A New Method for Explaining the Regression Output of a Multi-layer Perceptron Network with Real Valued Input Features

M.L.Vaughn, J.G.Franks
Cranfield University,
Defence Academy,
Swindon, UK
M.L.Vaughn@rmcs.cranfield.ac.uk

Abstract

In this paper, we present a new method for explaining the regression output of a MLP network with real valued input features. The application selected to demonstrate the method is the prediction of helicopter airframe load spectra from continuously valued flight parameter data. Two example regression outputs are studied - a high strain case and a low strain case. For each case the input channels are discovered that determine the output activation values. The underlying mechanism that drives the MLP regression output is determined.

INTRODUCTION

A full explanation facility has been developed by the first author [1-4] for interpreting the output on a case-by-case basis from any standard multi-layer perceptron (MLP) network that classifies binary input data in *n*-dimensional input space, using a sigmoidal activation function at the hidden layer and output layer neurons. The method represents a significant advance towards the goal of readily interpreting trained neural networks that solve real-world problems with a large number of input features [5].

This study adapts the MLP interpretation method developed for classification tasks to include MLP regression models that predict continuous variables from continuously valued input features. The need for such methods has been stated in [6] within the context of interpretation and knowledge extraction from trained artificial neural networks.

THE HELICOPTER MLP NETWORK

The aim in developing the helicopter MLP [7] was to predict component loads from easily measurable parameters without the use of strain gauges or other measuring equipment. The flight data was generated from flight tests using a Westland Lynx Helicopter. The total data consisted of 45 channels of flight parameter data and three channels of helicopter load spectra measured using strain gauges. The total flight data was in excess of four hours of flying time and represented 2 Gbytes of information. Each data channel was scaled between 0 and 1 before MLP development.

MLP Input and Output Channels

An 80 second sample was taken for MLP development [7], which included two cyclic 'pull-up' maneuvers and some non-transitional flight. A number of channels were omitted which were judged by experts not to affect the strain in the airframe or rotor-head. To simplify initial experiments, only one load channel was selected as the channel to be predicted. The final data set consisted of 21 channels of flight parameter data (F1 – F21) and one channel of load spectra (F0), measured using a strain gauge on the I-beam assembly, as shown in Table 1.

Table 1. MLP input channels and strain output channel

Input Channels	Description	Corr Coeff
F1	Collective Lever Position	+0.12
F2	Collective Servo Position	–0.07
F3	F/A Cyclic Stick Position	+0.10
F4	Lat Cyclic Stick Position	–0.05
F5	F/A Cyclic Servo Position	+0.79
F6	Lat Cyclic Servo Position	–0.01
F7	Tail Rotor Pitch Angle	–0.05
F8	Rudder Pedal Position	–0.13
F9	Barometric Altitude	–0.15
F10	Barometric Airspeed	+0.08
F11	Lateral 'G' at C of G	–0.04
F12	Main Rotor Speed	+0.04
F13	Outside Air Temperature	–0.00
F14	Pitch Attitude A/C System	+0.47
F15	Pitch Rate	+0.04
F16	Roll Attitude A/C System	–0.15
F17	Roll Rate	–0.32
F18	Aerodynamic Yaw	–0.06
F19	Normal 'G' at C of G	+0.53
F20	Yaw Rate	–0.18
F21	Main Rotor Hub Torque	–0.02
Output Channel	**Description**	
F0	Vertical Shear stbd STN 420A	

The correlation coefficient was evaluated between each of the input channels and the strain output channel F0, as shown in Table 1 [7]. It can be seen from Table 1 that the input channels with the highest correlation are F5, F14, F17 and F19.

MLP Selected for this Study

The first 40 seconds containing the first 'pull-up' maneuver were used for the MLP training. Data was sampled at 840 Hz and the 33,600 training vectors were presented in a random sequence during training. The following 40 seconds containing the second 'pull-up' maneuver were used for network testing with a further 33,600 test vectors.

A number of different MLP architectures were developed (using NeuralWare Professional II Plus™ on a Sun Sparc Workstation) with a range of hidden neurons from 5 to 25 and sigmoidal activation functions at hidden and output layer neurons. The MLP with 20 hidden neurons was found to have the minimum test mean absolute error of 0.021 at 250,000 cycles of training. The corresponding mean absolute error of the training data was 0.016. The 21-20-1 MLP was selected for this study.

