Using neural networks to predict high-risk flight environments from accident and incident data

Maynard, E & Harris, D

DOI 10.1080/10803548.2021.1877455
ISSN 1080-3548
ESSN 2376-9130

Publisher: Taylor and Francis

USING NEURAL NETWORKS TO PREDICT HIGH-RISK FLIGHT ENVIRONMENTS FROM ACCIDENT AND INCIDENT DATA

Elizabeth Maynard and Don Harris*

Institute for Future Cities and Transport

Coventry University

Coventry

CV1 5FB

United Kingdom

don.harris@coventry.ac.uk

+44 (0)24 7765 8052

* Author for correspondence

KEYWORDS

Neural Networks, Flight Safety, Flight Risk, Modelling, Safety Management

ABSTRACT

Pre-flight risk analysis tools (FRATs) aid pilots in evaluating risk arising from the flight environment. Current FRATs are subjective, based on linear analyses and subject matter expert interpretation of flight-factor/risk relationships. However, a 'flight system' is complex with non-linear relationships between variables and emergent outcomes. A neural network (NN) was trained to categorize high and low risk flight environments from factors such as the weather and pilot experience using data extracted from accident and incident reports. Negative outcomes were used as markers of risk level, with low severity outcomes representing low-risk environments, and high severity outcomes representing high-risk. Eighteen models with varied architectures were created and evaluated for convergence, generalization and stability. Classification results of the highest performing model indicated that NNs have the ability to learn and generalize to unseen accident and incident data, suggesting that they have the potential to offer an alternative to current risk analysis methods.

1 INTRODUCTION

In the aviation industry risk management has become increasingly formalized and regulated. Many U.S. Part 91 operators (covered by general operating and flight rules) and Part 135 operators (companies operating commuter and on-demand services) have voluntarily established risk assessment procedures [1]. In 2012, the International Civil Aviation Organization (ICAO) implemented Annex 19, requiring member states to mandate for the establishment of data-driven risk management processes for international operators of large and turbojet aircraft [2]. In 2014, the Federal Aviation Administration (FAA) implemented 14 CFR (Code of Federal Regulations) part 135.617 requiring helicopter air ambulance pilots to conduct formal pre-flight risk assessments [3]. In 2015, the FAA added part 5 to 14 CFR applicable to US part 121 (airline) and international GA (General Aviation) operators [4].

Current Flight Risk Assessment Tools (FRATs), such as that developed by the FAA [5] are typically based upon an SME (Subject Matter Expert) risk and hazard analysis of accident data and compute a 'risk number' from a linear combination of 'risk values' assigned to flight-environment (FE) factors related to the aircraft, aircrew, resource and atmospheric states expected during the flight. However, aviation operations are complex, with emergent outcomes arising from inter-related variables with relationships that are likely to be non-linear, and neither completely deterministic nor completely random [6]. As a result, the use of linear methods may not be the most appropriate approach.

In a comparative study of maintenance operations, linear and non-linear analysis methods were applied to the prediction of safety outcomes collected over 6.2 years [7]. While Poisson regression (linear) completely failed to predict safety outcomes, Neural Networks (NNs) (non-linear) were significantly more accurate. This suggests that the analysis method should reflect the properties of the system being analysed and that linear methods '*may be totally inappropriate if the underlying mechanism is nonlinear*' (p. 36) [8]. Therefore, for the prediction of flight risk, an emergent property of a complex system, it is expected that a non-linear analysis method would be more successful.

Aviation risk-analysis research is dominated by methods dependent on SME knowledge, often incorporated into Bayesian models. However, expert knowledge of highly complex systems is limited since threat-consequence relationships can be ambiguous [9-11] and unknown or unexpected relationships cannot be accounted for [11]. Furthermore, while Bayesian methods may result in high performing models if the prior probability is based on population statistics, in practice it is typically calculated from finite samples and/or SME knowledge. Therefore, the applicability of Bayesian networks is limited [12]. Fuzzy logic has also been applied to the analysis of flight environment factors to assess risk [e.g. 13, 14]. However, in concluding his discussion of fuzzy-logic methods Hadjimichael [13, p. 6516] stated, '*a more robust method of determining the "most causal" risk factors is necessary. This is a complex issue, as finding a meaningful and useful definition of "most causal" is a significant research challenge*'.

