

# COCKPIT CREW SAFETY PERFORMANCE PREDICTION BASED ON THE INTEGRATED MACHINE LEARNING MULTI-CLASS CLASSIFICATION MODELS AND MARKOV CHAIN

Naimeh BORJALILU , Fariborz JOLAI , Mahdieh TAVAKOLI 

*School of Industrial Engineering, College of Engineering, University of Tehran, Iran*


## Article History:

received 9 September 2022

accepted 25 January 2023

**Abstract.** The main tool of cockpit crew performance evaluation is the recorded flight data used for flight operations safety improvement since all certified airlines require implementation of a safety and quality management system. The safety performance of a flight has been a challenging issue in the aviation industry and plays an important role to acquire competitive benefits. In this study, an integrated multi-class classification machine learning models and Markov chain were developed for cockpit crew performance evaluation during their flights. At the outset, the main features related to a flight are identified based on the literature review, flight operations expert's statements, and the case study dataset (as numerical example). Afterwards, the flights' performance is evaluated as a target column based on four multi-class classification models (Decision Tree, Support Vector Machine, Neural Network, and Random Forest). The results showed that the random forest classifier has the greatest value in all evaluation metrics (i.e., accuracy = 0.90, precision = 0.91, recall = 0.97, and F1-score = 0.93). Therefore, this model can be used by the airline companies to predict flight crew performance before the flight in order to prevent or decrease flight safety risks.

**Keywords:** cockpit crew performance, flight evaluation prediction, multi-class classification, Markov chain, safety management system.

 Corresponding author. E-mail: [borjalilu@ut.ac.ir](mailto:borjalilu@ut.ac.ir)

## Introduction

In the aviation industry, serious incidents or accidents often occur, which have severe consequences. Furthermore, in the aviation industry, suitable strategies have successfully decreased the rate of accidents and incidents. The International Air Transport Association (IATA) has tried to analyze flight data for eventful flights monitoring and emphasizing the scopes of flight safety concerns (Rey et al., 2021). However, Flight Data Monitoring (FDM) has been used as a systematic, proactive and non-punitive tool to improve flight safety. By using the FDM system, the operator can make a comparison between their Standard Operating Procedures (SOPs) and the Pilot's flight activities. Aircraft Flight data in all phases of flight are driven and analyzed to identify and solve potential safety problems (Lan et al., 2012). Hence, FDA can be employed to detect non-standard or non-adherence procedures and weaknesses in aircraft performance (Wang et al., 2014). All aspects of a flight such as take-off, climb, cruise, descent, approach, landing, and finally cockpit crew performance can be monitored through FDA (International Civil Aviation Organization, 2013). Flight operations can be interrupted due to several scenarios; performance of cockpit crew members or flight safety issues such as fatigue or

any emergency condition. It is obvious that the aviation industry lacks a standardized, practical and an easily replicable protocol to consider the cockpit crew safety issues which can enhance the management of the cockpit crew safety performance. The aims of this study is to use an integrated machine learning and Markov chain model for assessing and predicting the cockpit crew performance in order to decrease operational crew risks before a flight. The main advantages of using the integrated model proposed in this paper are: i) prediction of the cockpit crew performance for decision making before the flight; based only on the historical data such as the flight crew assignment by the manager and (ii) the ability of the probability model to calculate changes of cockpit crew performance over a long time.

The remainder of this paper is organized as follows. Firstly, an overview of previous studies based on the fundamental role of the proposed model in advancing aviation-related research or other industries is provided. The methods and materials are proposed in next section, followed by the description of a case study with all the necessary elements of the integrated method in Section 2. The results with detailed explanations are presented in Section 4. Finally, this study and outlines the potential future directions.

## 1. Literature review

Many studies have tried to focus on the flight crew performance model with a predictive approach. For instance, Corker and Pisanich initialized a split-halves model to propose a complex human performance model in the environment of the flight deck. This model used one-half of the human performance data to predict the behavior of the remaining half by the help of a stochastic element. Corker and Pisanich (1995) and Yan and Tang (2007) proposed a heuristic model including the planning and real-time stages of integration to solve gate assignment problems under the stochastic flight delays. A gate assignment model for the stochastic flight delay, the reassignment rule, and two penalty adjustment methods were the main framework of the paper. Data from a Taiwanese international airport's operations were used to assess the framework as numerical tests. The importance of system analysis in flight performance was reviewed by Filippone (2008). This research emphasized the aerodynamics, aeroacoustics, propulsion, flight mechanics and operations, numerical optimization, stochastic methods and numerical analysis role by using a multi-disciplinary approach. Oreschko et al. (2012) modeled the stochastic nature of the individual sub-processes, empirical data from aircraft operators, airports, and ground handling companies to capture the influence of the stochastic arrival processes on the turnaround process by modeling all key parts of the turnaround stochastically. Martini et al. (2013) estimated the efficiency scores with two frontiers comparisons: a classical distance function with no undesirable output and a hyperbolic distance function estimation. In this research, a hyperbolic-stochastic approach was implemented and the flight performance margin boundary was identified by quantifying the human factors (Yang et al., 2014). As a systematic methodology, they initiated a computational pilot model and a pattern recognition method. The results of the simulation indicated that the flight performance can be improved by the quantitative human factors. In the model proposed, a multistage stochastic programming mechanism was used. Onan (2015) deployed a hybrid intelligent classification method to diagnose the breast cancer and the fuzzy-rough nearest neighbor algorithm was implemented in the classification phase of the model. An ensemble approach to feature selection was developed by Onan and Korukoğlu (2017) to aggregate the several individual features by a genetic algorithm implementation. Also, Onan et al. (2016a, 2016b) developed a learning algorithm using five ensemble methods (AdaBoost, Bagging, Dagging, Random Subspace and Majority Voting) to assess the effectiveness of the methods for statistical keyword extraction in conjunction with ensemble learning algorithms. Moreover, in another research, Onan (2016) studied the predictive capability of ensemble methods for web page classification. In another research, Onan et al. (2016a, 2016b) provided a strategic management, data mining disciplines and their basic concepts and applications. Different feature engineering schemes were analyzed by Onan (2018a, 2018b) in another paper by using

