Titre: Title:	Intelligent Adaptive Flight Training System- A Human Performance in the Loop for Real-Time Decision Making
Auteur: Author:	Jean-François Delisle
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Delisle, J.-F. (2022). Intelligent Adaptive Flight Training System- A Human Performance in the Loop for Real-Time Decision Making [Ph.D. thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10551/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10551/
Directeurs de recherche: Advisors:	Andrea Lodi
Programme: Program:	Doctorat en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

INTELLIGENT ADAPTIVE FLIGHT TRAINING SYSTEM - A HUMAN PERFORMANCE IN THE LOOP FOR REAL-TIME DECISION MAKING

JEAN-FRANÇOIS DELISLE

Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie industriel

Août 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée:

INTELLIGENT ADAPTIVE FLIGHT TRAINING SYSTEM - A HUMAN PERFORMANCE IN THE LOOP FOR REAL-TIME DECISION MAKING

Présentée par **Jean-François DELISLE**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment accepté par le jury d'examen constitué de :

Louis-Martin ROUSSEAU, président

Andrea LODI, membre et directeur de recherche

Pierre-Majorique LÉGER, membre

Susanne LAJOIE, membre externe

DEDICATION

To my family, for a safer, creative & open world

ACKNOWLEDGEMENTS

To people that have directly contributed to this research project:

Caroline Demers, *Conjointe*

Zoé Delisle, Alexis Delisle, Raphaël Delisle, *Enfants*

Louise Delisle(feu), Jean-Pierre Delisle, *parents.* Caroline Delisle, *Sœur*

Dr. Felix Gervais, Jean-François Gouger, Francis Durocher, Jean-Sébastien Moreau, Guillaume Mercier, Vincent Denommée, *Motivateurs*

Dr. Andrea Lodi, Mehdi Taobane, *Data Science for Real-Time Decision-Making research chair, Polytechnique de Montréal*

Marc St-Hilaire, Patricia Gilbert, Nick Giannias, Rony Karam, Houssam Alaouie, Andrew Fernie, Mark Soodeen, Aurelian Constantinescu, Thierry Paul, *Advanced Technology & Innovation (AT&I) department & Chief Technology Officer Office, CAE Inc.*

Benoit Menard, Guillaume Normandin, Yann-Cedric Garreau, Michel Lagacé, Pascal Labbé, Éric Manseau, André Babineau, *Global Engineering - Training Experience Department, CAE Inc.*

Dr. Maher Chaouachi, Arnaud Van de Voorde, Kevin Graham, Jaime Subiabre, Josée Laganière, Yannick Audet, Simon Riendeau, Mona Hallaq, Hugo Christiansen, *Digital Accelerator, CAE Inc.*

Paula Mazzaferro, Philippe Perey, Christophe Courtoy, *Defense & Security, CAE Inc.*

Ramzi Gharbi, Christopher Ranganathan, Maxime Théberge, *Civil Aviation, CAE Inc.*

At last, but not least, all co-authors and co-inventors: *Dac Toan Ho, Luc Gingras, Sébastien Malo, Marc-André Proulx, David Bowness, Stéphane Ouellet, Derek Linders, Dr. Andrea Lodi, Dr. Maher Chaouachi, Laurent Desmet, Melvyn Tan, Dr. Jian Qi, Dr. Navpreet Singh, Ben Winokur Bincy Baburaj Naranth, Karen Moore, Anthoine Dufour, Hamza Nabil, Theophile Demazure and a special thanks to Dr. Pierre Majorique Léger*

RÉSUMÉ

L’automatisation toujours plus grandissante des systèmes offre à l’humain de nouveaux défis dans l’opération de système complexe tel que le pilotage d'avion. Avec une conscience distribuée de la situation entre l’homme et la machine, nous parlons maintenant de système intelligent avec l’humain-dans-la-boucle. La conception des systèmes et la formation des pilotes doivent d’être élaborées en prenant compte de cette nouvelle réalité d’opération. La formation adaptative est une innovation majeure qui devrait améliorer la capacité d’apprentissage d’opération de système complexe tel que le vol d’avion.

Les simulateurs de vol offrent une plateforme immersive et un support pour la formation. En plus de favoriser l’apprentissage, des quantités importantes de données peuvent y être générées et collectées et ainsi alimenter des modèles d’intelligence artificielle qui permettent l’évaluation et l’adaptation de la formation. Ces modèles peuvent être complémentés par les données d’opérations d’entrainement, les profils de vol des avions simulés et les profils biométrique et psychométrique des pilotes et des instructeurs. L'intégration multiple de modèle intelligent à partir d'une variété de sources riches est au cœur de la recherche actuelle. Les différents modèles d’apprentissage machine ou humain évoluant selon des cycles de vie qui leur sont propres seront mis en chorégraphie afin permettre une évaluation des performances des pilotes et instructeurs. Cette évaluation consiste en la fondation permettant l’adaptation du système de formation afin de maximiser les objectifs d’apprentissage tout en augmentant la maturité de description, de diagnostic, de prédiction et de prescription des scénarios de formation en aviation.

En utilisant l'apprentissage machine, la neuroscience et l’optimisation mathématique, les pilotes, les instructeurs et les systèmes avions simulés peuvent être modélisés pour adapter l'environnement d'apprentissage. L'objectif est de maximiser les performances d'entraînement et d'augmenter la prise de décision en tenant compte des différents aspects des facteurs humains. La visualisation en temps réel dynamique, l'évaluation automatique des profils humains ainsi que les outils d’adaptation du système d’entrainement devraient améliorer le cycle de vie de la formation au pilotage tout en donnant à toutes les parties prenantes un aperçu plus approfondi des performances des acteurs, sur la voie de la numérisation au sein d'une variété de dispositifs immersifs virtuels et d'environnements d'apprentissage synthétiques.

ABSTRACT

The ever-increasing automation of systems offers humans new challenges in the operation of complex systems such as aircraft operations. With a distributed situation awareness between man and machine, we are now talking about intelligent systems with human-in-the-loop models. Systems design and pilot training must be developed considering this new reality of operations. Adaptive training is a major innovation that is expected to enhance the learning ability and complex system operation in such areas as aircraft flying.

Flight simulators offer an immersive platform and support for training. In addition to fostering learning, large amounts of data can be generated and collected and fed into artificial intelligence models that allow evaluation and adaptation of the training. These models can be complemented by training operation data, flight profiles of simulated aircraft and the learner profile of pilots and instructors using biometry and psychometry to objectively assess human performance.

Multiple intelligent model integration from a variety of rich sources that can provide automatic assessment and systems adaptation is at the heart of current research. The various models of machine or human learning that are evolving according to their own lifecycles will be choreographed to allow an evaluation of the performances of the pilots and instructors. This evaluation consists of the foundation to adapt the training system and maximize learning objectives while enhancing the maturity of description, diagnosis, prediction and prescription of aviation training scenarios.

Using machine learning, neuroscience and mathematical optimization, pilots, instructors and simulated aircraft systems can be modelled to fit the learning environment. The aim is to maximize training performance and augment decision-making by considering the different aspects of human factors using objective biometric and psychometric data to assess human behavior

Dynamic real-time visualization, automatic human profile assessment, and training system adaptation tools should improve the lifecycle of flight training while providing all stakeholders with a deeper insight into the performance of the players involved, on the road of digitalization within and variety of immersive virtual device and synthetic learning environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF SYMBOLS AND ABBREVIATIONS

ATC	Air Traffic Controller
ATM	Air Traffic Management
BDS	Brief Debrief System
BBN	Bayesian Belief Network
CBTA	Competency-Based Training and Assessment
CFI	Certified Flight Instructor
CNN	Convolutional Neural Network
EASA	European Union Aviation Safety Agency
ECG	Electrocardiogram
EEG	Electroencephalography
EMG	Facial Electromyography
ETR	Electronic Training Record
FAA	Federal Aviation Administration
FDR	Flight Data Recorder
FTO	Flight Training Organization
fMRI	Functional Magnetic Resonance Imaging
fNIR	Functional Near Infrared
IAFTS	Intelligent Adaptive Flight Training System
IATA	International Air Transport Association
ICAO	International Civil Aviation Organization
IOS	Instructor Operation Station
LCMS	Learning Content Management System
LMS	Learning Management System

LPS	Lesson Plan System
LRS	Learning Record System
FFS	Full Flight Simulator
MSE	Mean Square Error
MILP	Mixed Integer Linear Programming
PET	Positron Emission Tomography
SVM	Support Vector Machine
VR	Virtual Reality

CHAPTER 1 INTRODUCTION

1.1 Research context

The aviation market is undergoing a major transformation, from classical training to adaptive flight training and competency-based training and assessment (CBTA), where artificial intelligence capabilities will be at the heart of the new vision. This vision proposes a significant change to pilot's training methodologies.

The research project is evolving in the domain of flight training in an industry-academic partnership between Polytechnique of Montreal and CAE Inc. Most of the research resources are provided by CAE Inc., including experimentation infrastructure and flight training domain expertise.

The fundamentals at the heart of the project are based on Data Science Engineering to support Flight Training with Human Factor perspective. The data science here combines the disciplines of mathematical optimization, artificial intelligence, and big data computing. Data Science, at the heart of the capability provided by this project, will be enforced with flight simulation telemetry data, biometry data, and psychometry data to objectively assess pilots' performance during adaptive flight training.

The Data Science for Real-Time Decision-Making Research Chair led by Dr. Andrea Lodi is providing research methodology teaching and data science expertise.

CANADA
EXCELLENCE
RESEARCH
CHAIR

DATA SCIENCE
FOR REAL-TIME
DECISION-MAKING

1.2 Problem Statement

Commercial and military flight operators are experimenting issues on pilot recruitment, training, and retention. There is a need to optimize the efficiency of training programs to address the shortage of pilots, accelerate pilot learning, and reduce training cost. National defense organizations as well as commercial airlines are looking to increase flight safety. After the Germanwings Flight 9525, where the crash was caused by a pilot with suicidal tendencies, aviation regulator pays more attention to psychology and the mental state of pilots. The crash of Air France Flight 447, where inappropriate situational awareness by the crew showed that the increasing automation in aircraft cockpit brought new complexity on the skills required in the flight operation. An assessment of the mental state and the technical and non-technical abilities of pilots during the pilot training process can play a fundamental role in the safety of aviation organisation.

1.3 Research Objective

An Intelligent Adaptive Flight Training System (IAFTS) focus on providing tools to assess pilot performance, training recommendations, and various descriptive, diagnostic, predictive, and prescriptive insights from data analytics to support adaptive flight training. As shown in Figure 1.1, the analysis of the inputs of the actors, i.e., pilots, instructors, and flight training experts, will be at the centre of the research while artificial intelligence technologies data processing, system automation, optimization, and machine learning techniques will continually improve the quality of pilot performance within the learning environments such as flight simulators, aircraft, eLearning, and classroom.

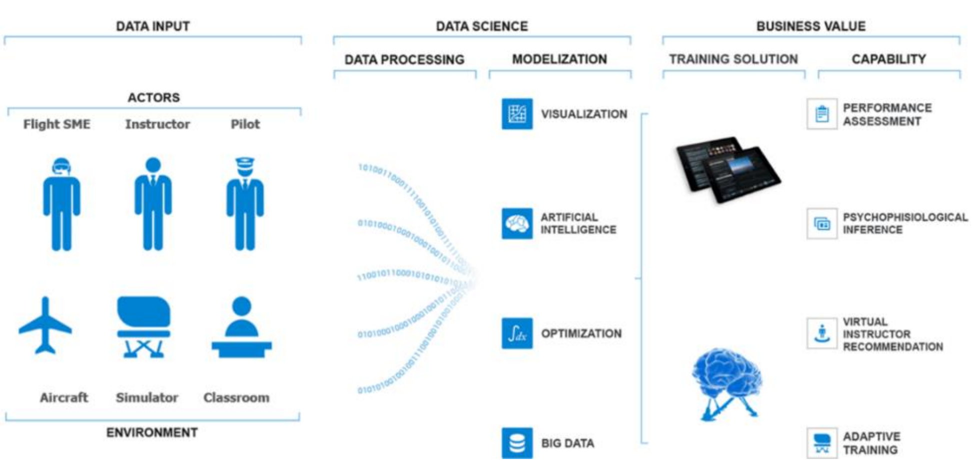


Figure 1.1 - Data Science for Adaptive Flight Training

The goal of the research is to discover the data science capabilities required to support an intelligent and adaptive flight training system and assess where artificial intelligence can enable and/or augment autonomous decision-making using a variety of data collected from a full-flight simulator, biometric sensors, and flight training operation data. More specific objectives are:

- Assess pilot performance against a standard operation procedure using machine learning techniques.
- Optimize the decision-making with adaptive flight training in real time based on assessment of pilot performance, instructor evaluation, and training scenarios.
- Analyze the behaviour of the pilots and instructors and model their cognition and decision-making processes with neuroscience and psychometry to enhance machine learning models.

We also aim that this research will serve as new insights for regulators such as the Federal Aviation Administration (FAA), Transport Canada, and the European Union Aviation Safety Agency (EASA), as well as aviation standards organization such as The International Air Transport Association (IATA) and the International Civil Aviation Organization (ICAO) that supports aviation with global standards and air transport policies.

1.3.1 Research question & Hypothesis

Based on the collection of data from learning environments, the following research questions focus on the automation of an adaptive flight training system and support users for real-time decision-making:

"How can we implement an artificial intelligence solution with various machine learning models using heterogeneous data such as biometry, psychometry, flight telemetry, and training operation data to provide autonomous real-time decision-making capability for adaptive flight training?"

We hypothesize that the introduction of intelligent agents integrated into a training process will allow for the maximization of pilot's learning, maximize teaching performance of the instructor, and optimize overall training program efficiency.

1.4 Contribution, originality, and impact

The introduction of an integrated intelligent agent in the flight training process is a novelty and their integration into real-time flight training environment that offers a wide variety of data at high frequency is a challenge that was successfully addressed.

We obtained an increase in the quality, objectivity, and maturity of the assessments by combining various models using various data sources in a multi-modal framework. As an example, we demonstrate that using human performance assessment combined with machine learning performance assessment will increase the objectivity and the quality of the assessment. We also demonstrate that the performance assessment and neuroscience can provide new insight to offer better human-machine interactions. We conceive a way to adapt a training scenario in real-time using flight simulation automation capability in order to sustain the pilot's engagement and motivation to learn. Alternatively, psychometric data collection can improve the pilot profile where the new metrics can be used to adapt the training program and provide insight to the instructor while doing his pedagogical duty.

1.4.1 Scientific Contribution

As a scientific contribution, we provide a methodology to solve the subjectivity of instructor's assessment with the combination of flight training expert knowledge and historic instructor evaluation to augment the labeling of machine learning models. We also found a methodology to bring objective data from a high variety of sensors to capture the pilot behavior and assess flight performance on both the technical skills and the soft skills using biometry and neuroscience. This methodology helped us to find a way to profile a pilot based on the mental model using psychometry that captures aptitude and preferences in aviation domain. We also identified how artificial intelligence can support autonomous adaptive flight training using to support instructor's pedagogy as a human-in-the-loop system. We finally identified how AI and data can augment training system technological capability in a leaning & competency framework and ensure maturity evolution within a Flight Training Organization (FTO) compliant to CBTA with the used of standards and global federation approach on a flight training center network.

1.4.2 Industrial contribution

This research took place in the context of an enterprise digital transformation program from a flight training and simulator manufacturing company, CAE Inc., where, by transforming a world-leading manufacturing company into a training-focused company, we stand a better chance of having that contribution paid off and driving innovation in the world's most important flight training organization.

This transformation is done through change management considering the maturity of the business. With a Capability Maturity Model (CMM) applied to Flight Training, we went into the direction of establishing a standard in order to provide regulators and aviation authorities a structured approach enabling and certifying an intelligent adaptive flight training system.

As a more direct impact on the flight training industry, augmenting the flight training with adaptive flight training system capability can improve the training delivery by increasing the instructor’s situation awareness and reduce pilot workload, and maximize the training and learning performance or knowledge transfer during flight simulation. This may allow an instructor to train more students to respond to the increasing pilot demand in the aviation industry.

We also demonstrated an engineering efficiency improvement by reducing engineer effort in systems expert’s manual rules creation and tuning that may be gargantuan regarding the number of aircraft simulation model to support. All automatic assessment rules, event detection rules, simulation control rules, and view selection rules need to be replicated in hundreds of aircraft models to become a training product that is sustainable to deploy at a large scale. Furthermore, there are multiple maneuvers for a multitude of airline policies. Having engineering to define specifically all rules may be a serious loss of time in comparison to a machine learning approach. So, our machine learning approach is a serious option to generalize system engineering in flight training.

With twelve granted patents, four published, and five in the publication process, we contribute to the flight training community patent portfolio, and we present concrete & applied solutions that were implemented in a real industry engineering process and deployed around the world in Asia, North America, and Europe. The development of our solution is aligned with the competency framework of the various training programs as well as considering their specificity in a variety of learning & training methodologies.

Finally, we contributed to enhancing the automation of flight training process with scalable method and apparatus to collect flight training data and deploy artificial intelligence solutions in a cold start context of machine learning, from ground zero, to provide new predictive measurement of human behavior operating immersive training devices.

1.5 General Methodology

The general methodology (Figure 1.2) consists of choosing analytics techniques that can enable building an intelligent flight training system that will receive a big variety of data type to provide autonomous decision-making for the main actors involved in flight training.

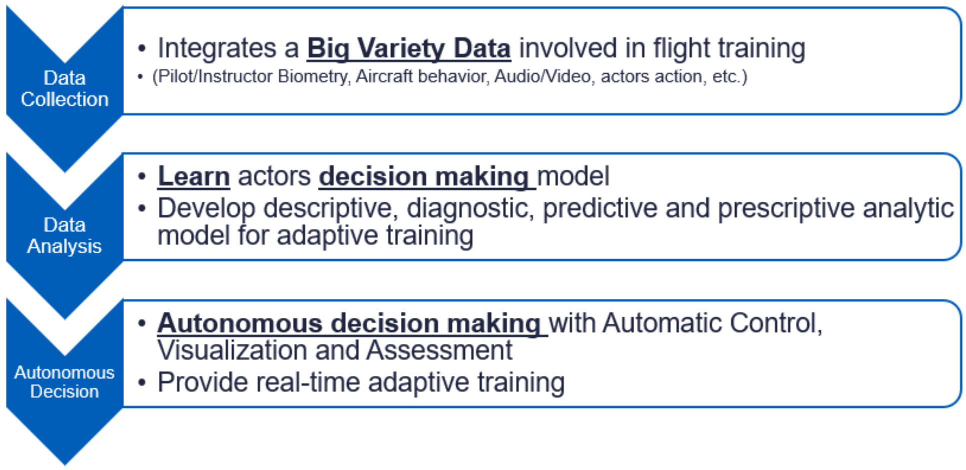


Figure 1.2 - General methodology

A huge variety of data have been used in this research such as human-machine interaction data, biometry, psychometry, flight simulator telemetry, instructor's assessment, training content, standard operation procedures, training scenarios and crew audio/video that are offered through sensors timeseries, training operational data store or multimedia files.

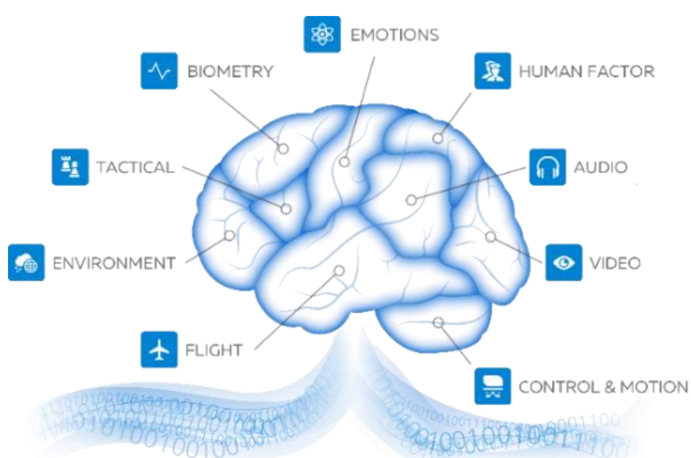


Figure 1.3 - Big Varity of Data

Simulated aircraft flight parameters, synthetic environmental conditions and training context data are some of the primary sources of data of the adaptive flight training process demonstrated at Figure 1.4. To identify the training context, data from the lesson plans system, standard procedure repository, scenario manager, training management system and learning management system are used to contribute to the descriptive, diagnostic, predictive or prescriptive models of performance analysis.

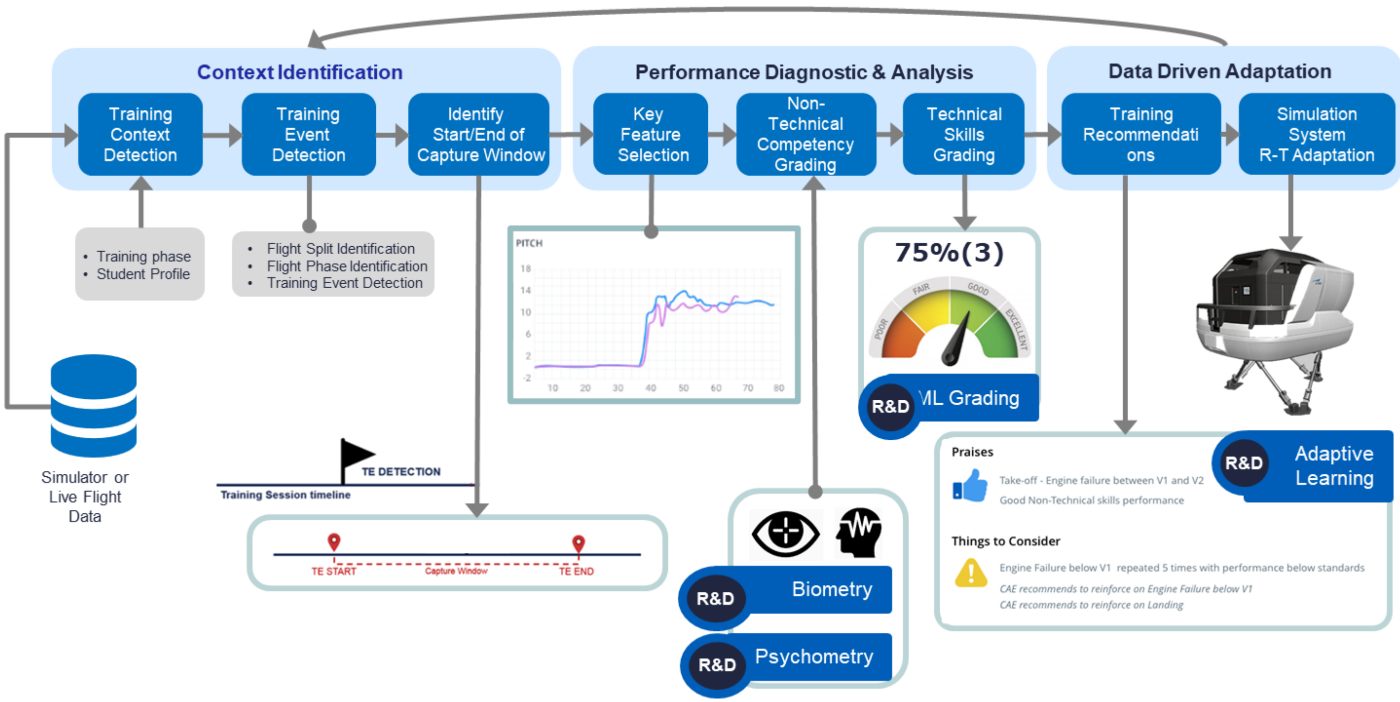


Figure 1.4 - Adaptive Flight Training Process

We first built an automatic pilot performance assessment capability using flight training expert and Instructor knowledge (Chapter 5). The deployment required simulation data collection and insertion with existing training session with e-grading capability in a digitalization process of an organization. The digitalization included the integration of information systems that can enable the profile of a pilot with his certifications and flight training history.

The pilot profile, composed of their flight history, human behavior and performance records (Figure 1.5), is the other important dataset built and used to feed the analytics engines.

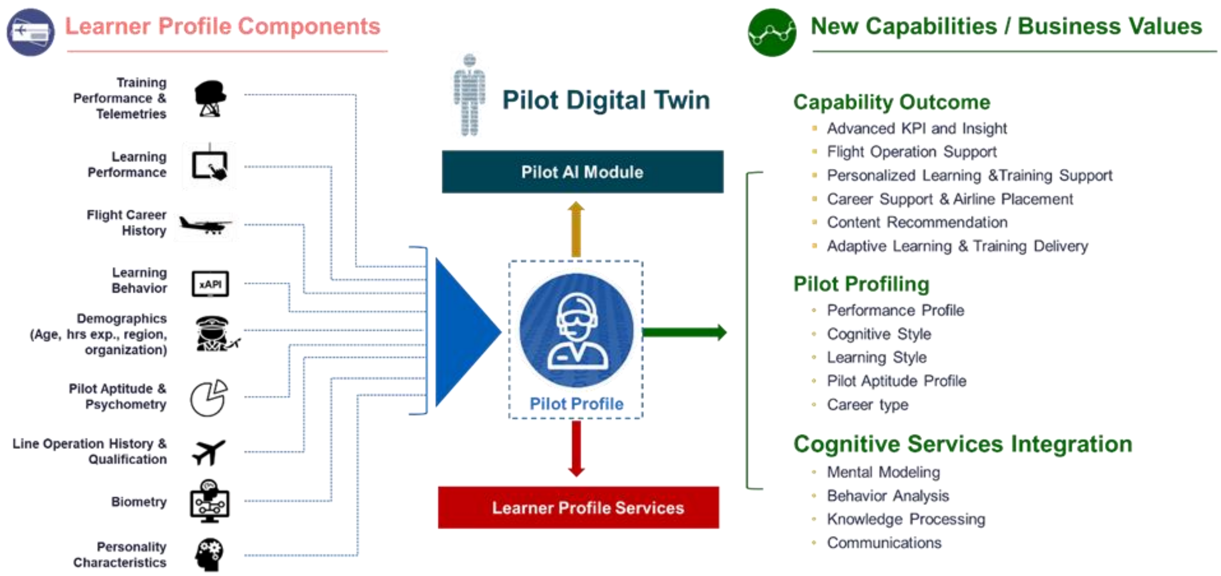


Figure 1.5 - Pilot Digital Twin Concept

Combined with the training program's learning objectives and the training content, the pilot profile is the essential data source for an AI solution (Figure 1.6) to generates insights and recommendations that can be broadcast through a variety of training support systems or dashboards.

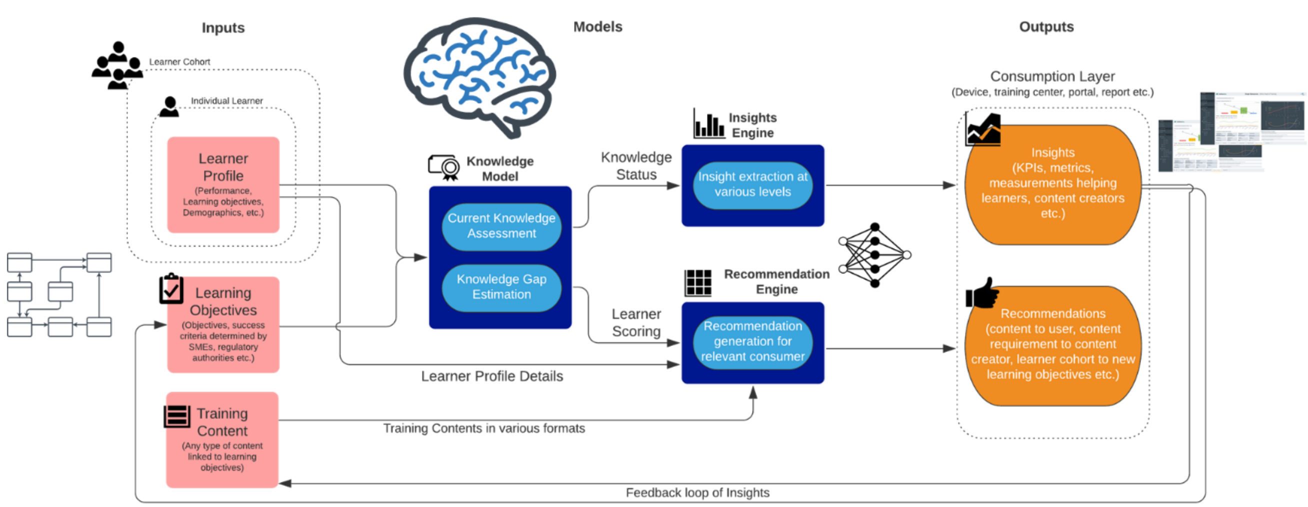


Figure 1.6 - High-level Architecture of a training analytics platform

To complete the pilot profile and to ensure a more robust human performance assessment, the introduction of more data around the human behavior was required. With biometric data capture, we can gather a new level of insight with an objective measure of the pilot using neuroscience. The introduction of neuroscience is a new trend where various research has been executed in the current research project on both commercial and military aviation sectors. A research review with over 300 papers have been surveyed on neuroscience and biometric sensors for the pilots and the instructors in aviation. This enabled partnership with university labs during the research project.

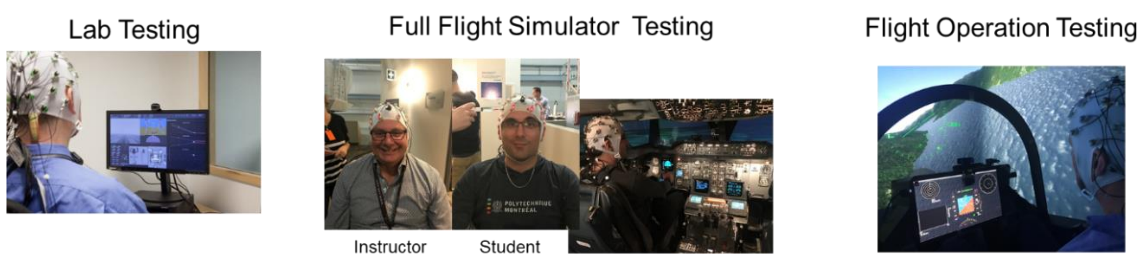


Figure 1.7 - Neuroscience Experimental Environments

We then moved to static lab with instructors that assessed multiple maneuvers recorded by an expert pilot in a B737MAX. This led us to explore the experimental protocol definition and understand the challenges around data synchronization and their playback. We then moved into a full-flight simulator of an Airbus 310 aircraft with a variety of biometric sensors. This experiment helped us to understand the challenge related to the operation environment and evaluate the robustness of the biometric sensors with the proximity of the cockpit system and potential electronic interference or movement interference. We were then ready for a full-flight simulation experiment (Chapter 8) with a cohort of novices to study the behavior of human in progressive maneuvers complexity during the initial training of a fastjet operation.

To complete the human performance analysis, we went in the direction of psychology (6.2). Using psychometry, we were able to bring a complete 360 of the pilot performance. With AI and clustering methodology combined with the flight performance history, we were able to use pilot aptitude testing data to assess both pilot’s cognitive model and preferences during an existing pilot selection process offered to civil commercial airlines.

With data accumulated from the multitude of training center during the research project, we were able to use the federated machine learning approach (Chapter 6) to scale from multiple dataset structure and build adaptive flight training model with recommendation system and performance prediction (Figure 1.8). This enables us to prepare the organisation for scalability and deployment methodology using multiple learning devices as part of the leaner journey of an adaptive flight raining program.

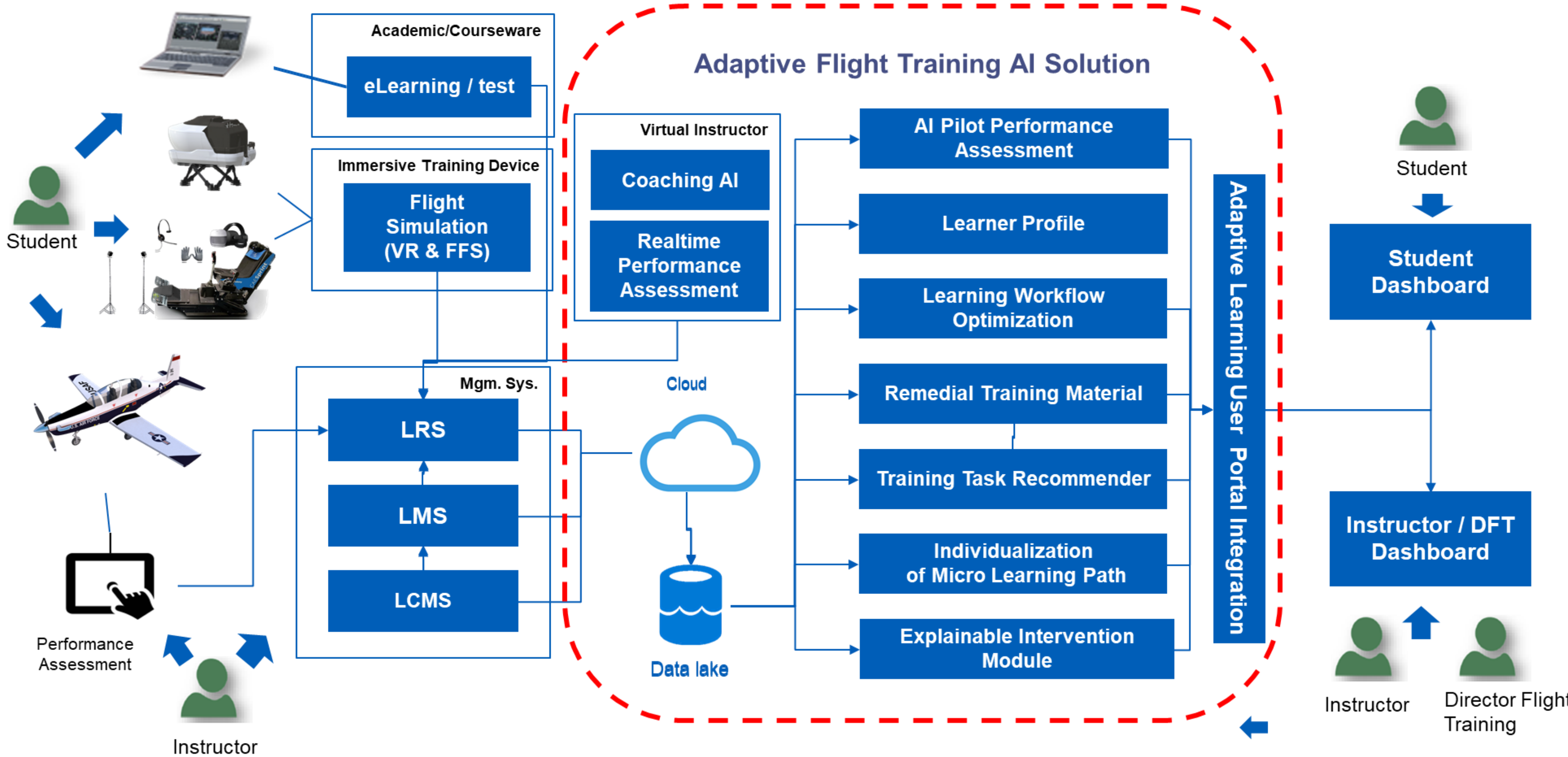


Figure 1.8 - Adaptive Flight Training AI Solution

CHAPTER 2 LITERATURE REVIEW

This project offers us a deep reflection on the integration of artificial intelligence, human factor, and flight training to ensure the maximum of training capacities through the various learning environments. It is around these three axes, which we will present the literature review.