The aim of this study is to explain how the helicopter MLP predicts the single strain channel, F0, from the 21 flight parameter data channels. This is achieved by interpreting two different cases – a high strain output value, and a low strain output value.

INTERPRETING THE HIGHEST STRAIN CASE

The training case with the maximum strain MLP output value of +0.7560 was the high strain case selected for interpretation. This was the training case with the maximum F0 target value of +0.7957 in the training set.

The interpretation begins at the MLP output neuron activation which, similar to a classifier output activation [1–4], is driven as high as possible by the MLP to meet the maximum target output. The high strain value of +0.7560 is sigmoid(+1.1309) where +1.1309 is the combined input sum from hidden neurons and the hidden bias to the single output neuron. The combined sum is made up of a positive part +1.9165 and a negative part –0.7760. The hidden bias contributes –0.0096.

Input Channels Driving the Positive Signal at the Regression Output Neuron

Following the same approach as the method that interprets classifier MLPs with binary inputs [1-4], the hidden layer feature detectors and the input channels are first discovered that drive the positive part of the combined input sum at the regression output high. This is then followed by the discovery of the input channels that drive the negative part of the combined input sum low.

Discovery of the Hidden Feature Detectors

In predicting a high strain output the MLP drives the *positive* part of the combined sum as *high* as possible to meet the maximum target output. This is achieved by the hidden neurons that provide all the *positive* part of the combined sum at the output neuron. The method defines these neurons as the *hidden layer feature detector* neurons [1–4].

Using sigmoidal activation functions, the feature detectors are hidden neurons connected with a *positive weight* to the regression output neuron. For the high strain case, 7 hidden neurons H_1, H_2, H_5, H_9, H_{13}, H_{19} and H_{20} are feature detectors, providing all the positive signal +1.9165 to the combined sum. It is of note that, for the high strain case, the only hidden neurons with activation >0.5 are H_1 and H_{19}, with activations of +0.5020 and +0.5383, respectively.

Ranking the Hidden Feature Detectors

Some positive feature detectors are more important to the MLP regression output than others. These are the positive feature detectors that contribute the highest positive combined input. In the high strain case, H_1, H_{19} and H_{13} together provide 97.4% of the total positive input at the strain output neuron, contributing 47.6%, 40.0% and 9.8% respectively.

Discovery of the Positive Input Channels

The positive input channels detected by each feature detector are the MLP inputs that provide the *positive* part of the combined input sum at each detector [1–4]. In this application, since all input channels have positive values, these are channels connected with positive weights to the detectors.

For example, the positive input channels detected by feature detector H_1 are channels F1, F3, F5, F6, F10, F11, F14, F15, F18, F19. All of the positive input channels detected by all 7 feature detector neurons are the 13 (out of 21) channels: F1, F3, F5, F6, F7, F8, F9, F10, F11, F14, F15, F18, F19.

Ranking the Positive Input Channels

The MLP is expected to use channels with high positive input values connected with high positive connection weights to drive the activation of the feature detectors, and subsequently the strain output, as high as possible. For this reason, the positive input channels can be ranked by forcing each channel in turn to its *minimum value* in the training set and presenting the new input vector to the MLP. The highest ranked channel is the one that produces the largest drop in activation at the MLP regression output neuron [1–4].

For the high strain case the highest ranked features are channels F5, F3, F14 and F15 with changes in the MLP output strain value of –18.0%, –11.5%, –6.5% and –3.2% respectively.

The Positive Input Channel Data Relationship

The positive input channel data relationship shows how the top ranked channels are related and provides the explanation of the positive activation of the highest strain case [1–4]. The data relationship is found by progressively *minimizing* each of the positive channels together *in ranked order*.

The results for the highest strain case are shown in Fig. 1, where the MLP strain value smoothly and gracefully degrades from a maximum of +0.7560 to a minimum of +0.4045. It can be seen from Fig. 1 that the positive activation in the highest strain case is driven mostly by the top 4 ranked channels F5, F3, F14 and F15 – and primarily by F5 and F3.