ensemble learning methods and in another research, he proposed an efficient approach of multiple classification to text categorization by using swarm optimized topic modeling. Recently, Delgado et al. (2019) developed a multistage stochastic programming model for the optimal allocation of a cargo considering the passenger's network to optimize profit, income and cost. Papadopoulos et al. (2019) reviewed and classified Markov models implementation in manufacturing systems. They studied the significant importance of timed models and their applications in manufacturing. Onan (2019a, 2019b) established a two-stage framework to extract a topic from the scientific literature. In this study, a method of conventional clustering (i.e., k-means, k-modes, kmeansCC, self-organizing maps and DIANA algorithm) was deployed by means of the iterative voting consensus. Onan (2019a, 2019b) proposed a consensus clustering-based under sampling approach to imbalance learning and suggested a deep learning based method to sarcasm identification. Onan (2020) established a sentiment classification scheme on reviews of instructor evaluation by pursuing the paradigm of deep learning and using a method based on recurrent neural network (RNN). Lyu and Liem (2020) formulated a model based on hybrid data-driven physics to estimate aircraft fuel consumption and flight operations data, using flight operational data to analyze aircraft performance at each phase of the flight. The model demonstrated the schematic of the algorithm for the flight performance analysis. Gharaibeh et al. (2020) tried to predict changes in land use using the Artificial Neural Network (ANN) along with the integration of CA-MC and improved the simulation capability of the Cellular Automata Markov Chain model (CA-MC). This research tried to predict changes in land use for the future land changes. In addition, a machine learning-based model with the uncertainty of renewable generation was proposed by Li et al. (2020). They combined Markov Chain Monte Carlo (MCMC) with Generative Adversarial Networks (GAN) for the development of numerous future scenarios to operate future states of the system. Samaee and Kobravi (2020) developed a predictive model for predicting the timing of tremor bursts, using nonlinear Markov and the maximum entropy algorithm. Hon et al. (2020) provided a multi-index prediction based on machine learning for aviation turbulence over the Asia-Pacific by the XGBoost algorithm to obtain aviation turbulence forecasts. Here, the models for numerical weather prediction (NWP) were used with a collection of conventional "turbulence indices". Toçoğlu and Onan (2020) provided a machine learning model for sentiment analysis in evaluation of students' in higher educational institutions. A machine learning model was provided by Yaakoubi et al. (2020) to produce a solution with the combination of several operation research advanced techniques to assemble and modify these clusters for solving crew pairing problem (CPP). A literature review was done by Shone et al. (2021) to show the essential role of stochastic models in advancing aviation-related research. They demonstrated how these methods could be implemented in other issues such

as: large-scale, dynamic, no stationary optimization by using a stochastic queuing theory to model aircraft queues incapacitated settings. An air transport network reliability was assessed in terms of meeting under the time constraint by Nhuyen and Lin (2021) for consideration of late arrivals under the time constraint and as an air transport network with stochastic-flow formulated under the time constraint with a depth-first search-based approach. Rey et al. (2021) developed a data-driven approach along with a method of data analysis and machine learning implementation to classify flights with safe or in-danger flight conditions. In this research, the likelihood of unstable approaches is applied in each flight and efforts were made to train a gradient boosted tree classifier. An integrated model including support vector machine, Markov chain and cellular automata for urban changes was developed by Okwuashi and Ndehedehe (2021) for the realistic simulation of urban land-use improvement. Zhou et al. (2021) reviewed all the main aspects about machine-learning models which contained input parameters and feature selection. In the research methods such as data pre-processing algorithms, output ensemble methods, model purposes and seven classes of machine learning models were surveyed to be proposed in future studies. Ashiku et al. (2021) developed a specialized analytical method and analyzed integrated technologies with large data to provide support decision making systems in healthcare settings through machine learning. Also, a sentiment analysis on a product was deployed by Onan (2021). The empirical analysis was assessed by using several weighting functions. In another research, Onan and Toçoğlu (2021) presented an architecture of three-layer stacked bidirectional long short-term memory to identify sarcastic text documents. In the literature review, there are some studies which used machine learning methods to predict the trend or the correlation between variables. For instance, Mokhtarimousavi and Mehrabi (2022) introduced an empirical analysis to study the potential unobserved heterogeneity of a flight delay and the impact of significant variables on it using two modeling approaches (SVM and ABC). A complete analysis for prediction of low-visibility events problems were carried out by Castillo-Botón et al. (2022) using regression and classification problems.

According to literature review study, there is a necessity in having an integrated model to assess airlines flight crew performance since there are some gaps in flight performance prediction studies. It can be seen in literature review study that a few studies used machine learning models to evaluate the flight performance. However, Rey et al. (2021) tried to categorize flights as safe or in-danger while there are only two classes for each flight (i.e., binary classification), but it was decided to consider five different levels for each flight performance (i.e., Multi-class classification). In addition, Lyu and Liem focused on estimating information about the aircraft fuel consumption and flight condition by using flight operational data using machine learning models (Lyu & Liem, 2020). In this study, the cockpit crew's performance was assessed through flight data, integrated machine learning models and Markov chain in

the aviation industry, for the first time. Furthermore, the contribution of this study can be described as follows:

- First, there is no study which considers all characteristics that impact on flight operations performance in a systematic evaluation process.
- Second, the machine learning multi-class classification models were not used in previous studies for predicting several levels of classes related to a flight.
- Last, in this study Markov chain is used in order to estimate the performance of cockpit crew during the flight, while previous studies never modeled cockpit crew performance during flight under uncertain conditions by using integrated machine learning classification models and Markov chain.

## 2. Methods and materials

Machine learning has been used for decision-making and its importance has increased in businesses. Methods to include machine learning are: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Van et al., 2022). Cockpit crew safety assessment and also using a model to predict their performance with available flight data (which is driven from A/C) is a challenging issue which can improve flight safety in aviation industry.