The current research proposal aims to augment flight training with artificial intelligence model with the human-in-the-loop learning process. Figure 2.1 presents the structure of the literature review.

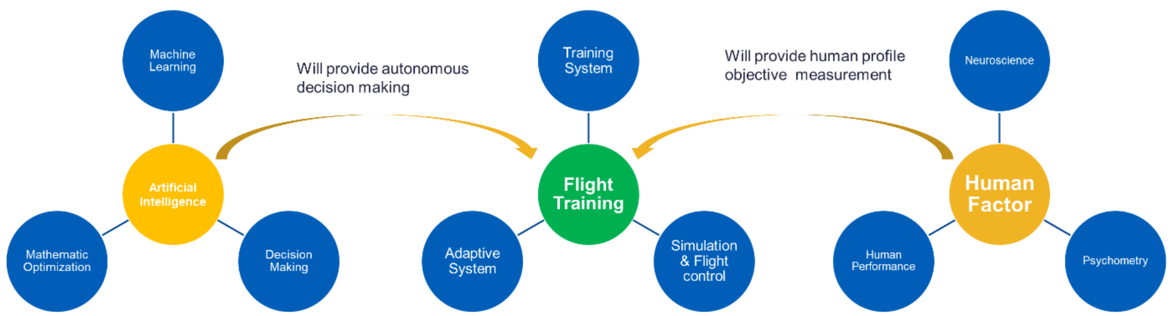


Figure 2.1 - Literature Review Concept and Key Words

This literature review will first introduce flight training essentials before proceeding with flight training system environments. In an evolved learning environment, these training systems can be intelligent, adaptive, and provide great educational support in the training process. Aircraft simulation is at the core of the training system and will also be reviewed to provide an overview of the main objective of improving flight safety.

With the increased demand for pilots and instructors (CAE, 2017; The Boeing Company, 2017), the need to augment instruction & learning performance calls for higher technology support. Artificial Intelligence is an obvious choice in the implementation of an intelligent adaptive system. Machine learning techniques will be reviewed in the second part of this literature review. Because of the variability of the aircraft type, sensors, and maneuvers in the curriculum, it is essential to review optimization techniques that can be scaled and that takes into consideration human-in-the loop decision-making in the machine learning process.

The last section of the literature review will focus on the human factor in the decision-making process. Our challenge leads us to use objective data to ensure robust AI decision-making. Neuroscience will bring evidence in the machine learning process by using biometric tools that will reinforce the intelligence of the training system and help understanding of the foundation of human performance in flight training.

2.1 Flight Training

Flight training means the orchestration of training events in aircraft or flight simulators in accordance with a training curriculum. The objective of flight training is to provide pilots or crew members an opportunity to acquire skills, attitudes and knowledge of performance against a standard operation procedure. A flight training session provides instruction for the teaching of the procedures to prepare the crew for their duties. A flight training curriculum corresponds to the collection of segments that includes many modules that define training events and associated conditions and can be arranged into lesson plans. The pilots are required to successfully meet the requirement of a curriculum segment to complete the course. The assessment is the measurement of a crew performance in the execution of a training event. Assessment criteria describe how the crew must perform the associated tasks. A criterion-Referenced Assessment is done against established standards or criteria.

2.1.1 Flight Training System

A full-flight simulator can support a training event with an immersive environment for the crew and flight training supporting systems for the instructors to maximize learning objectives of the training sessions of a training event. Instructor Operation Stations (IOS) allows the instructor to efficiently conduct a simulator training session. The IOS is comprised of a user interface that provides simulation control and monitoring capabilities on the various aircraft systems and synthetic environmental conditions. Lesson plans systems and electronic training records also provide, respectively, automation of the lesson plan flow to reduce the instructor's workload and grading of the crew performance to provide evidence for the assessment by the instructor.

(Branaghan et al., 2011) proposed an IOS user interface redesign based on domain expert knowledge representation. Using hierarchical card sorting techniques and hierarchical cluster analysis, the paper presents a methodology to reorganize menu structure of an IOS. This knowledge representation can be used in this research to provide structure in the training system to be more inline with adaptive flight training by providing dynamic user interface at the IOS.

2.1.2 Adaptive Learning Environment

2.1.2.1 Intelligent Tutoring System

(Chieu & Herbst, 2010) proposes a computer-based simulator for apprentice teachers to interact with virtual students to learn how to teach. A Bayesian network was used for modelling the teacher's decision-making. A Bayesian network was also used by (Remolina et al., 2004) to represent student models used by an Assessment Manager, module of Intelligent Flight Training architecture, to decide which skills the student should work on next in rotor-wings training in the U.S. Army. A student model for Intelligent Tutoring System can be extended to crew debriefing as presented by (Swigger, 1988) and his Intelligent Effective Reviewer, designed to assist in debriefing Air Force bomber crews. Because structure is required between training event and lesson plans, these techniques are considered to build adaptive flow control in a Lesson Plan System (LPS) that can provide automation of a flight simulator control.

(Bass, 1998) extended the model with the instructional model as part of an Intelligent Instructor Pilot Decision Support System's architecture as referred to in Figure 2.2. Expert Model also serves to assess the state of the simulation in line with the curriculum and the instructional model. These experts', instructional and student models are in line with the proposed research project.

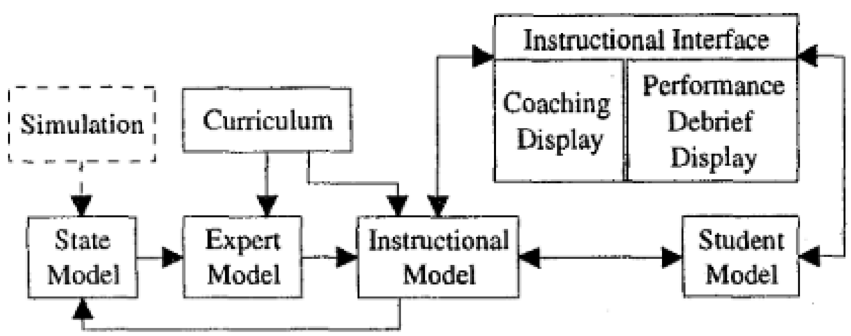


Figure 2.2 - IP Decision Support System Architecture (Bass, 1998)

(Ludwig & Ramachandran, 2005) goes further than using students' past performance to build intelligent tutoring systems. By using students' preferred learning style and stated-based adaptation, an Adaptive Instructional System can adapt to student anxiety by focusing on the student's attention

2.1.2.2 Learning Analysis

Similar to full-flight simulation, gaming provides a great environment for exploratory analysis of student learning behaviour. (Gibson & de Freitas, 2015) study a case to predict the assessment related to knowledge and skills acquisition from collected user actions in a game. The analysis consists of searching the patterns of action that can be correlated with the student's assessment of his or her performance. A similar approach was used by (Loh & Li, 2016), but the paper goes further by prescribing training based on the decision-making profile of the student. By using serious games for measuring skills acquired, an analysis can be made of the learning profiles of players such as fulfillers, who appear single-minded once they find a workable route, explorers who switch from one route to another, and quitters who tend to abandon the game. Serious games are excellent tools to analyze what a player would do in a work environment.

Learning Analytics can be combined with Educational Data Mining to enhance the impact of adaptive training as presented by (Papamitsiou & Economides, 2014). With an efficient review of previous research, the authors point out key studies analysis:

- Student behaviour modelling
- Prediction of performance
- Increase self-reflection and self-awareness
- Prediction of dropout and retention
- Improve feedback and assessment services
- Recommendation of resources

These analysis topics are relevant to flight training and will be used to highlight the capability of adaptive training in this research.

2.1.2.3 Aircraft Performance Analysis

Flight Training involves the learning aspect, but certainly involves an understanding of the aircraft itself. As we will see in chapter 4, an automatic performance assessment is usually done around a training event detection that offers us a time range of parameters that will be evaluated. This training event has similarities to safety or anomaly event detection. Aircraft performance analysis is key in the study of adaptive flight training. Focusing on detecting aircraft performance anomalies (Eric Chu, 2010b) uses a similar approach to machine learning by using fleet data to build a regression model to detect anomalies of new aircraft data, as shown in Figure 2.3. The paper approach is data-driven, without prior knowledge of the aircraft model. This approach will certainly help in the current research project to scale our solution for multiple aircraft types.

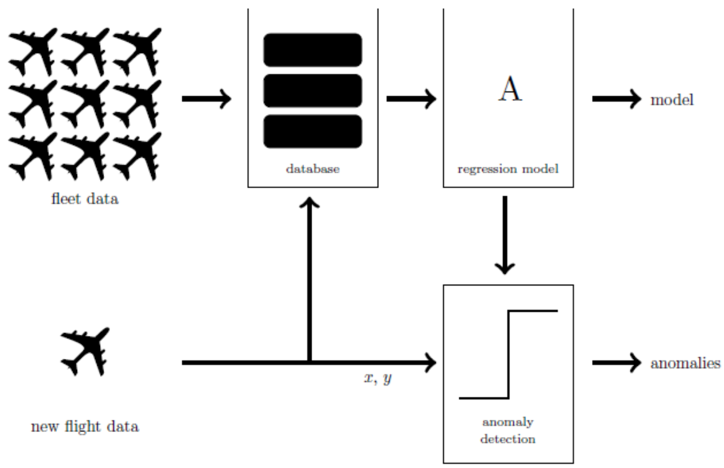


Figure 2.3 - Anomalies detection information flow chart

(Matthews et al., 2013) Also bring the methodology of discovering Anomalous Aviation Safety Events. By presenting an Aviation Safety Knowledge Discovery (AvSKD) process (Figure 2.4), the authors clearly identify where domain expert contribution will be necessary. The flow chart also presents a methodology for taking care of discrete data and continuous data like that needed in the current research.

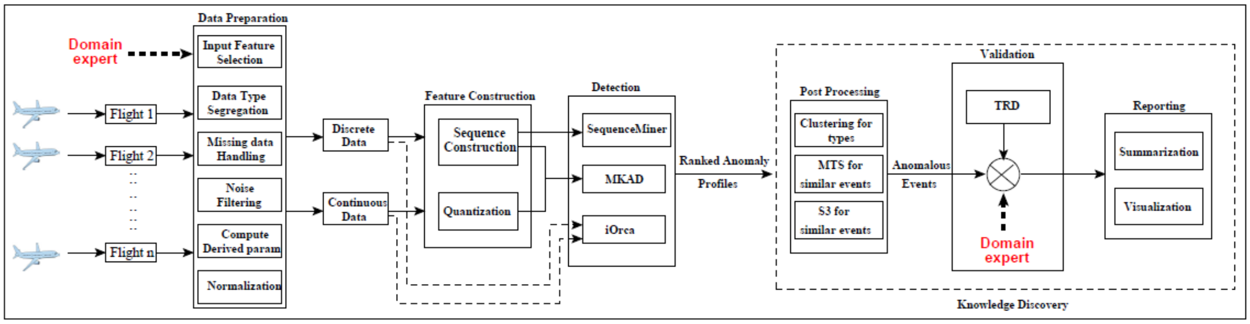


Figure 2.4 - Aviation Safety Knowledge Discovery (AvSKD) process for identifying abnormal aviation events.

The anomaly detection algorithms use a Multiple Kernel Anomaly Detection (MKAD) which is designed to run on heterogenous data sets that results in the presence of multiple attribute types, continuous and discrete. The MKAD is a one-class SVM model that constructs optimal hyperplane solved by the following optimization problem

$$\min Q = \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

$$subject\ to\ \ 0 \leq \alpha_i \leq \frac{1}{l_v}, \sum_{i} \alpha_i = 1, p \geq 0, \qquad 0 < v < 1$$

Where $where\ \alpha_i's$ are Lagrange multipliers, *l* is the number of examples, *v* a user specified parameter that defines the upper bound on the training error, and the lower bound on the fraction of training points that are support vector, *p* is a bias term and *K* is the kernel matrix

(Luxhøj, 2013) presents a model for predictive Safety Analytics of Complex Aerospace Systems, the Aviation System Risk Model (ASRM) that can be used to evaluate the causal factors that lead to unsafe states.

The ASRM uses a probabilistic approach of Bayesian Belief Network (BNN). Figure 2.5 presents the general structure of BNN with random variables represented as circles and decision nodes as rectangles.

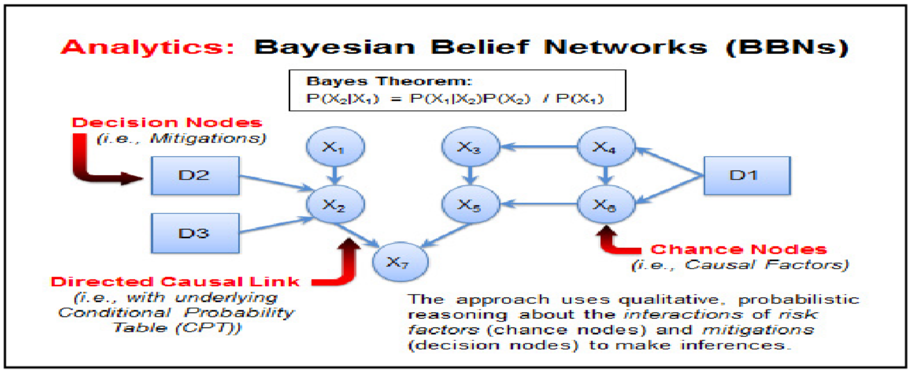


Figure 2.5 - General structure of a Bayesian Belief Networks (BBN)

During the current research project, Bayesian Network is considered as an alternative approach for detecting training events, representing them as a network where flow can be adapted as well as establishing a causal model to help explain the root cause of a pilot performance.

Bayesian Network was also used by (Chang et al., 2011) (Bayesian Network to Manage Learner Model in Context-Aware Adaptive System in Mobile Learning) to manage learner models in the Context-Aware Adaptive System. Figure 2.6 presents the layered architecture of their Context-Aware Mobile Learning Architecture.

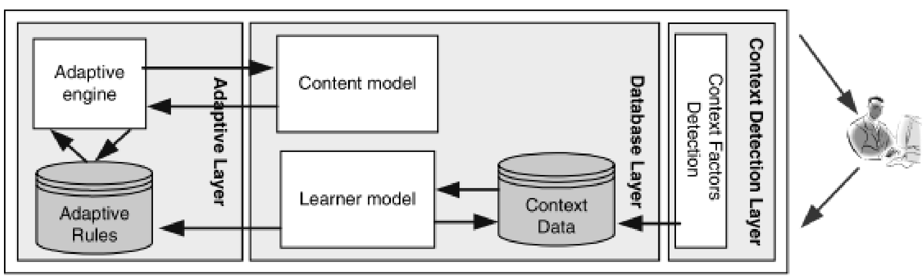


Figure 2.6 - Context-Aware Mobile Learning Architecture

The Bayesian network was used to quantify the level of knowledge understood by a learner and used this quantitative value as a basis for adaptive content selection. In a rule-based approach, the authors present two strategies of quantitative reasoning: *Diagnostic reasoning*, used in the case of misunderstanding of a concept, and *predictive reasoning* used to determine quantitively the value of the probability of the understanding level of a concept. In the present research project, such an adaptive engine and learner model may be considered as the first prototype before applying machine learning implementation.

2.2 Artificial Intelligence

2.2.1 Machine Learning Introduction

The goal of machine learning is to detect patterns in data to predict future data outcomes. To properly introduce machine learning concepts this section will be greatly inspired by (Hastie et al., 2013)

A set of variables, the inputs, can influence another set of variables, the outputs. Machine learning problems can be categorized in three main types:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

In supervised learning, the goal is to use the input to predict the values of the outputs. We can build model $f(x)$ that will predict the class of input X.

In unsupervised learning, we can build a density estimator that models the distribution of X to predict the density of $p(x)$. We can also try to discover the underlying structure of the data with clustering, dimensionality reduction and characteristic extraction that are known as descriptive models. The unsupervised learning process uses unlabelled data to train the descriptor.

In reinforcement learning, we can build an artificial agent that learns in order to decide which actions to execute in a changing environment to maximize the reward.

2.2.1.1 Supervised Learning

A predictor that handles classification problems is called a classifier. In classification problems, the response variable G is qualitative. The supervised learning process uses labelled data to train the classifier to eventually assign a class label to a future unlabelled observation X.

We can also build models that will predict expected value Y from an input X, $E[Y|X= x]$. We call that predictor a regressor that will handle regression problems.

Two methods are easy to introduce in supervised learning: Least Squares and Nearest Neighbour. The first is a parametric model that can have the advantage of being fast but make assumptions about the nature of the distribution. The second is a non-parametric model that has the advantage of being more flexible but is computationally difficult for large datasets.

2.2.1.1.1 Least Squares

Given a vector of inputs $X^T = (X_1, X_2, \ldots, X_p)$, we predict the output Y via the model.

$$\hat{Y} = \widehat{\beta_0} + \sum_{j=1}^{p} X_j \hat{\beta}_j \text{(Hastie et al., 2013)}$$

To fit the linear model to a set of training data, we pick the coefficient β that minimizes the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

Using an example from (Hastie et al., 2013), Figure 2.7 presents a scatterplot of training data in input *X1, X2* and where output variable *G* can take value *orange* and *blue* and where fitted values are converted to a fitted class variable $\hat{G}$ as follows:

$$\hat{G} = \begin{cases} \texttt{ORANGE} & \text{if } \hat{Y} > 0.5, \\ \texttt{BLUE} & \text{if } \hat{Y} \le 0.5. \end{cases}$$

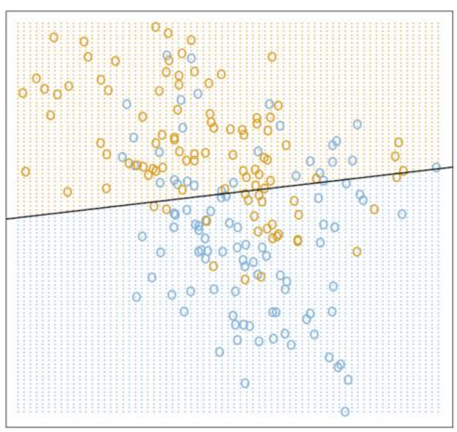


Figure 2.7 - Linear classification example (Hastie et al., 2013)

2.2.1.1.2 Nearest Neighbour

The Nearest Neighbour methods use observations in a training set closest in input space to x to form $\hat{Y}$. With $N_k(x)$ the neighborhood of x defined by the k closest points x_i in the training set, we have:

$$\hat{Y} = \frac{1}{k} + \sum_{x_i \in N_k(x)} y_i$$

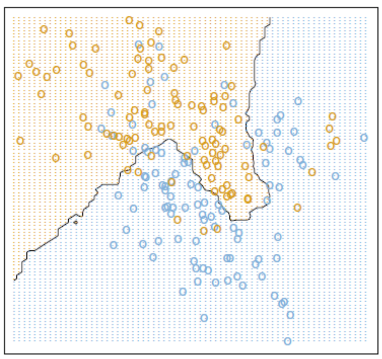


Figure 2.8 - Predictor with 15-Nearest Neighbour

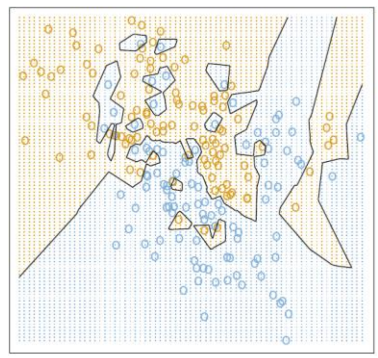


Figure 2.9 - Predictor with 1-Nearest Neighbour (Hastie et al., 2013)

This leads us to the most important question of the selection of k as a hyperparameter of the model. Figure 2.8 and Figure 2.9 present, respectively, prediction for $K = 15$ and $K = 1$, the number of Neighbours used in the training process. If low k seems a good predictor with the training set, the error becomes catastrophic with the test data set.

2.2.1.2 Training, Validation and Test Errors

The main goal of supervised learning is to predict a target from new inputs different than the one used in the training set. So, the objective is to train models to obtain a small **generalization** error.

Hence, the tuning of the model will be very important to make sure we have a good generalization error. Figure 2.10 shows an example of the error between train and test data for various values of K.

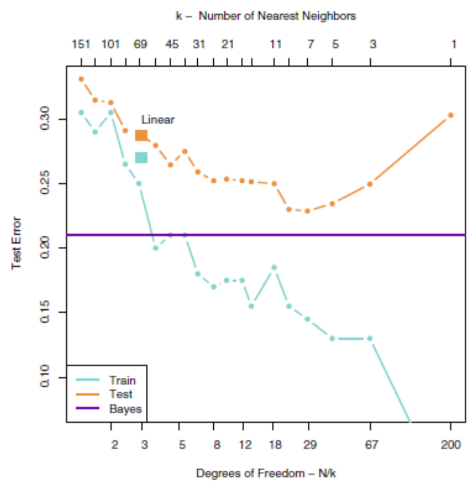


Figure 2.10 - Nearest Neighbour Misclassification curves for various K (Hastie et al., 2013)

A training error is calculated from the observations used in the training set. The test error is the overall error of predicting the output with new observation. The training error is often different from the test error. The training error can underestimate the test error depending on the complexity of the model and size of the data.

When a large set of test data is not necessarily available, a method of using a third data set called *validation set* can be used. In the validation set approach, the data is first trained in the training data set; found models are then used to predict responses in the validation set. The error from the validation set can provide a good estimate of the test error. MSE or Misclassification rates can be used for, respectively, a quantitative or a qualitative response.

K-fold cross validation is an approach to test error estimation by randomly dividing the data set into K sets, leaving one of the subsets out and training the model with the other K-1 combined subset. Prediction is obtained with the set previously left out. The process is done for each subset K. The cross-validation rate is obtained applying Mean Square Error on all K subsets:

$$CV_{(k)} = \sum_{k=1}^{k} \frac{n_k}{n} MSE_i \text{ where } MSE_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k}$$

and $\hat{y}_i$ is the fit for observation I, obtained from the data with part K removed. (James et al., 2013).

2.2.1.3 The Curse of Dimensionality

Nearest Neighbour can be good for small dimension p. However, the methods can be lousy when p is large because of the *curse of dimensionality*. In such a problem, the Nearest Neighbour may be far away in high dimensions. Because the problem needs a reasonable fraction of the data to average, this fraction neighboured in high dimensions may no longer be local, so we lose the spirit of local averaging. As shown by the Figure 2.11, the distance of neighbours augment significantly when *p* is large. Since it's hard to find near neighborhood in high dimensions and stay local, the need for having a more structured and parametric model becomes evident.

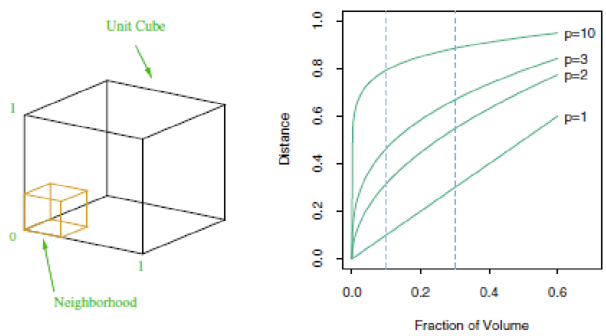


Figure 2.11 - Distance of neighbours in different dimensions

2.2.1.4 Linear regression

Linear regression is a very common model for regression and can be written as follows:

$$y(\mathrm{x}) = \mathrm{w}^T\mathrm{x} + \epsilon$$

Where $\mathrm{w}^T\mathrm{x}$ is the scalar product between the input vector and the weight vector, with ϵ as the residual error between the true response and the prediction.

2.2.1.5 Logistic regression

Because linear regression in classification might produce probabilities less than zero or bigger than one, logistic regression may be more appropriate to estimate *p(Y=1|x).*

Logistic regression uses the form $p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$. Where p(x) will have values between 0 and 1 and is called the s*igmoid* function.

Maximum Likelihood is used to estimate parameters and is defined as:

$$l(\beta_0\beta) = \prod_{i=1} p(x_i) \prod_{i=0} (1 - p(x_i))$$

In multi-class problems, linear regression is not appropriate because classes may be in interchangeable order. Here multiclass logistic regression or discriminant analysis is more appropriate.

2.2.1.6 Bias-Variance trade-off

Suppose we have fitted a model $\hat{f}$ (x) to some training data and let (x0, y0) be a test observation if true model is $y(\mathrm{x}) = \mathrm{w}^T\mathrm{x} + \epsilon$ with f(x) $= E[Y|X = x]$ *then*

$$E(y_0 - \hat{f}(x)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias(\hat{f}(x_0))\right]^2 + Var(\varepsilon)$$

Where $E(y_0 - \hat{f}(x)^2$is the expected prediction error, $Var\left(\hat{f}(x_0)\right)$ is the variance between different training sets, $\left[Bias(\hat{f}(x_0))\right]^2 = E\left[\hat{f}(x_0)\right] - f(x_0)$ where $E\left[\hat{f}(x_0)\right]$is the average prediction at x_0 across different training sets and f(x0) the true function. As shown in Figure 2.12, we can see that as the flexibility of $\hat{f}$(x) increases, the variance increases and the bias decreases. So, the flexibility (or complexity) of the model is a bias-variance trade-off and can lead to underfitting or overfitting. The goal of machine learning is to learn a model that can **generalize**. This trade-off analysis will be important to provide models that can provide generally acceptable error on test set.

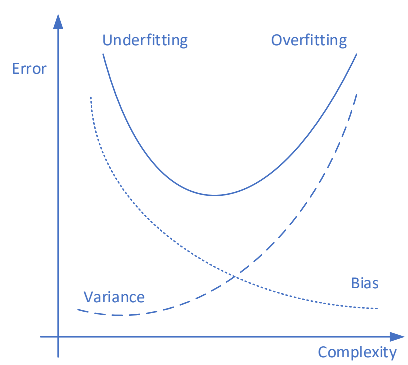


Figure 2.12 - Bias-Variance Trade-off

2.2.1.7 Unsupervised Learning

In unsupervised learning, we only observe input values X1, X2, …, Xp. The data is unlabelled, which offers a certain advantage knowing that labelling data sometimes requires human intervention. In unsupervised learning, we are looking for special structures in the data. Clustering, density modelling, and Principal Component Analysis are examples of methods in unsupervised learning.

2.2.1.8 Clustering.

Clustering goals is to find subgroups or clusters in a data set. The data is partitioned into groups where observations are quite like each other. With *K*, the number of clusters, the objective is to estimate the distribution *p(K|D).*

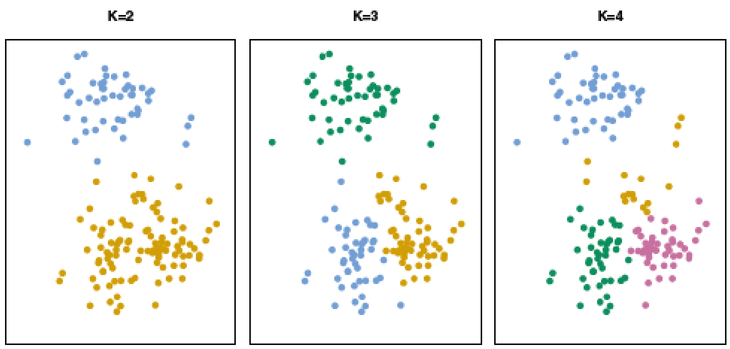


Figure 2.13 - *K*-means clustering with various values of *K* (James et al., 2013)

2.2.1.9 Deep Learning

Machine learning has advanced well in recent years, and is predominant with the advent of **deep learning** (Bengio, 2009). This method will be used in the project by first used multi-layer neural network and then test various architecture appropriate for timeseries such as Recurrent Neural Network or Convolutional Neural Network.

Multi-layer neural networks can be used to make a prediction or classification. As recalls by the author, deterministic transformations are computed in a feedforward way from the input x, through the hidden layers h^k, to the network output $h\ell$, which gets compared with a label y to obtain the loss $L(h\ell, y)$ to be minimized. (Bengio, 2009).

As presented Figure 2.14, layer k computes an output vector h^k using the output h^{k-1} of the previous layer, starting with the input $x = h^0$, $h^k = \tanh(b^k + W^k h^{k-1})$ with parameters b^k, a vector of offsets and W^k, a matrix of weights.

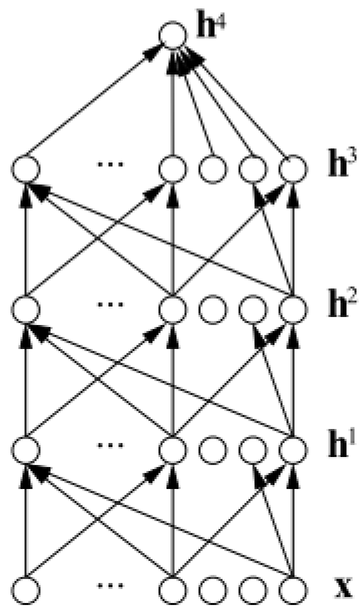


Figure 2.14 - Multi-Layer Neural Networks (Bengio, 2009).

2.2.1.10 Time Series Classification.

Because flight parameters analysis involves timeseries, various methodologies need to be identified to apply Time Series Classification (Smith-Jentsch et al.). Convolutional Neural Network (CNN), generally used in Computer Vision and Recurrent Neural Network (RNN) are Deep Learning approaches that can be explored in the research project in order to increase machine learning prediction using flight simulator data.

2.2.1.10.1 CNN for TSC

CNN as great success in image recognition. An approach presented by (Hatami et al., 2017) suggests using a transformation of a timeseries in 2D images and then uses CNN on it. As shown by the Figure 2.15, the authors propose a 2-stages CNN architecture for Time Series Classification (Smith-Jentsch et al.).

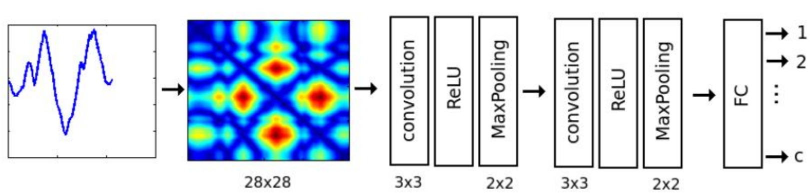


Figure 2.15 - 2-stages CNN architecture for Time Series Classification (Hatami et al., 2017)

2.2.1.10.2 RNN for TSC

As recalls by (Husken & Stagge, 2003), Recurrent Neural networks (RNN) can be used for the prediction and classification of timeseries. In their paper, the authors used an Elman-Network RNN topology that is composed of a feed-forward part and a memory part to perform the classification of timeseries.

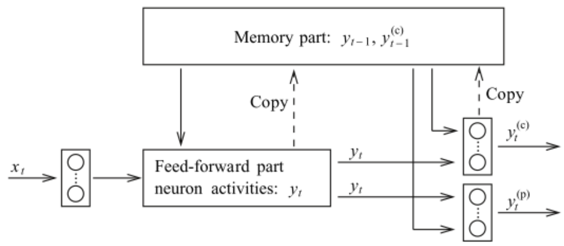


Figure 2.16 - Elman-Network RNN Topology

2.2.2 Discriminant Analysis

The Discriminant Analysis approach is to model the distribution of X in each class and then use the Bayes theorem to obtain p(y|x). By recalling the basic probability theories:

- Joint probabilities: $p(A,B) = p(A \wedge B) = p(A|B)p(B)$,
- Conditional probability: $p(A|B) = \frac{p(A,B)}{p(B)} if\ p(B) > 0$,
- Chain rule: $p(x_{1:n}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1) \ldots p(x_{1n}|x_{1:n-1})$
- Gaussian distribution: $\mathcal{N}(x|\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
- Laplace distribution: $Lap(x|b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$

the Bayes' theorem formulation is: $p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j))}$

To build a Bayes classifier from density estimations, we separate the training set in *m* subset containing all points of the same class *c* and we train a density estimator on each of them.

$$\hat{p}_c(x) \simeq p_{x|y}(x|c)$$

We then determine the a-priori probabilities of each class. $\hat{p}_c \simeq P_Y(c) = P(Y = c)$

Then we apply Bayes' rule to get a-posteriori probability at x.

$P_{Y|X}(c|x) = \frac{P_{X|Y}(x|c)P_Y(c)}{P_X(x)}$, where $P_{Y|X}(c|x)$ is the posterior and $P_Y(c)$ the prior.

Naïve Bayes classifiers, assuming conditional independence: $p(y, \mathbf{x}) = p(y) \prod_{j=1}^{n} p(x_j|y)$

2.2.3 **Mathematical Optimization**

Machine Learning involves mathematical optimization. Integer Programming can offer techniques in order to optimize feature selection, hyperparameters and possibly a consensus in the integration of multiple machine learning decisions.

2.2.3.1 Integer Programming introduction

A Mixed Integer Linear Programming problem (MILP) in the form:

$$\min \{c^T x: Ax \geq b, x \geq 0, x_j \in \mathbb{Z}, \forall j \in I\}$$

MILP problems belong to the NP-hard complexity class, meaning there is no known algorithm that can solve it in a polynomial time. To solve MILP for a large size problem, we can apply relaxation by removing the integrality requirement to obtain a Linear Programming problem that can be solved in polynomial time and has the form: $min \{c^T x: Ax \geq b, x \geq 0\}$. We then apply iteratively branch-and-bound and cutting plane algorithms. This section is heavily based on Andrea Lodi's lecture notes (Lodi, 2017).