In contrast, NNs do not require prior assumptions or expert knowledge of a system. A NN performs a holistic analysis of historical data to determine an overall relationship between multiple inputs and multiple outputs. NNs are based upon a simplified model of human neural networks, containing nodes, which represent input and output values (derived from a learning dataset), and 'hidden' layers which allow complex, non-linear relationships to be modelled. The essential feature of NNs is that they learn the relationship(s) between variables in a building process, and self-correct. They are trained by exposure to historical data presented in a supervised leaning set, with known inputs and outputs. Once data from the supervised learning set has passed through the model, the error between actual and expected output is calculated and the weights in the model are revised iteratively through a back-propagation process until the solution converges and the overall error rate falls below a pre-defined criterion. To validate the NN, the derived model is then supplied with input data from an unseen (hold-out) data set and the output predictions from the network are compared to the known actual outputs [15, 16].

In recent years NNs have emerged as a data mining technique for exploring complex relationships in large datasets [e.g. 16, 17]. NNs allow the simultaneous prediction of multiple outcomes from multiple inputs, and hence are ideal for predicting flight risk from FE-factors. They are particularly well suited to applications with noisy, missing, overlapping, non-linear and non-continuous data [18] and can also handle highly unstructured data typical of that derived from accident and incident reports. They provide a means of building an empirically describable and verifiable model, which allows for the incorporation of contextual information [19].

NNs have been applied to a diverse range of classification problems such as the stock market [20] and cancer survival [21]. In the transportation field, they have been successfully applied to specific complex prediction problems. As a decision-making tool for aircraft safety inspectors [22] evaluated the ability of a hybrid two-stage NN to analyse the relationships between aircraft operation and maintenance data, and service difficulty reporting (SDR) profiles. When compared with actual SDR numbers, 13 out of 19 NN models developed had R^2 values above 0.80. Classifications of marine accidents based on river stage, traffic level, utilizations, location, weather, and time were performed with 80% accuracy [23]. Predictions of the landing speed of McDonnell-Douglas MD80 aircraft in no/low-gust and high gust conditions based on airport topology, the environment, and flight and aircraft parameters were 95% correct [24]. Predictions of pilot decision-making in the resolution of a disruptive passenger incident [16] were 100% correct. Liu et al. [25] developed a NN to predict fatal accidents in GA operations from an analysis of FE-type factors and achieved a classification accuracy of over 78%. Harris & Li [26] developed a NN model based upon the theoretical model of error causation underpinning the Human Factors Analysis and Classification System [27, 28]. It was found that 74% of unsafe acts (errors) implicated in 523 military aviation accidents could be correctly predicted from their preconditions.

The aim of this study was to investigate the ability of a NN to classify flight environments according to level of risk. Since risk itself is '*multidimensional and nuanced'* (p. 1647) [29] and therefore difficult to define, accidents and incidents – classified as having high and low severity outcomes – were used as markers of risk level. The NN was trained to predict outcome severity using data extracted from historic accident and incident reports. Model development and training was based on the four-stage process described by Diallo [24]: selecting and defining the input and output nodes; data source selection and coding; building the model and model evaluation.

2 Method

2.1 Stage 1: Selecting and Defining the Input and Output Nodes

An initial set of 37 input variables was derived from the FAA's FRAT [5]; Aviation Safety Reporting System (ASRS) coding taxonomy [30]; the International Civil Aviation Organization - ICAO (1993) Human Factors Checklist [31]; and the SHEL(L) (Software-Hardware-Environment Liveware model [31, 32]. After a review of the National Transportation Safety Board (NTSB) and ASRS reports two further variables were included; 'High altitude' and 'busy/complex airspace', both of which were frequently present at the time of the accident or incident. After initial coding the number of variables was subsequently reduced as a result of missing data and infrequently occurring factors resulting in 29 input variables (see Table 1).

Commonly used FRATs typically categorize flight risk at two levels: high-risk that requires threat mitigation and management consultation, and low-risk that does not. Therefore, the outputs of the NN were chosen to similarly categorize flight outcomes at two levels. High-severity outcomes (moderate, major and catastrophic events) were considered ~~as~~ markers of high-risk environments; low-severity outcomes (negligible and minor events) were attributed to low-risk environments. These outputs were defined by a combination of the levels of severity definitions found in the ICAO [2] Safety Risk Severity Table and Example Severity Table (see Table 2).