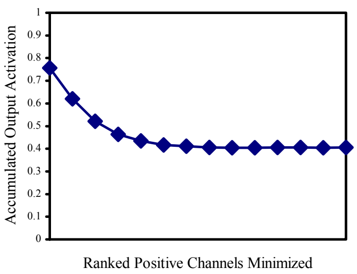

Fig. 1. Highest strain positive channel data relationship

Role of Not Feature Detector Hidden Neurons

In predicting a high strain output the MLP drives the *negative part* of the combined sum as *low* as possible so that the net sum is high enough to meet the maximum target output. The hidden neurons that provide the *negative part* of the combined sum at the output neuron are defined as the *not* feature detector neurons [1–4]. Using sigmoidal activation functions, these are hidden neurons connected with a *negative weight* to the output neuron. For the high strain case, 13 hidden neurons H_3, H_4, H_6, H_7, H_8, H_{10}, H_{11}, H_{12}, H_{14}, H_{15}, H_{16}, H_{17} and H_{18} are not feature detector neurons contributing –0.7760 to the combined sum, as previously stated.

Input Channels Reducing the Negative Signal at the Regression Output Neuron

The MLP drives the negative part of the combined sum at the regression output neuron as *low* as possible by *reducing* the activation of the not hidden feature detectors and by reducing the negative connection weights to the output neuron. For example, in the high strain case the activation of *all* not feature detectors is lower than all the feature detectors.

The reduction of a not feature detector activation is achieved in two ways, as follows.

By maximizing negative input at not feature detectors.

In the high strain case, it is discovered that *all input channels* are connected to the not feature detectors with negative weights. Since the MLP is expected to use *higher* channel values with higher *negative* weights to reduce the not feature detector activation, the channels are ranked by forcing each in turn to its *minimum value* in the training set and presenting the new input vector to the MLP. This is the same ranking procedure as used to find the ranked positive features.

For the highest strain case, the 9 highest ranked input channels reducing the not feature detector activation are discovered to be the same top ranked positive input channels. This demonstrates that the highest ranked input channels play a dual role in activating the hidden feature detectors and in de-activating the not feature detectors. The channel data relationship in this context is very similar to Fig. 1.

By minimizing positive input at not feature detectors

To reduce the not feature detector activation, the MLP is also expected to use *lower* channel values with lower *positive* weights. In this context, for the highest strain case, it is discovered that all input channels are used by the MLP for this purpose *except F3, F14* and *F15*, notably 3 of the top 4 ranked positive channels.

Since the MLP is expected to use *lower* channel values with lower *positive* weights, these channels are ranked by forcing each in turn to its *maximum value* in the training set and presenting the new input vector to the MLP. For the highest strain case, the de-activating channels *in highest ranked order* are found to be F17*, F21*, F16*, F8, F18, F20*, F2*, F12*, F4*, F7, F9, F13*, F5, F6, F1, F11, F19, F10, where starred channels are not also positive channels. However, all these channels have a minimal effect in this context since the output strain value of +0.7560 drops by a maximum value of –1.85% for highest ranked F17*.

In this context, the channel data relationship is found by progressively *maximizing* each of the channels together in ranked order. It is found that the accumulated output decreases smoothly from +0.7560 to a minimum of +0.7192 at F11*. This confirms that the de-activating channels in this context for the highest strain case have minimal effect, unlike results from binary input cases [1–4].

Summary of the Highest Strain Case Results

The analysis of the positive part of the combined input to the strain output neuron has discovered that 13 input channels (out of 21) positively activate 7 hidden layer feature detector neurons, which in turn positively activate the regression output neuron. The channel data relationship shown in Fig. 1 clearly demonstrates that channels F5, F3, F14 and F15 have a *major* control effect (in decreasing order) on the network during the 'pull-up' maneuver, but especially F5 and F3. When these two channels are minimized together, the strain output +0.7560 drops by 67% of the maximum drop in activation. The additional effect of F14 causes the activation to drop by 83%.

INTERPRETING THE LOWEST STRAIN CASE

The training input case with the lowest strain MLP output value of +0.5518 was the low strain case selected for interpretation. This was the training case with the minimum F0 target value of +0.5006 in the training set.

Before commencing the interpretation, it must be emphasised that the regression output of the low strain case is no longer similar to the high output of a classifier MLP, as it was for the high strain output. In this situation the MLP is aiming to drive the regression output as *low* as possible.

The interpretation begins, as before, at the MLP output neuron. The lowest strain value of +0.5518 is sigmoid(+0.2080), where +0.2080 is the combined input sum from hidden neurons and the hidden bias to the single output neuron. The combined sum is made up of a positive part +1.1966 and a negative part –0.9790. This represents a decrease of +0.7199 (37.56%) in positive activation and an increase of –0.2030 (26.16%) in negative activation compared with the high strain case. The hidden bias still contributes –0.0096.