### 2.1. Data collection and acquisition

The airline company's top managers and decision makers for flight operations issues need to assess their cockpit crew performance through the flight data which has been extracted through their flight. The flight data which are collected in this paper are based on one of the cockpit crew data sets. In this survey, the pilots were selected and their performance parameters were monitored.

These parameters include event description, FLT Hours, aircraft (A/C), Years' experience, Route (From-To), limitation, Maximum Value, and Average Gap. Finally, in the last column, the performance rate of each record is labeled into five classes (absolutely weak, weak, moderate, good, and excellent). The description of each item is as follows:

- **Pilot Code:** a specific code allocated to a pilot.
- **Event Description:** shows each flight event (Regarding FDD procedures) done by pilots and their specific description.
- **FLT Hours:** shows the total hours which each pilot flies during the FLT experience.
- **A/C Name:** shows the special type of the aircraft.
- **Experience yearsYears' experience:** shows the total number of years which each pilot has gained experience.
- **Route (From-To):** shows the initial and destination point of a route which each pilot flies.
- **Limit:** shows the different levels of exceedance from FDM procedures.
- **Maximum Value:** shows the maximum level of exceedance that occurs during the flight.

- **Average Gap:** shows the average gap of the event exceedance which occurs during a specific duration.
- **Rate:** the cockpit crew performance level.

## 2.2. Application of machine learning techniques

A detailed description of the proposed steps in this study is given in Figure 1. Datasets from flight operations data were developed trained and validated. There are 6 steps to develop the model in this study, which are:

- Flight Data set;
- Data cleaning;
- Data pre-processing;
- Multi-class models implementation;
- Multi-class Flight Performance;
- Markovian-based prediction.

### 2.2.1. Data cleaning and pre-processing

In the first step the features related to a flight are collected and cleaned in a dataset. Then, the dataset was pre-processed in order to be ready for model entry. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Data preprocessing involves the transformation of the raw dataset into an understandable format. Preprocessing data is a fundamental stage in data mining which improves data efficiency. The data preprocessing methods directly affect the outcomes of any analytic algorithm. Steps in Data Preprocessing are as follows:

- Gathering data;
- Importing the dataset & Libraries;
- Dealing with Missing Values;
- Dividing the dataset into Dependent & Independent variable;

- Dealing with Categorical values;
- Splitting the dataset into training and test sets;
- Feature Scaling.

### 2.2.2. Data classification

In the fourth step, four different multi class classification models including Decision Tree (DT), Support Vector Machine (SVM), Neural Network (NN), and Random Forest (RF) were developed and are described as follows:

#### 2.2.2.1. Multi-class classification

Classification could be considered a method of machine learning and can be used to learn how to assign a class label to an input. There are four approaches of classification as follows:

- **Binary Classification:** Problems that have two class labels such as identifying a spam email that has two tags, spam or non-spam (Fodeh & Tiwari, 2018).
- **Multi-class classification:** These are classification issues that have more than two class labels such as face classification, plant species classification and identification of optical characters (Moral-García et al., 2020).
- **Multi-label classification:** Tasks that predict two or more class tags for each instance. In the photo classification example, when a photo contains multiple components in an image, the model predicts multiple tags in the photo, such as people, bicycles, etc. (The figure above shows the difference between multi-class and multi-tag classifications) (Qian et al., 2021).
- **Imbalanced Classification:** Classification problems in which the number of samples in each class is unequally distributed. For example, in cancer screening tests, there are a large number of healthy people and a small number with cancer (Aljedani et al., 2021).

In this study, multi-class classification models were used since our target column has five different labels and each record in one of these classes was tried to classify.

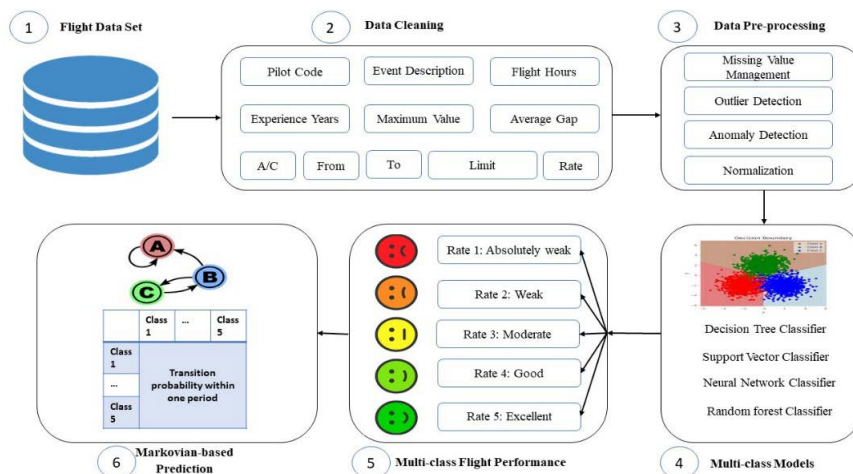


Figure 1. A detailed description of the proposed steps in this study

### 2.2.2.2. Neural network

Neural networks are a collection of neurons which recognize patterns by the human brain inspiration. Neural networks use algorithms of machine learning to classify the input data and provide the optimal output by data analysis in their hidden layers. These data can be a group of images, sounds, texts, etc. Different types of information can be classified based on their similarity using neural networks. The main advantages of neural networks are as mentioned:

- The ability of powerful learning and producing outputs not limited to the number of inputs.
- The ability of detecting faults and minimizing the errors.
- Its fast processing rate (Güven & Şimşir, 2020).

### 2.2.2.3. Support vector machine

A linear classifier is a primary principle applied by the Support Vector Machine (SVM), which can classify data and separate them linearly. However, SVM is developed to rectify non-linear problems with kernel tricks concept implementation. To obtain the best hyper plane which separates two classes in linear SVM and several classes in non-linear SVM is the main goal of SVM (Utami et al., 2021). Today, SVM can be applied to such areas as classification (clustering), regression, character recognition, and time series. Vapnik suggested it for classification problems in 1992. Minimizing structural risks is the main advantage of using SVM among the other machine learning methods. In addition, kernel functions can be applied in nonlinear data sets, and SVM can be adjusted for nonlinear problems by changing the kernel functions. The kernel function restructures a nonlinear input set into a linear input set. Therefore, it can give the best results in nonlinear regression problems (Güven & Şimşir, 2020).