Table 1 Definition of input variables

Inputs	**Operational Definition**
Single pilot	*Operations in which a second pilot is not present* *Operations in which a second pilot is present, but not trained or qualified in accident/incident aircraft type*
Inexperienced First Officer	*First Officer with less than 200 hours in accident/incident aircraft type*
Captain less than 200hrs in type	*Captain with less than 200 hours in accident/incident aircraft type*
Captain less than 100hrs in 90 days	*Captain with less than 100 hours in the 90 days prior to the accident/incident in any aircraft*
Low light/night	*Night, dawn, dusk or low light reported for entire or part of operation*
Weather below VFR	*Ceilings 3000ft* [915m] *or less and/or visibility 5 statute miles* [8 Km] *or less at either departure or destination airport*
Surface winds greater than 30kts [15.4 ms^{-1}]	*Winds (including gusts) reported greater than 30kts* [15.4 ms^{-1}] *at departure or destination at time of accident/incident*
Crosswinds greater than 15kts [7.7 ms^{-1}]	*Crosswinds (including gusts) reported greater than 15kts* [7.7 ms^{-1}] *at departure or destination at time of accident/incident*
Heavy rain	*Heavy rain reported at or in the vicinity of departure or destination airport*
Runway surface condition	*Runway surface other than dry concrete*
Uncontrolled airport	*No operating control tower at departure or destination airport*
No weather reporting	*No weather reporting system located at departure or destination airport*
Short runway	*Runway shorter than 5000ft* [1524 m] *Runway reported as short for aircraft type*

Airport elevation greater than 5000ft [1524 m]	*Airport elevation greater than 5000ft* [1524 m] *MSL*
Terrain	*Airports defined as mountainous airports per ICAO Doc 8168* *Rising terrain reported as a factor in the accident/incident* *Operations involving arrival or departure procedures over terrain with elevation changes of 3000ft* [915m] *or greater from airport elevation*
High altitude flight	*Operations at altitudes higher than Flight Level 250* [*Circa*. 7,620m]
Busy/complex airspace	*Flight in or under Class B airspace* *Flight at airports with more than two runways* *Reported high density traffic*
Non-precision approach	*Approach used does not have a glideslope*
Circling approach	*Approach used is not aligned with landing runway*
Visual approach	*No instrument approach is available at destination airport* *Instrument approach is available but pilots are cleared a visual approach in VMC conditions*
Icing conditions	*Reported flight conditions involve IMC flight and temperatures 10°C and below* *Ground conditions require de-icing and/or anti-icing prior to take-off* *Use of aircraft anti-icing and de-icing systems is reported*
Thunderstorms/wind shear	*Reported thunderstorms and/or wind shear* *Thunderstorms and/or wind shear reported on the METAR at departure or destination*
Turbulence	*Reported turbulence*
Reposition flight	*Operations, other than personal flights, that do not involve the carriage of revenue passengers or cargo*
International	*Flights that either originate or terminate outside the contiguous US, Hawaii or Alaska*
Training	*Flights involving the training of flight crew* *Initial operating experience* *Check rides and line checks*

Maintenance factors	*Deferred maintenance items on aircraft* *First flight of aircraft after maintenance has been performed* *Maintenance test or functional check flight*
Time pressure	*Flights involving a delay* *Medevac operations* *Time pressure reported by pilot*
Personal factors	*Reported sleep depravation* *Reported pilot personal stress or inter-crew stress* *Flight time greater than 5 hours in duty period* *Duty period greater than 12 hours*

* Note: All units are initially expressed as ICAO standard aeronautical units of measure; equivalent SI values are given in parentheses. IMC = Instrument Meteorological Conditions: METAR = Meteorological Aerodrome Report: MSL = Mean Sea Level: VMC = Visual Meteorological Conditions.

2.1 Stage 2: Data Source Selection and Coding

The performance of a NN depends on the input and output combinations contained in the training data [33]. Therefore, to predict flight-outcome severity, the data needed to encompass a range of events from insignificant incidents to catastrophic accidents. NTSB investigations involve accidents having substantial consequences for participants and equipment (aircraft and property damage, casualties and fatalities). Events reported to the ASRS programme typically involve minor incidents (e.g. altitude and air space deviations) which, although undesirable generally result in little or no consequence. Therefore, the appropriate range of data were obtained from a combination of NTSB and ASRS reports extracted from their online databases.

Table 2 Output variables and definitions taken from the Example Severity Table (ICAO, p2-App 2-3) and the Safety Risk Severity Table (ICAO, p2-29) [2]. Outcome severity is included in parentheses.

