commercial satellite monitoring and control systems. These use the maximum single deviation of the current spectrum from a representative spectrum mask to flag anomalies [1] [2]. We further compared the mean absolute prediction errors of the two models, as a more refined measure. For the LSTM model, this was around 8%, compared to 11% for the baseline model (Table II). While this difference of $\approx$ 3% is small in absolute terms, it represents

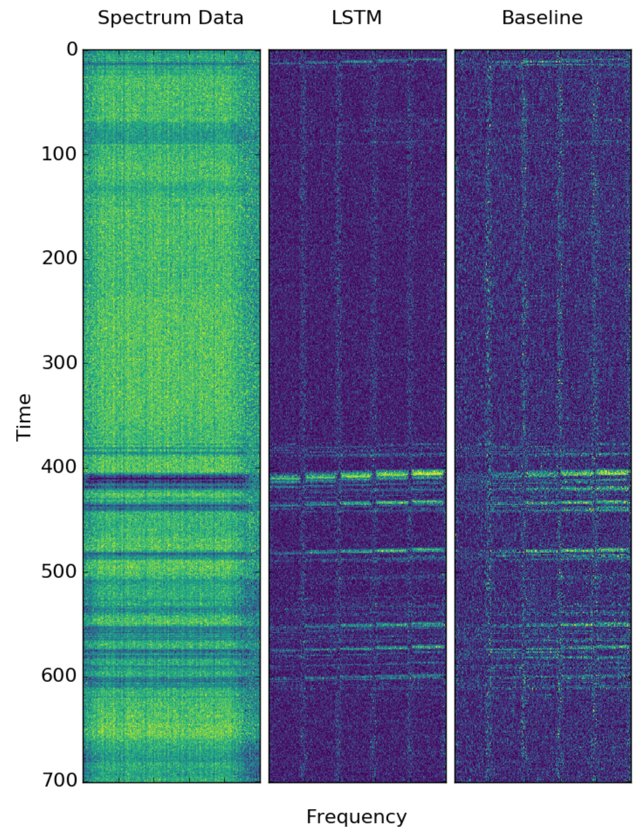

Fig. 2. Heat map of prediction errors for the LSTM and baseline models. The leftmost column shows the raw waterfall spectrum data. The middle column shows the LSTM prediction errrors at 1-5 levels of lookahead, from left to right. The final column shows the identical data for the baseline predictor. Lighter colours represent larger prediction errors. Note that lookahead columns have been compressed to fit the 5 levels to the diagram.

a relative improvement of approximately 30% for the LSTM network, and is statistically highly significant (two-sample $t(1402) = 19.5$, $p < 0.0001$). It is also in line with the LSTM network's 35% relative advantage seen in the maximum error results (Table I).

For the data corresponding to the absence of a carrier, the relative difference between the models was increased, rising to 37% (LSTM 8.7%, baseline 12%, Table III). An examination of the heat maps in Figures 2 and 3 indicates that the improvement in prediction errors is heterogeneous. It can be seen from these maps that the LSTM model makes lower prediction errors on the steady background and carrier signals relative to the baseline model. In contrast, larger and more abrupt changes in the signal are better predicted by the baseline model. This explains the sharper banding visible in the LSTM versus baseline predictions in Figure 2: the LSTM prediction errors are much larger in these regions.

Current satellite communications monitoring tools use deviation from a representative mean spectrum to detect anomalies [1] [2]. We have used the mean of the combined

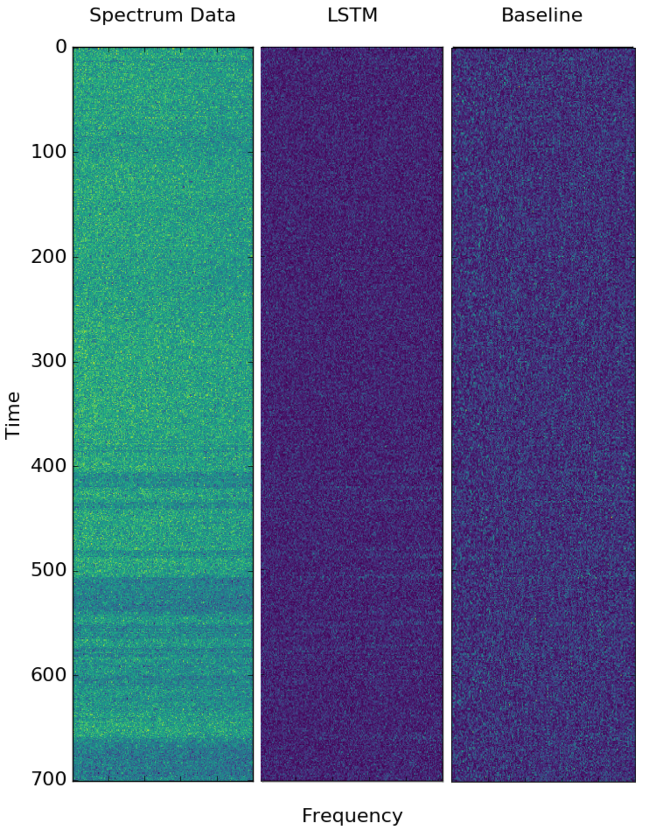

Fig. 3. Heat map of prediction errors for the LSTM and baseline models in the absence of a carrier. The leftmost column shows the raw signal data, with systematic changes in the background levels revealed by dark banding. The middle column shows the LSTM prediction errrors at 1-5 levels of lookahead. The final column shows the identical data for the baseline predictor. The lighter colours apparent in the last column indicate the highter prediction errors of the baseline predictor.