### 2.2.2.4. Decision tree

The decision tree algorithm is one of the most widely used algorithms of data mining. The decision tree algorithm is a predictive tool used for both regression and stratified models. When a tree is used for classification tasks, it is known as a decision tree classifier, and when it is used for regression activities, it is called a decision tree regression. In the decision tree structure, the prediction obtained from the tree appears in the form of a series of rules. Each path from the root to a leaf of the decision tree shows a rule, and finally the leaf is labeled with the class where the most records belong (Moshkov, 2021).

## 2.3. Assessment and prediction

In the next step, the cockpit crew performance is assessed and predicted through the three models in five different classes (Absolutely weak, weak, moderate, good, and excellent). Consequently, the steady state probability of flight performance changing over time is calculated by a Markovian-based model.

## Markov Chain

Markov chain as a probability model tries to depict the system evolution randomly in time. If the condition of a system is a set of discrete times, a discrete-time stochastic process occurs (Kulkarni, 2011). A stochastic process  $\{X_n, n = 0, 1, 2, \dots\}$  is considered as finite or countable values.  $P_{ij}$  as a fixed probability (non-negative) of the process in state  $i$  (with next in the state  $j$ ) is defined in Equation ((1)–(2)) mentioned below:

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} = P_{ij}. \quad (1)$$

$$P_{ij} \geq 0, i, j \geq 0; \sum_{j=0}^{\infty} P_{ij} = 1, i = 0, 1, \dots \quad (2)$$

The matrix of one-step transition probabilities  $P_{ij}$  define  $P$  as the matrix in Equation (3):

$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \dots & \dots \\ P_{10} & P_{11} & P_{12} & \dots & \dots \\ \cdot & \cdot & \cdot & \dots & \dots \\ \cdot & \cdot & \cdot & \dots & \dots \\ P_{i0} & P_{i1} & P_{i2} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}. \quad (3)$$

The  $P$  matrix denotes a one-step transition probability matrix or transition matrix for short, of the Discrete-Time Markov Chains (DTMC) in the Equation (3). Note that the rows correspond to the starting state and the columns to the ending state of a transition. Thus the probability of going from state 2 to state 3 in one step is stored in row number 2 and column number 3.  $P_{ij}^n$  denotes the  $n$ -step transition probabilities as Equation (4):

$$P_{ij}^n = P\{X_{n+k} = j | X_k = i\}. \quad (4)$$

The  $n$ -step transition probabilities computation is based on Equation (5): (Ross, 2014).

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m. \quad (5)$$

The limiting behavior of  $X_n$  as  $n$  tends to infinity is called the limiting or steady-state distribution and denotes it by Equations (6, 7):

$$\pi = \pi_1 \ \pi_2 \ \dots \ \pi_n; \quad (6)$$

$$\pi_j = \lim_{n \rightarrow \infty} P(X_n = j). \quad (7)$$

## 3. Findings and discussions

### 3.1. Background and methods

In this section, the multi-class classification models discussed in Section 3 are developed. In the initial stage, the data are collected from a flight database. This dataset has 1074 records and 11 features, 6 of which are numeric and the rest are categorical. The target (independent) column is the "Rate" which is referred to as the flight performance. Other columns are dependent ones except for the pilot code. Then, data cleaning and preprocessing were done. In this step, duplicated records were evaluated which there were no supplicated ones. Then, the Pearson's correlation

between the two-by-two numeric features, features and the target column were investigated. Figure 2 shows the heat map of correlations. Our study is not without limitations. There were no missing values in all features except for three records in the “To” column which were being excluded from the dataset. So, the records decreased from 1074 to 1071. The outliers were detected with Local Outlier Factor (LOF), but no records were detected as outliers. The pre-processes for each column are summarized in Table 1. Each Flight was performed by one pilot and in this paper, a unique code to each pilot was assigned. The safety performance of each pilot is assessed based on FDM output after each flight. In accordance with Aircraft SOPs (standard operations procedures), each flight data (which was downloaded from the aircraft), has some limitation (between the maximum and minimum value). The cockpit crew performance is evaluated and rated on levels 1 to 5, according to average gaps from the Max. This value was detected during the flight.

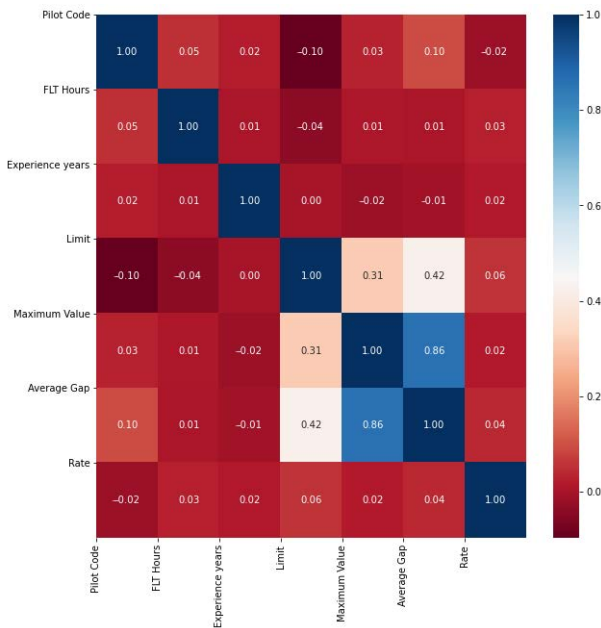


Figure 2. Heatmap of correlations (Pearson’s correlation between the two-by-two numeric features)