Outputs	Initial categorization	ICAO Definitions	
Low Severity Consequence	Insignificant (0.2)	*No significance to aircraft-related operational safety*	***Negligible*** *Few consequences*
	Minor (0.4)	*Degrades or affects normal aircraft operational procedures or performance*	***Nuisance*** *Operating limitations* *Use of emergency procedures* *Minor incident*
High Severity Consequence	Moderate (0.6)	*Partial loss of significant/major aircraft systems or results in abnormal application of flight operations procedures*	***Major*** *Significant reduction in safety margins, a reduction in the ability of the operators to cope with adverse operating conditions as a result of an increase in workload or as a result of conditions impairing their efficiency* *Serious incident* *Injury to persons*
	Major (0.8)	*Complete failure of significant/major aircraft systems or results in emergency application of flight operations procedures*	***Hazardous*** *A large reduction in safety margins, physical distress or a workload such that the operators cannot be relied upon to perform their tasks accurately or completely* *Serious injury* *Major equipment damage*
	Catastrophic (1.0)	*Loss of aircraft or lives*	***Catastrophic*** *Equipment destroyed* *Multiple deaths*

* Note: ICAO = International Civil Aviation Organization

This study focused on U.S. 14 CFR Part 135 (scheduled and non-scheduled) and Part 91 (general) operations involving two-engine turbofan, turbojet and turboprop airplanes certified under U.S. 14 CFR Part 25 (light and medium transport only) and Part 23 (normal category airplanes) [4]. Narrowing the focus to similar types of operation had the advantage of making certain factors approximately constant (e.g. level of training), thus reducing the number of input variables. Report selection was limited to these categories. Reports were rejected if they involved:

- Illegal activity e.g. use of illicit drugs.
- System failures whose cause or handling was not affected by factors encountered between engine start and engine shutdown.
- Military, government, skydiving and low altitude operations.
- Experimental aircraft.
- A Captain who did not hold a Commercial Pilot or Air Transport Pilot certificate.

The data extraction and coding process was undertaken in two steps. In step one all initial 39 input factors were coded. Categorical factors, e.g. single pilot, were recorded using a binary code with '1' indicating its presence and '0' indicating its absence. For scale data, e.g. runway length, the value was recorded. Outcome severity was coded in 0.2 increments, with 0.2 indicating an insignificant outcome and 1.0 indicating a catastrophic outcome (see Table 2). Depending on the nature of the event, the most appropriate of column one or column two of table 2 was used to determine outcome severity. Where necessary the NTSB and ASRS reports were supplemented with airport, terrain, navigation and weather data from SkyVector [34], AirNav [35] and Ogimet [36].

While NTSB reports include information describing the entire flight, ASRS reports are de-identified and typically describe only the circumstances surrounding the incident itself. Therefore, they frequently lack data such as departure and destination airport information and pilot flight time. Additionally, both NTSB and ASRS reports yielded low numbers of personal aircrew factors, such as fatigue and stress. To address these issues:

- Departure and destination airport variables selected in Stage 1 were combined and re-defined as a single 'airport' variable (e.g. uncontrolled airport) which was applied to either airport, provided the missing airport data was not directly associated with the accident or incident. With the exception of reports concerning en-route events, if both the departure and destination airport information was missing, the report was rejected.
- For reports concerning en-route incidents containing neither departure nor destination information and for which the airports were not directly associated with the incident, all airport factors were coded as 'absent'. This allowed the inclusion of high altitude and en-route incidents.
- 'First Officer (FO) less than 200 hours in type' and 'FO less than 100 hours in 90 days' were combined into a single variable 'Inexperienced FO'. If FO hours were not reported, this factor was coded as absent unless the narrative indicated the FO was inexperienced.

- Flight time accrued during the duty period, sleep deprivation, personal stress and intra-crew stress were combined into a single ‘Personal factors’ input.

Scalar values were then re-coded as binary categorical factors according to the discriminator indicated by its definition. The output factors were also re-coded as binary categorical factors, with 0.5 discriminating between low and high severity outcomes.

A total of 467 reports meeting the operations and airplane criteria were retrieved from the NTSB database, involving events that occurred between January 2005 and December 2015. Of these, 206 reports were rejected. A further 1,614 reports meeting the operations criteria were retrieved from the ASRS database, involving events that occurred between January 2010 and December 2015. Of these, 1,220 reports were rejected. As a result, a total of 600 reports were coded and analysed.

2.3 Stage 3: Building the Model

The Neural Network analysis tools in IBM SPSS version 24 were used to build and evaluate multiple models.