2.2.3.2 Branch & Bound and Cutting Plane

In the Cutting plane method, we first recall a MILP formulation as defined previously and the following two sets: $S := \{Ax \leq b, x \geq 0, x_j \in \mathbb{Z}, \forall j \in I\}$ and $P := \{Ax \leq b, x \geq 0\}$

The method consists of iteratively strengthening by solving $\max\{c^T x: x \in P\}$ $and\ get\ x^*$, if $x^* \in S$ Then Stop, solve the separation problem and add $\alpha x \leq \beta$ to P and resolve.

The branch-and-bound algorithm iteratively partitions the solution space into sub-MILP by creating two children by rounding the solution of the LP relaxation value of a variable x_j constrained to be integral.

$$x_j \leq \lfloor x_j^* \rfloor \ OR \ x_j \geq \lfloor x_j^* \rfloor + 1$$

We then apply the LP relaxation continuously on the resulting two children until the LP relaxation is directly integral or infeasible. At every step, two important decisions are made: The node selection, where we can choose to select either a best-bound first or a depth-first strategy, and the Variable selection, where we should choose which variable will be used to partition the current node.

2.2.3.3 MILP models

Multiple MILP models will be explored during this research to optimize decision-making. Few of well known of these models is introduced in this section.

2.2.3.3.1 The Facility Location Problem

Given *m* clients to serve, *n* facilities, for each facility *j*, f_j, is the cost of opening the facility *j* and for each client *i* and each facility *j*, c_{ij} is the cost of serving client *i* by facility *j*

Incapacitated Facility Location Problem aims to determine which facilities need to be open and which one among the open facilities serves each client with the objective of minimizing the overall cost of opening and solving. This problem can be a model with binary decision variables,

$$y_j = \begin{cases} 1, & if\ facility\ j\ is\ opened \\ 0, & otherwise \end{cases} \quad \text{and} \quad x_{ij} = \begin{cases} 1, & if\ client\ i\ is\ assigned\ to\ facility\ j \\ 0, & otherwise \end{cases}$$

With objective function:

$$\min \sum_{j=1}^{n} f_j y_j + \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\, x_{ij}$$

$$\text{Subject to: } \sum_{j=1}^{n} x_{ij} = 1,\ i = 1, \dots, m$$

$$x_{ij} \le y_j, \quad i = 1, \dots, m, j = 1, \dots, n$$

The Capacitated Facility Location problem is a variant of the Uncapacitated Facility Location problem where clients are associated with a demand d_i and each facility has a capacity b_j and an additional constraint:

$$\sum_{i=1}^{m} d_i\, x_{ij} \le\ b_j y_i, \quad j = 1, \dots, n$$

A special case of the Capacitated Facility Location problem is the Bin Packing Problem where service costs are null and facilities have equal opening cost and capacities, leading to a new objective function:

$$\min \sum_{j=1}^{n} y_j$$

2.2.3.3.2 Set Covering, Partitioning, Packing

A special case of the UFLP is the Set Covering Problem-SCP where cost can be either 0 or ∞ and have the following formulation:

$$min\ c^T x$$

$$Ax \ \geq \mathbf{1}$$

$$x \in \{0,1\}^n$$

$Where\ A \ \in \{0,1\}^{mxn}, and\ \mathbf{1} = all - 1\ vector.$

There are two variants of the Set Covering Problem, the Set Partitioning and Set Packing Problems that are obtained by replacing the SCP inequalities by '=' and '≤':

- Set Partitioning: $min\ c^T x$, subject to $Ax = \mathbf{1},\ x \in \{0,1\}^n$
- Set Packing: $min\ c^T x$, subject to $Ax \ \leq \mathbf{1},\ x \in \{0,1\}^n$

2.2.3.3.3 Stable Set

Consider a non-oriented graph *G – (V, E)* with weight p_j for each vertex $j \in V$, a Stable Set of *G* is a subset of vertices $S \subseteq V$ such that no edge connects two vertices in S directly. The MILP model is obtained using the following natural binary variables, objective function and constraints:

$$x_{ij} = \begin{cases} 1, & if\ client\ i\ is\ assigned\ to\ facility\ j \\ 0, & otherwise \end{cases}$$

$$\max \sum_{j \in V} p_j x_j$$

$$x_i + x_j \leq 1, \qquad (i,j) \in E$$

$$x_j \ \in \{0,1\}, \qquad j \in V$$

A stronger version can be obtained by using the notion of *clique*. A *clique* is a subset of vertices $K \subseteq V$ such that all pairs of vertices are directly connected by an edge. A clique is maximal if and only if it does not exist in another vertex in V\K directly connected to every vertex of K by an edge in E. With $\mathcal{K}$ the collection of all maximal cliques of *G*, a strong model is obtained by replacing $x_i + x_j \leq 1$ by $\sum_{j \in k} x_j \leq 1, k \in \mathcal{K}$.

2.2.3.3.4 Vertex Colouring

Vertex Colouring Problem calls for assigning a colour to vertices of an undirected graph *G = (V, E)* with *n* vertices and *m* edges, such that vertices directly connected by an edge receive different colours, and the number of used colour is minimum. With the following binary variables:

$$y_j = \begin{cases} 1, & if\ color\ is\ used \\ 0, & otherwise \end{cases} ;\ x_{ij} = \begin{cases} 1, & if\ vertex\ i\ is\ colored\ by\ color\ j \\ 0, & otherwise \end{cases}$$

The problem can be formulated as:

$$\min \sum_{j=1}^{n} y_j$$

$$\text{Subject to:} \textstyle\sum_{j=1}^{n} x_{ij} = 1,\ i \in V$$

$$x_{ij} + x_{hj} \leq y_j, \qquad (i,h) \in E, j = 1, \dots, n$$

$$y_j, x_{ij} \in \{0,1\}, \qquad i \in V, j = 1, \dots, n$$

As the stable set, this weak formulation can have an exponentially high number of constraints and variables and can be strengthened by replacing constraints $x_{ij} + x_{hj} \leq y_j$ with the constraints:

$$\sum_{i \in k} x_{ij} \leq y_j, k \in K, j = 1, \dots, N$$

2.2.3.3.5 Travelling Salesman

The Travelling Salesman Problem (TSP) calls for determining a tour of a complete graph G = (V, A) with cost $c_a = c_{(i,j)}$ for each arc $a = (i,j) \epsilon A$ in Asymmetric TSP where each vertex I ϵ V will be visited exactly once at minimum cost. The MILP model has the following binary variables:

$$x_a = \begin{cases} 1, & if\ arc\ a\ belongs\ to\ the\ tour \\ 0, & otherwise \end{cases}$$

With vertices having exactly one incoming and one outgoing arc for each vertex of G, the MILP model can be formulated as follows:

$$\min \sum_{a \in A} c_a x_a$$

$$\sum_{a \in \delta^-(i)}^{n} x_a = 1,\ i \in V\ ;\quad \sum_{a \in \delta^+(i)}^{n} x_a = 1,\ i \in V$$

$$x_a \ \in \{0,1\}, \qquad a \in A$$

The last two constraints correspond to the so-called *degrees constraints* and correspond to the assignment problem. *Subtour Elimination Constraints* needs to be added to ensure that the solution is a unique and complete tour as reads as follows:

$$\sum_{a \in A(s)} x_a \le |S| - 1, \qquad SV, 2 \le |S| \le |V| - 2$$

The Symmetric Travelling Salesman problem is a variant of the ATSP for an undirected graph *G = (V, E)* with cost $c_e = c_{(i, j)}$ for each edge $e = (i,j) \in E$. The problem is formulated as follows:

$$x_e = \begin{cases} 1, & if\ edge\ e\ belongs\ to\ the\ tour \\ 0, & otherwise \end{cases}$$

$$\min \sum_{e \in E} c_e x_e$$

$$\sum_{a \in \delta^-(i)}^{n} x_e = 2, \qquad i \in V$$

$$\sum_{a \in \delta^+(i)}^{n} x_e \ \le |S| - 1, \qquad s \subseteq V, 2 \le |S| \le |V| - 2$$

$$x_e \ \in \{0,1\}, \qquad e \in E$$

2.2.3.4 Maximum Consensus

The Maximum Consensus problem is an instance of Maximum Feasible Subsystem problems and can be formulated as a MILP problem.

As presented by (Speciale et al., 2017), maximum consensus is a NP-hard combinatorial optimization problem. Maximum consensus seeks to find a parameter estimation criterion giving a set of measurements $M = \{ai, bi\}, i = 1, \dots N$ and a threshold ϵ. The objective function is:

$$\max_{x,\varsigma \subseteq M} |\varsigma| \quad \text{, subject to } \gamma i(x) \leq \epsilon, \forall\, Mi \in \varsigma, A(x) \geqslant 0$$

2.2.4 Machine Learning model integration

With different types of machine learning such as supervised, unsupervised, semi-supervised or reinforcement learning, the data scientists have a range of techniques to train their Artificial Intelligence software. As the aim of this project is to integrate multiple data sources within multiple models, it is pertinent to consider the consensus work of multiple machine learning processes.

An effective way to make these different methods work together is possible with **Consensus Maximization,** which is presented by (Jing et al., 2013); they propose this approach by combining several supervised and unsupervised learning models. Because the current research may be faced with partially labelled data, semi-supervised may be a good option. Figure 2.17 well describes the positioning of consensus maximization, like majority voting of a supervised learning approach.

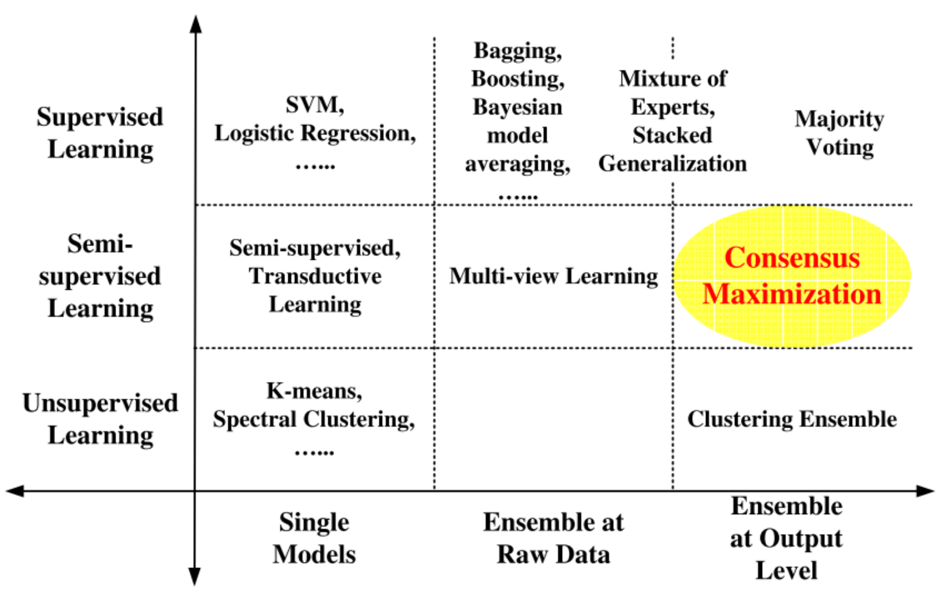


Figure 2.17 - Position of Consensus Maximization

The combining of supervised learning models is also presented by (Bosoni et al., 2016) for the building of models that provide better predictors of health outcomes. (Bekolay et al., 2013) highlights the limitations of previous approaches to combining supervised and unsupervised learning. The paper presents an approach that remains functional during **online learning**, which is very relevant to this research. Essentially, his approach uses a spiking neural networks and Semantic Pointer Architecture method, which was used to create a functional model of the brain by (Eliasmith et al., 2012).

Supervised learning can also be combined with reinforcement learning. (Ye et al., 2003) present this combination by proposing a fuzzy controller with supervised learning assisted by reinforcement learning to improve obstacle avoidance of motorized robots. The field of video games also shows methods for leveraging supervised and reinforcement learning. (Miyashita et al., 2017) presents a method that allows the Artificial Intelligence of a video game to learn how to play as a human from a player's performance database. During the game, reinforcement learning allows the AI to beat the human with the objective of adapting the game to beat the opponent. As this can be frustrating for the player, the game, which promotes the player's entertainment, then uses a player's performance database to learn to play like a human, thus providing a level of competitiveness equal to that of a human opponent. This concept therefore seems to be very much aligned with the objective of this research project to offer an adaptive training to realistic piloting while considering the human skills in learning AI.

As presented by (Settles, 2012), **Active Learning** is an interactive learning method that allows for the development of new hypotheses in a continuous interactive learning process. This form of learning can involve an external agent to label non-tagged data in passive learning. This technique is also used in the detection of credit card fraud where an investigator will label transactions in almost-real-time. (Fabrizio Carcillo, 2017) presents this case with an effective reminder of the method of intelligently selecting the data to be tagged by a third-party agent that will be used in the retraining process. **Active learning in harmonization with semi-supervised learning** is presented by (Zhu et al., 2003) by combining them under a Gaussian random field model where the labelled or non-labelled data is represented as nodes and a weighted graph.

Still with the aim of integrating several models under data heterogeneity, (Dong et al., 2015) includes the principle of **online consensus maximization,** which adds to the real-time aspect that interests us in the current research.

As explored in our research project, this consensus maximization can be solved by Mixed-Integer Linear Programming, thus leading us to the mathematical optimization section.

The deep learning model presents a multitude of neurone layers in which a neurone's activation function can be activated based on a 0-1 maximization problem modelled by Mixed Integer Linear Programming-MILP. (Fischetti & Jo, 2017) presents a feasibility study of this maximization problem.

Mixed-Integer programming can also be useful for **machine learning optimization**. (Maldonado et al., 2014) present a performance optimization of Support Vector Machine-SVM algorithms by **MILP formulation on features selection**. The feature selection approach was also optimized by (Jr. & Rubin, 2003) that presents a branch-and-bound feature selection as a mixed-integer programming approach to multi-class data classification problems.

If mixed integer programming can help machine learning, the other way around is still valid. Decisions within a **continuous branch and bound algorithm can be optimized using machine learning**. (Khalil, 2016) presented this approach using an on-the-fly variable selection.

Mixed integer linear programming-based machine learning was also applied to cancer treatment by (Poos et al., 2016). This last paper also presents a use of cross-validation algorithms in a computational workflow that offers a prediction regulator for different validation data, resulting in a better performance in the identification of cancer cases.

Mixed Integer Linear Programming can also be used in cooperation with **control of autonomous multi-agent systems,** as presented by (Earl & D'Andrea, 2002). The latter uses a similar approach to (Richards & How, 2002) who uses it for aircraft trajectory planning.

Since machine learning is about minimization of error, as in supervised learning, (Lodi et al., 2016) propose a MIQP formulation for the ramp loss model of a machine vector support algorithm.

The MILP has been able to optimize the field of aviation at the level of the operation of airlines. **Cumulative Assignment Problem** presented by (Lodi et al., 1999) is also an interesting technique that can be useful in flight crew and instructor assignment scheduling if the concept is also used for flight training.

(Jiefeng Xu, 2006) presents an algorithm tabu search algorithm to schedule a flight instructor for a flight training session. There is also (Jacobs, 2014), which presents a Flight Training Scheduler FTS using a mixed-integer linear optimization program that pushes the student towards a training syllabus.

It is therefore natural to integrate the MILP with AI for even more capacity in the field of aviation, which offers ever more complex and increasingly autonomous systems.

2.2.5 Artificial intelligence in the aerospace sector

The field of aerospace has a lot of technology that is supported by intelligent systems. Air Traffic Systems, Unmanned Aerial Vehicles (UAVs), Autonomous UAVs and spacecraft provide us with a complex environment that drives us to use intelligent assistance to manage their operations while maintaining flight safety concerns.

The autonomy of vehicles is also used in the field of space. (D'Angelo et al., 2017) presents the use of **machine learning in autonomous Spacecraft** via Markov Decision Process Formulation using Dynamic Programming.

Neural networks can be applied to flight control. (Ippolito et al., 2007) proposes a **flight controller augmented with a neural network** that is nourished with data collected from cameras and sensors in order to allow online learning. (Aydogan Savran, 2006) also presents a smart approach to **adaptive aircraft flight control using neural networks** in a high-performance aircraft to model the dynamic behaviour of an F16. A dynamic identification model is developed using a multi-layer neural network combined with a Proportional Integral Derivative–PID controller to provide adaptive flight control.

The latest machine learning techniques such as deep learning or reinforcement learning are also used in aviation. (Guan et al., 2016) uses deep learning in the trajectory prediction of Air Traffic Management (ATM) systems. Historical aircraft flight data are used as a training data set to construct predictive trajectory models.

Reinforcement learning is presented by (Balakrishna et al., 2008) with an approach to estimate taxi-out phasing to reduce fuel consumption, gas emissions and operational costs. Using the Federal Aviation Authority's Aviation System Performance Metrics (ASPM) database, a Markov Decision Process (MDP) algorithm was tested at the John F. Kennedy International Airport.

These intelligent autonomous systems are generally coupled with ground stations to control their flight plan and ensure communications. Aircraft are also controlled by the air traffic controller network. The pilots must learn to enter into collaboration with the air traffic controllers to guarantee the safety of the air traffic. These ATCs are also supported by artificial intelligence. (Kulkarni, 2015) presents the automation of Air Traffic Control systems using neural networks. It uses a Backpropagation Network algorithm to add the necessary intelligence to the ATC system and reduce the workload of the air traffic controller.

This intelligent and autonomous capability drives us to analyze the role of humans and their performance in error detection and flight control input errors when interacting with complex systems. (Cox et al., 2000) presents a real-time algorithm based on neural networks to detect the Pilot Induced Oscillations (PIO). These PIOs are a succession of oscillations resulting from an undesirable interaction between the pilot and the aircraft and represent a significant danger that can cause accidents. In a context of distributed control, to understand the interaction between machine and man becomes essential.

2.3 Human Factor

2.3.1 Neuroscience

2.3.1.1 Introduction to Neuroscience (Riedl & Léger, 2016)

The nervous system and the brain are the source of human behaviour and are studied by the science of neuropsychology. Cognitive neuroscience focuses on the study of higher functions such as learning, memory, language, emotion, attention and their neurobiological determinants.

Neuroergonomics studies the brain and behaviour in a work setting by relating the neurobiological basis of mental function and physical performance to technology and tasks to be executed. Neuroergonomics also contributes to understanding neuroadaptive interface, mental workload, multitasking, human error, learning, stress, fatigue and skills acquisition, all very important factors in aviation and flight training.

Biological state analysis can predict behaviour and can be used to adapt the system. As an example, user interface can be adapted in real time based on biometry indicating cognitive and affective states. These biometry indicators can also provide user biofeedback to help self-regulation in the performance of actions. Systems with such capability are called Neuro-adaptive information systems, which are systems that recognize the physiological state of the user and that adapt in real time (Riedl et al., 2014).

2.3.1.2 The human nervous system (Riedl & Léger, 2016)

The nervous system is composed of the central nervous system, the brain and the spinal cord, as well as the peripheral nervous system (PNS). Information detected by the receptors is brought via the spinal cord to the brain, which processes the information. The brain then sends commands via the PNS that can be of somatic or autonomous type. The autonomous nervous system is important to consider in neuroscience research because it is associated with emotions, mental workload or stress. The ANS is responsible for the involuntary functions of the human body and can be composed of the sympathetic nervous system, in action when a threat is perceived, and the parasympathetic, responsible for the body at rest.

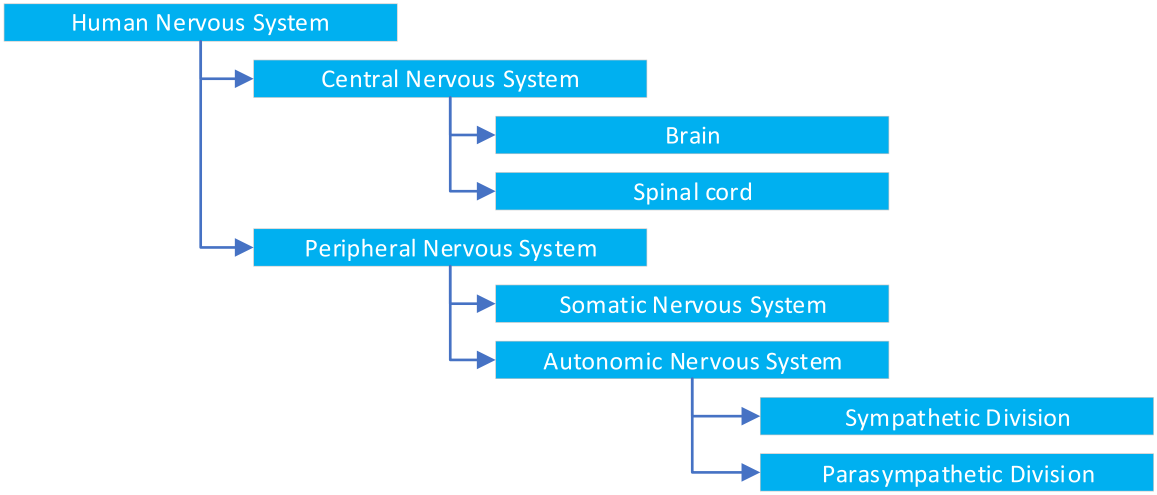


Figure 2.18 - Human Nervous System

2.3.1.3 Neurophysiological tools (Dimoka et al., 2012).

Neuroscience provides evidence on the psychophysiological state of the subject. Self reports can be biased by unconscious or deliberately inaccurate data resulting from the desire for social acceptance. Neuroscience tools can track and measure unconscious processes of the human body and basic perceptions. With continuous real-time data, they can provide analytical precision and allow the temporal ordering of constructs and the inference of causal relationships.

However, the cost of neuroscience tools is high and they are not very accessible. The intrusiveness of most of them creates an artificial environment that limits validity. The data extraction and analysis are labour-intensive and provide various responses to stimuli from the various baselines of a user's profile. Because a neurophysiological measure can be linked to several theoretical constructs, the results can be difficult to interpret and multiple measures may be required to avoid mono-operationalization bias. These technical factors are only one facet in comparison to the ethical issues, such as the manipulation of behaviour, that neuroscience can give rise to.

During the execution of the task, a subject can enter into cognitive overload when too much information and inference difficulty arises. In the design of the user interface, neuroscience can help reduce cognitive overload. By measuring cognitive overload in the brain, systems can be tested and refined to enhance user capabilities.

Various psychophysiological tools are common to measure the peripheral nervous system and hormone system:

- Eye tracking: eye pupil location and movement,
- Skin Conductance Response: sweat in eccrine glands,
- Facial Electromyography (EMG): electrical impulse caused by muscle fibres,
- Electrocardiogram (ECG): electrical activity of the heart on the skin.

Brain imaging tools are also available:

- Functional Magnetic Resonance Imaging (fMRI): neural activity by changes in blood flow,
- Positron Emission Tomography (PET): metabolic activity by radioactive isotopes,
- Electroencephalography (EEG): electrical brain activity on the scalp,
- Magnetoencephalography (Strait & Scheutz): changes in the magnetic fields of brain activity,
- Functional Near Infrared (fNIR): changes in the oxy-hemoglobin and deoxy-hemoglobin concentration in the cortical tissue.

Cost, intrusiveness and accuracy properties of these neuroscience tools can be factors in choosing which one, or which combination, of them to use. Accuracy of tools can be defined in terms of temporal and spatial resolution. As recalls by (Riedl & Léger, 2016), temporal resolution is a property of a measurement tool that describes the time between the beginning of the stimulus perception and the measurement of the signal by the tool. The spatial resolution describes how precisely activity in the bran can be localized.

As an example, EEG can offer millisecond temporal resolution but very poor spatial resolution. fMRIs offer spatial resolution of a few millimetres but have poor temporal resolution with few second and the tool is invasive

2.3.1.4 Neuroscience Research Methodology (Riedl et al., 2014)

Methodological quality is the most important factor for assessing the psychophysiological state of users. Reasons for adequately defining the construct of a conceptual domain are:

- Helps clarify what the construct does,
- Reduces the probability of inadequate or contaminated construct's indicators,
- reduces invalid conclusions of construct relationship.

During a research experiment, these methodology themes need to be considered:

- Reliability: proportion of total variance due to random error,
- Sensitivity: values differentiation along a continuum,
- Objectivity: results independent of investigators and replication,
- Intrusiveness: inference with the user tasks,
- Validity: instruments measure the construct that it is supposed to measure,
- Diagnosticity: construct measurement precision as opposed to other constructs.

2.3.1.5 Affective computing (Picard, 2010)

Emotions are not irrational; they are involved in decision-making and action selection. They are essential for the human to act rationally and they play a major role in perception, memory and attention in decision-making. They also help to handle complex unpredictable inputs in real time.

The Artificial Intelligence research focus is on verbal and mathematical intelligence, but emotions may be just as important. Successfully building intelligent machines will require the incorporation of emotions. With neuroscience, researchers can elucidate how emotion works and how to use this knowledge to build intelligent systems.

2.3.1.6 Physiological computing (Fairclough, 2008)

Physiological computing systems employ real-time measures of psychophysiology. Six fundamental issues of these systems are identified:

- Complexity of the psychophysiological inference,

- Validating the psychophysiological inference,
- Representing the psychophysiological state of the user,
- Designing explicit and implicit system interventions,
- Defining the bio-cybernetic loop that controls system adaptation,
- Ethical implications.

Adaptive systems must respond proactively and implicitly. Preferences of the user can also be learned and bring a shift from master-slave to collaborative relationships that require awareness of the user in real-time.

Among adaptive responses of the system to an undesirable user state are offering assistance, adapting the level of challenge to sustain task engagement, and incorporating an emotional display to reinforce emotions.

Psychophysiological inference tries to map physiological measures onto psychological states. Sensitivity of the psychophysiological inference is a vital attribute to enable a physiological computer system to respond in a timely and appropriate fashion to change user state.

Representation of the user is an important aspect of the system design that determines the range of adaptive strategies available to a bio-cybernetic loop and the level of intelligence exhibited by the system. The bio-cybernetic loop encompasses the decision-making process underlying software adaptation and incorporates the psychophysiological inference implicit in the quantification of those trigger points used to activate the rules. In a safety-critical system such as autonomous function in the aircraft, it is the top priority for the loop by preserving effectiveness to minimize the risk of accident.

2.3.1.7 Neuroergonomics (Parasuraman, 2003)

Neuroergonomics merge notions of ergonomics and neuroscience to understand the brain function underlying human performance. With the increasing number of automated systems and the emergence of brain imaging technologies, there are more interests for human factors analysis in the workspace, especially in Joint Cognitive Systems where cognitive science studies the mind interacting with systems.

Adaptive systems are well studied in neuroergonomic research. In adaptive systems, the work is shared between human and machine. Because the division of tasks between the human and the machine is not predefined during the design phase of system engineering and context dependent during the operation of the system, a proper design methodology is required; neuroergonomics-design can support this. As an example, automated support is a requirement when the operator mental workload is high, and not required when the operator becomes disengaged from the system.

When an individual commits an error, a neural mechanism is activated. The error-related negativity (ERN), related to perceived accuracy, allows identification, prediction and prevention of errors in real-time.

Skill acquisition can also be analyzed with neuroscience. Procedural learning is generally thought to progress through a series of stages that can be better understood through functional brain imaging techniques. As an example, fMRI can be used to examine the acquisition of motor skills and give evidence of the multiple stages of learning.

These use cases of system adaptation are in line with aviation needs to provide better safety in the pilots' operation of the flight.

2.3.2 Psychophysiological inference

Since current research involves multiple measures of cognitive state using multiple methods, it will be relevant to use proven methodologies to evaluate the validity of psychophysiological constructs. Multi-Trait-Multi-Method presented by (Ortiz de Guinea et al., 2013) consists of a table of correlations arranged in a way to facilitate the interpretation of construct validity. The method uses a combination of neurophysiological instruments and self-reported instruments to assess the validity of cognitive load, engagement and arousal, states that are also going to be studied in the current research.

2.3.3 **Neuroscience in aviation**

(Schnell et al., 2008) presents a framework to assess task performance of trainees based on physiological measures in flight training. The Quality of Training Effectiveness Assessment (QTEA) concept, as summarized in Figure 2.19, has the objective of allowing instructors to assess a student in real-time using sensors that can measure the cognitive and physiological workload.

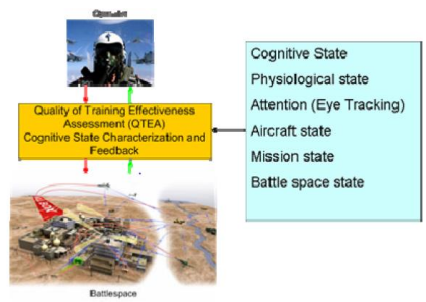


Figure 2.19 - QTEA concept

The framework also proposes a scenario adaptation in real-time. One of the paper's goals was to use neurocognitive and physiological measures in conjunction with flight performance measures to determine the effect of the simulation in flight training. However, data collected were insufficient to provide a conclusion. By using a thermal camera, EKG & High-res body temperature, Skin Galvanic Response, Pulse Oximetry, High Density EEG, Respiration Belt and Eye tracker, the paper presented the technical readiness of these various sensors with advantages and disadvantages. The architecture described the flow of collected data from sensors as shown in Figure 2.20. However, this research didn't provide enough data to valid hypothesis of the architecture. This research may contribute to this avenue.

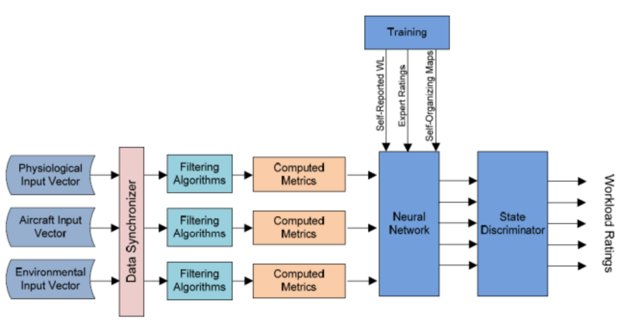


Figure 2.20 - Simplified Flow of Input Vectors from Sensors to Classifier (Schnell et al., 2008)

2.3.4 Neuroscience in an Intelligent Adaptive System

As recalled by (Ming Hou, 2014), research suggests that real-time integration of the operator's psychophysiological and performance state provides a true human-in-the-loop system. A human-in-the-loop-based design of an Intelligent Adaptive System can provide operator workload and fatigue reduction by assigning tasks to the machine during high-stress, and reengagement of the operator during the under-load to improve situation awareness. The author also points out that technological advances in Artificial Intelligence and augmented cognition can support the implementation of this kind of intelligent system. A Technology Readiness Level model to categorize usage of psychophysiological measurement tools is proposed to identify diagnostic and augmentation maturity. The conceptual framework presents the usage of fNIR, EDR, EEG, ECG and other biometric sensors to support diagnostics capability to augment the capability of an Intelligent Adaptive System.

2.3.5 Psychometry

Successful flight requires not only flight skills, but also the ability to work well as a team. In the article by (Hedge et al., 2000), a project developed and validated Crew Resource Management (CRM) skills test for Air Force transport pilots. The CRM emphasized the importance of certain knowledge, skills and abilities (KSA) related to the crew rather than continuing to measure personality or related non-cognitive traits using the more traditional personality inventory methodology.

The Situational Test of Aircrew Response Styles (STARS) is a CRM skills test designed to measure problem-solving, decision-making, knowledge, communication, crew management, and interpersonal effectiveness attributes most critical for success as an Air Force aircraft commander. The STARS involve four primary steps: situation generation, response options generation, item reviews, and response option scaling. The results of the validation study showed a significant relation between performance on the CRM skills test and aircraft commander job performance.

In a book intended to be a reference to pilot selection, (Bor et al., 2020) offers us a range of different skills and methods that can be used for pilot selection and evaluation. It aims to describe how the recruitment, selection and psychological assessment of pilots can be carried out. The authors also seek to address some key topics, such as remotely piloted aircraft, pilot retirement, and personality assessment.

The authors provide us with a reference book intended for recruiters, trainers, human factor engineers and psychologists involved in the selection of pilots. Researchers are trying to determine whether the pilot selection process adds value and ensures that the best candidates are hired. They also offer us an in-depth discussion on selection systems and their effectiveness in currently used approaches and comment on current trends in the aviation industry.

From the same author, (King, 2014) attempts to clarify the topic and the assessment of personality in aviation. The article conceptualizes pilot selection as a two-step process. The first, “Select-In” methods are psychological tests measuring traits that have been found to be desirable for the job task analysis (JTA). They assess a candidate's level of knowledge, skills, abilities, and other characteristics (KSAO). The second, the “Select-Out” methods are an assessment of psychopathology used to assess psychiatric fitness. The challenges of personality disorders are explored, and an integrative approach, with a practitioner combining selection and selection methods, is discussed. Details of the Americans With Disabilities Act (ADA) with respect to psychiatric disorders are explored in the article. Personality disorders are treated in much the same way in the US civil and military aviation, with the elimination of aviation duties due to safety disruptive behaviour.

Examining participation in complex professions, such as aviation, requires specialized aptitude tests. A job task analysis (JTA) is first performed to determine the desirable KSAOs for the specific occupation. Then, measurement methods are identified, or, if they do not already exist, are developed to assess the KSAOs. Some of the desirable qualities identified by a JTA may be personality traits. These personality characteristics are increasingly analyzed in terms of emotional stability, extraversion, openness to experience, conscientiousness, and agreeableness. These five characteristics form the Big Five of the organization of the personality (Tupes & Christal, 1961)

Psychological testing can be used as a screening method to determine whether or not a more comprehensive psychological or psychiatric examination is warranted. The authors provide, for example, the case of the assessment of psychopathology in candidate air traffic control specialists (ATCS). This case provides an example of a selection procedure in the selection of aeronautical personnel. The FAA established the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) (J. N. Butcher et al., 1989) as the clinical cut-off limit triggering a comprehensive assessment of mental health at the 95th percentile of the ATCS population. The MMPI-2 is not really a personality test. Rather, it should be viewed as a measure of psychopathology.

MMPI-2-RF is a personality questionnaire for diagnostic, descriptive and therapeutic purposes: it identifies the psychological dynamics of the subject, such as psychopathological disorders and personality disorders to plan treatment and appropriate care.

The Minnesota Multiphasic Personality Inventory (MMPI) and MMPI-2 have been widely used in programs to select personnel for positions that require good psychological adjustment and responsibility, such as police, firefighters, air traffic controller and airline flight crews.