Table 1. Pre-processes carried out for each column

Row	Feature	Description	Pre-process
1	Pilot Code	100 distinct values related to 100 pilots	–
2	Event Description	51 distinct types	Dummy encoding to 50 binary columns
3	Flight Hours	Between 100–2000 hours	Z-normalization
4	Experience Years	Between 2–25 years	Z-normalization
5	Maximum Value	Between –4000 to 14000	Z-normalization
6	Average Gap	Between –4000 to 13000	Z-normalization
7	A/C	5 distinct types	Dummy encoding to 4 binary columns
8	From	26 distinct types	Dummy encoding to 25 binary columns
9	To	27 distinct types	Dummy encoding to 26 binary columns
10	Limit	Between –2500–10000	Z-normalization
11	Rate	1 to 5 levels	–

The following description provides groups of FDM events (According to AMC1 ORO.AOC.130-Annex III) which were used in Table 2:

- Rejected take-off;
- Take-off pitch;
- Unstick speeds;
- Height loss in climb-out;
- Slow climb-out;
- Climb-out speeds;
- High rate of descent;
- Missed approach;
- Low approach;
- Glideslope;
- Approach power;
- Approach speeds;
- Landing flap;
- Landing pitch;
- Bank angles;
- Normal acceleration;
- Abnormal configuration;
- Ground proximity warning;
- Airborne collision avoidance system (ACAS II) warning;
- Margin to stall/buffet;
- Aircraft flight manual limitations.

Each group includes some events (For example, Landing pitch includes a high and low pitch attitude on landing). To check if the “Rate” column is balanced or not, the number of records in each class is compared and shown in Figure 3.

Table 2. Pre-processes carried out for each column

Class	Train records numbers	Test records number
1	381	163
2	120	52
3	110	47
4	81	35
5	57	25
Total	724	322



**Figure 3.** Comparison between the numbers of records in each class

Based on Figure 3, the first class has a large amount compared to the other classes. Furthermore, 30% of the dataset is considered as a test set (322 records) and the rest as a train set (742 records) based on the “Class Balance” library to select train and test data on the basis of each class amount. The train and test records in each class are shown in Table 2. For each classification model, “GridSearch” is used to evaluate different combinations of hyper parameters which were used in each model. All the parameters containing the best ones with the highest accuracy are summarized in Table 3.

K-cross-validation has been applied for the dataset to classify it into  $k$  folds. In this method, each fold is applied once as a validation, and the remaining ( $k - 1$ ) folds form the training set. Also, parameters are adjusted during training. Finally, the test data are used to evaluate the generalization performance of the predictor (Yamaguchi et al., 2022). In this paper, the K-cross validation was carried out, where  $K = 5$ . This means that the data set with the subset for selected features was divided into five nearly equal parts and the distribution of the samples per subclass in each part of the data set was kept almost the same as the distribution in the total data set used for training of

the model. From among the five sets, four were used for training, and one set was kept for testing as shown below in Figure 4. Then, the class-wise true positives ( $TP_i$ ), true negatives ( $TN_i$ ), false positives ( $FP_i$ ), and false negatives ( $FN_i$ ) are measured to compute the multiclass classifier performance. For the purpose of accuracy, Precision, Recall, and F1-score of the model computation, and parameters mentioned are used according to Equations ((8)–(11)):

$$Accuracy = \frac{TP_i}{\sum_{i=1}^l TP_i + FP_i + TN_i + FN_i}; \quad (8)$$

$$Precision = \frac{\sum_{i=1}^l TP_i}{\sum_{i=1}^l TP_i + FN_i} \quad (9)$$

$$Recall = \frac{\sum_{i=1}^l TP_i}{\sum_{i=1}^l TP_i + FP_i}; \quad (10)$$

$$F1 - Score = 2 * \left( \frac{Precision * Recall}{Precision + Recall} \right). \quad (11)$$

N/A	Set1	Set2	Set3	Set4	Set5
M1	Testing	Training	Training	Training	Training
M2	Training	Testing	Training	Training	Training
M3	Training	Training	Testing	Training	Training
M4	Training	Training	Training	Testing	Training
M5	Training	Training	Training	Training	Testing

**Figure 4.** Scheme demonstration for the set used for training and testing

**Table 3.** Different parameters and the best ones with the highest accuracy

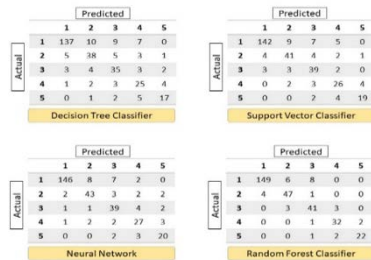
Model	Parameters	Best parameters
Decision Tree Classifier	param = {'estimator__criterion': ['gini', 'entropy'], 'estimator__max_depth': [30,40,50,60,70,80], 'estimator__min_samples_split': [30,40,50,60,70,80], 'estimator__min_samples_leaf': [20,30,40]}	{'estimator__criterion': 'entropy', 'estimator__max_depth': 30, 'estimator__min_samples_leaf': 30, 'estimator__min_samples_split': 80}
Support Vector Classifier	param = {'estimator__kernel': ['linear', 'rbf', 'sigmoid'], 'estimator__gamma': [0.001, 0.01, 0.1, 1, 10], 'estimator__C': [0.01, 0.1, 1, 10, 100]}	{'estimator__kernel': 'rbf', 'estimator__gamma': 0.01, 'estimator__C': 10}
Neural Network	param = {"estimator__activation": ["relu", "logistic", "tanh", "identity"], "estimator__hidden_layer_sizes": [(10),(20), (20,30)], "estimator__max_iter": [10, 50, 100, 200], "estimator__solver": ["sgd", "adam", "lbfgs"], "estimator__learning_rate_init": [0.01, 0.001, 0.0001, 0.025]}	{"estimator__activation": "logistic", "estimator__hidden_layer_sizes": (20,30), "estimator__max_iter": 100, "estimator__solver": "lbfgs", "estimator__learning_rate_init": 0.01}
Random Forest Classifier	param = {'estimator__bootstrap': [True, False], 'estimator__max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None], 'estimator__max_features': ['auto', 'sqrt'], 'estimator__min_samples_leaf': [1, 2, 4], 'estimator__min_samples_split': [2, 5, 10], 'estimator__n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}	{'estimator__bootstrap': True, 'estimator__max_depth': 40, 'estimator__max_features': 'sqrt', 'estimator__min_samples_leaf': 2, 'estimator__min_samples_split': 5, 'estimator__n_estimators': 600}

The confusion matrix of each classification model is presented in Figure 5. For the confusion matrix, all the values in the diagonal line from the top left to the bottom right are correctly classified data samples.