Severity of outcome was entered as the dependent variable. The data set was split into three partitions: ‘training’, ‘testing’ and ‘holdout’. Accident/incident data sets were randomly assigned to each partition, however the same ratio of NTSB to ASRS reports was maintained in each partition (see Table 3).

Table 3 Assignment of data sets

Partition	Proportion of data (%)	Number of reports	
		NTSB	ASRS
Training	60.2	157	203
Testing	29.8	78	102
Holdout	10	26	34

* Note: ASRS – Aviation Safety Reporting System; NTSB = National Transportation Safety Board

Selecting the initial parameters for a NN architecture (such as the number of hidden layers and/or the number of nodes in each hidden layer) has been described as both '*an empirical procedure with guidelines*' (p.5) [7] and '*more of an art than a science*' (p. 42) [8]. While rules-of-thumb guide NN design, the optimum combination of layers, nodes, activation functions and training settings for a particular problem is generally determined through trial and error. Therefore, multiple models were developed using either the automatic architecture option or the custom selections. Models designed using the custom options varied in choice of activation function of the output layer (identity, softmax, hyperbolic tangent or sigmoid) and the number of nodes in the hidden layer (5, 14, 29). The hyperbolic tangent function was selected as the activation function for the hidden nodes for all models.

The gradient-descent training algorithm was used for all models. The maximum number of steps without an error was set at three. The minimum relative change in training error was set to 0.0001 and the training error ratio was 0.001.

Each NN model analysed the data ten times, five times with the inputs ordered as shown in Table 1, and five times with the inputs reversed. The mean, median and standard deviation for the percentage of correct predictions were calculated from the 10 runs of each model. Mean importance values for each variable were calculated from the values reported for each run.

The best performing model was selected based on a combination of highest mean accuracies for the hold-out sample, lowest standard deviations and greatest area under the ROC (Receiver Operator Characteristic) curve.

3 RESULTS

Twenty NN models were created. When tested on the hold-out partition, fifteen models had average classification accuracies of over 60% for both low and high severity outcomes. Three models had accuracies of over 65%. The highest performing model predicted the outcomes of unseen data with an average accuracy of 69.8% (SD = 2.8%) for high-severity outcomes, and 65.5% (SD = 8.4%) for low severity outcomes with the highest performing run having a prediction accuracy of 71.1% for high severity outcomes and 77.3% for low severity outcomes. The mean area under the ROC curves was 0.843 for both low and high severity outcomes.

The average classification results of this model and the results of the highest performing (HP) run are shown in Tables 4 and 5. The architecture of the highest performing model had 58 nodes in the input layer; 29 nodes in the hidden layer and two nodes in the output layer. Mean variable rankings, importance values and normalized mean importance values can be found in Table 6.

Table 4 Average classification accuracy for the highest performing model.

		Classification Accuracy	
Data Sample	**Flight-outcome (severity)**	**Mean (%)**	**Standard Deviation (%)**
Training	Low	75.3	3.6
	High	77.4	1.6
Testing	Low	75.1	4.7
	High	75.1	4.1
Hold-out	Low	65.5	8.4
	High	69.8	2.8

Table 5 Classification accuracy of the highest performing run of the highest performing model.

Sample			Predicted: Low Severity	Predicted: High Severity	Percentage Correct
Training	Observed	Low Severity	131	39	77.1
		High Severity	45	146	76.4
		Overall			76.7
Testing	Observed	Low Severity	55	17	76.4
		High Severity	28	79	73.8
		Overall			74.9
Holdout	Observed	Low Severity	17	5	77.3
		High Severity	11	27	71.1
		Overall			73.3