The objectives of a study presented by (James N. Butcher, 1994) were to provide descriptive information on the use of MMPI-2 in the psychological assessment of an airline pilot and to examine the effects of the new MMPI-2 standards on airline pilot profiles. They were also careful to examine the difference in the responses of airline pilots compared with other non-clinical subjects taking the MMPI-2 in a different setting. They examined the effects of the new standards on MMPI-2 validity and clinical scale scores and the effects of defensive testing on the MMPI-2 scores of candidate pilots. They also presented descriptive information on the MMPI-2 scales and the results of a factor analysis of the MMPI-2 scores of the pilot candidates compared with those of the normative MMPI-2 sample.

The results showed that the MMPI-2 standards are probably more suitable for characterizing nonclinical candidate groups than the original MMPI standards. The different factor structure of the profiles of airline pilot candidates compared to those of other normal individuals suggests that the different response pattern must be considered in interpreting the profiles. Defensive character and impulsiveness are likely to be significant considerations in the MMPI-2 scores of airline pilot candidates.

In (Causse et al., 2011), the authors examined the relationship between pilot’s cognitive status, personality traits, and experience with regard to their flying performance to determine which criteria are the most predictive of the flying performance. They focused on the three low-level executive functions (EFs), shifting, inhibition, and updating, and on linking their efficiency to flight performance and decisions-making during the landing.

Twenty-four pilots underwent neuropsychological tests. The cognitive assessment encompassed the three basic executive functions (Miyake et al., 2000), reasoning, and psychomotor velocity. The flight scenario was set up in cooperation with flight instructors. The personal characteristics were age, flight experience, and levels of impulsivity.

The flight performance assessment was founded on the flight path deviations (FPD), a widely used indicator of the primary flight performance. Multivariate regression was used to determine the influence of the independent variables on FPD. The ability of independent variables to predict the performance was tested by an all-possible-subset regression. Pearson correlations were computed among all considered independent variables and between the FPD and the independent variables. Reasoning, updating in working memory, and flight experience was predictive of the flight performance.

In (Barron et al., 2017), 3,140 USAF-manned aircraft pilots and 330 Remotely Piloted Aircraft (RPA) pilots were surveyed to determine whether pilot selection metrics are predictive of long-term pilot performance. Flight performance was documented in an Officer Performance Reports (OPR) after completion of flight training. Participants completed the Air Force Officer Qualification Test (AFOQT) consisting of 11 cognitive tests and a five-factor model (FFM) measure.

The analysis showed that neuroticism and introversion are significantly related to poor pilot training outcomes. The cognitive skills, knowledge, and personality traits that predict the professional performance of USAF-piloted aircraft pilots are also predictive of the professional performance of RPA pilots. The subtests that contribute to the AFOQT pilot composite were significant predictors of early career job performance for both manned aircraft and RPA pilots. The results demonstrate that predictive relationships were observed for manned aircraft and stationed RPA pilots

CHAPTER 3 GENERAL ORGANIZATION OF THE THESIS AND COHERENCE OF THE ARTICLES IN RELATION TO THE RESEARCH GOALS

3.1 Thesis Organization

As indicated by Figure 3.1, the current thesis is composed of a chapter on the patent publications made is the context of this research, and five chapters, each summarizing the five papers publishing the experimentation results of the research and the conference presentations where the research got the opportunity to present results to the expert's community. Discussion and Conclusion will close the thesis followed by the References and Annexes sections.

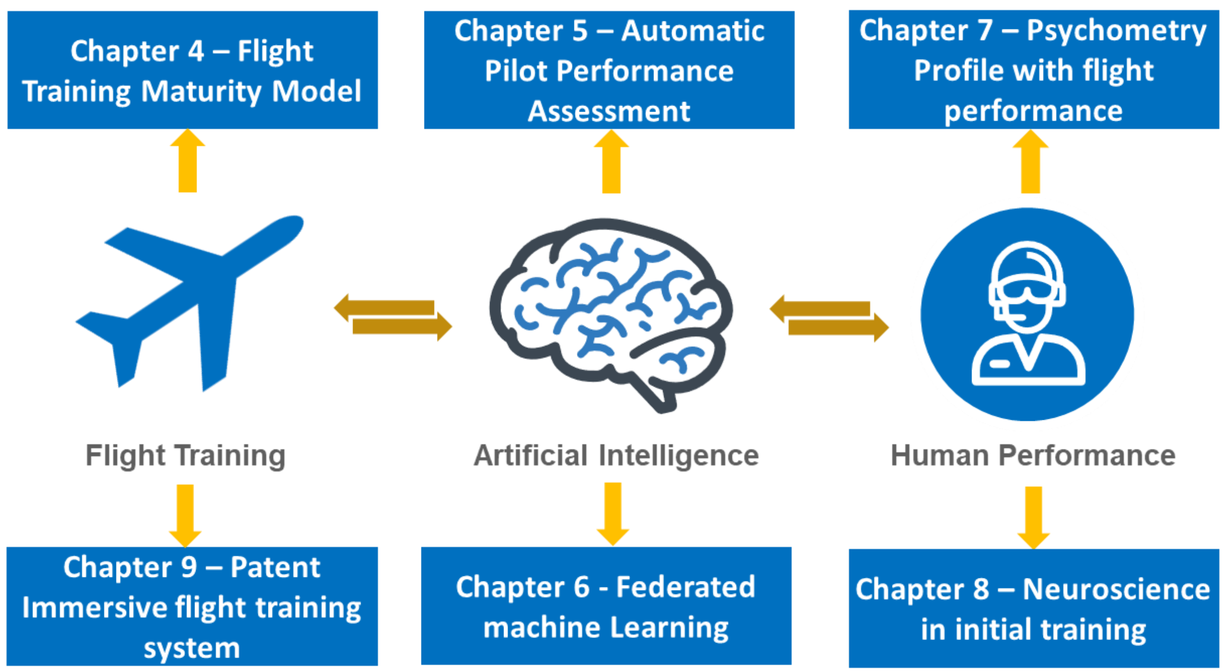


Figure 3.1 - Thesis Organization

The following chapters will present the **five scientific papers** that will describe how data science and machine learning capability considering human performance in the loop can be used to augment the efficiency of an Intelligent Flight Training System (IFTS).

Chapter 4 will present the paper on a maturity model for flight training. This paper will set the foundation of the process to integrate the various adaptive flight training solutions in the operations of a flight training program. Then, various types of data will be introduced through the next papers. With the paper presented in Chapter 5, we introduce the flight simulation data recordings and the instructor assessment to build an automatic objective assessment capability using machine learning to assess pilot performance. In Chapter 6 we augment this capability with multiple training devices from multiple training center in a federated machine learning organization to support adaptive flight training. In 6.2 and Chapter 8, we respectively add psychometry and biometry data to bring objective measurements to the human behavior during flight training. The pilot aptitudes clustering associated with cadet flight training performance can bring an interesting aspect to flight training recommendations and adaptations. This will be complemented by neuroscience in immersive flight training device measuring cognitive load and pilot gaze scanning behavior analysis.

Within the paper chapters, we describe the conferences where results were presented to the worldwide community. The defense/military flight training community was addressed in Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) the most important conference in the world for this community. The European Aviation Training Summit (EATS) was given the opportunity to reach the commercial aviation community. The Institut de Valorisation des Données (IVADO) offered us a connection with the data and artificial intelligence community.

Chapter 9 will indicate how this research thesis directly contribute to the design of an intelligent adaptive flight training system deployed in the aerospace industry and where advancement is published through **sixteen patent publications.**

CHAPTER 4 ARTICLE 1: A CAPABILITY MATURITY MODEL FOR FLIGHT TRAINING

A Capability Maturity Model for Flight Training, Jean-François Delisle, Stéphane Ouellet, Derek Linders, Published in *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2018 proceedings,* November 2018

4.1 **Abstract**

The aviation industry invests heavily in technology to ensure training quality and compliance with flight safety regulations. At the same time, the steady growth of air travel is driving the demand for pilots. Maintaining pilot throughput without sacrificing the quality of training is a major challenge facing the industry. A methodology for improving training could potentially benefit flight training organizations by identifying areas for process improvement.

We propose a Flight Training Maturity Model (FTMM) derived from the Capability Maturity Model Integrated (CMMI) published by the Software Engineering Institute (SEI) at Carnegie Mellon University. Initially developed to improve the performance of military software projects, the CMMI is a reference model designed to evaluate and improve the software development lifecycle process. Maturity models are not new and they exist for many disciplines. Relevant models include (Wagenstein, PMI, 2006) in education and training (Marshall 2012) in eLearning, the Federal Aviation Administration Integrated Capability Maturity Model (FAA-iCMM) and (Ibrahim, 2000) in system engineering. We believe the flight training process can benefit from an adaptive training approach based on competencies. Adding a competency-based approach to the maturity model was studied by (Gillies & Howard, 2003) for change management in healthcare, but can benefit the aviation sector as well.

In our paper, we derive from the CMMI's five-level maturity scale an analytical framework for flight training that defines key performance indicators (KPIs) for each key process area, in order to evaluate organizations. The KPIs target the training processes, participants, learning environments, and technologies involved at each maturity level. Using existing training center programs and industry standards, such as the Analysis-Design-Development-Implementation-Evaluation (ADDIE) instructional system design framework and the FAA–Industry Training

Standards (FITS), we identify process areas that can be compared with valid KPIs for corresponding maturity levels of different flight training organizations.

4.2 Introduction

The aviation industry invests heavily in technology to ensure training quality and compliance with flight safety regulations. At the same time, the steady growth of air travel is driving the demand for pilots. Rapid fleet expansion and high pilot retirement rates create a need to develop new aircrews, more than in any previous decade. Over 50% of the pilots who will fly the world's aircraft in 10 years have not yet started to train (CAE, 2017). This record demand will challenge current pilot-recruitment channels and development programs. In turn, new and innovative pilot career pathways and training systems will be required to meet the industry's crewing needs and ever-evolving safety standards.

Maintaining pilot throughput without sacrificing the quality of training is a major challenge facing the industry. To address this challenge, we believe that a flexible and independent process model of technology platforms, organizational structures, and educational frameworks is needed to guide the improvement of flight training. The model can also encourage the development of effective and adaptive flight training technologies and support competency-based and evidence-based flight training methodologies.

The discipline of software engineering can be used as an example for developing such a model. Software engineers have implemented a method, the Capability Maturity Model Integrated (CMMI) (Chrissis et al., 2011), to guide improvements in their work, increase the quality of the software produced, and avoid traditional and chaotic ad hoc practices. The CMMI is a method to assess the extent to which an organization has processes in place to develop high-quality software. Given that the CMMI model has a proven track record of ensuring quality in software engineering, can it be used for flight training?

The aim of flight training is to prepare pilots for activities that are certainly as complex as those involved in software development, which have been considered deserving of a maturity model. Flight training involves preparing humans to operate complex critical systems. Process maturity is important and can make a difference in the quality of the trained pilots and can decrease the learning time required for novices to acquire the required competency level. Instructor and student interactions influence learning and teaching. A flight maturity model can benefit both students and instructors by supporting appropriate and objective metrics and assessment framework. The model developed in this document is a first step in the process of developing a Flight Training Maturity Model (FTMM) that can be applied continuously to help stakeholders identify the weaknesses and strengths of their flight training process lifecycle. A Flight Training Organization (FTO) can be ranked at a certain maturity level and can improve over time and reach the next level of maturity.

An FTMM is the opportunity to set the basis for ongoing discussion within the flight training community to identify the key practice areas of the different maturity levels, a discussion that is necessary for achieving globally standardized improvements in the flight training industry.

Enhancing training in real-time through adaptive learning techniques and technologies may allow for the implementation of a Canadian Forces Individual Training and Education System (CFITES) or United States Air Force (USAF) training lifecycle with closed-loop pretty much in real-time, potentially allowing for a significant improvement in either quality of learning or speed of learning (improve time-to-competency). We believe that a formal model can support flight training improvement activities and can be developed as part of a collaborative research program by members of the flight training community.

4.2.1 Methodology and Summary

This document will review the essence of the CMMI and propose a short review of other maturity models inspired by the CMMI, before presenting our view of the Flight Training Maturity Model. Inspired by the maturity model design principle framework presented by (Jens Pöppelbuß, 2011), our model aims to be used descriptively and prescriptively. To initiate adoption and validation, we present a mapping of the FTMM with flight training standards. Our methodology uses subsets of elements of the mapping process presented by (Pino et al., 2009) and the systematic mapping study of (Wendler, 2012). Canada CFITES (Blake C.W. Martin, 2016) and USAF are the two flight training standards described in this paper. The standards, handbooks, vision and directions that inspired the definition of our model are the ADDIE framework, US FAA–Industry Training Standards (FITS), US FAA *Pilot's Handbook* (FAA, 2016), US FAA training & testing (FAA, 2018), Australia RAAF (Neill, 2003) and UK *Defense Direction and Guidance for Training and Education* (JSP 822) (GOV.UK, 2017). The last section will focus on new technological & methodological capabilities that can enhance flight training but that need a maturity model to be managed properly and require a Technical Readiness Level (TRL) in line with the flight training maturity level.

4.3 Models and frameworks

4.3.1 The Capability Maturity Model Integrated

The CMMI aims to provide objective metrics to assess quality and promote improvement in software development processes. The CMMI classifies software development organizations according to five hierarchical levels of maturity as presented in Table 4.1. Each level defines an ability to produce software quality through key processes area and practices that articulate what needs to be done to achieve the quality level. Each level builds on the previous ones and adds functionalities to the process. The advantage of the model is that it identifies key process areas, as defined by professionals of the domain.

Table 4.1: The Capability Maturity Model Integrated

Maturity Level	**Description**
1- Initial	Informal and poorly controlled process. Success relies on an individual's performance and not on the use of established processes. The processes are mostly chaotic and ad hoc, and lead to unpredictable product quality.
2- Repeatable	A repeatable process is one that is planned and executed and has the basic infrastructure in place to support it. It produces controlled outputs, and is monitored, controlled, reviewed, and evaluated for adherence to policies. A project management system is in place to track costs, schedules, and functionalities.
3- Defined	A defined process is managed, documented, standardized, and integrated. Processes are tailored according to the organization's tailoring guidelines.

Table 4.2: The Capability Maturity Model Integrated (cont'd)

Maturity Level	Description
4- *Managed*	A quantitatively managed process is a defined process that is overseen using quantitative measurements of the process and product quality. Processes at this level are predictable.
5- *Optimized*	Continuous process improvement is effective in the organization and based on an understanding of the common causes of the variations inherent to the process, and it is improved through incremental organizational, process, and technological improvements.

4.3.2 **Learning Models and Frameworks**

The ADDIE framework breaks down the instructional development cycle into phases with defined deliverables, that is, an outcome that feeds into the subsequent step. The ADDIE model is somewhat generic and has been extended by the USAF into the 5-Step Model and by the Canadian Department of Defense into its CFITES model.

- *Analysis:* The analysis phase clarifies and establish, the instructional problem, the instructional goals and objectives in an identified learning environment. Outputs from this phase include training needs analysis and requirements analysis.
- *Design:* The design phase deals with learning objectives, assessment instruments, exercises, content, subject matter analysis, lesson planning, and media selection. Outputs from this phase include training plans, training blueprints, design documents, and implementation plans.
- *Development : In* the development phase, developers create and assemble content assets to the specifications created during the design phase. Programmers work to develop and/or integrate technologies. Testers perform debugging procedures. Outputs from this phase include storyboards, rapid prototypes, and courseware.

- *Implementation:* The implementation phase focus on procedures for training facilitators and learners. Outputs from this phase include content migration to Learning Management System (LMS), preparation of LMS hardware/software, and training delivery.
- *Evaluation:* The evaluation phase consists of tests designed to provide feedback for process and content improvement. Outputs from this phase include formative evaluations, summative evaluations, and field trials.

4.3.3 **Maturity Models in Learning, Education, and Aviation**

Maturity models are not new and they exist for many disciplines. Among the various models, some are process improvement models in disciplines related to training, such as education, e-learning, and curriculum design.

With the *Computing Education Maturity Model (CEMM),* (Lutteroth et al., 2007) propose a model whose aim is to rate the maturity of educational organizations, where the course is the basic building block of the CEMM and where the maturity level is determined by the maturity of the course. With his *e-Learning Maturity Model – eLMM, (Marshall S., 2018)* adapts the CMMI to provide a roadmap to improve organizational processes associated with e-learning. The author considers the value of adapting the CMMI as a guide to improve: course-level adoption, institutional-level adoption and integration of e-learning. (Gartner Group, 2004) also defined an *e-learning maturity model* to provide a framework for evaluating the current state of e-learning programs.

In a C*urriculum Design for Higher Education,* (Thong Chee Ling, 2012) agrees that there are many maturity models built to improve the quality of the design process, but highlights the lack of a maturity model for curriculum designers. With a model that can tell designers at what maturity level they are and guide them to complete the process before moving on to the next level, the quality of the design process will increase. (Dennis Drinka, 2008) also provide a maturity model for curriculum redesign for a Management Information Systems degree. The *Online Course Design Maturity Model* (OCDMM) (Neuhauser, 2004) serves as a tool in planning online courses and assessing them for improvement. It incorporates all the best practices in online course design. Technological additions to raise the quality of the courses are done continuously in designing online courses. However, the limitation identified in this model is that it lacks the use of basic design principles. (Wagenstein, 2006) presents usage of CMMI as a framework in training and education for creating, managing, and measuring an organizational training program, but the author does not show a clear mapping or adaptation to training and education practices.

With the *FAA-iCMM (Ibrahim, 2000)*, the aviation sector used the maturity model approach. By integrating the systems engineering CMM, the software acquisition CMM, and the CMM for software, the FAA processes used to manage, acquire, and engineer systems, products, and services have been successfully guided to systematic improvement. However, this model is not used for aviation practices, but only for its IT/engineering processes.

4.4 The flight training maturity model

The Flight Training Maturity Model is an extension of the CMMI framework, customized specifically for the needs of Flight Training Organization. The model is presented in Table 4.3. As we felt that the "Initial" CMMI phase was too broad for our purposes, we have broken it down into two subphases: "Initial" (Level 0) and "Basic" (Level 1). This allows for a distinction to be made between schools that are minimal and in-the-moment, using only pre-existing packaged instructional assets, with no customization for the student, and addressing only basic legislative standards on record keeping. The use of "Level 0" avoids confusion when referencing other five-stage maturity models and allows for comparisons to them.

Table 4.3:. The Flight Training Maturity Model

Maturity Level	Description	Process Areas
Level 0 – Initial	There are no standards, no monitoring, no formal communication, and no retained records, and performance is not evaluated.	• No process area defined
Level 1 – Basic	There are very few standards; success depends on individual performance; monitoring is limited to accreditation compliance; there is neither formal communication nor retained records, and performance is not evaluated.	• Conducted instructional program

Table 4.4:. The Flight Training Maturity Model (cont'd)

Maturity Level	**Description**	**Process Areas**
Level 2 - Managed	Training plans are established for each program. Discipline is being established and implemented through development plans, quality assurance, and curriculum management. Training is tailored to flight tasks. This is the level where most aviation schools find themselves: operationally focused, turning out the same courses over and over. Major process areas are documented, and student results are consistently recorded. Changes in course delivery or content will typically be triggered by outside authorities and will include updated checklists, diagrams, and exercises only when ordered to change.	• Flight training requirement management • Flight training planning, control, and monitoring • Flight training quality assurance • Versioned training and learning content in configuration management • Flight training delivery • Measurement and analysis
Level 3 - Defined	At this level, a strategic plan exists and process improvement planning is in line with the business strategy of the flight training organization. Training requirements are identified and in line with flight performance objectives. Flight instructors are well trained and understand their own performance and have defined improvement plans. The quality referential is standardized by defined practices. Business objectives and performance goals to strive toward are set. Safety, efficiency, support, and student success goals are set; metrics are reported on in order to assess training efficacy.	• Flight training organizational process definition • Integrated learning and training environment • Risk management • Flight training services development with continuity • Training system availability and incident management • Training service management • Flight training development process

Table 4.5:. The Flight Training Maturity Model (cont'd)

Maturity Level	**Description**	**Process Areas**
Level 4 – Quantitatively Managed	Flight training programs are driven by quantitative objectives of assessed training quality and training process. Training content is modified based on training requirements gathered from metrics and analytics. Metrics are benchmarked and tracked for KPIs both on individual and collective student and staff performance. Periodic, formal reviews allow for analysis and improvement cycle.	• Flight training measurement and quantitatively managed • Training process performance auditing and analysis • Organization and role performance auditing and analysis • Training requirement analytics on metrics • Collective efficiency measurements • Talent management • Flight training performance benchmarking
Level 5 - Optimized	Processes that are quantitatively managed for training management are constantly optimized to anticipate training needs and potential technology advances. Training can be tailored for individuals while still fulfilling rigorous and systematic training objectives. Training and operational results are continuously fed into advanced analytics systems that perform root-cause analyses, continuous improvement efforts (kaizen), automated remediation (virtual tutor), and full closed-loop operational analysis and prescriptions for improvement.	• FTO performance management • Causal analysis and closed-loop resolution • Training lifecycle optimization • Training process improvement impact prediction

4.4.1 **Mapping the CFITES to the FTMM**

The Canadian Air Force has done a thorough job establishing excellence in training, through the CFITES. It should be noted that the same standard is used to develop group training as well. The CFITES is the Canadian implementation of the Defense System Approach to Training (DSAT). It implements the ADDIE instruction system design in great detail.

To verify the usefulness of establishing a Flight Training Maturity Model, it is a useful exercise to compare it to prescribed military training standards. For reasons of criticality and mission success because many military jobs are matters of life or death, military institutions have been leaders in cognitive science research, and many of them have instituted training standards and associated processes that are mandatory in the development of military training. The CFITES requires training organizations and training developers to implement the following training definition and development steps:

1- Needs Assessment: Of interest is the fact that this stem from the assessment of mission performance, discrepancies, or job analysis. Many solutions can be considered to improve performance: the need assessment in such a context would be the identification of training need, a situation where training is the aspect to improve to correct a performance gap which may be an individual or collective (team or training cohort) deficit.

2- Analysis of Instructional Requirements: This analysis may include several analyses, such as job task analysis, Difficulty/Importance/Frequency (DIF) analysis, as well as the discipline of formulating performance objectives and educational objectives (often referred to as Enabling Objectives).

3- Design of Instructional Programs: The design of instructional programs involves more detailed analyses, such as target population analysis and formalizing the Knowledge/Skills/Abilities (KSAs) to be acquired through the training program. This also involves a large array of activities, including training-media analysis and cost analysis, to determine the best approach for training, including how to assess the students. For large training undertaking, this phase may involve instructional design, engineering design of training systems, or putting in place the business cases justifying the training programs.

4- Development of Instructional Programs: This phase concerns the actual development of training, be it new training systems and technologies, or courseware in general. This phase also includes acquisition processes when solutions are not internally developed but rather acquired. This phase concludes with the testing and validation phases required for the acceptance and trial of the new instructional program.

5- Conduct of Instructional Programs: This phase generally concerns delivery of instructions, including the implementation of instructional methods. The CFITES also focuses attention on other aspects, such as on-the-job training, monitoring or learning; and monitoring the instruction activities.

6- Evaluation of Learners: The evaluation of learners involves tasks such as determining the content and standards, which are generally based on a qualification standard. It involves constructing the tests, preparing instructions for the tests, and piloting the tests in both academic and practical contexts (evaluating the performance of a task). Assessment plans and evaluation tests are discussed as well. It should be noted that, as learning technologies have improved and led to increased use of simulation and serious gaming, the vast possibilities and realism of training environments have not simplified assessments. So far, assessment has not kept pace with technology. In fact, to some degree, assessment has become more subjective as the complexity of training has increased.

7- Validation of Instructional Programs: The validation of instructional programs is even more demanding than evaluating learners. This process involves planning and scoping the validation activities.

8- Managing Individual Training and Education Projects: With this process, the Royal Canadian Air Force (RCAF) is institutionalizing the processes required to perform project management in the context of developing and delivering training and education projects. The FTMM could indeed contribute to this process by providing a tool to assess the organization’s maturity and help focus management efforts. This process deals with integrated testing and evaluation and with the positioning of the training and education project with the overall management of integrated logistics support.

9- Evaluation of the Instructional Program: This process is mostly about evaluating the actual training organization and training activities. The FTMM can serve as a tool in assessing a training organization.

10- Prior Learning Assessment: This process provides guidance in the evaluation of personnel learning, which may alleviate training needs and qualify personnel for a job. It may be one process used as part of training needs assessments. This differs from the Evaluation of Learners in that it assesses pre-existing knowledge or gaps, as opposed to assessing the efficacy of the training delivered.

11- Evaluation and Validation Techniques: This section of the CFITES essentially provides guidance on the use of techniques to evaluate and validate training. It precisely describes data collection methods; the design and preparation of questionnaires; observation and interview techniques; and mathematical data analysis.

Achieving the CFITES objectives is a challenge and, most of the time, part of the CFITES is overlooked due to a lack of capability by the organization to achieve it with its current resources. Traceability of the contribution of the Capability Maturity Model to achieving the CFITES objective can be established. The Table 4.6 below shows the potential for a capability model to act as a tool to analyze a training organization's maturity with CFITES process.

Table 4.6:. CFITES process maturity mapping.

CFITES Process Achieved	**Maturity Level**	**Evidence and Processes**
Needs Assessment	2	At this level, the organization achieves the requirements for management of KPIs in a training context, which should serve to achieve this CFITES process.
Analysis of Instructional Requirements	2	At this level, the organization achieves the requirements for management of KPIs in a training context, which should serve to achieve this CFITES process.

Table 4.7:. CFITES process maturity mapping (cont'd)

CFITES Process Achieved	**Maturity Level**	**Evidence and Processes**
Design of Instructional Program	2.	This is related to development itself, which can be evaluated with CMMI.
Development of Instructional Program	2	This is related to development itself, which can be evaluated with CMMI.
Conduct of Instructional Program	3	Thoroughly conducting a training program in accordance with CFITES guidance is likely to require an FTO maturity of 3.
Evaluation of Learner	3-4	It’s at Level 4 that quantitative management of training outcomes is achieved, which is required to properly achieve this CFITES process.
Validation of Instructional Program	4	It’s at Level 4 that quantitative management of training outcomes is achieved, which is required to properly achieve this CFITES process. Training performance KPIs are required to achieve this step.
Managing Individual Training and Education Program	3	In line with level 3 - Training service management
Evaluation of Instructional Program	3-4	It’s at Level 4 that quantitative management of training outcomes is achieved, which is required to properly achieve this CFITES process. Training performance KPIs are required to achieve this step.
Evaluation and Validation Techniques	3	Many techniques defined here will now become possible or be accelerated by new technologies. It’s possible that new methods to evaluate and validate training will arise from the big data analytics.

4.4.2 **Mapping the US Standard to FTMM**

US military organizations have been leaders in establishing instruction system design methodologies. The USAF AFMAN 36-2234 provides a thorough description of the system approach to training. The sequential process phases/areas are these development phases: Analysis → Design → Development → Implementation. These development phases are all articulated around the evaluation process. The evaluation process is the central process of evaluating the results and performance of the organization during the development of an instructional program. A proper evaluation does not only rely on tests and on the verification of deliverables, but also on measuring and evaluating quality through the development of a progressive increase in complexity and realism in the validation of the instructional program. Evaluation starts during the analysis phase and continues via various methodologies during the development phase. The evaluation would start with evaluation plans, move toward test plans and scenarios, and eventually, conclude with testing and tryouts during the implementation phase.

The continuous processes involved in a training organization's operations are stated to be *Management, Support, Administration, and Delivery*, the last one being the organization's central purpose, i.e., instructing and training a target trainee population. The appropriate application of these processes as a closed-loop training cycle fosters quality and continuous improvements. It should be noted that the USAF SAT processes, just like the CFITES, do not require that FTO Maturity Level 5 be reached. Attaining Level 5 maturity is more like a force multiplier, whereby every training process will benefit and be enhanced in continuous (or quasi-real time) based on how fast the feedback on lessons learned or improvements are fed into the training system. The Table 4.8 below shows the potential for a capability model to act as a tool to analyze a training organization's maturity with USAF process.

Table 4.8:. USAF process maturity mapping

USAF SAT	**Maturity**	**Evidence and Processes**
Analysis	2	At this level, the organization achieves the requirements for management of KPIs, which should help with this phase in particular.
Design	2	At this level, the organization achieves the requirements for management of KPIs, which should help with this phase in particular
Development	2	This is related to development itself, which can be evaluated with CMMI.
Implementation	2	This is related to development itself, which can be evaluated with CMMI.
Evaluation	3	It is likely that full compliance to meet the spirit of the USAF SAT Evaluation phase requires an FTO maturity of 3.
Management	3-4	It’s at Level 4 that the quantitative management of training outcomes is achieved to support training management decisions.
Administration	4	Same comment as above.
Support	3	N.A. It should be noted that this process may be the soft/weak aspect of CMMI and the FTO maturity model. In general, the full scope of support activities is the subject of integrated logistic support standards.
Delivery	3-4	It’s at Level 4 that the quantitative management of training outcomes is achieved, which is required to properly achieve the delivery process. When closed-loop technologies and methodologies are deployed to provide evidence-based training or adaptive learning, the delivery of instruction should be positively impacted and demonstrated through higher learning or faster learning.

4.5 Technology and methodology at high flight-training maturity

Our goal is to create a process model that encourages the development of effective educational technology resources, independent of technical platforms and training methodologies. The last section will focus on new technological capabilities that can enhance flight training, but need a maturity model to be managed properly, and require a Technical Readiness Level (TRL) (Source: DoD (2010), *Defense Acquisition Guidebook*) in line with the flight training maturity level. The TRLs make up "*basic principles observed and reported (Level 1)"*, up to Level 9, where the *"actual system needs to be proven through successful mission operations",* which can be mapped to technologies by the flight training organization to maintain its training delivery capability in a sustainable way.

4.5.1 Flight Training Data Science

Flight training data science involves collecting and analyzing data, where the challenge is collecting relevant data. The collection and analysis of data here concern actual job and mission performance, such that the methodology for, and capability to even obtain the data is a key issue. It is in this area that today's instrumentation and connectivity bring the possibility of large data collection (big data analysis) and of being able to truly close the loop. The advent of big data analysis, the automated collection and storage of massive amounts of data, and advances in data analytics will add to the tools for performing evaluations and validations, or significantly change the processes involved (or, at a minimum, accelerate them). There will be added capability to interpret results, discover trends that used to be impossible to see, and even affect training systems in real-time.

Data analysis can provide an overview of the learning and training experience and can provide additional value with the presence of intelligent and autonomous capability, closing the loop between training and flight training operation and within learning environment. Automatic causal analysis, real-time reporting, training validation, automatic student assessment and performance evaluation are data analysis capabilities that can provide new training methodologies such as evidence-based training.

As an example, the purpose of the CFITES *Validation of Instructional Programs* step purpose is to close the loop of the training cycle to generate the next cycle of training improvements, taking into account the results from the last iteration of training. The process involves documenting, reporting on training efficiency based on mission performance, and recommending improvements. Traditionally, this process has required a great deal of data collection and analysis over a long period, and the process may have often been neglected or abandoned for feasibility reasons. (The process is a long one, and if the operational environment changes quickly enough, the evaluation may lose its relevance.) The enhancement of our ability to process and analyze data can significantly accelerate this process. Analytics, artificial intelligence, or even just knowledge-based systems promise to accelerate this analysis, allowing recommendations to be developed while data is still fresh and the context is still relevant.

If future technologies allow for the implementation of adaptive learning technology, it’s possible that student assessments and training validation will be performed in real-time. We can foresee simulation/game-based training that will not only assess students in real-time but change the parameters of the game to adapt to trainee performance and learning curve (for example, adding cognitive load, sensory distractions, and increasing exercise tempo as elements are mastered), or at least orient the next exercise toward the element the trainee needs to improve.

Flight training involves a learning aspect, but also certainly involves an understanding of the aircraft itself. An automatic performance assessment is usually done around a training event detection that offers us a time range of aircraft parameters that will be evaluated. This training event has similarities to safety or anomaly event detection and aircraft performance analysis that are key in the study of adaptive flight training. Focusing on detecting aircraft performance anomalies, (Eric Chu, 2010a) uses a similar approach to machine learning by using fleet data to build a regression model to detect anomalies in new aircraft data.

(Luxhøj, 2013) presents a model for predictive safety analytics of complex aerospace systems, the Aviation System Risk Model (ASRM), which can be used to evaluate the causal factors that lead to unsafe states. The ASRM uses a probabilistic approach to the Bayesian Belief Network (BNN). A Bayesian network was also used by (Maiga Chang, 2011) to manage learner models in a context-aware adaptive system and to quantify the level of knowledge understood by a learner, and this quantitative value was used as the basis for adaptive content selection. The authors present two quantitative reasoning strategies: *diagnostic reasoning*, used in the case where a concept is misunderstood, and *predictive reasoning*, used to quantitatively determine the value of the probability of the understanding level of a concept. Learning profiling helps prevent this risk by using a sensing → assessing → recommending loop to tune training to be most appropriate for a given student.

Although the use of automatic sensing and analysis of learner action and aircraft state is not in itself an indicator of advanced training maturity (advanced training devices may be used in a lower maturity training center) many of the technology building blocks (LMS, Learning Records Store (LRS), exercise planning, after-action review tools) enable the processes and knowledge required to enable higher maturity within a training program.

4.5.2 Competency-Based & Evidence-Based Training

Some experts state that the aviation industry is moving away from an hours-based system to a competency-based training system (Suzanne K. Kearns, 2016). They present an alternate means of compliance that can result in shorter, less expensive and more efficient training programs. They also address implementation challenges by presenting the role of the International Civil Aviation Organization (ICAO) regulations. Adding a competency-based approach to the maturity model was studied by (Gillies & Howard, 2003) for change management in healthcare, but can benefit the aviation sector as well. Evidence-based training (EBT) is characterized by developing and assessing the overall capability of a trainee across a range of competencies rather than by measuring the performance of individual events or maneuvers. (IATA, 2013), (ICAO, 2013). In order to provide EBT, it is necessary to assess the competencies of trainees. Grading systems are key in competency assessment and are selected and developed including all parties involved in an EBT program inline with the development of a competency framework.

4.5.3 Distant, Collective and Embedded Learning

Training can be provided without requiring physical attendance at a specific training center (Harper, 2004). This has the benefit of offloading the instructor prior to a training session. Distance training may be delivered in a synchronous way where the student may enroll for real-time lectures and labs or asynchronous where the student accesses content on their own time, at their own location. Collaborative learning is based on the idea that learning is a social act, where learning occurs due to the learners' interactions while working on the task (Strother, 2002). This leads us to collective efficacy and shows as a high level of maturity in the Innovation Maturity Model (Patrick Corsi, 2015). In a collaborative learning environment, learners are challenged both socially and emotionally, as they listen to different perspectives and are required to articulate and defend their ideas. Collaborative learning provides group goals and individual accountability (Slavin, 2006). At a high maturity level work and learning activities are embedded and can be perceived as a unique behavior.