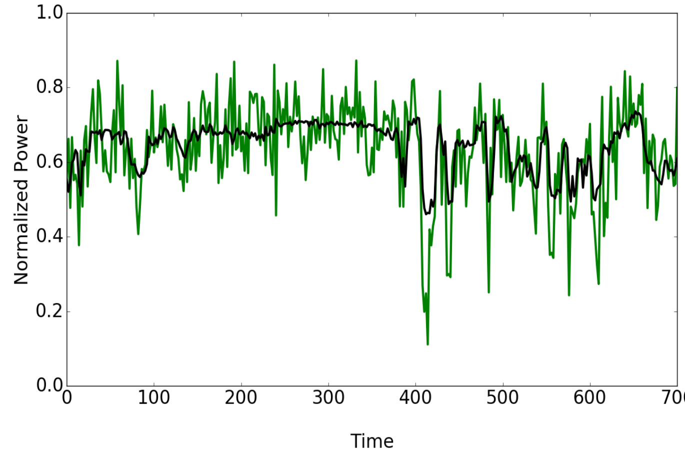

Fig. 4. A representative single frequency trace shown in green, compared to its predicted values by the trained LSTM model, shown in black. The network has learned to ignore much of the noise in its predictions. Note that each time increment represents a 30 minute interval.

carrier training data, in addition to a moving average over the previous ten spectra, as further comparative prediction models. The mean model has an absolute prediction error of 0.103 ± 0.07, which not surprisingly is similar at all levels of lookahead. The moving average model does somewhat better, ranging from 0.092 ± 0.06 at level one lookahead to 0.103 ± 0.064 at level five lookahead. The LSTM prediction errors are 29.7% and 16% lower than these two models, respectively, at one level of lookahead. Independent t-tests confirmed that the prediction errors of the LSTM were significantly lower than both of these models ($p \ll 0.0001$ in all comparisons).

When we examine the predictions qualitatively, it is obvious that the baseline model, by predicting that the next trace will be identical to the last trace, is effectively sustaining the noise from that previous trace. This exaggerates its prediction errors in line with the level of noise. In contrast, the LSTM-based predictor takes as input the spectra of several time-steps, allowing it to learn to suppress noise by filtering and smoothing (Figure 4). This is most pronounced when we compare the models on data without a carrier, where the LSTM predictions are approximately 37% better than those of the static baseline model (Table III). In the carrier data, it can be seen that the baseline model quickly adapts to dramatic changes in a carrier's behaviour, simply predicting that any changes will persist. The LSTM model by comparison, continues to predict its trained expectations. As a result, the LSTM more accurately predicts persistent common patterns, but is less accurate for large and novel changes. In effect, the LSTM provides much greater contrast between the ongoing patterns and any changes to these than the baseline model does. This is exactly the kind of behaviour we want from an anomaly detection system. By not adapting to large and unusual changes quickly, as the baseline model does, the LSTM produces larger prediction errors for such events. These errors can then be used to flag anomalies.

Figure 5 shows the number of anomalies detected at incremental thresholds for both the LSTM and baseline models. Here, an anomaly is simply defined as a prediction error value outside the threshold of the detector. The LSTM network consistently flags fewer anomalies at any given threshold. This is a direct consequence of the models' greater prediction accuracy: lower average prediction errors mean fewer errors will exceed an arbitrary threshold. By having a lower threshold, we can improve the sensitivity of the model while retaining its specificity [12]. Although the baseline model flags more anomalies, its lower prediction accuracy will coincide with more false positives.

IV. Discussion

We have shown here that an LSTM network is able to estimate future data points in complex multidimensional time series data derived from satellite frequency spectra with some accuracy. A fully trained network can successfully be used to flag anomalies in unseen data, using a prediction error threshold. This approach takes noise levels into consideration,