Furthermore, for each model, the precision, recall, and F1-score are calculated based on the equations mentioned above and the results are shown in Table 4.

**Table 4.** The precision, recall, and F1-score of each model calculation

Model	Accuracy	Precision	Recall	F1-score
Decision Tree Classifier	0.78	0.84	0.93	0.88
Support Vector Classifier	0.82	0.87	0.95	0.90
Neural Network	0.85	0.89	0.97	0.92
Random Forest Classifier	0.90	0.91	0.97	0.93



**Figure 5.** Confusion matrix of each classification model

### 4. Results

Based on Table 4, among the classification models, the Random Forest (RF) classifier has the highest accuracy, precision, recall, and F1-score. In addition, the evaluation metrics score of this model is desirable and the managers can trust its results in their decision making. Afterwards, the probability of changing the levels can be calculated based on Markov chain. To do this, the transition matrix P (through the average of historical flight data – pilot crew performance) as discussed in Section 3.3 is stated as Equation (12):

$$P = \begin{matrix} & \begin{matrix} \text{Rate : 1} & \text{Rate : 2} & \text{Rate : 3} & \text{Rate : 4} & \text{Rate : 5} \end{matrix} \\ \begin{matrix} \text{Rate : 1} \\ \text{Rate : 2} \\ \text{Rate : 3} \\ \text{Rate : 4} \\ \text{Rate : 5} \end{matrix} & \begin{bmatrix} 0.19 & 0.39 & 0.25 & 0.09 & 0.34 \\ 0.15 & 0.37 & 0.21 & 0.18 & 0.30 \\ 0.17 & 0.29 & 0.42 & 0.05 & 0.24 \\ 0.18 & 0.28 & 0.32 & 0.07 & 0.25 \\ 0.20 & 0.20 & 0.28 & 0.12 & 0.47 \end{bmatrix} \end{matrix} \tag{12}$$

Based on matrices  $W_{need}$  and P, the priority matrices for the initial period and the next periods are calculated as below:

$$P^{(0)} = [0.165 \ 0.207 \ 0.320 \ 0.120 \ 0.245];$$

$$P^{(1)} = [0.151 \ 0.224 \ 0.29 \ 0.128 \ 0.256];$$

$$P^{(2)} = [0.140 \ 0.229 \ 0.232 \ 0.129 \ 0.259];$$

$$P^{(3)} = [0.149 \ 0.229 \ 0.233 \ 0.129 \ 0.265] \ i \geq 3.$$

The limiting distribution for flight crew performance (Rate “1 to 5” levels: referred to the flight performance) is obtained as Equation (13).

$$\pi = \begin{matrix} \text{Rate : 1} \\ \text{Rate : 2} \\ \text{Rate : 3} \\ \text{Rate : 4} \\ \text{Rate : 5} \end{matrix} \begin{bmatrix} 0.149 \\ 0.229 \\ 0.233 \\ 0.129 \\ 0.260 \end{bmatrix} \tag{13}$$

### Conclusions

In the aviation business environment, using the proper model to establish beneficial safety policies is one of important issues for airlines all over the world. As a solution for this issue, expert viewpoints, categorizing and prioritizing approaches for the cockpit crew performance assessment can be used. The selection of a cockpit crew is a very critical aspect of flight operations. In this regard, the safety performance of flights have been a challenging issue in the aviation industry and plays an important role in acquiring competitive benefits. This study was aimed to use machine learning models and Markov chains in order to assess the cockpit crew performance before the flight and help their selection process by use of their historical data. So, the managers are able to arrange the cockpit crew for each flight considering how the main attributes affect the flight performance level. In this study, a new model for the evaluation of cockpit crew performance during their flights while considering major flight types was developed. The results suggested that the random forest classifier is one of the successful and accurate models applied in the field of flight data analysis for aviation industry with prediction of a flight performance level with accuracy = 0.90, precision = 0.91, recall = 0.97, and F1-score = 0.93. All evaluation metrics are high enough and can be used by managers and decision-makers. In this study, K-cross-validation was employed to classify data into 30% of the dataset as a test set (322 records) and the rest as a train set (742 records) based on the “Class Balance” library. As a case study of how the classifier could be used in combination with existing flight data for the purposes of aviation knowledge generation, an integrated multi-class classification machine learning model and a Markov chain for cockpit crew performance evaluation during their flight were offered, for the first time. The main contribution of this study is considering all the characteristics which impact on flight operation’s performance in a systematic evaluation process, while using the machine learning multi-class classification models to predict several levels of classes related to a prediction of the cockpit crew performance over time by using the Markov chain. The results of this study also suggest that application of the deep learning model and multi-class classification models to FDA data not only offers improved flight data classification but also provides a framework for cockpit crew selection or flight scheduling decision systems. Airline companies can employ this model to predict flight crew performance before the flight



in order to prevent or decrease flight safety risks. Other studies can consider more features and use other classification models, especially the probabilistic ones. In addition, the hidden Markov chain and also the semi-hidden Markov chain could be continued by researchers in future.