4.6 Conclusion

This document can be considered the kickoff to a FTMM working group and an entry point into the flight training industry. We argue that the model provides a set of useful guides, but that its use should be considered within the context of each organization's learning environment and its approach to flight training.

A working group may recommend the use of a third-party organization familiar with both CMMI and the FTMM to provide objective assessment of organizational maturity, completeness of execution, and recommendations for improvement. Similar to ISO and CMMI auditors, this organization could assess with professional detachment and without conflicts of interest. Description of assessment, identification of candidate organizations, and development of a charter would be a required next step in the striking of this organization.

We are currently performing pilot projects advancing our pilot training. Early lessons learned, which could assist the working group include:

- Empower People - Attribute process ownership to employees. Listen to and evaluate dissent.
- Process Performance Models - Focus on specific areas. Establish realistic but challenging objectives.
- Prototype First, Automate After - All prototypes are done in tools that can be easily customizable.
- Appropriate Tools - Invest in process management tools and automatic data collection and reporting tools

Next steps will include maturity assessments; identifying gaps and the desired end maturity level. The pilot projects will be assessed and initial results, benefits, and lessons learned will feed into an industry-facing working group which will benefit from this initial adoption effort and will strive to refine the framework, processes and tools to provide a ready-to-use roadmap for organizations to identify their current and desired maturity training levels.

4.7 Acknowledgements

Acknowledgements to: Nick Giannias, Senior Technical Fellow, Chief Architect, CAE Inc. and Andrea Lodi, Research Director, Polytechnique de Montréal

4.8 Conference

Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2018 Proceedings, peer reviewed. Status: Published

Peer review process by I/ITSEC Policy & Standards committee for both abstract and paper selection with a birddog assignment for the review facilitation.

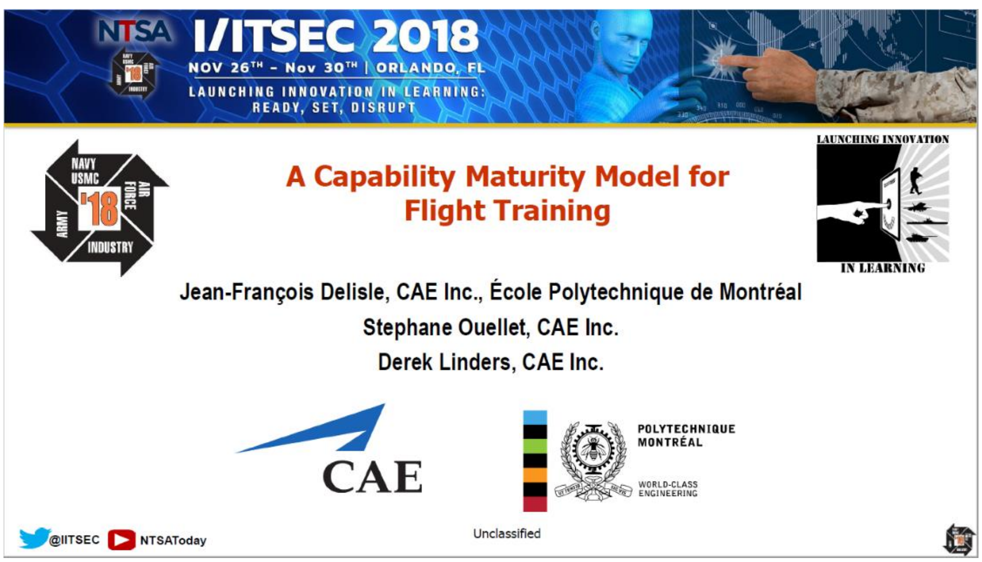


Figure 4.1 - I/ITSEC 2018 Conference

CHAPTER 5 ARTICLE 2: PILOT PERFORMANCE ASSESSMENT USING A HYBRID EXPERT SYSTEM AND MACHINE LEARNING FOR AN AUTOMATIC OBJECTIVE ASSESSMENT IN FLIGHT SIMULATION

Pilot performance assessment using a hybrid expert system and machine learning for an automatic objective assessment in flight simulation, Jean-François Delisle, Andrea Lodi, Maher Chaouachi, Melvyn Tan, Laurent Desmet, Submitted in *PLOS One*, August 2022

5.1 Abstract

An automatic pilot assessment capability using machine learning algorithms that can inform a flight instructor during a flight training session in full flight simulators is proposed in this paper. The current research explores a hybrid expert system and machine learning capability to assess pilot performance in flight simulation. Hybrid rule-based and machine learning algorithms are considered in the approach. Assessing a pilot's performance during a flight training session is a capability that can considerably improve the effectiveness of a training session and help the flight instructor provide better instructions and feedback. In this paper, we investigate an efficient way to build an automatic objective assessment engine, that provides a performance index that uses both knowledge of subject matter experts and instructors to train the artificial intelligence capability. By using multi-labels that have the same meaning but come from different sources of knowledge, we demonstrate that an automatic assessment engine is able to reduce the subjectivity of the instructor and optimize the time of the rules creation, tuning, and testing effort for the expert system development. In addition, we show that this hybrid approach increases the accuracy and precision of the assessment of pilot maneuvers during training sessions by using a consensus methodology that blends the multiple sources of knowledge.

5.2 Introduction

With the ongoing evolution of flight training systems and their increasing complexity, it is necessary to have a robust approach to assist the instructor in the evaluation of pilot performance during flight training. Nevertheless, the evaluation of pilot performance by flight instructors has many drawbacks. This task can be labour-intensive and, furthermore, a human evaluator may not precisely evaluate all the flight parameters due to the limitations of human observational capabilities and the positioning of the instructor who is typically sitting behind the trainees, that considerably hinders their ability to see all the details. In addition, it is possible that the instructor has bias or simply uses a different evaluation standard compared to their peers. Flight training involves the orchestration of training events in aircraft or flight simulators in accordance with a training curriculum. The flight training's objective is to provide pilots or crewmembers with an opportunity to acquire skills, attitudes, and knowledge of the standard operation procedure. The assessment is the measurement of a crew performance in the execution of a training event. Assessment criteria describe how the crew must perform the associated tasks. A Criterion-Referenced Assessment is done against established standards or criteria. A flight training session provides instruction, through practice of the procedures to prepare the crew for their duties. A flight training curriculum corresponds to the collection of segments that includes many modules that define training events and associated conditions, that can be arranged into lesson plans. The pilots are required to successfully meet the requirement of a curriculum segment to complete the course.

In this work, we aim to provide a new system that will objectively assess pilot competencies in real-time and provide insightful and objective performance assessment of how pilots are performing during their training. An experiment was realized, during which data from several full flight training sessions was collected and processed in order to build a hybrid system that uses rule-based expert system and machine learning algorithms to automatically assess pilots' performance

5.2.1 Problem Statement

During a typical training session, the instructor is evaluating the pilot by direct observation. Since the instructor is located behind the pilot and doing multiple tasks at the same time (e.g., launching and controlling the simulation parameters, giving instructions, taking notes, etc.) and may have obstructed the field of view, there is a substantial risk that they are not noticing all students' errors and missing the root cause of the pilot's actions and inaccuracies.

In addition, multiple grading schemes exist in the industry, some training centres are grading using a binary performance classifier (satisfactory/unsatisfactory) and some are using an ordinal scale (1 to 4 or 1 to 5). Grading a pilot performance is accomplished in a subjective manner and may be heavily influenced by instructor bias caused by factors external to the training. This problem could potentially lead to a more significant issue in the training program as it might be the cause of a non-standardization of global evaluations since grading can vary from one instructor to another, which, in other words, means that two pilots could have different evaluations while executing the same procedure.

For a machine learning capability, using an instructor grade to train the system means using a subjective label and then fails the purpose of adding an automatic and objective assessment capability. There is a need to use a more reliable labeling strategy.

Furthermore, training machine learning systems with instructor grading as the target label may be a concern for a new deployment of a new training program as we are confronted with a cold start issue. More precisely, to build the automatic assessment, a data collection phase is required, that impacts the deployment of such capabilities in real-world operation. Therefore, there is a need for a deployment strategy that considers the progressive collection of labels as source of truth.

With a classic software engineering approach using a rule-based automatic assessment capability, a lot of pre-engineering efforts are required to implement all the various possibilities of performance. Rule creation and tuning cost are high when scaling for multiple aircrafts and multiple training programs. There is a need to optimize the engineering phase with a more scalable technique using machine learning and transfer learning.

During a training session, only a few sequences are of interest to evaluate. The instructor needs to have an elevated level of attention to capture all events of interest and maintain his situation awareness all the time. With an automatic assessment capability, the problem is the same since the software will require to compute the entire training session data to perform the evaluation. A segment of the flight data is much larger than the segment of interest for the evaluation of the performance to be detected. There is a need to have a software capability that identifies the training session segment to be used for an evaluator, machine, or human. The algorithm must be able to identify the exact start and end time of the data segment of the flight sequence.

5.2.2 Objective & Hypothesis

Automatically assessing a pilot's performance during a flight simulator training session can enhance the performance of flight instructors and provide objectivity during flight training assessment. Considering a dataset of labelled flight segments, the algorithm should learn how to assess pilot performance on any new unlabelled dataset with proper precision and accuracy metrics, provide objectivity, and interpretability in the assessment process.

We aim to explain the results of a pilot performance assessment by giving proper flight parameters that have importance into the decision-making process with explanation and understandability of a pilot assessment. In addition, we intend to offer interpretability of the machine learning algorithms so we can understand the decision-making process of the algorithm and how the model will behave in a production environment. We also aim to explain and interpret the results of deep-learning algorithms that can assess pilot performance from a flight data segment that correspond to a flight training maneuver. By using multiple sources of truth, we are aiming to determine how we can integrate a few machine learning predictions together and apply consensus method, such as weighted majority voting to improve machine learning prediction accuracy and allow model management in cold-start approaches.

In this paper, we will explore various machine learning techniques to assess pilot performance against a standard operation procedure. We will also create a Performance Index that standardizes the grading scheme and enabled the portability into multiple grading schemes. (Supporting 1-4 and 1-5 grading scale as well as satisfactory/unsatisfactory binary scale). We will discuss the machine learning performance of an automatic grading capability against instructor grades and rule-based engine grades. We hypothesize that a combination of machine learning algorithms can increase accuracy by applying consensus functions such as weighted majority voting, and taking into consideration the various models maturity. This combination will increase performance accuracy and will provide better explainability of the machine learning model.

With the proposed artificial intelligence solution architecture, we believe that we can reduce the flight engineer’s workload. Manually code all rules to detect and assess pilot maneuvers for various aircraft models and training curricula is a huge task. We hypothesis that a machine learning approach that will learn from historical data can reduce the software and data engineering effort for large-scale deployment.

5.2.3 Paper Structure

After a literature review in section 2, this paper will present the research methodology and will reveal the strategy combining system expert and machine learning at section 3. It is through a description of the data collection and modelling process that we will introduce the end-to-end data flow to be able to provide performance assessment.

Several steps are required to assess flight performance. From the raw flight telemetry of the flight simulator, various phases of data extraction and pre-processing are required. Flight phase and Training Event detection is a key step toward an objective assessment capability, we will present the identification of the flight segment extraction strategy and how machine learning can improve detection. It will be followed by the description of the automatic assessment process using Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN) to compute the automatic objective assessment of the pilot performance. From the identified challenge, we will present a strategy to optimize the machine learning process with the addition of a consensus strategy leveraging the multiple sources of knowledge, dealing with unbalanced datasets, and the instructor bias.

We will present the results of the implementation of machine learning models to automatically detect the occurrence of a training event during an aviation assessment in section 5.5 with the analysis of the results. The approach is based on machine learning and aims to provide measurements to help decision-making to the human instructor and the artificial intelligence that can be a function of a synthetic instructor and his ability to evaluate the pilot's performance.

The section 4 will also present the appropriate metrics and KPIs that can inform about learning performance and the maturity of a machine learning model against an instructor and rule-based grading using the elaborated strategy. The discussion and future work will be presented in section 5, where the scalability for multiple training program and the competency assessment against the ICAO Evidence-Based Training (EBT) framework will be highlighted as the next step to this research.

5.3 Literature Review

As investigated by (Iqbal et al., 2017), grade-based prediction aims to help students improve their performance and allow them to get the needed help from instructors. The authors presented a Restricted Boltzmann Machines (RBM) technique to analyze the academic performance of university students and predict their performance. (Rechkoski et al., 2018) also presented an estimate of students' course grades to help them make decisions in order to achieve better results and obtain a degree in a timely and comprehensive manner. This paper provides an assessment of score prediction using collaborative filtering methods.

The evaluation of the performance of computer-assisted pilots has been a subject of research for several years in the field of aviation. (Stein, 1984) examined a method to provide a performance index developed analytically by several subject matter experts, to capture observer opinion to an automated performance measurement. Developed at the Federal Aviation Administration (FAA) Technical Center, the experiment's goal was to determine whether a new automated measurement system could differentiate performance on a flight simulator.

Evaluating pilot performance is certainly accessible using data from actual aircraft. The authors of (Rantanen et al., 2007) describe measures of pilot performance that can be derived from data of the Flight Data Recorder (FDR). Data from a Beechcraft BE23 Sundowner aircraft is useful for training and evaluating the objective performance of pilots when precise quantitative data cannot be obtained. Standard deviation, root mean square error, number of deviations, time out of tolerance, and mean time to exceed tolerance are the used measurements.

The work by (Stevens-Adams et al., 2010) automatically evaluated student performance based on observed examples of good and poor performance in the assessment of tactical air engagement scenarios. The study provided a rigorous empirical evaluation demonstrating the improvement in training that can be achieved with automated assessment technology, in addition to instructor feedback.

With (LeVie, 2016), the National Aeronautics and Space Administration (NASA) conducted a literature review to determine and identify quantitative standards for evaluating disruption recovery performance. This study contains current recovery procedures for military and commercial aviation and includes parameters for evaluating pilot performance in the context of upset prevention and recovery (UPRT) training in flight simulators.

The authors of (Chu et al., 2010) proposed an approach allowing precise detection of aircraft performance anomalies in cruise flight data. Detection is based on a model learnt from historical data of a fleet of aircrafts. Using the historical data, an average model is created based on the nominal vehicle operating data. No prior knowledge of the aircraft model is used, except knowledge of the inputs and outputs of the dynamic model. The flight dynamics model is empirically determined and identified from a set of flight data. Anomalies can be detected as deviations from the model. For a variety of cruising flight conditions with and without turbulence, the authors validated the approach using a Flight Operations Quality Assurance (FOQA) dataset generated by a NASA flight simulator, where flight performance is monitored to ensure optimal operations and also to detect anomalies. They identified a regression model that maps flight conditions and aircraft control inputs. Anomalies are detected as outliers that exceed the dispersion caused by turbulence and modelling error. The detection method is related to the control of the multivariate statistical process.

In the paper of (Wang et al., 2015), the flight data history of thousands of aircraft flights are analyzed to provide a pattern recognition method based on the feature matching. This is adopted for automatic identification by analyzing the flight attitude of the aircraft and identifying the type of maneuver from the operational flight data.

(Oza et al., 2003) from the NASA Ames Research Center used in-flight data from two helicopters, an AH1 Cobra and an OH58c Kiowa, performing two designated sequences of steady-state maneuvers. The authors proposed a method where the offset between the actual flight maneuver in progress and the maneuver predicted by a classifier is a strong indicator of the presence of a fault. Nevertheless, in order to reduce high false alarm rates, it is important to understand the source of variability present in the flight environment.

The authors of (Bryan Matthews, 2013) proposed a multivariate time-series search algorithm to search for anomaly patterns discovered in high-dimensional Commercial FOQA datasets. The process can identify operationally significant events due to environmental, mechanical, and human issues. The anomalies discovered were validated by a team of experts. The automated knowledge discovery process aimed to complement exceedance analysis done by humans, that fails to uncover previously unknown aviation safety incidents.

5.4 Methodology

5.4.1 Data collection

A data collection phase was performed during flight training sessions executed by qualified professional pilots. The recording system saved more than 700 flight parameters (i.e., vertical speed, pitch angle, flap position, etc.) in the form of a time series with a frequency varying between 10 Hz and 60 Hz. Each training session is structured into training sequences called training events. More than 50 types of training events could be performed by the pilots during a training session such as: crosswind landing, missed approach procedures-go around, abnormal and emergency procedures-reactive wind shear on take-off, and Low Visibility Operation (LVO) with Runway Visual Range (RVR) 150-400m. The selection of the training events to be performed was made by the instructor according to a predefined lesson plan.

After each training event execution, the instructor grades the pilot's technical performance according to an ordinal 1 to 4 scale. It is defined as follows: Grade 1 – Fail, Grade 2 - Pass but room for improvement, Grade 3 - Meets expectations and Grade 4 - Exceeds expectations. During the training session, the instructors use an electronic grading system that automatically saves pilots' performance assessments on each performed training event. Each individual assessment is called a scorecard. From a three years data collection process, a total of 2,154 training event assessment scorecards were completed during 484 training sessions with an average of 4 different training events performed per training session. Overall, 210 different pilots, and 21 instructors were involved in this data collection.

5.4.2 Challenges related to the dataset

The main challenge in this research is that we are dealing with high multidimensional (over 700 features) time series classification problems. Each time series is a 2-minute sequence that corresponds to the performance of a single training event. Moreover, flight parameters that will be used in the artificial intelligence models are highly generally correlated (ex: altitude and airspeed are highly correlated with the landing maneuver). This has a consequence that requires multiple models' integration and is heavy on the explainability requirements. We have high multidimensional, both numerical and categorical features, ~700 collected, that leads to more complicated processing and that can be heavy on feature selection model too.

Since the Flight Training industry is highly regulated, and our study population were certified commercial pilots, we had an Unbalanced Multiclass Classification problem. In an ordinal 1-4 scheme as the grading of the instructor, we had a lot of 2 and 3 representing the normal and pass behavior. Because 1 is a failure of the maneuver execution and 4 an exceptional performance rarely reached by pilots, we have very few 1 and 4 grades.

We are in the presence of Multiple Target Label with the same meaning but not in accordance, since one is coming from rules defined by a subject matter expert during the engineering phase of the cold-start approach, and the second is using grades by the instructor during official and regulated flight training sessions. Both rule-based and instructor grading are not guaranteed to be accurate and in accordance with each other.

An expected difficulty is that the pattern to be detected may not have the same length as the set of datasets of the training set. Flight pattern recognition algorithms must be able to recognize patterns from different time ranges. This imposes large data collection to be able to recognize flight pattern profiles. That results in an important effort in data quality, data cleaning, and data pre-processing due to the overall size of the data. It also brings an important effort in real-time processing of time series data by preserving the quality of the flight simulation model that replicates at high frequency the flight model of an A320 aircraft manufactured by Airbus.

5.4.3 Solution Strategy

The general strategy is to use a hybrid system of expert and machine learning algorithms that can assess pilot performance from a flight training session data segment that corresponds to a flight training maneuver, as part of a flight training session.

As presented in Figure 5.1, the data collection will begin with a local agent and a gateway already installed on a flight simulator that transfers data into a cloud platform data storage. The data is transformed into a Data Frame that is used in order to identify the capture window of a training event. Flight Phase and Training Event Detection algorithms are used to identify the segment of interest corresponding to a flight data time range. This will be used by the assessment module to compute a performance index based on key flight parameters involved in the maneuver.

A flight training event sample is selected to associate the pilot maneuver data with a training context. This data will be evaluated (labelled) manually by an instructor, while automatically labelled by an expert system. This will be used to train machine learning models.

Figure 5.1 - End-to-end flow of a performance assessment on a training event

5.4.3.1 System Expert

The system expert methodology consists of creating training event rules and grading rules that are developed with subject-matter experts (SME) and interpreted by a software service. The rules are based on aircraft manufacturer standard operation procedures (SOP) and regulatory guidance to obtain the automatic performance assessment. Criteria and tolerances are commonly provided by the FAA and form the basis of instructor pilot evaluations. An existing expert system automatically detects training events and assesses pilot performance against a collection of standard maneuvers, in order to provide labelled data from flight SME knowledge. (Rule-based engine grading)

5.4.3.2 The Hybrid Machine Learning and System Expert Architecture

The machine learning models are deployed in an environment that has access to the flight simulator data during flight training in real-time. As presented in Figure 5.3, flight simulation data collection and storage capability for the machine learning training process and runtime classification are available in a cloud platform.

An electronic grading application that lists lesson plan steps and allows grading of a pilot maneuver is used by the instructor to state the instructor grade. The system expert is deployed and records rule-based grades from the same flight training data. We stored machine learning grading in a central storage side-by-side rule-based and instructor grading.

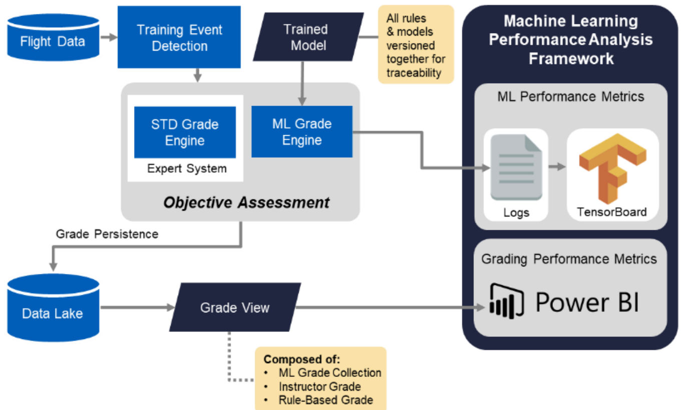


Figure 5.2 - Hybrid solution architecture from data collection to performance metrics dashboards

Both labels are then available to a machine learning service that trains multiple performance assessment models, that can predict the pilot performance and the grading.

Machine learning has advanced a lot in recent years, and is predominant with the advent of deep learning (Bengio, 2009). This method will be adopted in the first using multi-layer neural networks project and then tested in various architecture appropriate for time series data, such as Convolutional Neural Networks (CNN). Because flight parameter analysis involves time series, various methodologies are need to be identified to apply Time Series Classification (Smith-Jentsch et al.). CNNs, generally used in Computer Vision, is a Deep Learning approach that will be explored in order to increase machine learning prediction using flight simulator data.

Hybrid architectures are common in order to optimize solutions, add defined constraints, and respond to cold-start issues, such as (Cercone et al., 1999) who presented a solution by combining Rule-Induction and Case-Based Reasoning in the machine learning framework. The researchers present a variety of machine learning techniques to improve the quality of the solution by combining rule induction methods with case-based reasoning techniques to improve performance compared to more traditional single-representation architectures.

The author's article (Oladimeji et al., 2015) presented us with a hybrid approach that involves the use of the k-means algorithm with neural networks during supervised machine learning that extracts patterns and detects hidden trends in complex data.

The following algorithms are used in parallel to optimize both accuracy and explainability depending on their respective advantages.

- *Support Vector Machine (SVM):* This method strives to find the boundary that best separates two different classes. It does so by identifying the extreme points of the dataset that are close to the opposite class, and a support vector is then drawn between these extreme points. A linear separator is then established between the two classes. Out of all the separation options, the model chooses the option that yields the largest distance between the two support vectors.
- *Deep Neural Network (DNN):* Designed after the structure of human brain neurons, they contain hidden layers that allow the model to identify different combination of features to better carry out the required task (e.g., classification).
- *Convolutional Neural Network (CNN):* Like its name suggests, it contains convolutional layers that have filters that can identify patterns. Multiple layers can allow for identification of more complicated patterns/features. The key difference with the DNN approach is that convolution of operations are performed here. As such, the model input shapes for CNN would be different from that of DNN.
- *Decision Tree- Extreme Gradient Boosting (XGBoost):* It is a decision tree algorithm that incorporates the advancement in this area, such as having multiple trees (bagging), randomly chosen features for each tree (random forest), using the output from one tree to the next for better performance (boosting), and using gradient descent to minimize errors (gradient boosting). Compared with gradient boosting, Extreme Gradient Boosting has multiple features built-in that enhance its performance. For example, it has a built-in cross-validation feature; to help the model better learn from the data. L1 and L2 regularization are also used to prevent the model from overfitting. In terms of efficiency, this model is capable of parallel processing that helps decrease the time required to build it.

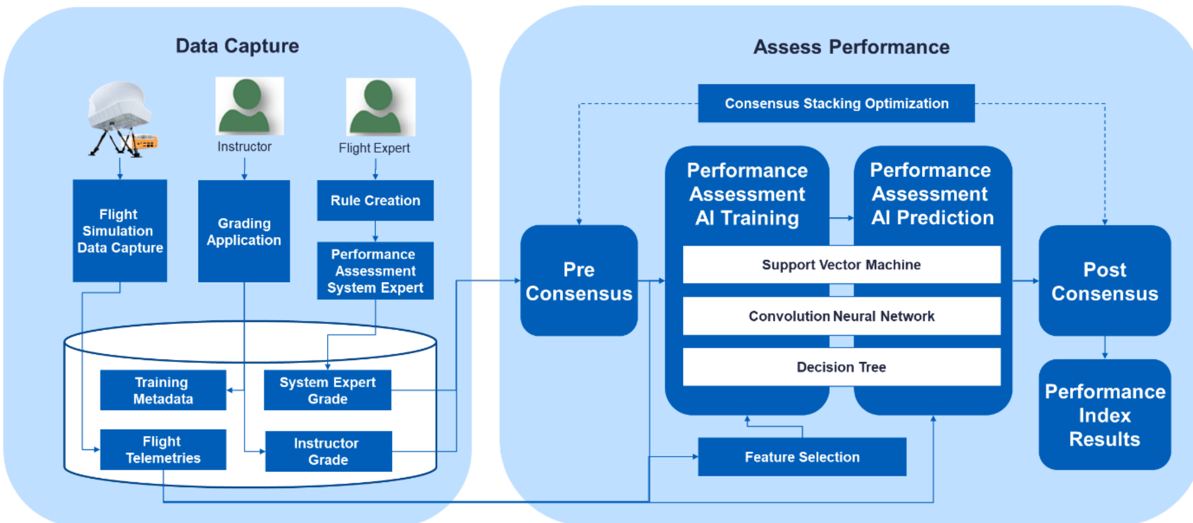


Figure 5.3 - Hybrid Machine Learning and System Expert Performance Assessment Architecture

In order to prepare the data for use by the machine learning model, several steps are taken. First, the data is cleaned during pre-processing. The dataset is then separated into train, validation and testsets and we normalized as appropriate on the train set. During the training process, hyperparameter tuning is conducted, as well as regularization (L1, L2 or Elastic Net), to prevent overfitting.

We hypothesize that a feature selection module that learns which parameter has importance in the grading of a flight maneuver will contribute to scaling for multiple training events. The feature selection model will identify the important features involved in a specific task to scale the models for multiple maneuvers.

5.4.3.3 Flight Phase and Training Event Detection

We identify time segments that correspond to training events and are the times of interest for pilot evaluation. By displaying them in the appropriate media, it allows instructors to save time in order to consult information for evaluation purposes or to demonstrate different concepts to students for pedagogical purposes in addition to feeding automated and real-time engines. Rule-based inference defined by a subject matter expert and analyzed by software services using a classic approach to detecting it is not always robust enough to fully capture all the possibilities and variations that define the context of an event. A particular difficulty lies in determining a good capture window, i.e. being able to predict with precision, within a time limit imposed by the rhythm of the training session, when an event begins and when it is finished. The uncertainty resides around the identification of each event for an increased accuracy rate even for a high maneuver complexity rate. We hypothesize that machine learning techniques can be used to increase detection capability using flight telemetry and some derived data.

Flight Phases divide a flight into separate sections with distinct characteristics, such as Climb, Cruise, or Descent. Coming from a pre-calculated flight parameters dataset, Flight phase is then an important marker to support training event detection.

Recognizing flight phase and training event patterns automatically requires a lot of pre-engineering effort using a rule-based approach. Despite the important amount of data and flight engineering experience, recognizing the flight pattern requires a lot of authoring, tuning, and testing. This results in high cost, as well as low portability between aircraft types and between pattern types. Machine Learning Training Event Detection is proposed to complete the flow from the flight data segment to the end of the grading process.

The more complex the maneuver, the harder it is to define the rules allowing the detection of training events by a classic rules-based software system. Moreover, a rules-based system is sometimes unable to detect certain events that have a high degree of complexity. We developed a machine learning model architecture to classify the presence of a training event in a training session evaluation. A Convolutional Neural Network (CNN) and XGBoost models were developed with instructor’s evaluation metadata and time windows generated from the raw telemetry data. We trained and tested a machine learning model to identify the presence of a training event individually for each relevant training event.

5.4.3.4 Multi-Label Optimization Using Consensus as Ensemble Method

Since the quantity of data and their statistical property are variable, the maturity of the various machine learning models is also variable. By applying weighted consensus method on machine learning models according to their level of maturity; it forms a strategy where the combination of a variety of model’s maturity is not leveling down the result. This strategy also allows combining multiple sources of knowledge, from flight training subject matter expert and certified flight instructor, as a human in the loop process that enhances the objectivity of the pilot performance assessment.

We first used a Pre-Consensus method using the multiple target labels, each target with the same meaning (a pilot performance assessment grade), but not from the same source of knowledge (Rule-based objective assessment engine vs. instructor grades). With these two sources of labels, we applied the pre-consensus to obtain a new target label that will be used also in the training of the machine learning performance assessment models. We also trained with the perfect agreement set of assessment by the objective assessment rule-based engine and instructor grades. This gives us either flawless performance or catastrophic outlier. We then combined with another data selection by removing exceptions and a separated set of perfect matches and worst matches in the assessment.

After training, validation and test of the machine-learning algorithms, we apply a Post-Consensus methodology.

It is the key strategy to manage the maturity of the models as a form of ensemble method that reduces the “Black box” side effect inherent to a deep learning model and can prevent providing good explanation for the flight instructor of the pilot performance. This is particularly useful during a cold-start approach where low maturity algorithms that might require different amounts of data can have less impact and degradation of the results compared to mature models. Various methods such as majority voting vs. weighted majority voting, where the weight is a maturity index composed of accuracy, lost, and F1-Score are used to optimize the results.

5.4.3.5 Dealing with unbalanced data

As shown previously, in terms of the grade labels, there is a lack of balanced data in the datasets. While there are tools from the Imbalanced Learn library, like over/under sampling and synthetic methods (e.g., SMOTE - Synthetic Minority Oversampling Technique), these did not yield very satisfactory results. Hence, there were a few different approaches that were combined to make up the strategy for treating the unbalanced data. We are using precision, recall & F1 score for the metrics, splitting the dataset for training to yield a less unbalanced dataset and using class weight to penalize over-represented classes. For hyperparameter tuning, we used Optuna (Akiba et al., 2019) a hyperparameter optimization software framework.

1) Using precision, recall, & F1 score for the metrics

As there is an unbalanced dataset, using precision and recall would be appropriate to evaluate the performance of the models before treatment for unbalanced data and after. As the F1 score incorporates both the precision and recall, this would be a convenient value to use for comparison.

Using these metrics for the classification cases would be straightforward. However, for the regression case, the model's output would be a continuous grade value from 0 to 100%. In order to use precision, recall, and the F1 score, the model's continuous output would need to be reclassified back into the four discrete categories (1/2/3/4, or 25/50/75/100%). The way this was done is described in Table 5.1

Table 5.1:. Continuous grade reclassification

Model's continuous output value	**Categorized grade value**
<37.5%	25%
>=37.5% & <62.5%	50%
>=62.5% & <87.5%	75%
>=87.5%	100%

The values found in the left column of Table 5.1 were obtained from the midpoints between each of the four discrete grade values. For example, the midpoint between 25% and 50% is 37.5%.

2) Splitting the dataset for training to yield a less unbalanced dataset

To split data for training, considering the unbalanced dataset, the majority and minority classes in terms of data points are found from the training dataset. Now the numbers of data points in the majority classes are divided by the number of data points int the minority class, to yield the number of folds. Also, the training dataset is now split into two sections – one containing only the majority class data points ("majority dataset"), and the other containing all other points ("rest of datasets").

In each of the folds, the "rest of datasets" is combined with a fraction of the "majority dataset". This fraction is proportionate to the number of folds. For example, if the number of folds is 5, then one fifth of the "majority dataset" is joined to the "rest of datasets" to yield the new training dataset. The best dataset is then chosen for training, by using the metric of the average F1 score for all classes. Once again, the F1 score calculation is straightforward for the classification case, but not for the regression case. The same approach as 1) to reclassify the model's continuous output value is used here.

3) Using class weights and Optuna for hyperparameter tuning

Class weights are also used to address the class imbalance. In order to determine the size of class weights to use, the Optuna hyperparameter tuning framework is used. Float values between 0.1 and 20 are suggested for the four different grade classes. The average F1 scores across all classes are used as the metric for Optuna to optimize and obtain the maximum possible value. The best dataset from step 2) is used here to find the best class weights to use.

If the dataset is being treated for instructor bias, the original instructor grade is used to determine which class weights to assign to that specific data point.

5.4.3.6 Instructor bias model

With multiple instructors, it is possible for some to be more lenient with grading, and others to be stricter. To remove this bias and standardize the scores, the average grade given by each instructor is calculated. Then, the average grade for all instructors is established. The average grade for each instructor is compared against the overall average for all instructors. This difference in value is measured against the overall average, to generate a percent difference.

To calculate the adjusted grade, the original grade (1-4) is first multiplied by 25, to yield a maximum score of 100%. The percent difference is applied to this grade, so the final score is capped at 100%. This means that a student who receives a grade of 4 by a strict instructor should technically have a percent grade above 100%, but this cap prevents this from happening. Note that this treatment of bias is applied to the regression model, since its continuous nature allows for such tweaking of grades. The following set of equations summarizes this approach.

$$avg\_all\_instructor\_grade = \frac{\sum avg\_instructor\_grade}{number\_instructors}$$

$$instructor\_diff = \frac{avg\_all\_instructor\_grade - avg\_instructor\_grade}{avg\_all\_instructor\_grade}$$

Note that instructor bias treatment is only applicable for instructor grades, and only in the regression case. From the above equations, the grades will be adjusted by a certain percentage. As classification requires that the labels of data (grades in this case) are discrete – 1, 2, 3, or 4, applying this grade adjustment would no longer yield discrete grade values.