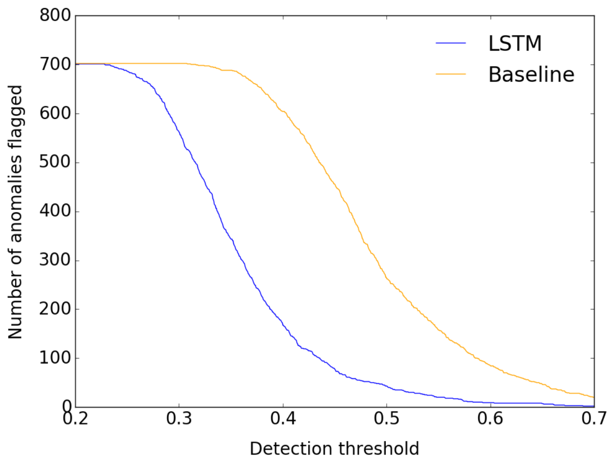

Fig. 5. Graph comparing the number of anomalies detected at a given threshold level for the LSTM and baseline models. The baseline model has a much higher average prediction error than the LSTM, so that more prediction errors will fall above any given threshold. The more accurate predictions of the LSTM allow a lower threshold for anomaly detection. This results in greater specificity and fewer false positives.

and adapts to the temporal dynamics and multivariate nature of the data. The model outperforms both simple threshold and moving average-based detectors, as well as a minimal static predictor, by significant margins. The relative improvement of the LSTM network ranged from 16% against the moving average, to $\approx$ 30%-35% against threshold-style and static predictors. The main advantage of our approach is that the model can be trained with only normal, unlabelled data, and that it is able to detect novel never-before-seen anomalies. While all of the models are able to rapidly classify spectrum traces as normal or anomalous, the LSTM network takes significantly longer to train than the simpler models.

We assume here that better prediction performance results in improved anomaly detection. This is because all classification models necessarily make a trade off between sensitivity (recall) and specificity (precision) [12]. Increasing the sensitivity of a model normally results in lower precision, such as false positives, and vice versa. If, for example, we lowered the threshold of our LSTM without improving prediction accuracy, sensitivity would be improved at the expense of specificity. By improving prediction accuracy, however, we can lower the threshold of what is considered abnormal, but still retain specificity. This is a result of narrowing the bounds that define normal behaviour with our refined predictions. These narrower bounds are more sensitive to genuine outliers, while still tracking and defining the normal range of behaviours. As a specific example, a 30% improvement in prediction accuracy enables the detection threshold to be lowered by an equivalent amount. For the LSTM detector in Figure 5, lowering its threshold by 30%, from 0.4 to 0.28, results in a greater than five-fold increase in the number of anomalies flagged. Since this increased detection is a consequence of the greater prediction accuracy, the number of false positives will remain unaffected.

One issue with the current data set is that the transponder data do not exhibit the kind of rich temporal behaviour patterns that are most amenable to machine learning systems [11]. In essence, we are analysing a relatively noisy system in which a carrier may be in only one of two states: present, or absent. The decision to raise or lower a carrier is normally an arbitrary decision by the operators that demonstrates no clear pattern. The network is thus unable to anticipate such behaviours. Furthermore, the relatively coarse-grained time samples, taken at 30 minute intervals, do not permit us to analyze the effects of rapidly-occuring events.

In future work, we intend to examine more dynamic spectra at a much finer temporal resolution. By selecting spectra that display more complex behaviours over time, the LSTM approach can be evaluated to its fullest potential. We further intend to explore the prediction horizons of LSTM networks in order to determine how many time-steps can be anticipated by such systems in practice.

Acknowledgment

This work was funded by a Competitive Evaluation Research Agreement (CERA) grant to the University of South Australia by the Defence Science and Technology Group. Lachlan Gunn was at the University of South Australia when this work was performed.

References

[1] Kratos Communications. Monics Net: Advanced and Scalable Carrier Monitoring and Interference Identification Solution. Accessed 2018-05-17. [Online]. Available: http://www.kratoscomms.com/~/media/communications/pdf/factsheet-monicsnet.pdf

[2] Kratos Communications. Compass: Manage, Monitor and Control Mission Critical Networks. Accessed 2018-05-17. [Online]. Available: http://www.kratoscomms.com/~/media/communications/pdf/factsheet-compassmonitorandcontrol.pdf

[3] E. Arbon and P. Smet, ser. International Communications Satellite Systems Conferences (ICSSC). American Institute of Aeronautics and Astronautics, Sep 2015, ch. Anomaly Detection in Satellite Communications Networks using Support Vector Machines, 0. [Online]. Available: https://doi.org/10.2514/6.2015-4321

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15:1–15:58, 2009.

[5] S. Ahmad, A. Lavin, S. Purdy, and Z. Aghaa, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.

[6] T. J. O'Shea, T. C. Clancy, and R. W. McGwier, "Recurrent neural radio anomaly detection," *CoRR*, vol. abs/1611.00301, 2016. [Online]. Available: http://arxiv.org/abs/1611.00301

[7] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 04 2015.

[8] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," *CoRR*, vol. abs/1607.00148, 2016. [Online]. Available: http://arxiv.org/abs/1607.00148

[9] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *CoRR*, vol. abs/1503.04069, 2015. [Online]. Available: http://arxiv.org/abs/1503.04069

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[12] M. Buckland and F. Gey, “The relationship between recall and precision,” *J. Am. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, Jan. 1994. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L