## Conflict of interests

Hereby, the author(s) declare that there have been no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

## References

- Aljedani, N., Alotaibi, R., & Taileb, M. (2021). HMATC: Hierarchical multi-label Arabic text classification model using machine learning. *Egyptian Informatics Journal*, 22(3), 225–237. <https://doi.org/10.1016/j.eij.2020.08.004>
- Ashiku, L., Al-Amin, M., Madria, S., & Dagli, C. (2021). Machine learning models and big data tools for evaluating kidney acceptance. *Procedia Computer Science*, 185(June), 177–184. <https://doi.org/10.1016/j.procs.2021.05.019>
- Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutierrez, P. A., Deo, R. C., & Salcedo-Sanz, S. (2022). Machine learning regression and classification methods for fog events prediction. *Atmospheric Research*, 272, 106157. <https://doi.org/10.1016/j.atmosres.2022.106157>
- Corker, K. M., & Pisanich, G. M. (1995). Analysis and modeling of flight crew performance in automated air traffic management systems. *IFAC Proceedings Volumes*, 28(15), 547–552. [https://doi.org/10.1016/S1474-6670\(17\)45289-X](https://doi.org/10.1016/S1474-6670(17)45289-X)
- Delgado, F., Trincado, R., & Pagnoncelli, B. K. (2019). A multistage stochastic programming model for the network air cargo allocation under capacity uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, 131(November), 292–307. <https://doi.org/10.1016/j.tre.2019.09.011>
- EASA. (2019). *AMC1 ORO.AOC.130 – Annex III*. <https://www.easa.europa.eu/>
- European Aviation Safety Agency (EASA). (2016). *Developing Standardised Fdm-Based Indicators Focus* (2 December, pp. 1–55). [https://www.easa.europa.eu/sites/default/files/dfu/EAFDM\\_standardised\\_FDM-based\\_indicators\\_v2\\_Ed2017.pdf](https://www.easa.europa.eu/sites/default/files/dfu/EAFDM_standardised_FDM-based_indicators_v2_Ed2017.pdf)
- Filippone, A. (2008). Comprehensive analysis of transport aircraft flight performance. *Progress in Aerospace Sciences*, 44(3), 192–236. <https://doi.org/10.1016/j.paerosci.2007.10.005>
- Fodeh, S. J., & Tiwari, A. (2018). Exploiting MEDLINE for gene molecular function prediction via NMF based multi-label classification. *Journal of Biomedical Informatics*, 86(August 2017), 160–166. <https://doi.org/10.1016/j.jbi.2018.08.009>
- Gharaibeh, A., Shaamala, A., Obeidat, R., & Al-Kofahi, S. (2020). Improving land-use change modeling by integrating ANN with Cellular Automata-Markov Chain model. *Heliyon*, 6(9), e05092. <https://doi.org/10.1016/j.heliyon.2020.e05092>
- Güven, İ., & Şimşir, F. (2020). Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods. *Computers and Industrial Engineering*, 147. <https://doi.org/10.1016/j.cie.2020.106678>
- Hon, K. K., Ng, C. W., & Chan, P. W. (2020). Machine learning based multi-index prediction of aviation turbulence over the Asia-Pacific. *Machine Learning with Applications*, 2, 100008. <https://doi.org/10.1016/j.mlwa.2020.100008>
- International Civil Aviation Organization. (2013). *Operation of Aircraft – Fatigue. Excerpts of Fatigue Management Related Provisions from: Annex 6 to the Convention of International Civil Aviation, February*. <https://www.icao.int/safety/fatiguemanagement/FRMS%20Tools/Amendment%2037%20for%20FRMS%20SARPS%20en%29.pdf>
- Kulkarni, V. G. (2011). Brownian motion. In *Introduction to modeling and analysis of stochastic systems* (pp. 247–280). Springer. [https://doi.org/10.1007/978-1-4419-1772-0\\_7](https://doi.org/10.1007/978-1-4419-1772-0_7)
- Lan, C. E., Wu, K., & Yu, J. (2012). Flight characteristics analysis based on QAR data of a jet transport during landing at a high-altitude airport. *Chinese Journal of Aeronautics*, 25(1), 13–24. [https://doi.org/10.1016/S1000-9361\(11\)60357-9](https://doi.org/10.1016/S1000-9361(11)60357-9)
- Li, H., Wang, W., Fan, L., Li, Q., & Chen, X. (2020). A novel hybrid MCDM model for machine tool selection using fuzzy DEMATEL, entropy weighting and later defuzzification VIKOR. *Applied Soft Computing*, 91, 106207. <https://doi.org/10.1016/j.asoc.2020.106207>
- Lyu, Y., & Liem, R. P. (2020). Flight performance analysis with data-driven mission parameterization: Mapping flight operational data to aircraft performance analysis. *Transportation Engineering*, 2(September), 100035. <https://doi.org/10.1016/j.treng.2020.100035>
- Martini, G., Scotti, D., & Volta, N. (2013). Including local air pollution in airport efficiency assessment: A hyperbolic-stochastic approach. *Transportation Research Part D: Transport and Environment*, 24(2007), 27–36. <https://doi.org/10.1016/j.trd.2013.05.002>
- Mokhtarimousavi, S., & Mehrabi, A. (2022). Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. *International Journal of Transportation Science and Technology*, 12(1), 230–244. <https://doi.org/10.1016/j.ijst.2022.01.007>
- Moral-García, S., Mantas, C. J., Castellano, J. G., & Abellán, J. (2020). Non-parametric predictive inference for solving multi-label classification. *Applied Soft Computing Journal*, 88. <https://doi.org/10.1016/j.asoc.2019.106011>
- Moshkov, M. (2021). On the depth of decision trees over infinite 1-homogeneous binary information systems. *Array*, 10(March), 100060. <https://doi.org/10.1016/j.array.2021.100060>
- Nguyen, T. P., & Lin, Y. K. (2021). Reliability assessment of a stochastic air transport network with late arrivals. *Computers and Industrial Engineering*, 151(January). <https://doi.org/10.1016/j.cie.2020.106956>
- Okwuashi, O., & Ndehedehe, C. E. (2021). Integrating machine learning with Markov chain and cellular automata models for modelling urban land use change. *Remote Sensing Applications: Society and Environment*, 27(January). <https://doi.org/10.1016/j.rsase.2020.100461>
- Onan, A., Korukoğlu, S., & Bulut, H. (2016a). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57, 232–247. <https://doi.org/10.1016/j.eswa.2016.03.045>
- Onan, A., Bal, V., & Yanar Bayam, B. (2016b). The use of data mining for strategic management: A case study on mining association rules in student information system. *Croatian Journal of Education: Hrvatski časopis za odgoj i obrazovanje*, 18(1), 41–70. <https://doi.org/10.15516/cje.v18i1.1471>
- Onan, A. (2019). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, 7, 145614–145633. <https://doi.org/10.1109/ACCESS.2019.2945911>
- Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, 43(1), 25–38. <https://doi.org/10.1177/0165551515613226>

- Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 33(23), e5909. <https://doi.org/10.1002/cpe.5909>
- Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Systems with Applications*, 42(20), 6844–6852. <https://doi.org/10.1016/j.eswa.2015.05.006>
- Onan, A. (2020). Mining opinions from instructor evaluation reviews: A deep learning approach. *Computer Applications in Engineering Education*, 28(1), 117–138. <https://doi.org/10.1002/cae.22179>
- Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science*, 42(2), 150–165. <https://doi.org/10.1177/0165551515591724>
- Onan, A. (2018a). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, 44(1), 28–47. <https://doi.org/10.1177/0165551516677911>
- Onan, A. (2018b). Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Computational and Mathematical Methods in Medicine*, 2018. <https://doi.org/10.1155/2018/2497471>
- Onan, A. (2019a). Consensus clustering-based undersampling approach to imbalanced learning. *Scientific Programming*, 2019. <https://doi.org/10.1155/2019/5901087>
- Onan, A. (2019b). Topic-enriched word embeddings for sarcasm identification. In R. Silhavy, *Software Engineering Methods in Intelligent Algorithms: Proceedings of 8th Computer Science On-line Conference 2019* (Vol. 984, pp. 293–304). Springer International Publishing. [https://doi.org/10.1007/978-3-030-19807-7\\_29](https://doi.org/10.1007/978-3-030-19807-7_29)
- Onan, A., & Toçođlu, M. A. (2021). A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access*, 9, 7701–7722. <https://doi.org/10.1109/ACCESS.2021.3049734>
- Oreshko, B., Kunze, T., Schultz, M., Fricke, H., Kumar, V., & Sherry, L. (2012). Turnaround prediction with stochastic process times and airport specific delay pattern airport delays. In *The 5<sup>th</sup> International Conference on Research in Air Transportation*, 9. ResearchGate.
- Papadopoulos, C. T., Li, J., & O'Kelly, M. E. J. (2019). A classification and review of timed Markov models of manufacturing systems. *Computers and Industrial Engineering*, 128(November 2018), 219–244. <https://doi.org/10.1016/j.cie.2018.12.019>
- Qian, W., Xiong, C., & Wang, Y. (2021). A ranking-based feature selection for multi-label classification with fuzzy relative discernibility. *Applied Soft Computing*, 102. <https://doi.org/10.1016/j.asoc.2020.106995>
- Rey, M., Aloise, D., Soumis, F., & Piegueu, R. (2021). A data-driven model for safety risk identification from flight data analysis. *Transportation Engineering*, 5, 100087. <https://doi.org/10.1016/j.treng.2021.100087>
- Ross, S. M. (2014). *Introduction to probability models*. Academic Press. <https://doi.org/10.1016/B978-0-12-407948-9.00001-3>
- Samaee, S., & Kobravi, H. R. (2020). Predicting the occurrence of wrist tremor based on electromyography using a hidden Markov model and entropy based learning algorithm. *Biomedical Signal Processing and Control*, 57(March). <https://doi.org/10.1016/j.bspc.2019.101739>
- Shone, R., Glazebrook, K., & Zografos, K. G. (2021). Applications of stochastic modeling in air traffic management: Methods, challenges and opportunities for solving air traffic problems under uncertainty. *European Journal of Operational Research*, 292(1), 1–26. <https://doi.org/10.1016/j.ejor.2020.10.039>
- Toçođlu, M. A., & Onan, A. (2020, July). Sentiment analysis on students' evaluation of higher educational institutions. In *International Conference on Intelligent and Fuzzy Systems* (pp. 1693–1700). *INFUS 2020: Intelligent and Fuzzy Techniques: Smart and Innovative Solutions*. Springer. [https://doi.org/10.1007/978-3-030-51156-2\\_197](https://doi.org/10.1007/978-3-030-51156-2_197)
- Utami, N. A., Maharani, W., & Atastina, I. (2021). Personality classification of Facebook users according to big five personality using SVM (Support Vector Machine) method. *Procedia Computer Science*, 179(2020), 177–184. <https://doi.org/10.1016/j.procs.2020.12.023>
- Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, 144, 93–106. <https://doi.org/10.1016/j.jbusres.2022.01.076>
- Wang, L., Wu, C., & Sun, R. (2014). An analysis of flight Quick Access Recorder (QAR) data and its applications in preventing landing incidents. *Reliability Engineering and System Safety*, 127, 86–96. <https://doi.org/10.1016/j.res.2014.03.013>
- Yaakoubi, Y., Soumis, F., & Lacoste-Julien, S. (2020). Machine learning in airline crew pairing to construct initial clusters for dynamic constraint aggregation. *EURO Journal on Transportation and Logistics*, 9(4), 100020. <https://doi.org/10.1016/j.ejtl.2020.100020>
- Yamaguchi, S., Nakashima, H., Moriwaki, Y., Terada, T., & Shimizu, K. (2022). Prediction of protein mononucleotide binding sites using AlphaFold2 and machine learning. *Computational Biology and Chemistry*, 107744. <https://doi.org/10.1016/j.compbiolchem.2022.107744>
- Yan, S., & Tang, C.-H. (2007). A heuristic approach for airport gate assignments for stochastic flight delays. *European Journal of Operational Research*, 180(2), 547–567. <https://doi.org/10.1016/j.ejor.2006.05.002>
- Yang, C., Yin, T., Zhao, W., Huang, D., & Fu, S. (2014). Human factors quantification via boundary identification of flight performance margin. *Chinese Journal of Aeronautics*, 27(4), 977–985. <https://doi.org/10.1016/j.cja.2014.03.016>
- Zhou, Y., Liu, Y., Wang, D., Liu, X., & Wang, Y. (2021). A review on global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Conversion and Management*, 235(13), 113960. <https://doi.org/10.1016/j.enconman.2021.113960>