Table 6 Variable importance values

Variable	Mean Importance	Standard Deviation	Normalized Mean Importance
Busy/complex airspace	0.059	0.012	100.0
Uncontrolled airport	0.055	0.012	93.9
Visual approach	0.054	0.008	91.2
Runway surface condition	0.049	0.013	82.4
Maintenance factors	0.049	0.017	82.2
Single pilot	0.046	0.010	77.3
Surface winds greater than 30kts [15.4 ms^{-1}]	0.045	0.017	75.4
Circling approach	0.040	0.016	66.9
Thunderstorms/wind shear	0.039	0.014	65.4
Weather below Visual Flight Rules minima	0.036	0.012	60.7
Training	0.035	0.013	59.7
Personal factors	0.035	0.017	59.0
Crosswinds greater than 15kts [7.7 ms^{-1}]	0.035	0.007	58.8
Non-precision approach	0.034	0.010	57.8
Inexperienced First Officer	0.033	0.007	55.6
Short runway	0.031	0.013	53.1
Icing conditions	0.030	0.009	50.8
Heavy rain	0.030	0.005	50.5
Airport elevation greater than 5000ft [1524 m]	0.030	0.011	50.0
Time pressure	0.029	0.009	49.2
High altitude flight	0.026	0.005	44.7
International	0.026	0.005	43.2
Reposition flight	0.025	0.008	41.7
Captain less than 200hrs in type	0.024	0.007	40.8
Turbulence	0.024	0.007	39.8
No weather reporting	0.023	0.007	39.5
Terrain	0.023	0.007	38.6
Low light/night	0.020	0.003	34.4
Captain less than 100hrs in 90 days	0.019	0.002	31.5

* Note: All units are initially expressed as ICAO standard aeronautical units of measure; equivalent SI values are given in parentheses.

4 DISCUSSION

Adya & Colopy [37] suggest a 'good' prediction model should demonstrate good training sample performance, be able to accurately predict outcomes from unseen data (good generalization) and should be stable (produce consistent results). When measured against these standards, the results indicate NNs have the potential for becoming 'good' models for categorising flight environments. The training and test results of the highest performing model show that the model was able to learn, and the hold-out accuracy of the highest performing run suggests that the model has the potential to generalize well to unseen data (see Table 6). The areas under the ROC curves indicate the NN has a better than 0.8 probability of correctly categorising a random event and is, therefore, a 'good' discriminator of low and high-risk flight environments. However, the standard deviations and hold-out results suggest that the model may be sensitive to data-set and variable input order, particularly for low-severity outcomes but less so for more critical high-severity outcomes.

Colopy, Adya & Armstrong [38] suggest that a NN should be validated by a comparison of out-of-sample performance with other well-accepted models. However, in this case, selecting such models against which to make an informative comparison proved to be problematic. Typical pre-flight risk analyses (e.g. those similar to the FAA's FRAT) are generally tailored to the specific needs of an organisation. They are highly subjective, depending on the choice of inputs, the scores assigned, and the operators' tolerance of risk. Additionally, there is no universally accepted measure of high and low risk.

However, A comparison with the GA-fatalities NN, although limited by fundamental differences such as type of aviation operation and inputs analysed, was possible to some degree. While the average training and test results for the highest performing model were similar to the accuracy achieved by Liu et al. [25] the average hold-out accuracy was a little less favourable. This may be explained though by the broader range of causal factors analysed by the GA network which, while perhaps being good indicators of risk (e.g. phase of flight), would not be suitable for pre-flight risk analysis.

With no well-accepted risk assessment model against which to validate the NN, the results were bench marked against the prediction accuracies of two other complex systems, the weather and the stock market. A NN developed to predict the height of the 500hPa pressure level in the northern hemisphere achieved an average monthly prediction accuracy of between 0.70 and 0.95 [39]. In contrast, the stock market, which is highly reactive to human emotions and decisions, has an average prediction accuracy of between 50.8% and 65.79% [20]. These studies indicate first that the ability to predict the outcomes of any complex system is limited, and second, the more it involves human agents, the less predictable it gets. However~~,~~ research also indicates that this effect may be moderated by agent training and the resolution of the analysis. A NN applied to a single aviation decision made by highly trained airline pilots predicted outcomes with 100% accuracy [16]. In contrast the GA-fatalities NN [25] applied to a more macroscopic element of the aviation system involving pilots who were likely, in general, less well trained, achieved only 78.9% accuracy. These results imply that, for systems involving human agents, higher resolution (more specific problems), better training and less freedom of choice increases predictability.

The flight environment involves a mix of human, machine, resources and the atmosphere. The predictability of these components runs on a continuum from highly predicable (the machine), to somewhat predictable (the atmosphere and resources) to less predictable (the human). However, the humans in the professional aviation system ~~are~~ (pilots, air traffic controllers, maintainers, etc.) are well trained and their range of choices is restrained by regulation. Against this context, the results obtained in this study, ~~are~~ better than stock market prediction but inferior to the prediction of weather events, ~~hence~~ are considered reasonable for an initial study of a NN flight system analysis. However, it is acknowledged that, before a NN can be applied to real-world pre-flight risk assessment tasks, where high-risk “misses” could have significant negative consequences, further refinements are required.