As before, the input channels that contribute to the positive part of the combined input sum at the regression output neuron are discovered first, followed by the input channels that contribute to the negative part of the combined sum.

Discovery of the Hidden Feature Detectors

For the lowest strain case, *exactly the same* 7 hidden neurons H_1, H_2, H_5, H_9, H_{13}, H_{19} and H_{20} are discovered to be feature detector neurons, providing all the positive signal +1.1966 to the combined sum at the regression output.

This was an unexpected result, meaning that the MLP must *decrease* the activation of the feature detectors and *increase* the activation of the not feature detectors to produce the low strain regression output.

Ranking the Hidden Positive Feature Detectors

In the lowest strain case, the highest ranked feature detectors are H_1, H_{19} and H_{13}, exactly the same ranked detectors as for the high strain case. These neurons together provide 96.2% of the total positive input at the strain output neuron, contributing 44.7%, 39.2% and 12.6% respectively. It is of note that now the activation of all hidden neurons is significantly below 0.5. (Feature detector H_{19} has the highest activation of +0.3297.)

Discovery of the Positive Input Channels

For the lowest strain case, the input channels connected to the 7 feature detector neurons with positive weights are *exactly the same* 13 channels as for the high strain case: F1, F3, F5, F6, F7, F8, F9, F10, F11, F14, F15, F18, F19. Again, this was an unexpected result.

Ranking the Positive Input Channels

In the lowest strain case, we have the opposite situation to the highest strain case. Here, the MLP is expected to use channels with *lower* positive input values connected with positive connection weights to decrease the activation of the positive feature detectors, and subsequently the strain output, as *low* as possible to meet the lowest target output.

This means that the positive channels for the lowest strain case are ranked by forcing each positive channel in turn to its *maximum value* in the training set and presenting the new input vector to the MLP. The highest ranked positive channel now produces the largest *increase* in activation at the MLP output neuron, in contrast to the largest *decrease* for the highest strain case (and classification applications [1-4]).

For the lowest strain case the highest ranked positive features are channels F5, F15, F3 and F14 with changes in the MLP output strain value of +23.9%, +14.0%, +11.4% and +6.7% respectively. This is similar to the ranked order of the top 4 channels in the high strain case except now F15 has risen to 2nd place.

The Positive Channel Data Relationship

For the lowest strain case, the data relationship is found by progressively *maximizing* each of the positive channels together in ranked order. The results for the lowest strain case are shown in Fig. 2, where the MLP strain value smoothly and gracefully degrades from a minimum of +0.5518 to a maximum of +0.8342.

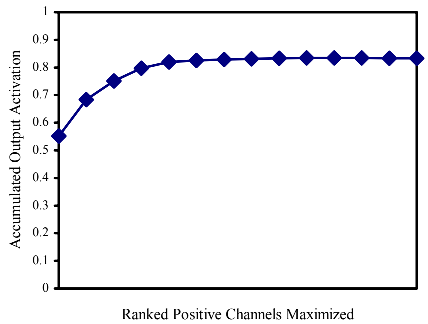

Fig. 2. Lowest strain positive channel data relationship

Input Channels Contributing to the Negative Signal at the Regression Output Neuron

For the low strain case, *exactly the same* 13 hidden neurons as for the high strain case are not feature detector neurons viz: H_3, H_4, H_6, H_7, H_8, H_{10}, H_{11}, H_{12}, H_{14}, H_{15}, H_{16}, H_{17} and H_{18}. These neurons contribute –0.9790 to the combined sum, which represents an increase in negative activation of 26.2%, as previously stated. As explained above, for the lowest strain case, the MLP increases the activation of the not hidden feature detectors so as to reduce the activation of the regression output neuron. This increase is achieved in two ways:

By minimizing negative input at not feature detectors.

To achieve this, the MLP is expected to use lower channel values with lower negative weights. As for the high strain case, it is discovered that *all* input channels are connected to the not feature detectors with negative weights. In this context, channels are ranked by forcing each in turn to its *maximum value* in the training set.

The channels in highest ranked order are exactly the same as the highest ranked positive channels, showing that these channels play a dual role in activating the hidden feature detectors and in de-activating the not feature detectors. The channel data relationship is found by progressively *maximizing* each of the channels together in ranked order. As before, the relationship is very similar to the lowest strain positive relationship shown in Fig. 2.