5.5 Experimentation results

In this section we describe the obtained results of the experimentation of training event detection and performance prediction of the algorithms presented in the previous chapter. The analytics of this paper focused on the engine failure event at takeoff between the V1 and V2 takeoff speeds (EASA 2.5.2 GEN).

5.5.1 Descriptive Analytics

We extracted this training event from the data generated between 2018 and 2020, and we were able to extract 4,454 instances of this event. These instances are conducted in 2,634 training sessions, on 22 training simulator devices. There are 1,921 pilots that have completed this particular training events, and 97 instructors that have graded pilots executed it.

The Figure 5.4 and Figure 5.5 show the distribution of grades for both the objective assessment grade and instructor grade datasets. The objective assessment dataset has about half of the data points being a 3, with the grade of 2 as second most common.

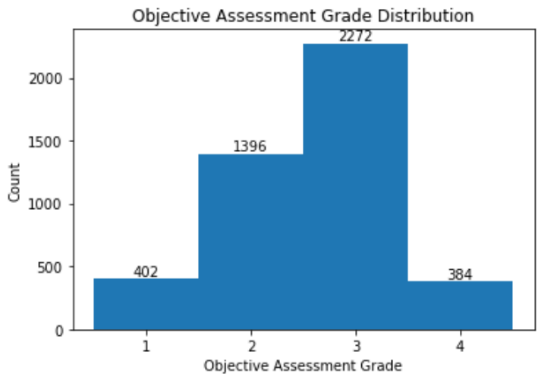


Figure 5.4 - Objective Assessment Rule-based system grade distribution & Statistics

Objective Assessment Grade Dataset	
Count	4454
Mean	2.59
Standard Deviation	0.77
Min	1
25%	2
50%	3
75%	3
Max	4

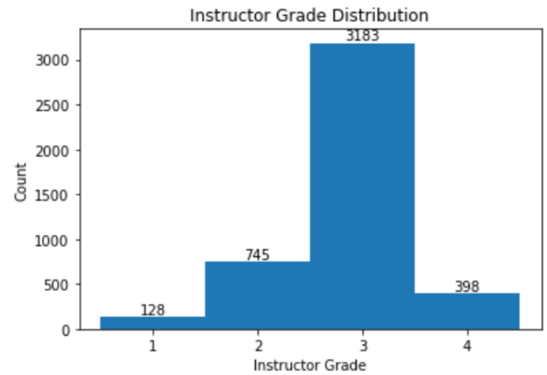


Figure 5.5 - Instructor grade distribution & Statistics

Instructor Grade Dataset	
Count	4454
Mean	2.86
Standard Deviation	0.59
Min	1
25%	3
50%	3
75%	3
Max	4

To compare the objective assessment (rule-based) and instructor grades, the difference between these two grades is calculated. Out of the 4,454 cases, 3,132 of them are exactly the same (70% of cases). In these cases, the rule-based engine is accurate in predicting the instructor grade. To break this 70% down further by time of training events, the Table 5.2 illustrates this percentage by time period.

Table 5.2:. Rule-based vs instructor grade agreement per training cycle (6 months)

Period	**Jul 2018 – Dec 2018**	**Jan 2019 – Jun 2019**	**Jul 2019 – Dec 2019**	**Jan 2020 – Jun 2020**	**Jul 2020 – Dec 2020**
% of cases with no difference between rule-based and instructor grades	63%	60%	70%	71%	83%

Looking at the trends in the above table, with later time periods, there is an overall increase in the percentage of cases with the same grade in both the rule-based and instructor-based scenarios.

In the event when the objective assessment does not predict the instructor grade accurately, 26% of the cases involve the situation where the instructor gave a higher grade than the rule-based engine. This means that the objective assessment is stricter in terms of evaluation. The remaining 4% of cases have the instructor grades being stricter. The distribution can be seen at Figure 5.6.

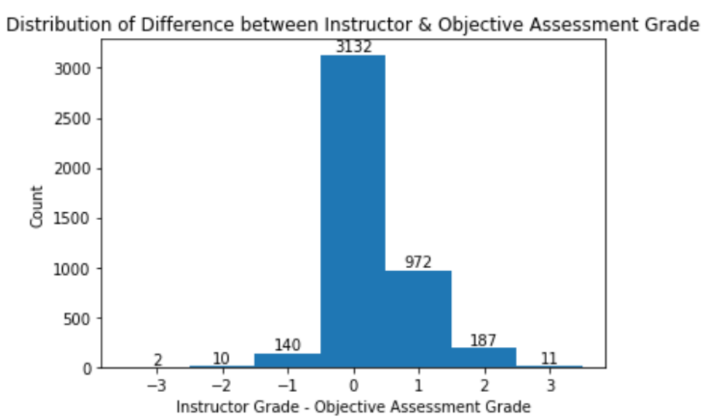


Figure 5.6 - Distribution of Difference between instructor and objective assessment engine grades

5.5.2 **Training Event Detection Results**

A training event average length is 7 minutes for a maximum of 58 minutes. Figure 5.7 presents the distribution of detected training events using rule-based software services.

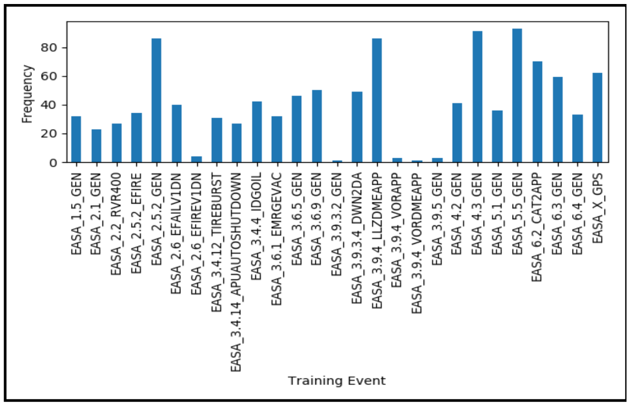


Figure 5.7 - Distribution of events detected using Rule-Based

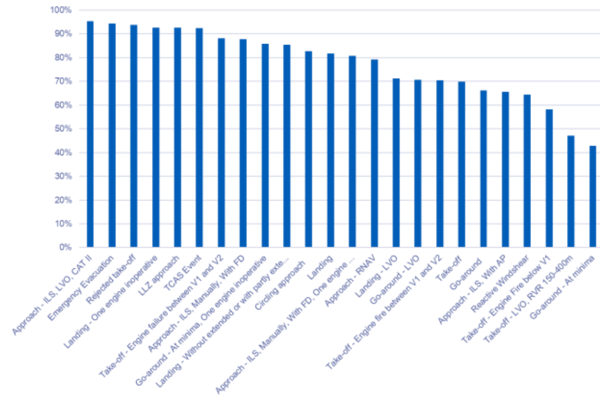


Figure 5.8 - Accuracy of Rule-Based event detection

Figure 5.8 presents the accuracy of rule-based event detection engine. The accuracy is based as a rate on the instructor's need to manually create the training event standardized by regulators and executed during a training session.

As the different sensors that collect the data are set at different frequencies, the first step before feeding the data to the model is to change the data from the sensors to the same frequency. For ease of running the scripts in terms of data volume, a subset of the more relevant sensors is chosen. A neural network was initially chosen as the model type. However, the performance was not spectacular, and for that reason a tree-based network (XGBoost) was considered, to see how the two models compare. This change in the model also allowed for a shorter script runtime. However, the XGBoost model requires that the sensor input be changed to a tabular format. For that reason, the general statistics for each sensor (mean, standard deviation, max value, min value) is extracted, as well as the fundamental frequencies of the Fourier transform of each sensor. Lastly, with a limited number of samples, a data generator is also used to create more data samples.

Except for a few outliers where the size of the dataset was insufficient as indicate in Figure 5.7, test F1-Score of over 90% of the detection rate was obtained using machine learning, as shown by Figure 5.9.

At that rate, involving the contribution of an expert to analyze the 10% gap has become very much effective and can provide a great new characteristic to be included in a flight safety analysis and can be subject to review with aviation standards organization and regulators.

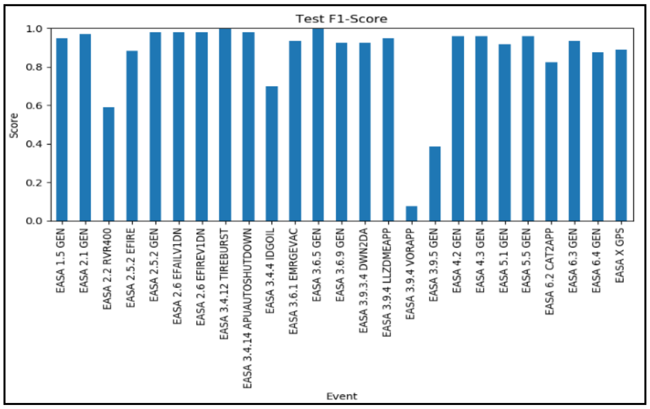


Figure 5.9 - Test F1-Score of machine learning training event detection engine

Figure 5.10 shows the results of training event detection using machine learning. Despite the addition of a machine learning model data generator, the results of CNN classification on telemetries as a time series recognition method still performed poorly given the small size of the dataset per maneuver type.

The tree model with engineering features to identify training events as opposed to deep learning was seen as more appropriate. It was found that an XGBoost classifier with designed features was better suited. The tree-based algorithm used was gradient boosting in its implementation of XGBoost. Feature extraction was applied to the raw sensor data in order to represent the data fed into the model as a collection of feature values as opposed to a collection of sensors time series.

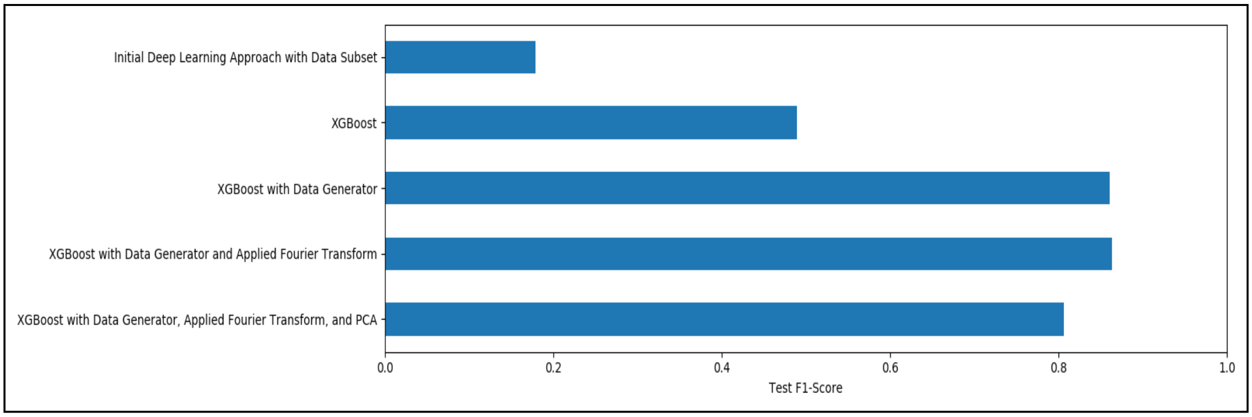


Figure 5.10 - F1-Score of Training Event Detection machine learning algorithms

5.5.3 Training Event Start-End Capture Window

We used a binary cross entropy for the detection and a Mean Absolute Error (MAE) for the regression on Window of 120 seconds. A deviation of 10 from the median of the Gaussian indicates a temporal standard deviation of 1 second. By analyzing the results of Figure 5.11, we can see almost 95% chance of having the correct prediction between +/-6 sec, 68% for +/- 3 sec. For an event that easily lasts 80 seconds, this makes a relative error of 5% compared to the duration of the event. However, with the low volume of data available per specific maneuvers there is a potential of overfitting.

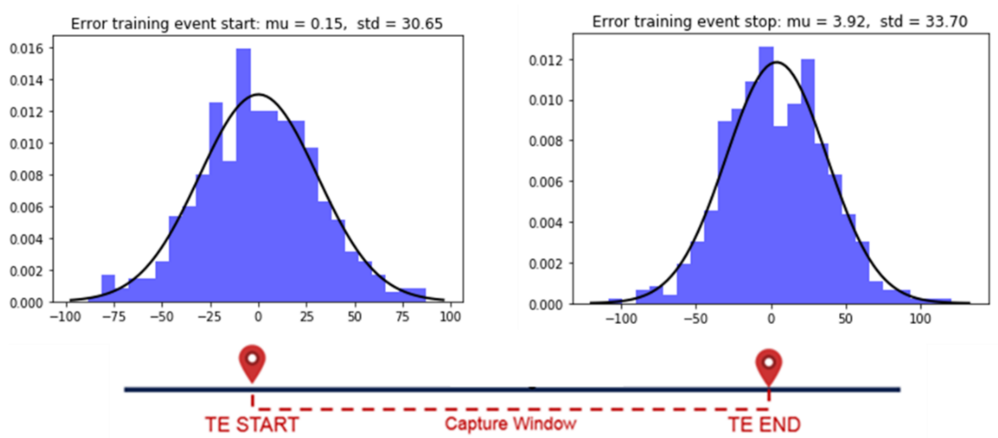


Figure 5.11 - Training Event Start and End time detection results to form the capture window

5.5.4 **Performance Assessment Results**

Figure 5.12 presents the accuracy of the machine models over data accumulation of pilot assessment data points. It presents the main SVM results for three supervised learning models that learned using instructor grading, rule base grading as well as the consensus of both instructor and rule-based grading at 50% weight each.

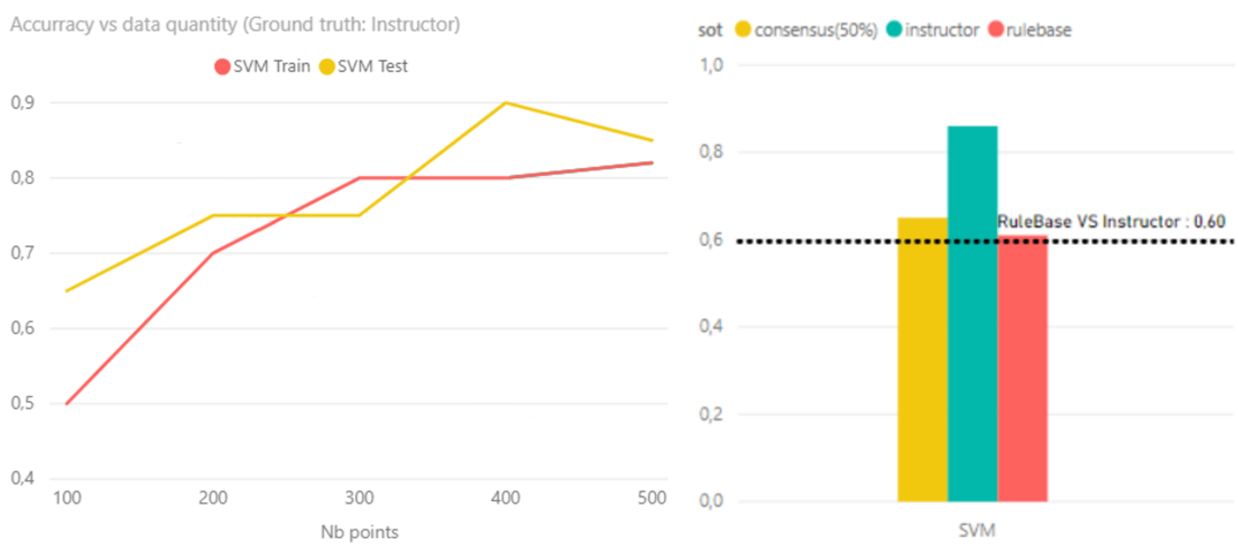


Figure 5.12 - SVM performance over data accumulation and compared to instructors and rule-based

We can see that machine learning SVM algorithm can grade technical performance like an instructor with the accuracy of ~85% vs. ~60% when using a rule-based assessment. We can determine how many maneuvers are required to reach the desired accuracy. We see that we can reach the 85% with around 500 data points using the machine-learning system by comparing with instructor grade labels. It also demonstrates that the AI model can replicate the rule-based at 99% of accuracy, making this architecture a good candidate for a cold start approach at the deployment phase.

5.5.4.1 Dealing with unbalanced data

The Figure 5.13, Figure 5.14 and Figure 5.15 illustrate the results comparing the baseline values with that of the balanced ones. The metric used here is the average F1 score over the four-grade classes. As regression results are shown here, note that the model's continuous output value was classified into one of the categorized grade values. Here, it would be ideal to see an increase in the average F1 score after balancing.

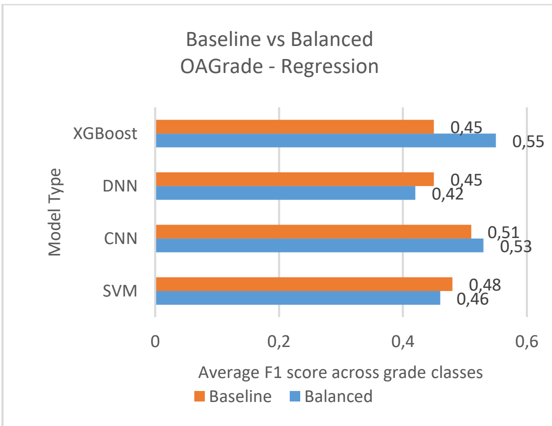


Figure 5.13 - Baseline vs Balanced OAGrade - Regression

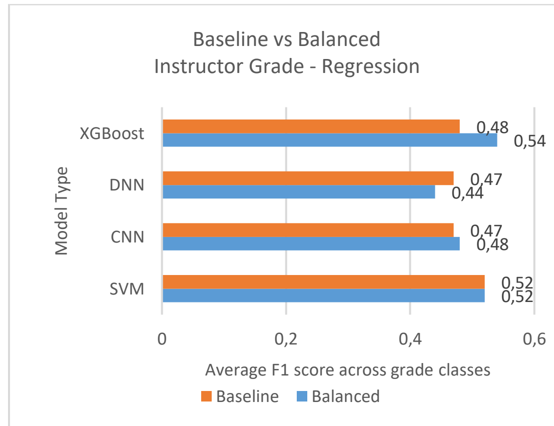


Figure 5.14 - Baseline vs Balanced Instructor Grade - Regression

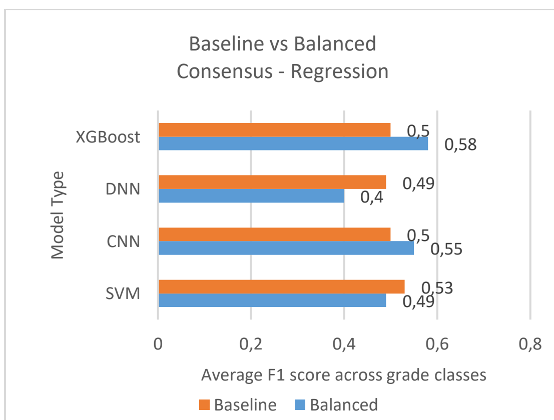


Figure 5.15 - Baseline vs Balanced Consensus - Regression

From the results presented at Figure 5.13, Figure 5.14 and Figure 5.15, there is an increase in the average F1 score in 7 out of the 12 cases (4 different models x 3 grade types) shown above. For the 5 cases that did not show an improvement in the F1 score after balancing, 3 of them are related to the DNN model, while 2 of them are related to the SVM model. Nevertheless, it is important to know that the results above do not tell the whole story, since they do not reflect the distribution of F1 scores across the classes after balancing. Consider, for example, the Instructor grades with the DNN model as shows at

.

Table 5.3:. Baseline and Balanced DNN results.

	DNN InstructorGrade bias_untreated Regression Baseline				**DNN InstructorGrade bias_untreated Regression Balanced**			
Results	25	50	75	100	25	50	75	100
Precision	0.814815	0.410714	0.755332	0.0909091	0.705882	0.240602	0.769841	0.131068
Recall	0.814815	0.150327	0.943574	0.0136986	0.888889	0.627451	0.304075	0.369863
F1	0.814815	0.220096	0.839024	0.0238095	0.786885	0.347826	0.435955	0.193548
Sample count	27	153	638	73	27	153	638	73

On the left are the baseline results, while the right shows the balanced results. Here, we see that there is a considerable improvement in the F1 score for both the 50 and 100 grade categories, at the expense of the 75 group. There is a negligible decrease in the 25 group. One of the goals for balancing is to ensure the model pays more attention to the minority classes, instead of the majority of 75 group. This is done by splitting the dataset into two - the majority and the rest of datasets and including a portion of the majority to the rest of the dataset for training. Therefore, while there is an overall decrease in the average F1 score across the grade categories, it is clear that the model is paying more attention to the other minority grade categories. This was the overall purpose of balancing the dataset.

5.5.4.2 Performance Index Calculation Using Regression on Multiple Labels and Pre-Consensus

To convert from a classification-based model to a regression-based one, there are a few modifications done. First, the label data is multiplied by 25. This converts the class-based data from 1-4 to a continuous one, from 0 to 100%. For XGBoost and SVM, the models are changed to use their regression instead of classification counterparts (XGBRegressor instead of XGBClassifier, SVR instead of SVC). For CNN, instead of having a softmax activation function in the last layer (as in the classification case), there is no activation function for the regression case. RELU was considered and may be used in later iterations. This allows the CNN model output to be continuous, instead of categorical.

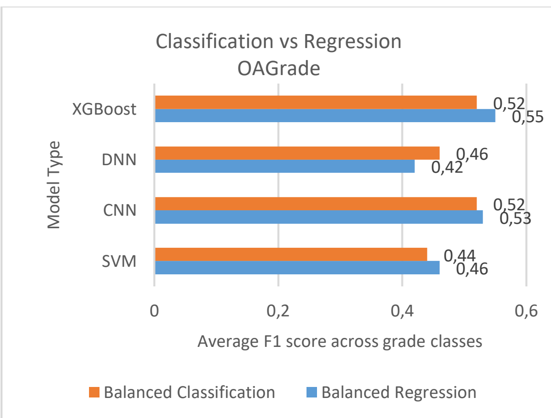


Figure 5.16 - Classification vs Regression OAGrade

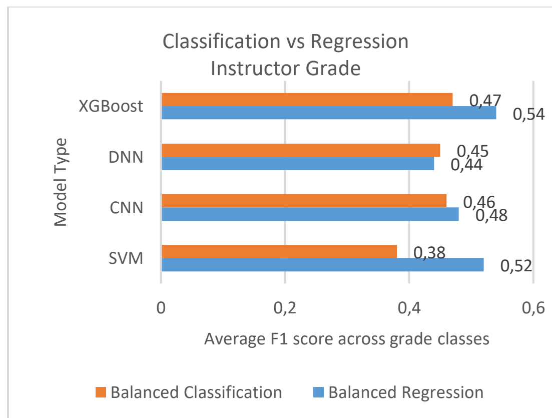


Figure 5.17 - Classification vs Regression Instructor Grade

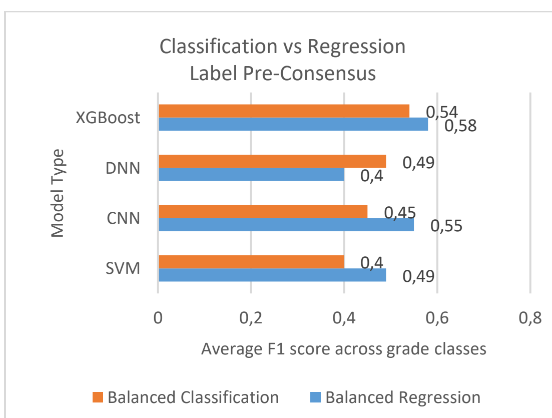


Figure 5.18 - Classification vs Regression Pre-Consensus

Comparing the performance in the classification and regression cases, out of the 12 cases (4 models x 3 grade types), there is a higher average F1 score in the regression case for 9 out of the 12 cases. For the remaining 3 cases where the regression case yields a lower F1 score than the classification case, all of them are related to the DNN model.

A possible reason for a higher F1 score (and better performance) in the regression case is that there is more flexibility for the model to be more accurate when quantifying the performance of the pilot. Nevertheless, when making this conclusion, one needs to acknowledge that there is also an extra step of reclassifying the continuous grades into discrete values in the regression case.

5.5.5 Consensus and model management

The Figure 5.19, Figure 5.20 and Figure 5.21 provide post consensus results of the machine learning regression's models for instructor, rule-based and pre-consensus target labels.

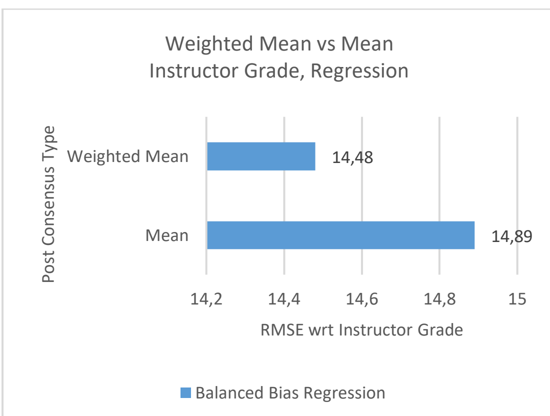


Figure 5.19 - Weighted Mean vs Mean Instructor Grade, Regression

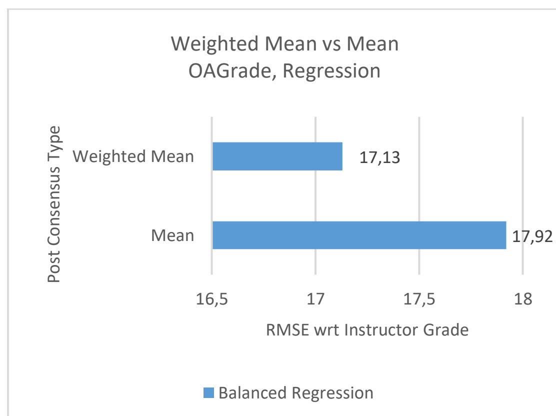


Figure 5.20 - Weighted Mean vs Mean OAGrade, Regression

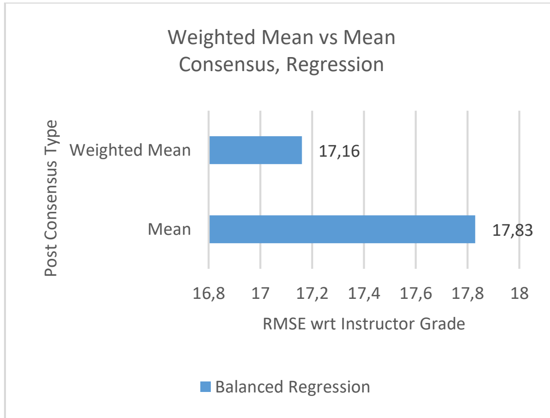


Figure 5.21 - Weighted Mean vs Mean Consensus, Regression

In the above graphs, weighted mean contains the following weights for the different models:

CNN – 0.4, DNN – 0.1, XGB – 0.25, SVM – 0.25.

Using the instructor grades for the model input would yield a better prediction of the instructor grades, as given by the lower RMSE values.

In terms of the sample size required for decent model performance, the following graphs show the RMSE values as a function of the number of samples used for building the baseline regression model. For the three different grades (OA, Instructor, and Consensus), the CNN model requires a minimum of 300 samples, before achieving a low RMSE value that is comparable with that of the other models. This minimum size is not observed with the SVM or XGBoost models; both offer a low and relatively steady RMSE from a sample size of 100.

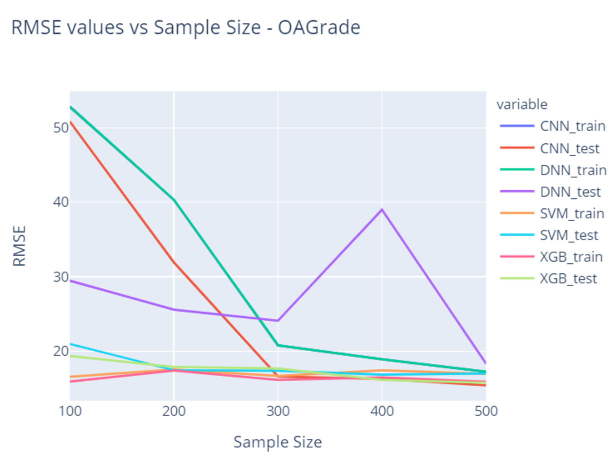


Figure 5.22 - RMSE values vs. Sample size - Objective Assessment engine grade

Figure 5.23 - RMSE values vs. Sample size - Pre-consensus grade

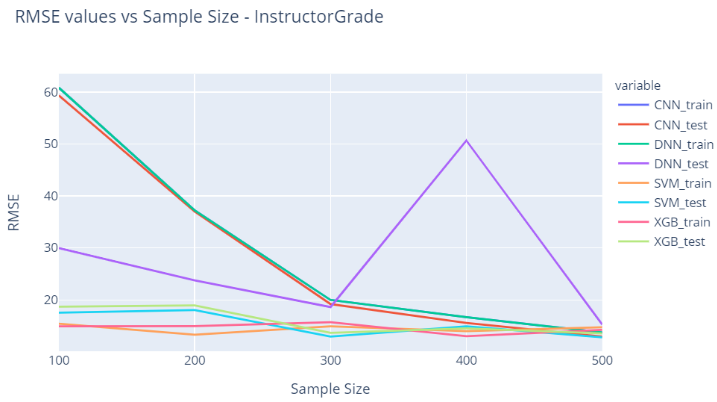


Figure 5.24 - RMSE values vs. Sample size - Instructor grade

5.5.6 Instructor bias results

To apply the bias treatment, the bias treated instructor grades are given to the model as inputs. This model will then balance the data as described in the unbalanced data section. In order to evaluate the performance of the models, the RMSE values for the balanced cases and the bias treated balanced cases are compared. Instead of using precision and recall as a metric like in the balanced treatment section, the RMSE is used. The reason for this is that the bias treatment adjusts the instructor grade by a certain percentage. This percentage might not yield a distinct difference, when we reclassify the results back into the four grade categories to calculate the precision and recall. However, the RMSE should allow us to identify if there is a clear improvement in the model's performance after treating for instructor bias. A lower RMSE would mean a more accurate model.

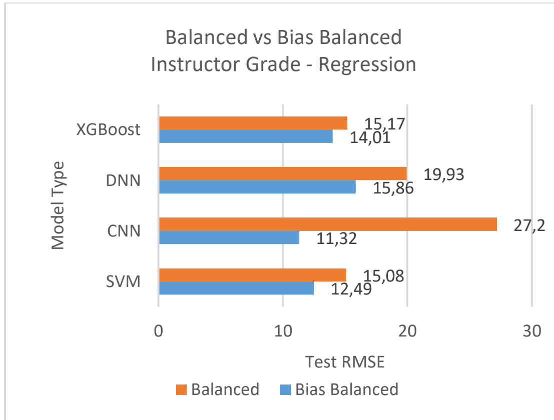


Figure 5.25 - Balanced vs Bias Balanced Instructor Grade - Regression

From the results above, there is an improvement in the RMSE score from all four model types, when the instructor bias treatment is applied to the balanced dataset.

5.6 Conclusion

5.6.1 Discussion

This research was developed to help instructors deliver training in accordance with airline standard operating procedures (SOPs). Key benefits of the system include the ability to objectively assess pilot skills in real time and provide insightful training analytics in the eventual support of a Threats and Error Management (TEM) and competency-based frameworks for EBT support.

An automatic capability provided to instructors can help them validate the results and proceed with the official assessment of the pilot performance and takes actions in the delivery of the training. Rule-based system is specifically built for aircraft models and maneuvers, and not easily transferable to new instances of the aforementioned cases. An automatic rule-based assessment engine cannot scale and cover every possible variation that could impact a human expert's assessment of student performance. The typical rule-based inference to detect it is not always robust enough to fully capture all the possibilities and variations that a flight maneuver might be subject to. The uncertainty here lies in how well we can fully assess performance by appreciating the simulated conditions and context. With artificial intelligence, we can make the data speak at a level unattainable by rules in an effective way.

The machine learning strategy presented in this paper was successfully integrated to rule-based to provide a hybrid solution that can support machine learning models training and cold-start deployment considering multiple sources of knowledge. The usage of a form of ensemble method demonstrates that the explainability of the models as well as an enhance model management method can provide a valid solution for a robust automatic flight performance assessment capability.

5.6.2 Future work

Reinforcement Learning is another machine learning technic that can be explored. With instructor in the loop that can provide feedback to the engine, it can improve the grading in runtime in a Man-Machine Teaming interaction in a distributed situation awareness context of the pilot performance. Semi-supervised learning can be also an option where consensus maximization with parameter estimation can be applied to consider that data may be missing (Missing instructor grades for few training events)

Transfer learning can be applied to be able to reuse learning model to more than one task to be automatically assessed. Transfer Learning techniques can be used to generalize the models for multiple maneuvers and with a limited size of data sample. Federated machine learning has horizontal, vertical, and transfer learning capability that can certainly be used in future work in order to combined multiple aircraft model and enable the scalability and cost reduction for the deployment on multiple training program from multiple training centre. The approach can also be used for Machine Learning Model Transferability to multiple maneuvers that provide strong difference in the flight profile (Straight Approach, Landing, Approach with go-around, etc.). This will address Scalability for multiple aircrafts and multiple training curriculums. The Gap Analysis Model will learn the difference between the Instructor and the rule-based grading to scale multiple aircraft where rule-based engine will not be present and where the instructor grading and the Gap Analysis Model will be used together to obtain similar results from an aircraft that will have both grades.

The system generated insights intended to help the instructor understand the performance of the pilots. A key element of performance evaluation is the identification of the root causes that led to flight exceedances. The root cause may not be explicitly captured in the raw simulated data. For example, a lack of pilot's situational awareness or decision-making can lead to a late reaction time or wrong course of action. The use of the neuroscience can then take on all its interest. A model for integrating technical skills and Non-technical skills in assessing pilots' performance can be used as mentioned by (Mavin 2010). Standardization of the grading can also be done using biometrics of the instructor to capture situation awareness during the evaluation of the pilot.

Finally, this research focused on the assessment of the pilot's technical skills that have specified performance parameters. In order to cover a full assessment, we should increase soft skills as well as crew resource management (CRM) to support Advanced Qualification Program (AQP) and Line Oriented Flight Training (LOFT)

5.7 Acknowledgements

Mehdi Taobane, Project Manager, DS4DM, Polytechnique de Montreal, Marc St-Hilaire, CTO, CAE Inc.; Mark Soodeen, AT&I Director, CAE Inc, Houssam Alaouie, AT&I Director, CAE Inc.; Patricia Gilbert, Project Manager, CAE Inc.; Anthoine Dufour, Data Scientist, CAE Inc., Yang Meng, Student, Trent University, Dac Toan Ho, Software Engineer, CAE Inc., Marc-André Proulx, Software Architect, CAE Inc., Samir Sahli, AI Advisor Microsoft, Jean-Mathieu Deschêne, Data Engineer, CAE Inc., Alejandro Sanchez, Data Scientist, Lixar Inc.