5 RECOMMENDATIONS AND CONCLUSION

Given the ambiguous nature of risk and the complexity of the flight system, the development and validation of accurate and meaningful pre-flight risk analysis tools is challenging. However, the findings of this exploratory study indicate that NNs are both suited to and have the potential for such a task. The classification accuracies, although limited by system complexity, are commensurate with those of other complex systems and similar research.

In addition to categorising flight environments, NNs potentially offer a means of identifying the most significant predictors of flight risk through a sensitivity analysis of the variables (see table 6). The trends, similarities and differences of the variable rankings found in this study and between other studies may provide insight into the underlying features of the aviation system.

It is recommended that future research be directed down three paths. The first aims to improve the convergence and stability of the NN by addressing the variable, missing data, data source, and coding problems. The current research can only be considered as a proof of concept - the current data set is quite small. As a result, the second recommendation is to expand the dataset to include accident and incident reports from other countries which can be used in the process of training, validation and testing of NN. Building on this, the third recommendation focuses on improving accuracy and generalization through NN design, allowing a more inclusive analysis of the flight system encompassing supervision and organisational elements. Suggestions for such methods include a 'zooming out' approach involving the analysis of individual system elements through multiple and multi-stage NNs and the output of risk indicators that feed forward into a flight-risk analysis (combining the methodologies applied by Luxhøj & Williams [22] and Hsiao et al. [7]). It is expected that progress made through such research could lead towards the development of functional, objective, and informative flight-risk analysis tools to aid pilot decision-making.

REFERENCES

1. Cory C. Building a Viable Flight Risk Assessment Process in Business Jet Operations: Selecting a Risk Assessment Tool, Setting Baselines, Trigger and Mitigation Points. Collegiate Aviation Review. 2013;13(2):39-57.
2. Safety Management Manual Doc 9859. 3rd Ed. International Civil Aviation Organization, 2012. Available from: https://www.icao.int/sam/documents/rst-smsssp-13/smm_3rd_ed_advance.pdf. [Accessed 2 January 2021].
3. Federal Register, Volume 79, No.25: Helicopter Air Ambulance, Commercial Helicopter, and Part 91 Helicopter Operations; Final Rule. Federal Aviation Administration, 2014. Available from: https://www.gpo.gov/fdsys/pkg/FR-2014-02-21/pdf/2014-03689.pdf. [Accessed 2 January 2021].
4. Electronic Code of Federal Regulations, Title 14, Chapter I, Subchapter A, Part 5. Federal Aviation Administration, 2017. Available from: http://www.ecfr.gov/cgi-bin/text-idx?SID=31a1ff34c5fb23bfab62b105fa038ceb&mc=true&node=pt14.1.5&rgn=div5. [Accessed 2 January 2021].
5. Flight Risk Assessment Tool (InFO 07015). Federal Aviation Administration, 2007. Available from: https://www.faa.gov/other_visit/aviation_industry/airline_operators/airline_safety/info/all_infos/media/2007/info07015.pdf. [Accessed 2 January 2021].
6. Sayama H. Introduction to the Modeling and Analysis of Complex Systems. Geneseo, NY: Open SUNY Textbooks; 2015.

7. Hsiao Y, Drury C, Wu, C. et al. Predictive models of safety based on audit findings: Part 1: Model development and reliability. Appl Ergon 2013;44(2):261-273.

8. Zhang G, Patuwo E, Hu MY. Forecasting with Artificial Neural Networks: The State of the Art. Int J Forecast. 1998;14(1):35-62.

9. Macrae C. Making risks visible: Identifying and interpreting threats to airline flight safety. J Occup Organ Psychol. 2009;82(2):273-293.

10. Booker P. Experts, Bayesian Belief Networks, rare events and aviation risk estimates. Saf Sci. 2011;49(8-9):1142-1155.

11. Sackman H. Delphi Assessment: Expert Opinion, Forecasting and Group Process Report no. AD-786 878. Santa Monica, CA: Rand Corporation; 1974.

12. Weiss SM, Kulikowski CA. Computers Systems that Learn. San Mateo, CA: Morgan Kaufmann Publishers, Inc.; 1991.

13. Hadjimichael M. A fuzzy expert system for aviation risk assessment. Expert Syst Appl. 2009; 36(3/2):6512-6519.

14. Cheng CB, Shyur HJ, Kuo YS. Implementation of a Flight Operations Risk Assessment System and Identification of Critical Risk Factors. Sci Iran. 2014;21(6): 2387-2398.