By maximizing positive input at not feature detectors.

To achieve this, the MLP is expected to use higher channel values with higher positive weights. As for the highest strain case, it is discovered that all input channels *except F3, F14* and *F15* are used for this purpose in the low strain case, notably 3 of the top 4 ranked positive features.

In this context, the channels are ranked by forcing each in turn to its *minimum value* in the training set. The channels in highest ranked order are found to be F17*, F16*, F18, F8, F21*, F4*, F20*, F2*, F12*, F13*, F9, F10, F19, F7, F1, F11, F5, F6, where starred channels are not also positive channels. The output strain value of +0.5518 increases by a maximum value of +2.7%, for channel F17, the highest ranked input channel.

The channel data relationship is found by progressively *minimizing* each of the negative features together in ranked order. It is found that the accumulated output increases smoothly from +0.5518 to a maximum of +0.5882 at F19, confirming that the negative input channels in this context have minimal effect, as for the high strain case.

Summarizing the Lowest Strain Case Results

The analysis of the positive part of the combined input to the strain output neuron has discovered that the same 13 input channels as for the high strain case contribute *all* the positive part of the combined input sum to the strain output neuron. These 13 channels positively activate the same 7 hidden layer positive feature detector neurons as for the high strain case, which in turn positively activate the regression output neuron. It is highly significant that this is *exactly the same result* as for the highest strain case.

The positive channel data relationship graph shown in Fig. 2 clearly demonstrates that channels F5, F15, F3 and F14 have most influence (in decreasing order) on the low strain output, but especially F5 and F15. When these two channels are minimized together, the low strain output +0.5518 increases by +71% of the maximum increase in activation. The additional minimization of F3 causes an increase of +87% in activation.

MLP MECHANISM FOR PREDICTING THE STRAIN

Mechanism for Predicting the Highest Strain

Interpreting the highest strain output has shown that the helicopter MLP uses primarily 3 hidden positive feature detector neurons H_1, H_{13} and H_{19} to drive the *positive* part of the output combined sum as high as possible (to +0.7560) to meet the maximum target output of +0.7957. This is achieved by 13 (out of 21) input channels but primarily by the top ranked channels F5 and F3, as explained below.

The MLP uses the 13 hidden not feature detector neurons H_3, H_4, H_6, H_7, H_8, H_{10}, H_{11}, H_{12}, H_{14}, H_{15}, H_{16}, H_{17} and H_{18} to drive the *negative* part of the combined sum as low as possible to meet the maximum target output. This is achieved by all 21 input channels but primarily by the same top ranked input channels that drive the positive part of the regression output combined sum.

Mechanism for Predicting the Lowest Strain

Interpreting the lowest strain output has shown that the MLP uses the *same* positive feature detectors, the *same* not feature detectors and the *same* top ranked input channels to drive the positive part of the output combined sum as *low* as possible (to +0.5518) to meet the minimum target output of +0.5006.

The same top ranked input channels decrease the positive part of the high strain combined sum by 38% (through the positive feature detectors) and increase the negative part by 26% (through the not feature detectors) to lower the strain output to its minimum value of +0.5518. This is achieved primarily by the top ranked channels F5 and F3, as explained in the following section.

Explaining the Top Ranked Input Channels

For the highest strain case, the ranked order of the top ranked 4 channels is F5, F3, F14, and F15. In this case, the value of F5 is almost at its maximum in the training set, as shown in Table 2. It lies 0.20% of the F5 training set range below the maximum F5 value.

In the lowest strain case, the value of F5 lies 10.3% of the range above the minimum value in the training set. This means that F5 uses 89.5% of the F5 training set max/min range between the highest strain and lowest strain cases. F5 is possibly the highest ranked input channel in the high strain case because it has the *maximum decrease* (+0.1824) of *all* input channels between the highest and lowest strain values. It is of interest that F5 has the highest positive cor- relation (+0.79) with F0 in the training set, as shown in Table 1.