CHAPTER 6 ARTICLE 3: FEDERATED MACHINE LEARNING IN ADAPTIVE FLIGHT TRAINING

Federated Machine Learning in Adaptive Flight Training, Jean-Francois Delisle, Navpreet Singh, Jian Qi, submitted *in AIAA Journal*, March 2022

6.1 Abstract

The aim of this paper is to provide an adaptive learning methodology powered by artificial intelligence using federated machine learning in order to improve the efficiency of flight training, while respecting the constraints of data-access security of military context and privacy of pilots. We propose using federated machine learning as a recommendation adaptive flight training system for initial training for piloting T6. An adaptive learning system aims to help pilots improve their learning performance by providing relevant information and recommendations based on training data, assessment results, and simulation training performance. The adaptive learning capacity is tailored to the students' learning experience, according to their learning styles and preferences in order to provide an optimal path through the initial pilot training program. It provides an adaptive approach, which we utilize to adapt the training according to the type of the learning environment. The adaptive learning system recommends individualized learning paths based on the students' performance in several learning environments, such as academic exams, simulator training sessions, and aircraft flights.

6.2 Introduction

Federated machine learning (FML) is a new approach, which enables us to respond to different challenges related to secure data sharing and privacy in artificial intelligence practice. Some cases require the integration of data from multiple systems to take advantage of the added value of federated machine learning. Data is generally managed in silos within the industry in different areas such as finance, healthcare, etc. The aerospace industry is no exception and even more of a challenge when dealing with military aviation.

The recommendation system is one of the newest capacities in artificial intelligence for different sectors such as finance, entertainment, etc. In the context of optimizing the pilots' learning in aviation training, it is obvious that a collection of data with regards to the different learning environments becomes necessary, while considering the learning path in a holistic view. By collecting data from eLearning, flight simulation, and real flight activities, data analysis can objectively and efficiently improve training, reduce costs, and increase training quality, while anticipating the training needs of the next generation of learners, as well as increase our ability to meet future training needs.

The aim of this paper is to provide an adaptive flight training methodology powered by artificial intelligence using federated machine learning to improve the efficiency of the flight training, while respecting the constraints of data-access security of military context and privacy of pilots.

We propose a federated machine learning technique to be used as a recommendation system in adaptive flight training in the context of an initial training in piloting T6. An adaptive flight training system aims to help pilots improve learning performance by providing relevant information and recommendations based on training data, assessment results, and simulation training performance. The adaptive flight training capacity is tailored to the students' learning experience according to their learning styles and preferences in order to provide the optimal path through an initial pilot training program. It provides an individualized approach that we utilize to adapt the training according to the types of learning environment and the associated training device. An adaptive flight training system recommends individualized learning paths based on the students' performance in several learning environments, such as academic exams, simulator trainings, and real flights.

6.2.1 **Adaptive Flight Training**

AI-based adaptive flight training orchestrates learning sequences and recommends the optimal method of delivering individualized educational content to students based on their preferred learning styles in a training program. In the context of pilot training, the results of instructor-led training, simulated and real flights training session, and ground school activities are used to customize learning activities and resources so that a learner can complete the training program with optimal efficiency. This gives the instructor more time to focus on coaching students on the less tangible aspects of flight training.

Adaptive training is a known concept for more than 50 years as presented by (Kelley, 1969). This concept was also quickly adopted in aviation and applied in flight simulation by (Paul W. Caro, 1969). Automation came slightly later by (Brown et al., 1975) and the US Air Force (USAF) evaluated the training effectiveness of an automatic flight training system in F-4 Combat Crew Training. The US Army Research Institute continues in this vein with (Ludwig & Ramachandran, 2005), which augments the Intelligent Tutoring System (ITS) using an individual's traits, conditions, and behaviors to provide an Adaptive Instructional System (AIS).

6.2.2 Recommender System

Recommender system is now is a mature technology part of our day-to-day lives and used in many industry sectors. (Ricci et al., 2011) presented a handbook demonstrating the main principles of recommendation systems, such as Content-based Recommender Systems, Neighborhood-based Recommendation Methods, Collaborative Filtering, Constraint-based Recommenders, and Context-Aware Recommender Systems. There is also (Portugal et al., 2018), which proposed a systematic review of machine learning algorithms to identify trends in the use or research of machine learning algorithms in recommender systems with 121 articles on Content-based filtering, Collaborative filtering, and Hybrid filtering.

Several challenges can be involved in recommendation systems such as preference modeling, personalization, real-time learning, and protection of private data. In addition to data distribution and privacy issues that we are addressing in this paper, cold start issue, corresponding to the problem of initial deployment without data, is also a recurring problem in artificial intelligence solution.

Cold start in hybrid recommendation system is frequently used by combining collaborative filtering and content filtering as provided by (S. Yang et al., 2020). There is also (Volkovs et al., 2017) who addressed this issue using DropoutNet, a latent neural network.

6.2.3 Federated Machine Learning

Originally introduced by Google (McMahan et al., 2017), federated machine learning is a machine learning technique that allows the sharing of models formed from varied data sources and stored in a decentralized manner.

Studies on privacy by (Zheng et al., 2020) also suggested federated machine learning can allow different owners to collaborate together in order to obtain the advantages of an optimal model learned by using each other's strengths without actually exchanging data.

As described by (Q. Yang et al., 2019), there are three main types of federated machine learning, Horizontal Federated Learning, Vertical Federated Learning, and Transfer Federated Learning.

Horizontal Federated Learning: This technique is used when the datasets have the same feature space, but different samples. For example, two banks in different cities may have different user groups and the intersection between the groups may be small, but the feature spaces may be same, since the banks' business is similar. To improve a model that predicts credit defaults, all banks can provide their datasets to a horizontal federated learning model.

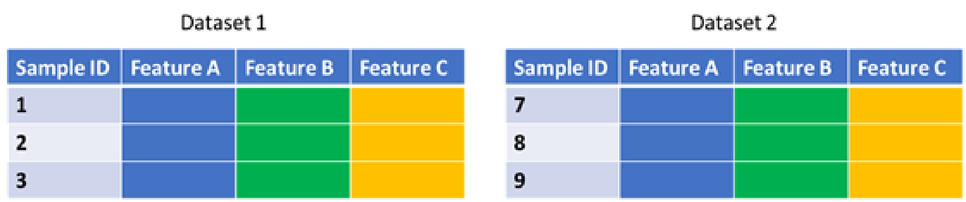


Figure 6.1 - Horizontal federated Learning: Different Samples, with similar feature space

Vertical Federated Learning: This technique is used when the datasets have different feature space, but same samples. For example, a bank and an e-commerce company in the same city may have a big intersection in their user groups. But since their businesses are quite different, their feature spaces are going to be vastly different. To predict whether a person is going to purchase a new product, vertical federated learning can learn about the person's financial information from the bank's dataset and their purchasing preferences from the e-commerce's dataset.

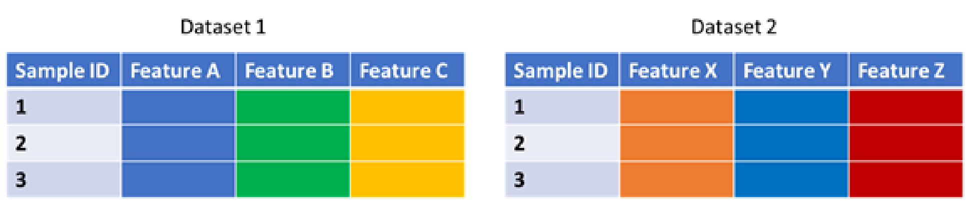


Figure 6.2 - Vertical federated Learning: Same samples, with different feature space

Transfer Federated Learning: This technique is used when the datasets have different feature space and different samples. For example, a bank in a given region and an e-commerce company in another region. The datasets from both companies will have different user groups, and because the businesses of the two companies are vastly different, the feature spaces will also be different.

Dataset 1

Sample ID	Feature A	Feature B	Feature C
1			
2			
3			

Dataset 2

Sample ID	Feature X	Feature Y	Feature Z
7			
8			
9			

Figure 6.3 - Transfer federated Learning: Different samples, with different feature space

In an FML task, each node initializes their own training using the common model retrieved from the server and builds a new model using local data. The parameters of the new model are returned to the server, which then averages the results with the other nodes.

To generalize, the goal of the model is to minimize the following objective function (Wu et al., 2020):

$$\min L(w), where\ L(w) = \sum_{k=1}^{n} p_k L_k(w)$$

where, n is the total number of datasets or node available, $p_k \geq 0$ and $\Sigma p_k = 1$ and L_k is the objective function of the model for the k^{th} dataset or the k^{th} node. The server receives the model updates from the local nodes in the form of the model weights, w^u and averages the weights as:

$$w = \sum \frac{\mathrm{n_u}}{n} w^u$$

where w^u are the weights of the u^{th} node, n_u is the sample size of the u^{th} node and n is the total sample size of the global data (i.e., the sum of the sample size of all the local nodes).

6.2.4 **Federated Machine Learning in aviation**

Federated machine learning has been studied in the aviation industry for unmanned aerial vehicles (UAVs) (Lahmeri et al., 2021). The authors provide a comprehensive overview of some potential applications of AI in networks for UAVs with a holistic literature review. They also highlighted some potential future applications of AI and its federated learning applications for UAV networks.

They reported some work done in federated machine learning for UAV-based networks such as path control, content caching, Flying Ad-Hoc Network (FANET) security, and UAV detection.

In (S. Chen et al., 2020), the authors proposed an approach to identify high-value passenger identification research based on federated learning and logistic regression model. They are providing a method to increase user privacy protection, as well as data security for airlines

The author (Sun et al., 2020) also identifies Federated Machine Learning as part of an Air-Ground Integrated Vehicle Network (AGVN) by proposing a predictable and intelligent architecture operating in a decentralized and geographically separated topologies. AGVN activity is expected to provide services to governments, businesses, and consumers in isolation. In this scenario, data privacy is a big concern for training models. Emerging methodologies such as blockchain and federated learning can be valuable for AGVN.

These two technologies are also offered by (Sharma et al., 2020) to use in the defense for distributed computing. The authors proposed to take advantage of the functionalities of blockchain technology and on-device federated learning in the defense IoT network. The proposed model introduces an approach, that limits training data by limiting it to only local nodes, while training the model based on a global view of the network. They also showcase the structure of the block in the blockchain network with detailed multi-layered computational workflow and experience classifying images using federated learning in various scenarios.

6.2.5 Summary

This paper will first restate the problem and the experimentation strategy by exposing the data used in an adaptive flight training solution. Each module of an adaptive flight training system will be described, from the pilot performance clustering to the system recommendation. We will present the statistics about the data used for the experimentation and machine learning algorithms. The Federated Machine Learning approach used from multiple sources of data in an initial training program will be presented, as well as and the cold start approach using data generation to comply with privacy regulation in military context.

The results of the performance assessment of a training center dataset will be presented against the results of the federated machine learning training phase from three datasets. Segmentation of the pilot performance and block prediction results will also be presented.

We will conclude with future perspective around the application of micro-learning strategy and learning workflow optimization.

6.3 Methodology

6.3.1 Problem Statement

We are aiming to tackle a machine learning problem in a distributed data context where different training centers would like to improve the accuracy of their training recommendations. The adaptive flight training capacity aim to tailor the student's learning experience according to their performance and to provide an optimal learning path through an initial pilot training program. We have data access and security constraints to preserve privacy between training centers. The amount of data samples and the cold start of an artificial intelligence solution are also constraints that we are attempting to tackle.

Because pilot training is very expensive, there is a need to optimize the effectiveness of a training programs to address the pilot shortage and reduce the training costs. Efficiency is measured by flight training performance, reduction in required flight hours, reduction in the time it takes to reach the first solo flight, and reduction in time till graduation.

For the instructor, there is a need to optimize the training delivery, so that they can focus on individual needs, provide early intervention, and provide just in time actionable recommendations. For the organization, there is a need to improve safety, reduce costs, and increase the quality of training objectively and efficiently. It is also essential to improve safety during training with a focus on training operations and to provide training delivery that cares about learner development rather than focusing on training outcomes that strictly emphases on evaluation results.

We hypothesize that Federated Machine Learning can enable training of an adaptive flight training artificial intelligence solution by using data from multiple training centers. We also hypothesize that the combination of various data sources can increase AI performance and accelerate the deployment of an adaptive flight training solution, while considering the cold start issue in machine learning.

6.3.2 Experimentation protocol

Our experimental environment covers a multitude of learning environments such as flight preparation, computer training, classroom lectures, classroom exercises with peers, quizzes and exams, and self-study with books and reference material. Immersive environments are also a part of the learning journey from self-paced training using virtual reality simulators and instructor-led lessons in full-motion simulators, to instructor-led training on aircrafts.

Using the hierarchy of skills documented in the learning management system (LMS), the adaptive learning system recommends individualized learning paths based on the student's performances and preferences (selected by the trainee or inferred from performance metrics) in several learning environments, such as academic exams, simulator training, and real flights. The system recommends additional study materials and course paths. The artificial intelligence engine also ingests the course curriculum, which allows the AI to recommend a learning path through lessons and maneuvers for the student. The artificial intelligence engine makes recommendations based on the information available in the Learning Record Store (LRS).

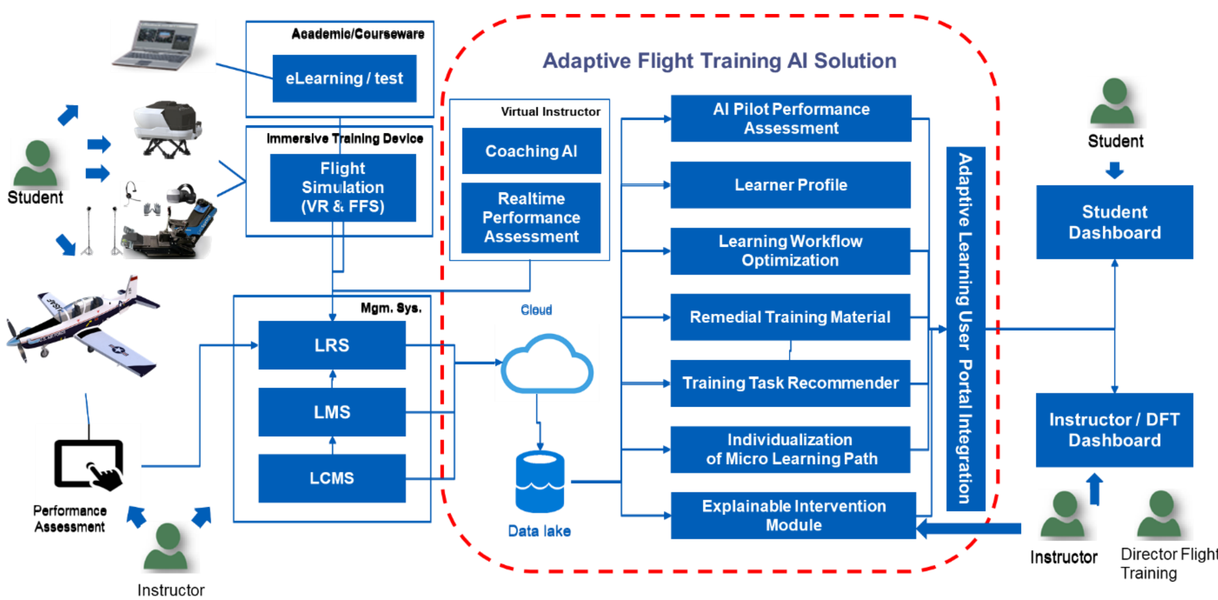


Figure 6.4 - AI Adaptive Flight Training Architecture

6.3.2.1 Pilot Performance Assessment

The pilot learning performance analysis module provides learning status within the program and allows students to get a view of their self-progress. Instructors can also view the learning path for different groups of pilots. For a training manager, this could be a useful indicator of how well the training program is training pilots. The overall assessment is based on the 8 ICAO competencies, which serves as the basis for micro-learning

6.3.2.2 Pilot Profile

The pilot profile module provides a 360° view of the student. The pilot grouping consists of finding and identify the performance models and learning behavior. This module therefore applies to a segment of students into performance and preference categories. It groups students into categories based on their performance and shows where a student stands in relation to others. Knowing which group a pilot is most closely associated with can help the student with a more effective journey through the training program. This can help tailor a training approach or style when teaching this student.

6.3.2.3 Training Task Recommender

The AI training task recommender recommends the next program activity for an individual pilot to maximize learning efficiency and time it takes for a student to complete all required activities of a training program. This module uses the performance history and driver profile to provide key performance indicators. The training task recommender uses the skills as a contribution to its recommendation. The AI algorithm recommends tasks from performance predictions at course, block, and lesson levels using collaborative filtering, a neural network approach, Bayesian knowledge tracing (BKT), deep knowledge tracing (DKT), directional graphing in a hybrid AI, and expert system-based approach.

6.3.2.4 Learning Workflow Optimization

This module makes it possible to recommend a progressive sequence of activities in the pilot training program in order to optimize the learning path. The optimized sequence is based on the historical activity performance of the individual pilot and on the optimal path. The optimization of the AI learning workflow provides an optimized sequence recommendation of lessons in the program to complete it faster and more efficiently. It provides a list of optimal learning flows using hybrid analysis and an AI-driven approach based on the Training Task Recommender. It separates students into an optimized course, a standard course, or a remedial course. It also predicts completion or transition dates for a cohort, then analyzes trainer-led lesson scores to indicate which areas need improvement or are working well. Finally, it identifies if there is a delay in a student's progress .

The recommendations can also optimize learning environments by varying theoretical courses, most courses in VR, and by plane. The most effective completion of the program must consider time and fluency. There is a little point in completing courses quickly, if the knowledge is not retained and applied successfully by the student. By recommending courses to students based on time and mastery, we should see an improvement in student performance when flying a real airplane. A student may notice that their knowledge in an early academic subject deteriorating as their education progresses. As they prepare for their final exams and check out the rides, it is recommended that they repeat some of this material in order to have the best chance of success.

6.3.2.5 Individualization of Micro-Learning Path

Hybrid AI-based modules produce individualized micro-learning at the learning objective (LO) level. It makes it possible to identify if there is any delay in a student's progress in terms of knowledge or skills. Based on performance and knowledge gap, this module suggests courses that can be taken out of sequence for a faster transition to aviation training, or to help the student get back on track to meet the group’s standard. Adapting the method of delivering learning to better suit the learner by recommending course material maximizes the success of the training. This module can also be used by instructional designers to help them decide what micro-learning content is needed and its effectiveness..The training task recommendation could be extended to make recommendations on micro-learning content during the program to increases the required competencies based on Knowledge, Skills and Attitude (KSA) competency decompositions.

6.3.2.6 AI Explainability and Intervention

The explainability and intervention module provides detailed information on the recommendations made and will provide the impact of the recommendations to the training program individually and globally. Students, instructors, and training managers must believe and trust that any AI component in the system is a useful contributor to the training mission and confirms that the AI components are contributing in a positive way. An instructor has the ability to modify the default sequence of lessons in the training program through its instructional intervention features. With data and performance visualization, this module reinforces all engines iteratively with user input, whether it is the student making learning requests or the instructor applying instructional interventions, such as an instructor seeking to speed up a particular student's learning so that they can keep pace with their classmates. It can also be the contribution to the educational aspect of the instructor’s actions, as well as the safety manager, who included training policies for compliance with the safety of flight operations.

6.3.3 Data Description

6.3.3.1 Flight Training Data Set

The federated organization datasets consist of three datasets as presented by Figure 6.5. One of the datasets comes from the Canadian military flight training center, the second comes from a United States civil training center and the third dataset uses simulated data using statistical properties from a military flight training center program, where raw data source is not directly accessed. All data is used while preserving security and export control regulation. The datasets include grades for students in different lessons or units of their respective program.

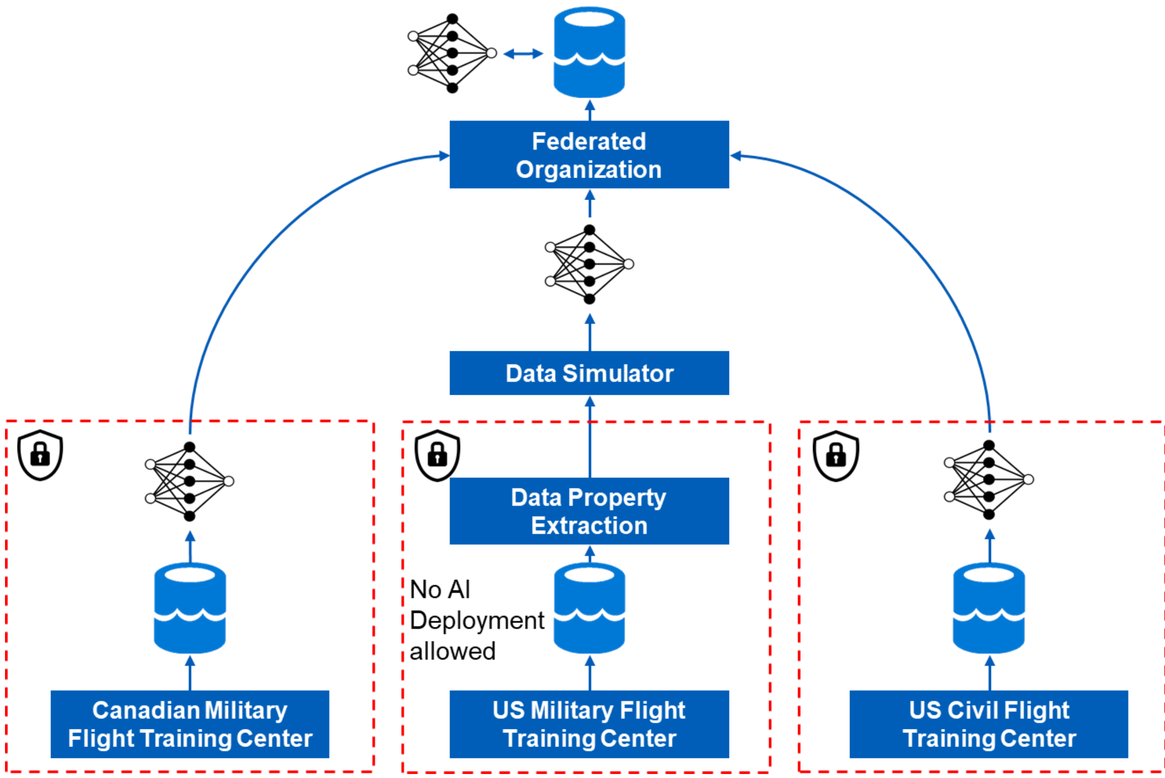


Figure 6.5 - Federated Data Organization

The Adaptive Learning system uses data from multiple learning environments where data are collected. Academic completion & examination, learning performance, simulator performance grading, flight performance grading, and instructor comments are regulated under a competency hierarchy framework, which define the success criteria for each learning objective.

The Canadian dataset consists of 926 students with 1268 different courses. Students are graded by tasks, which are a sub-category of the courses. Each student must complete multiple tasks for which they are graded on a scale of 1-5. Each row of the dataset consists of an identification number for the student and their grades for all the tasks they have completed. The civil US dataset consists of 760 students with 69 different courses. These courses are further broken into 86 different units for which the students are graded. The grading scheme for the civil dataset is labeled as SU, which means the student is graded as Satisfactory or Unsatisfactory. To maintain consistency, we have converted the grading scheme of the Canadian military dataset from a scale of 5 to Satisfactory/Unsatisfactory. Based on this new grading format, the distribution of the grades is as follows:

Table 6.1:. Distribution of grades for the datasets

Dataset	Satisfactory	Unsatisfactory
Canadian Military Data Set	98166	60160
Civil US Flight Training Center	25074	39526

The datasets include other related features like Instructor ID and Course ID along with the Student Identification number and the graded Unit/Task, where the distribution of these features is as shown below:

Table 6.2:. Distribution of grades for the datasets Number of unique features in the datasets

Dataset	Student IDs	Course IDs	Instructor IDs	Unit/Task
Canadian Military Data Set	926	1268	645	171
US Civil Flight Training Center	760	69	449	86

6.3.3.2 Simulated Dataset

It uses multinomial sampling to select students' grades for each block difficulty. Since there are 3 blocks thus there are 3 different levels of difficulty. Also the samples are taken based on the student's skill level, of which there are 4 levels, thus affecting the parameters of the multinomial distribution. Certain courses correlate, which means the grades between lessons are related. This is achieved using a graph method where the correlation is modelled via edges between lessons (each lesson is a node). The edges are placed based on hard pre-requisites (from curriculum) and soft pre-requisites (heuristics). The normalized graph adjacency matrix is used to multiply the grade vector to provide a weighted average output of the related observations. This effectively conditions future observations based on past grades.

By using the Watts-Strogatz random graph algorithm (Watts & Strogatz, 1998) to plot the lessons. Each graph node is a lesson, while each edge between nodes denotes a correlation between lessons. The algorithm follows two basic steps.

1. Creates a lattice ring with N (number of lessons) nodes of mean degree 2K (correlation between lessons). Each node is connected to its K nearest neighbors on each side.
2. For each edge in the graph, rewires the target node with some probability β. The rewired edge cannot self-loop or cannot be duplicated.

If β = 0, the graph is a perfect ring lattice, since no edges are rewired. However, if β = 1, the graph is a completely random graph. Setting the value of β to a value between 0 and 1 gives the desired graph.

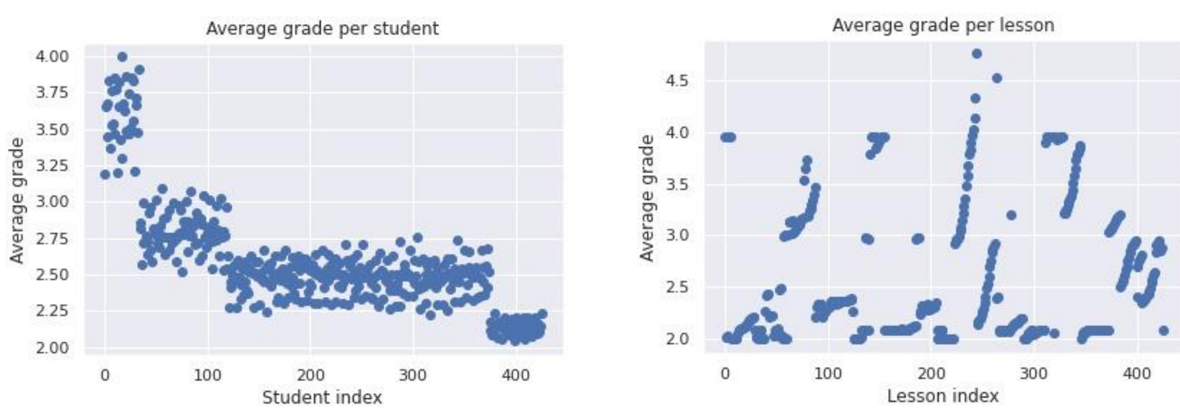


Figure 6.6 - Average grade distribution per student and per lesson

Table 6.3:. Grade distribution in the simulated datasets

Dataset	Satisfactory	Unsatisfactory
Simulated Data	594068	153316

Table 6.4:. Number of unique features in the datasets

Dataset	Student IDs	Instructor IDs	Lesson IDs
Simulated Data	180	18	426

The recommender system makes recommendation of a lesson for each individual student, based on their past performance. The system uses a neural network model Figure 6.7 to predict the probability of each grade on the scale of 1-4.

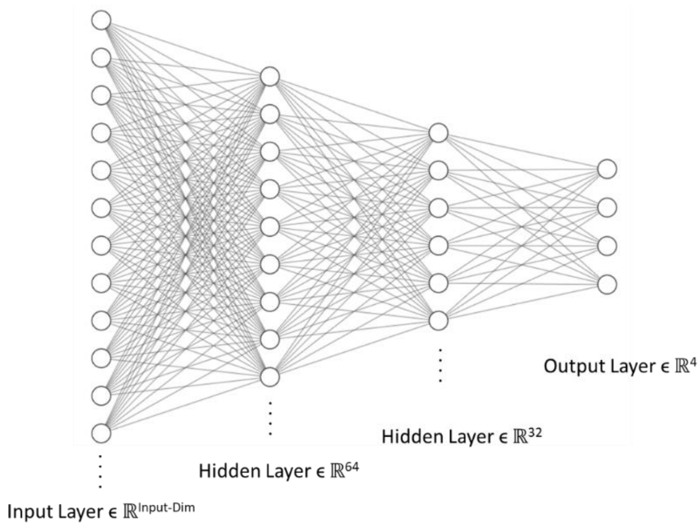


Figure 6.7 - Neural Network architecture for Task Recommender

The input format of the model is carefully curated in order to facilitate predictions. The input features consist of lessons in the training program. We create additional features for the model by using tagging information about the courses (as shown in Figure 6.8). The model is trained by using the training set, which includes the student's score for all lessons in the program. For example, suppose a training program consists of only 4 lessons as shown in the upper table in Figure 6.8, the training set has the information about the grades of 3 students in all lessons. We convert the upper table of information into the lower table in Figure 6.8 by adding a tag for the next lesson to make a prediction by utilizing the performance of the already taken lessons. By using this format, we can assist the neural network at making better predictions. When we need to make a recommendation, we use the student's past performance to predict their performance in all the upcoming lessons and select the one that will help them to graduate in the shortest amount of time.

Student ID	Lesson 1	Lesson 2	Lesson 3	Lesson 4
120001	4	3	5	2
120002	3	2	4	1
120003	5	5	2	3

Student ID	Lesson 1	Lesson 2	Lesson 3	Lesson 4	Taking Lesson 1	Taking Lesson 2	Taking Lesson 3	Taking Lesson 4
120001	4	0	0	0	0	1	0	0
120001	4	3	0	0	0	0	1	0
120001	4	3	5	0	0	0	0	1
120002	3	0	0	0	0	1	0	0
120002	3	2	0	0	0	0	1	0
120002	3	2	4	0	0	0	0	1

Figure 6.8 - Format of the input to the Neural Network model

6.3.4 **Preprocessing for Federated Learning**

Among the three different types of federated learning based on data distribution, we use the horizontal federated learning approach where the dataset features are similar, but the sample ids are different and non-overlapping. Since our datasets include data about three different pilot training programs, the number of total courses and lessons across all three datasets are different therefore, resulting in a different feature space. To facilitate the use of a single neural network model for all three local nodes of the federated learning approach, we reduce the dimensions of all three datasets to a single feature length. This is achieved by using autoencoders. By fixing the size of the bottleneck layer of the autoencoder to 100, we can reduce the dimension of all three datasets to 100. This eases the implementation of the federated learning approach.

6.3.5 **Federated Machine Learning**

We use the federated learning algorithm to aggregate the updates received by the server from the individual dataset nodes. The individual dataset nodes send these updates after completing the local training for the specified number of epochs. The algorithm starts by initializing the weights of the neural network model as w_0. The weights can be initialized randomly or by using the default Glorot kernel initializer. Glorot initializer was selected as the weight initializer because this initializer draws samples from a uniform distribution within a specified range, which depends on the size of the input tensors.(Glorot & Bengio, 2010). At communication round r, the server sends the initialized weights (w_r) to each of the individual dataset nodes. Each node splits its dataset into minibatches of size B and train the neural network using w_0 and their respective dataset to update the weights from w_r to w_{r+1} utilizing Stochastic Gradient Descent (SGD). The number of epochs to perform the local training is fixed to E. Each individual node sends the updated weights w^k_{r+1}, where k is the kth node, to the central server. The server then averages the weights across the nodes to compute the new model weights w_{r+1}. The following pseudocode explains the algorithm followed for implementing federated learning. (F. Chen et al., 2018)

Algorithm Federated Learning: The three data nodes are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate of the local optimizer.

ServerUpdate:

Initialize w_0 (*glorot_normal*)

for each communication round r = 1, 2, … do

    for each client k, in parallel do

        w^k_{r+1} ← **NodeUpdate**(k, w_r)

    endfor

    $w_{r+1} \leftarrow \sum_{k=0}^{n} \left(\frac{n_k}{n}\right) w^k_{r+1}$

endfor

NodeUpdate (k,w): //on node k

β ← (split the data of the node into batches of size B)

for each local epoch i from 1 to E do

    for batch $b \in \beta$ do

        $w \leftarrow w - \eta\Delta\ell(w;b)$

        endfor

    endfor

    Send w to server.

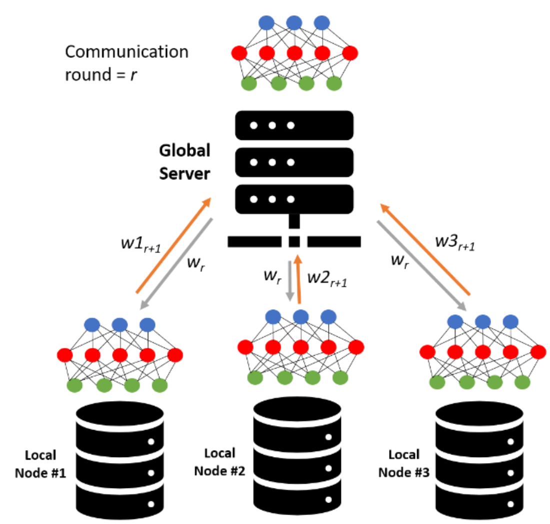


Figure 6.9 - Federated Machine Learning Algorith

6.4 Experimentation results

6.4.1 Performance Assessment

Pilot Performance assessment is conducted to analyze the performance progression of learning, overall performance, milestone completion status, syllabus completion status, and performance indicators of students, in order to provide descriptive analysis and student's learning performance in terms of multiple aspects.