15. Garson DG. Neural Networks: An Introductory Guide for Social Scientists. London; Sage: 1998.

16. Duggan SJ, Harris D. Modelling Naturalistic Decision Making using an Artificial Neural Network: Pilot's responses to a disruptive passenger incident. Hum Fac Aero Saf. 2001;1(2): 145-166.

17. Hair JF, Anderson RE, Tatham RL, et al. Multivariate Data Analysis. 5th Ed. Upper Saddle River, NJ: Prentice-Hall; 1998.

18. Moore B. 1988. ART1 and Pattern Clustering. In: D Touretzky, G Hinton, T Sejnowski editors. Proceedings of the 1988 Connectionist Model Summer School. San Mateo CA: Morgan Kaufmann; 1998. p. 174-185.

19. Haykin S. Neural Networks: A Comprehensive Foundation. 2nd Ed. Upper Saddle River NJ: Prentice Hall: 1999.

20. Kim K, Han I. Genetic Algorithms Approach to Feature Discretization in Artificial Neural Networks for the Prediction of Stock Price Index. Expert Syst Appl. 2000;19(2):125-132.

21. Burke HB, Goodman PH, Rosen DB et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer. 1997;79(4):857-862.

22. Luxhøj JT, Williams TP. Integrated decision support for aviation safety inspectors. Finite Elem Anal Des. 1996;23(2-4):381-403.

23. Hashemi RR, Le Bland LA, Rucks CT, et al. A Neural Network for Transportation Safety Modeling. Expert Syst Appl. 1995;9(3):247-256.

24. Diallo ON. A predictive aircraft landing speed model using neural network. Paper presented at: 31st Digital Avionics Systems Conference; 2012 14-18 October. NASA, Moffett Field, CA.

25. Liu D, Nickens T, Hardy L, et al. Effect of HFACS and Non-HFACS-Related Factors on Fatalities in General Aviation Accidents Using Neural Networks. Int J Aviat Psychol. 2013;23(2):153-168.

26. Harris D, Li W-C. Using Neural Networks to Predict HFACS Unsafe Acts from their Pre-Conditions. Ergonomics. 2019;62(2):181-191

27. Shappell SA, Wiegmann DA. The Human Factors Analysis and Classification System–HFACS (DOT/FAA/AM-00/7). Washington DC: FAA, Office of Aviation Medicine; 2000.

28. Wiegmann DA, Shappell SA. A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System. Aldershot, UK: Ashgate; 2003.

29. Haimes YY. On the Complex Definition of Risk: A Systems-Based Approach. Risk Anal. 2009;29(12):1647-1654.

30. ASRS Coding Taxonomy. National Aeronautics and Space Administration, n.d. Available from: https://asrs.arc.nasa.gov/search/dbol/databasecoding.html. [Accessed 6 January 2021].

31. Human Factors Digest No. 7, Circular 240-AN/144. International Civil Aviation Organization, 1993. Available from: http://www.skybrary.aero/bookshelf/books/2037.pdf. [Accessed 6 January 2021].

32. Edwards E. Man and Machine: Systems for Safety. In: Proceedings of British Airline Pilots Association Technical Symposium. London: British Airline Pilots Association; 1972. (p. 21-36).

33. Ung ST, Williams V, Bonsall S, et al. Test case based risk predictions using artificial neural network. J Saf Res. 2006;37(3):245-260.

34. SkyVector Aeronautical Charts. Skyvector, 2017. Available from: www.skyvector.com [Accessed 6 January 2021].

35. AirNav.com: The Pilots Window into a World of Aviation Information. AirNav, n.d. Available from: http://www.airnav.com [Accessed 6 January 2021].

36. Ogimet: Professional Information about Meteorological Conditions in the World. 2017. Available from: http://www.ogimet.com/metars.phtml.en [Accessed 6 January 2021].

37. Adya M, Collopy F. How Effective are Neural Networks at Forecasting and Prediction? A Review and Evaluation. J Forecast. 1998;17(5-6):481-498.

38. Collopy F, Adya M, Armstrong JS. Principles for Examining Predictive Validity: The Case of Information Systems Spending Forecasts. Inf Syst Res. 1994;5(2):170-179.

39. NCEP/EMC Global Model Experimental Forecast Performance Statistics. 2017. Available from: http://www.emc.ncep.noaa.gov/gmb/STATS_vsdb [Accessed 6 January 2021].