Table 2. Max/min training set values and high/low case values for 4 top ranked input channels

input channel	high case value	low case value	maxi- mum value	minimum value
F3	0.5167	0.3493	0.5285	0.2350
F5	0.5628	0.3804	0.5632	0.3594
F14	0.5403	0.5631	0.7897	0.1835
F15	0.3528	0.3947	0.8131	0.1913

The value of F3 in the high strain case lies 4% of the training set range below its maximum value and lies 38.9% of the training set range above its minimum value in the low strain case. F3 is possibly the second highest ranked input channel in the high strain case because it has the *second* maximum decrease (+0.1674) of all input channels between the highest and lowest strain values.

It is highly significant that for the top 7 ranked channels in the high strain case F5, F3, F6, F19 and F11 have the maximum decrease of all input channels in the *same* descending ranked order between the highest and lowest strain values.

The high strain data relationship in Fig. 1 shows that when F5 and F3 are *together* switched to their minimum training set values the output strain channel F0 value drops from +0.7560 to +0.5208, a value *below* the minimum F0 lowest training set strain value of +0.5518. This indicates that channels F5 and F3 are the *main controlling channels* for the difference in the highest and lowest strain output values.

For the lowest strain case the ranked order of the top ranked 4 channels is F5, F15, F3 and F14. This is the same order of ranking for the highest strain case except for F15, which is now in 2nd place. It is possible that this ranking position is anomalously high due to F15 having the highest increase to its maximum value during the ranking process. This is confirmed by finding the ranked order from a weighted percentage contribution of the positive features to the positive combined input sum to the 3 top ranked positive feature detectors H_1, H_{19} and H_{13}. For both the highest and lowest strain cases the ranked order is then F5, F3, F14, and F15.

SUMMARY AND CONCLUSIONS

The interpretation of the highest and lowest strain cases has shown that the helicopter MLP uses the same mechanism to predict both the high strain and low strain output values. The MLP uses the same hidden feature detector neurons, the same hidden not feature detector neurons and the same top ranked input channels to lower the strain output from a maximum of +0.7560 to a minimum of +0.5518.

The top ranked input channels decrease the positive part of the high strain combined sum by 38% and increase the negative part by 26% to lower the strain output to its minimum value, and vice versa. This is achieved primarily by the top ranked channels F5 and F3 which are the main channels controlling the difference in the high and low strain output values. F5 and F3 are possibly the highest ranked channels because they have the maximum decrease between the highest and lowest strain values.

The analysis has shown that the top ranked input channels have a dual role. As well as driving the positive part of the combined sum at the regression output neuron sufficiently high they also drive the negative part sufficiently low.

It is expected that a MLP with more than one regression output neuron will have a different set of hidden feature detector neurons associated with each regression output. It is expected that the MLP will use a similar mechanism to that observed in this study in predicting each output value.

REFERENCES

[1] Vaughn M.L., Cavill S.J., Taylor S.J., Foy M.A., & Fogg A.J.B. "A full explanation facility for a MLP network that classifies low-back-pain patients and for predicting its reliability". In Abraham A., Koppen M. (eds*),* *Recent Advances in Intelligent Paradigms and Applications*, Physica Verlag, 2002.

[2] Vaughn M.L., Cavill S.J., Taylor S.J., Foy M.A., & Fogg A.J.B. "Direct explanations for the development and use of a multi-layer perceptron network that classifies low-back-pain patients", Special Issue International Journal of Neural Systems, Vol 11 No. 4, 335–347, 2001.

[3] Vaughn M.L. "Derivation of the weight constraints for direct knowledge discovery from the multilayer perceptron network". *Neural Networks*, Vol 12, pp. 1259 – 1271, 1999.

[4] Vaughn M.L., Cavill S.J., Taylor S.J., Foy M.A., Fogg A.J.B. (2000). "Direct explanations and knowledge extraction from a multilayer perceptron network that performs low back pain classification". In S.Wermter and R.Sun (Eds.), *Hybrid Neural Systems*. Springer, pp. 270 – 285.

[5] Craven M.W., & Shavlik J.W. "Using sampling and queries to extract rules from trained neural networks, in machine learning". *Proceedings of the Eleventh International Conference on Machine Learning*, Amherst, MA, USA. Morgan Kaufmann, pp. 73 – 80, 1994.

[6] Tickle A. B.., Andrews, R., Golea, M., & Diederich, J. "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks". *IEEE Transactions on Neural Networks, 9, pp. 1057 – 1068,* 1998.

[7] A.C. Lison. *The prediction of helicopter airframe load spectra from flight parameter data using artificial neural networks*. Cranfield University (RMCS), MSc Dissertation, 1994.