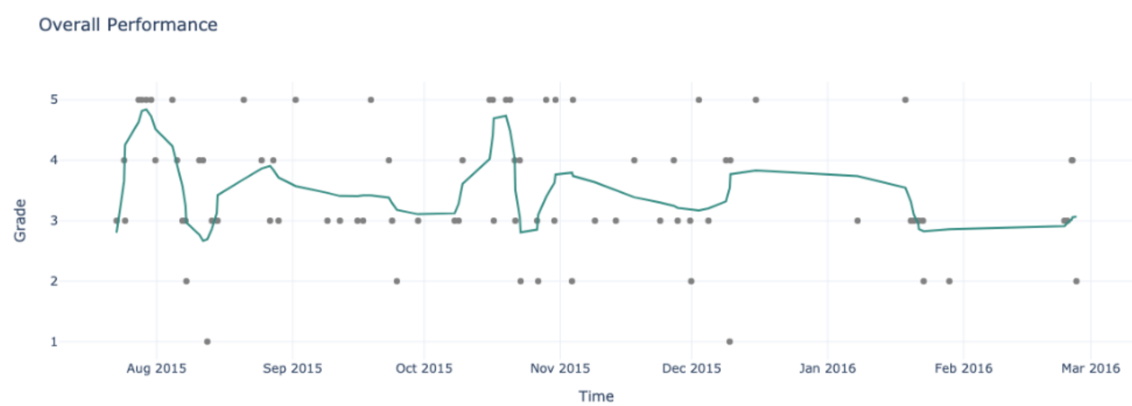


Figure 6.10 - Variation of individual student's lesson grade along with the approximated trend

6.4.2 Learner Profile

Pilot segmentation utilizes data-driven AI clustering algorithm to create student profiles, identify the pattern of each profile in terms of learning performance and behavior and then potentially provide actionable recommendation on the cohort level. Specifically, we use tSNE for dimension reduction and K-means for clustering to generate the learners' profiles.

Figure 6.11 - Learner profile/segment in terms of performance and progression indices

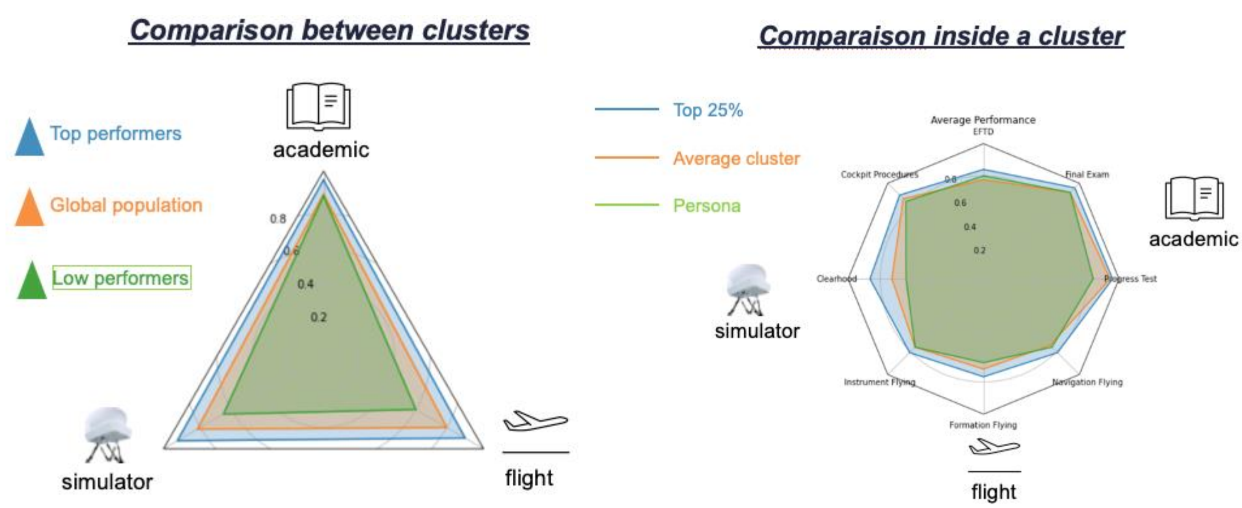


Figure 6.12 - Comparisons of performance among clusters and within cluster

6.4.3 **Training Task Recommender**

Block Prediction: This model aims at predicting the number of lessons a student is expected to fail in the upcoming block of lessons based on their performance in the previous block of lessons. Figure 6.13 shows the performance of a student in the three different blocks of lessons in the flying category, namely Clearhood flying, Instrument flying, and Navigation flying. X-axis depicts the grading scheme, i.e., a grading scale of 1-5, and the Y-axis shows the number of lessons in a block for each grade. Block prediction uses XGBoost model to predict the number of lessons a student will fail in the upcoming block of flying lessons, i.e., Formation flying. The student in the example is failing 6 lessons in the upcoming block and the model has predicted the number of lesson failures as 5.97. The model can predict the number of lesson failures with a mean error rate of 0.92. This shows that the model is highly accurate. This model can be used to improve the performance of a student by letting them know in advance if they are expected to fail a high number of lessons in the upcoming block. The student may start putting in extra efforts to pass the lessons in the upcoming block of lessons. It can also help the instructor by alerting them about a student if they are expected to fail a high number of lessons in the upcoming block.

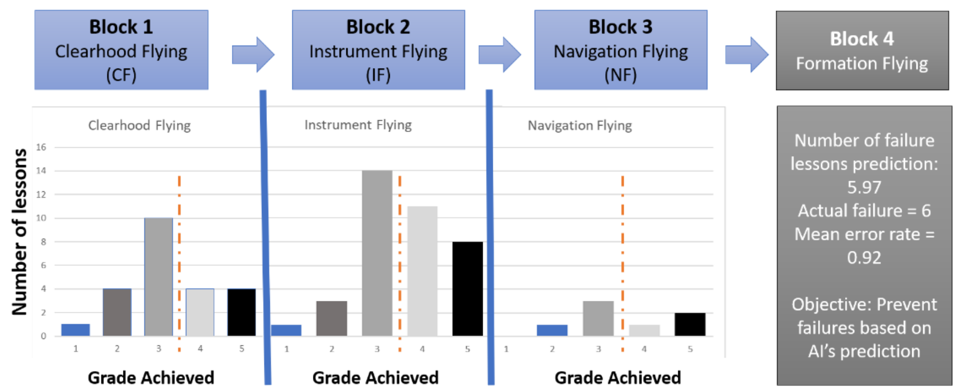


Figure 6.13 - Block prediction

6.4.4 **Federated Learning Results**

Table 5 shows the accuracy of the global model tested on all three datasets after 100 communication rounds of the federated learning. Since we are using the same model for all three local datasets, the datasets are encoded using an autoencoder to convert them into a feature space of 100. Performing this step reduces the accuracy slightly, but it is not very significant (as can be seen in the table), thus we decided to proceed with this approach. The accuracy after 100 rounds of communication of the federated learning approach is listed in the third column.

Table 6.5:. Accuracy of the global model for all three datasets

Data set	Accuracy on the Neural Network	Accuracy after encoding and using same neural network	Accuracy after 100 rounds of federated learning
Canadian Military Data Set	85.5%	84.0%	79.3%
Civil US Flight Training Center	64.6%	61.2%	61.1%
Simulated Data	98.6%	98.3%	98.0%

Figure 6.14 shows the accuracy of the global model tested with the test data from all three different datasets i.e., Civil US Flight Training Center, Canadian Military Data Set, and Simulated, after 100 communication rounds of federated learning. The plot displays the initial low accuracy for each data set during the first few rounds of federated learning. This behavior could be attributed to the fact the local models are fitted better to their corresponding data initially. But with each progressing communication round, the accuracy improves. This can be credited to the weights and bias shared by the local models with the global model and the global model sending an average of the same to each local model. This helps generalizing the weights and bias for all three local models. And after communication round 40, the accuracy seems to stabilize.

Figure 6.15 shows the accuracy of the global model, when tested on the cumulative test data from all three datasets. The test data from all three datasets are joined into a single dataset and the samples are shuffled before testing the global model. The accuracy seems to be higher than the individual Canadian Military Data Set and Civil US Flight Training Center datasets (as shown in Figure 6.15). This can be attributed to the fact that the number of samples in the simulated dataset is quite higher compared to the Canadian Military Data Set and Civil US Flight Training Center datasets.

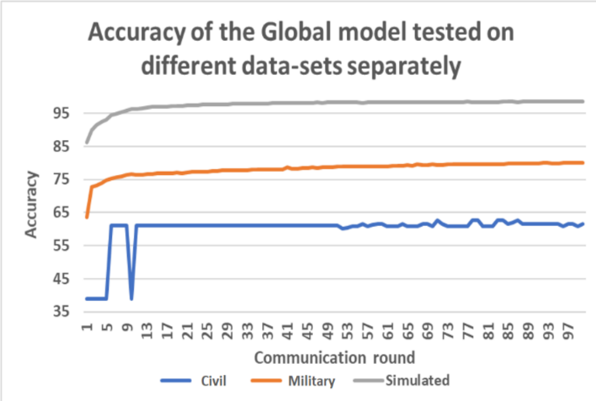


Figure 6.14 - Accuracy of the Global model on separate data

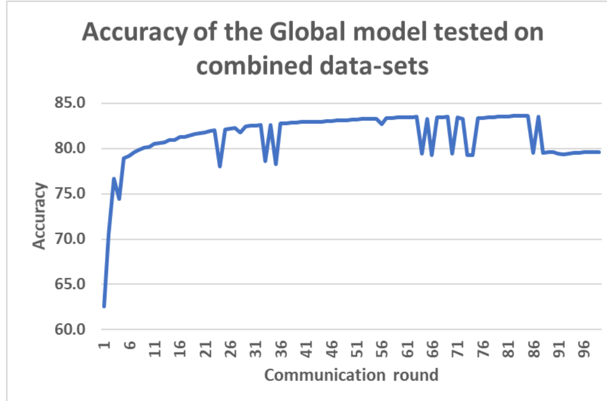


Figure 6.15 - Accuracy of the Global model on combined data

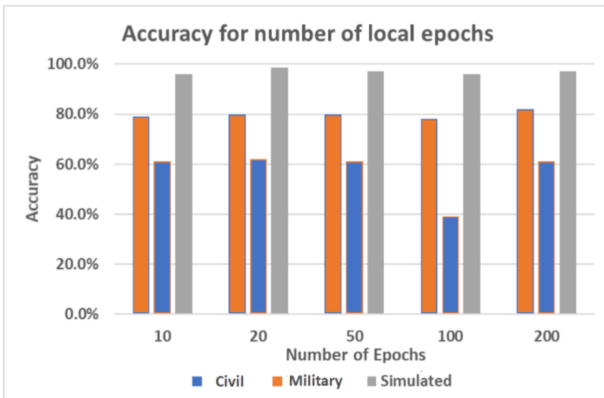


Figure 6.16 - Accuracy for different number of epochs for the local node

Figure 6.16 shows the accuracy of the federated learning approach by varying the number of epochs for the local model. This dictates the extent to which the local models generalize on their respective dataset. As can be seen from the figure, 20 local epochs gives the best performance. Number of local epochs is not the only hyper parameter that can be tuned. Another important hyper parameter is the local optimizer.

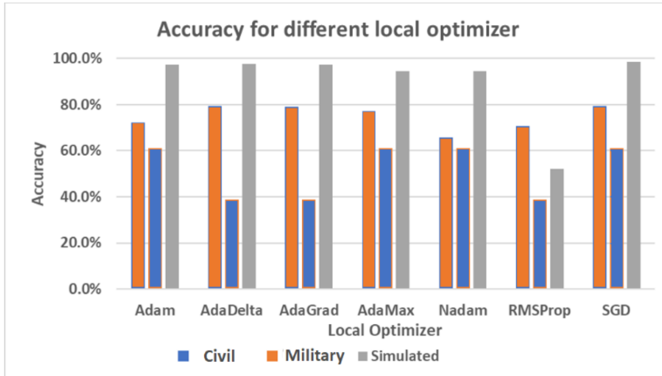


Figure 6.17 - Accuracy for different choice of local optimizer

Figure 6.17 displays the results obtained by using different local optimizers. Different optimizers have different techniques to converge to the global minima. The different optimizers tested are Adam, AdaDelta, AdaGrad, AdaMax, Nadam, RMSProp, and SGD. Using the Stochastic Gradient Descent (SGD) optimizer for the local models gives the best accuracy compared to the other optimizers. Although AdaDelta and AdaGrad give similar results for the Canadian Military Data Set and Simulated datasets, their performance when used on the Civil US Flight Training Center dataset is highly degraded. This experiment helped us to conclude that SGD is the best optimizer to be used locally.

6.5 Conclusion

We provided a federated machine learning solution to be used for prediction and recommendations in adaptive flight training as part of an aviation training. We have improved this intelligent flight training solution while respecting the security constraints of accessing military data and the confidentiality of pilots.

We faced issues of data quantity and quality in machine learning, as well as privacy concerns, in a cold start context of a new deployment of an artificial intelligence solution. FML offered us a solution to address these issues with an adaptive flight training methodology powered by artificial intelligence using a methodology to preserve data locally, while training global artificial intelligence.

As future work, we are looking to improve this ecosystem with more automation and deep adaptation of learning path.

The remedial course functionality could be extended to include the creation of personalized courses. An AI model would combine existing content into a custom lesson based on a subset of the existing tutorial. The adaptive learning system would also continuously refine the learning journey based on student's interactions and instructor's feedback.

While training recommendations are generally based more on instructor's observations and feedback, the goal should be to automate these recommendations as much as possible by using machine learning as the knowledge base of instructor intervention grows. By utilizing a smart pedagogical intervention module with human-in-the-loop, the need for manual intervention is expected to decrease over time as the adaptive learning model is properly trained.

The Deep Knowledge Tracing seems to be among the most accurate algorithm, however could not be tested on simulated data with the addition of metadata. This has now been implemented; it would be better to see how the algorithm works on more realistic data. Implement a parallel model on the metadata associated with each learning objective and the corresponding competence and attitude attributes (KSA).

In the rest of these experiments, we will use this concept in adaptive learning with a micro-learning approach while using Deep Knowledge Tracing. We will validate the Deep Knowledge Tracing approach in a federated learning architecture across multiple learning medias using multiple learning platforms to learn the flow optimization models.

6.6 Acknowledgements

Acknowledgements to CAE Inc. collaborators: Philippe Perey, Head of Technology, Defense and Security, Christophe Courtoy, Product Manager, Defense and Security, Laurent Desmet, Data Scientist, Soizic Bourgeois, Data Analyst, Adam Dementov, R&D Software Engineer, Ben Winokur, R&D Software Engineer, Antonios Valkanas

and Andrea Lodi, Research Director, Polytechnique de Montreal

CHAPTER 7 ARTICLE 4: PREDICTING PILOT TRAINING PERFORMANCE USING PSYCHOMETRY AND FLIGHT PERFORMANCE DATA

Predicting Pilot Training Performance Using Psychometry and Flight Performance Data, Jean-François Delisle, Bincy Baburaj Narath, Karen Moore, Submitted in *The International Journal of Aerospace Psychology*, July 2022

7.1 Abstract

We talk about pilots having the "Right stuff", and significant investment is made by both air operator certificates (AOCs) and approved training organizations/flight training organizations (ATO/FTOs) in assessing entrants to the aviation sector to try to find those people who have the "right stuff" for a successful career as a pilot.

IATA/ICAO has a clear competency framework for pilots, with a "competency" being defined as the knowledge, skills and attributes needed for an individual to carry out the role effectively; in other words, their behaviour, what and how they do something. Psychological theory around personality assessment states that understanding a person's preferences, their natural behaviours, predicts greater success in job performance where there is a good match between personality and the required competencies. Assessment at both ab initio and Direct Entry (DE) levels should be aligned with the IATA/ICAO competencies for greatest predictive validity. This is the theory, but does it hold up in practice?

A recent data analysis and artificial intelligence development of training assessments against personality and cognitive test results conducted by CAE and Symbiotics showed some interesting results relating to personality preferences and success in assessments during training, suggesting a need to take the personal style of cadets into account in training design.

This paper presents a study on pilot performance and pilot aptitude testing using AI methodology to help the pilot selection process and enable better individualized adaptive training support for our cadet during their initial training program.

7.2 Introduction

A cadet pilot trains to become a pilot through a 72-week training program to gain the necessary skills that will allow them to obtain a commercial pilot licence (CPL) with instrument rating (IR). They then move onto multi-engine (ME) aircraft, and subsequently successfully operate as a first officer on multi-pilot and multi-engine airplanes in commercial air transport.

Learner profiling using artificial intelligence capability can automatize decision-making in cadet selection based on pilot performance and pilot psychometric test data. This can offer better financing support, more relevant curricula, less remedial training, no cadet abandoning the program and better pilot pairing assignment.

A learner profile is a great tool to provide insight on pilot performance. It can be used to predict and explain a cadet's probability of success, improve overall success of cadets and communicate success potential. It can help improve admission decisions and increase access to pilot training for cadets. Appropriately used by training managers and instructors, it can individualize support based on results, and identify students who will need to repeat lessons. With the learner profile well integrated into a flight training management system, the instructor can adapt his teaching style based on this profile. By improving the overall success of cadets, we will also contribute to increased flight safety levels.

These advantages can significantly impact the business of a Flight Training Organization by reducing drop-out rates, reduced training cost, improved student satisfaction and improved airline satisfaction rates regarding pilot training. In a modern flight training process, learner profiling can be used to adapt the training and provide just-in-time recommendations to cadets, instructors and training managers. One of the foundational elements of the adaptation process is the ability to segment the learners and predict cadet performance in a competency area based on the psychometric test.

7.2.1 Objective & Hypothesis

This paper aims to consider the impact of using artificial intelligence for the prediction of cadet learning and flight training performance. Pilot psychometric testing data will be considered in relation to flight training performance data. The main objective of this study was to answer the question: can we correlate Pilot Psychometric Test Results with their performance during Flight Training to build a model that can predict performance?

It intends to consider the difference that information analyzed through clustering can provide compared with the traditional view of looking at each individual, even through the lens of normative data. The analysis should evidence that profiles can be segmented based on the outputs of psychometric testing and training performance, allowing future applicants to be more effectively selected and supported in the training program. The cadet sourcing process will be accelerated through the use of more refined screening models.

Once cadets enter the training program, the segmentation of profiles will support adaptive learning by suggesting additional learning content and individualized support services based on the psychometric results. Instructor/cadet pairs will be more efficiently matched through deeper understanding of the learner profile, further enhancing the learner support provided, as well as providing adaptive learning by suggesting additional learning content and individualized support service based on aptitude and performance results. Individualized support can be based on results (instructor support). Other questions addressed include: Can we identify strengths and weaknesses of those students? Based on the aptitude tests, how likely is the student to succeed in the cadet program? How likely is the student to fail?

New insights to the strengths and weaknesses associated with different profiles will permit more accurate prediction of licence success, which will facilitate the provision of cadet financing. Given funding is one of the greatest barriers to accessing pilot training, leading to homogenous cohorts, extended provision of financing will increase the diversity of cadets, while ensuring that the improved predictability of success will ensure cadets do not incur unsustainable debt.

7.2.2 **Literature Review**

The application of clustering analysis to flight training data was researched 30 years ago (Chidester et al., 1991). The objective of the research was to isolate groups of pilots according to personality dimensions linked to performance and to document the limits of the impact of crew coordination training. Three different profiles were identified through the analysis of personality scale clusters.

These clusters predicted attitude change after crew coordination training. They discussed bivariate correlations between personality dimensions and performance criteria, suggesting that personality may explain variance in performance.

The data were collected during the evaluation of a recurring US Air Force Military Airlift Command in Crew Resource Management (CRM) training program with 540 officers asked to complete the Cockpit Management Attitudes Questionnaire (CMAQ), a subset of the Extended Personal Attributes Questionnaire (EPAQ) (Spence et al., 1979) and the Work and Family Orientation Questionnaire (WOFO) (R. L. Helmreich, 1978). Cluster validity had been run and research appeared to have identified that membership in these clusters has consequences for training effectiveness. Clusters are reliable since they can be replicated. They can predict responses to CRM training, and differ across personality dimensions.

Successful flight requires not only flight skills, but also the ability to work well as a team. In the article by (Hedge et al., 2000), a project developed and validated Crew Resource Management (CRM) skills test for Air Force transport pilots. The CRM emphasized the importance of certain knowledge, skills and abilities (KSA) related to the crew rather than continuing to measure personality or related non-cognitive traits using the more traditional personality inventory methodology.

The Situational Test of Aircrew Response Styles (STARS) is a CRM skills test designed to measure problem-solving, decision-making, knowledge, communication, crew management, and interpersonal effectiveness attributes most critical for success as an Air Force aircraft commander. The STARS involve four primary steps: situation generation, response options generation, item reviews, and response option scaling. The results of the validation study showed a significant relation between performance on the CRM skills test and aircraft commander job performance.

Nearly 25 years ago, Richard S. Jensen, who was editor-in-chief of The International Journal of Aviation Psychology for 7 years, discussed in an editorial (Jensen, 1997) on the human factor, crew resource management and aviation decision-making in the field of aviation psychology. This editorial highlighted that aviation psychologists were thinking about what the crew should look like, that it was necessary to have an in-depth knowledge of people as perceived by cockpit designers and underline the importance of selecting and training pilots to the level of expertise expected by the airlines. In order to determine what makes a pilot an expert and what would be required to develop a training program that brings the competent pilot to the expert level, the author offered us a review of the language around human factors and aviation psychology. Research conducted at Ohio State University tells us that there are five main components in an expert aviator's decision-making. These five components are Experience, Risk Management, Dynamic Problem-Solving, Crew Resource Management, and Attention Control.

There are few studies that aggregate studies examining a pilot's personality and its impact on performance. (Maria E. Chaparro, 2020)) presents a review of the literature on pilot personality traits, and a summary of the trends in pilot personality traits in the different categories of commercial, student and military pilots. Their review of the literature was conducted to identify research that used personality inventories to examine the personality traits of pilots. The Five-Factor Model (FFM) was chosen as the benchmark against which to aggregate the results associated with the range of personality measures used to assess the personality of pilots. The five traits that are represented in the FFM are neuroticism, extroversion, openness to experience, agreeableness and awareness. From the results of the studies in the database, a map was created for each of the benchmark personality factors. In this mapping, the results of each of the studies were categorized as indicating that the pilots scored higher, lower or equal to a comparison population. The research questions attempted to answer whether the pilot's personality traits are different from those of the general population, whether there are differences in the personality traits of commercial pilots, the military and pilots in training, and whether there are differences in the personality traits of female and male pilots.

The effective selection and training of pilots play a fundamental role in the safety and efficiency of airlines. Aviation regulators require airlines to pay attention to the selection of pilots for their flying abilities and their ability to integrate into an organization. They also ensure that policies are in place to ensure that mental health and stability, personality and cognitive performance are factored into the process of pilot assessment and recruitment.

In a book intended to be a reference to pilot selection, (Bor et al., 2020) offers us a range of different skills and methods that can be used for pilot selection and evaluation. It aims to describe how the recruitment, selection and psychological assessment of pilots can be carried out. The authors also seek to address some key topics, such as remotely piloted aircraft, pilot retirement, and personality assessment.

The authors provide us with a reference book intended for recruiters, trainers, human factor engineers and psychologists involved in the selection of pilots. Researchers are trying to determine whether the pilot selection process adds value and ensures that the best candidates are hired. They also offer us an in-depth discussion on selection systems and their effectiveness in currently used approaches and comment on current trends in the aviation industry.

From the same author, (King, 2014) attempts to clarify the topic and the assessment of personality in aviation. The article conceptualizes pilot selection as a two-step process. The first, “Select-In” methods are psychological tests measuring traits that have been found to be desirable for job task analysis (JTA). They assess a candidate's level of knowledge, skills, abilities, and other characteristics (KSAO). The second, the “Select-Out” methods are an assessment of psychopathology used to assess psychiatric fitness. The challenges of personality disorders are explored, and an integrative approach, with a practitioner combining selection and selection methods, is discussed. Details of the Americans With Disabilities Act (ADA) with respect to psychiatric disorders are explored in the article. Personality disorders are treated in much the same way in the US civil and military aviation, with the elimination of aviation duties due to safety disruptive behaviour.

Examining participation in complex professions, such as aviation, requires specialized aptitude tests. A job task analysis (JTA) is first performed to determine the desirable KSAOs for the specific occupation. Then, measurement methods are identified, or, if they do not already exist, are developed to assess the KSAOs. Some of the desirable qualities identified by a JTA may be personality traits. These personality characteristics are increasingly analyzed in terms of emotional stability, extraversion, openness to experience, conscientiousness, and agreeableness. These five characteristics form the Big Five of the organization of the personality (Tupes & Christal, 1961)

Psychological testing can be used as a screening method to determine whether or not a more comprehensive psychological or psychiatric examination is warranted. The authors provide, for example, the case of the assessment of psychopathology in candidate air traffic control specialists (ATCS). This case provides an example of a selection procedure in the selection of aeronautical personnel. The FAA established the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) (J. N. Butcher et al., 1989) as the clinical cut-off limit triggering a comprehensive assessment of mental health at the 95th percentile of the ATCS population. The MMPI-2 is not really a personality test. Rather, it should be viewed as a measure of psychopathology.

MMPI-2-RF is a personality questionnaire for diagnostic, descriptive and therapeutic purposes: it identifies the psychological dynamics of the subject, such as psychopathological disorders and personality disorders to plan treatment and appropriate care.

The Minnesota Multiphasic Personality Inventory (MMPI) and MMPI-2 have been widely used in programs to select personnel for positions that require good psychological adjustment and responsibility, such as police, firefighters, air traffic controller and airline flight crews.

The objectives of a study presented by (James N. Butcher, 1994) were to provide descriptive information on the use of MMPI-2 in the psychological assessment of an airline pilot and to examine the effects of the new MMPI-2 standards on airline pilot profiles. They were also careful to examine the difference in the responses of airline pilots compared with other non-clinical subjects taking the MMPI-2 in a different setting. They examined the effects of the new standards on MMPI-2 validity and clinical scale scores and the effects of defensive testing on the MMPI-2 scores of candidate pilots. They also presented descriptive information on the MMPI-2 scales and the results of a factor analysis of the MMPI-2 scores of the pilot candidates compared with those of the normative MMPI-2 sample.

The results showed that the MMPI-2 standards are probably more suitable for characterizing nonclinical candidate groups than the original MMPI standards. The different factor structure of the profiles of airline pilot candidates compared to those of other normal individuals suggests that the different response pattern must be considered in interpreting the profiles. Defensive character and impulsiveness are likely to be significant considerations in the MMPI-2 scores of airline pilot candidates.

In (Causse et al., 2011), the authors examined the relationship between pilot’s cognitive status, personality traits, and experience with regard to their flying performance to determine which criteria are the most predictive of the flying performance. They focused on the three low-level executive functions (EFs), shifting, inhibition, and updating, and on linking their efficiency to flight performance and decisions-making during the landing.

Twenty-four pilots underwent neuropsychological tests. The cognitive assessment encompassed the three basic executive functions (Miyake et al., 2000), reasoning, and psychomotor velocity. The flight scenario was set up in cooperation with flight instructors. The personal characteristics were age, flight experience, and levels of impulsivity.

The flight performance assessment was founded on the flight path deviations (FPD), a widely used indicator of the primary flight performance. Multivariate regression was used to determine the influence of the independent variables on FPD. The ability of independent variables to predict the performance was tested by an all-possible-subset regression. Pearson correlations were computed among all considered independent variables and between the FPD and the independent variables. Reasoning, updating in working memory, and flight experience was predictive of the flight performance.

In (Barron et al., 2017), 3,140 USAF-manned aircraft pilots and 330 Remotely Piloted Aircraft (RPA) pilots were surveyed to determine whether pilot selection metrics are predictive of long-term pilot performance. Flight performance was documented in an Officer Performance Reports (OPR) after completion of flight training. Participants completed the Air Force Officer Qualification Test (AFOQT) consisting of 11 cognitive tests and a five-factor model (FFM) measure.

The analysis showed that neuroticism and introversion are significantly related to poor pilot training outcomes. The cognitive skills, knowledge, and personality traits that predict the professional performance of USAF-piloted aircraft pilots are also predictive of the professional performance of RPA pilots. The subtests that contribute to the AFOQT pilot composite were significant predictors of early career job performance for both manned aircraft and RPA pilots. The results demonstrate that predictive relationships were observed for manned aircraft and stationed RPA pilots.

7.2.3 **Summary**

In the next section, this paper will explain the methodology of the research. It will describe the basic profile of a typical cadet, the flight training assessment process, the data used in the testing, as well as the machine-learning algorithm that provided us with the clusters used in the profiling analysis. We then used a statistical method to normalize the grades of the enhance the robustness of the solution.

The third chapter will present the results of the data analytics. It will provide explanations for the algorithms and the distribution of flight training performance against psychometric profiles.

The paper will conclude with future work that can be pursued in order to increase the impact on cadet selection and flight training support.

7.3 Methodology

In this section, we will present the methodology of the testing. We developed artificial intelligence capability to classify cadet candidates based on desire/preferences to be a pilot (known or unknown) and learning aptitudes.

We describe the data-acquisition process, looking at the historical assessment data and subsequent grading from their flight training data and pilot psychometric testing, which are used as the main source for AI modeling.

The analytic data set is used for clustering algorithm selection and assigning data (e.g., cadets) to groups with common features, for example, condensing hundreds or thousands of assessments per cadet into a single profile that captures the key attributes that predict actions required for success, using an autoencoder.

7.3.1 Cadet Profile

To enrich the individual profiles, a general portrait of the cadet is first provided below using a general questionnaire, allowing to produce the persona of a cadet.

- **Comfortable and open to new technologies:** They are young and constantly use digital products in their daily lives. Most of them do not have experience using an e-learning platform, but they use other media platforms for their learning needs. They are open to using new technologies.

- **Resourceful when looking for information:** Cadets have the reflex of going on the Internet when they need any information, for example, they will try searching for videos on YouTube to better understand a subject.
- **Open to adapting:** They are ready to adapt their learning style to a different learning strategy.
- **Want to be treated as a client:** They pay a large sum of money to be in the program at CAE. However, CAE is not the only training school they reviewed before making their selection, so they consider themselves as clients and have high expectations.
- **No aviation experience:** The vast majority of cadets have no aviation experience and have never seen a cockpit in their life.
- **Need to feel supported:** Many have left the family home for the first time, and some are studying very far from their families and friends. Their experience with independence is limited, and they rely on community support to help them through their journey. This is why they will often get together to study.
- **Diverse cultural profile:** The cadet population has a very diverse cultural profile, because students come from different countries around the world. Many of the cadets have never been in contact with so many different cultures.
- **Happy to share their story:** They are very generous when the time comes to share their experience as a cadet, which can be useful for future cadets.

7.3.2 Flight Training & Assessment Process

Cadets first have to successfully go through Ground School, covering all the theoretical elements relating to aviation. Successful completion of 14 exams at the end of this stage enables cadets to move on to the Flight Training stage. At the beginning of the training, students start a course, and then start a lesson. After each flight/simulator session, the instructor assesses the lesson and awards a grade (unit grade: Unsatisfactory (U), Satisfactory (s) or Incomplete (I)).

Line items are graded using absolute grading. Instructor grades are predefined line items within the lessons between 1 and 4. Expected proficiency is defined for each line item under a lesson by a group of trainers who build the syllabus in order to meet the requirements of national authorities, as well as the requirements of the customer airline.

If the line item is not meeting expectations, the instructor is obliged to select a U/M Reason Code that corresponds to the most problematic area for the line-item performance. The instructor also adds an additional comment about the area to improve, and then completes the assessment. The student repeats the lesson with the remaining incomplete or failed line items. The instructor may ask students to repeat line items if assessed to be beneficial.

7.3.3 Data Description

There are two main data sets:

- Psychometric test data from Symbiotics
- FTO Performance data

The following courses were included in the analysis:

- “CASA PPL” – only one phase in a Pilot's training (the first certificate out of 3)
- “17 PPL” – only one phase in a Pilot's training (the first certificate out of 3)
- “AKMPL-Core1 v1.1” – competency-based course
- “AKMPL-Core2 v1.0” - competency-based course
- “SWA CPC” - competency-based course
- “SWA CPC 2.0” - competency-based course
- “MPL BAS FLT 1.0” – AirAsia Basic course, KL—competency-based course

7.3.4 **Data Acquisition Process**

Using a rigorous data privacy compliancy process, the testing used data from a database with over 6,000 records from pilot psychometric testing, and over 66,000 records from pilots' flight training records. After pre-processing, which selected & unselected pilots from both databases, a clean analytic dataset is created of 618 anonymous students. After selection of relevant courses for this study, a data set of 188 students will be consolidated.

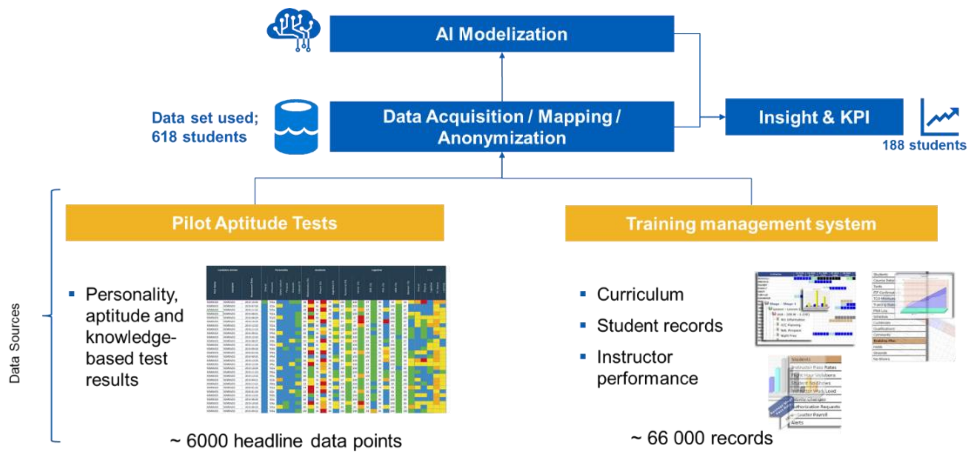


Figure 7.1 - Data flow and anonymization process.

7.3.4.1 Flight Training Data Set

The Flight Training Dataset is a collection of records of training curricula and student performance, including students course location, definition, objectives, associated standard at various stages, student historical performance, and expected performance through various courses, lessons and line items, instructor comments and grading.

Each course is divided into stages, which have lessons. Each lesson is further divided into units, which have multiple-line items within each unit. As the students complete these line items in each course, the associated grades are recorded. The line-item records of the students who have taken the psychometric tests are extracted and enhanced with extra information from other estimated time of arrival (ETA) data sets for the purpose of our analysis.

An overall description of the performance data of the students taken psychometric test is tabulated as below.

Table 7.1:. Data Statistics

Entity	No. Of Records
Courses	7
Stages	11
Units	306
Line Items	1,106
Students	271
Instructors	56
Students Registrations	358

The flight performance assessment data model of the analysis is captured in the Entity-Relationship Diagram (ERD) below.

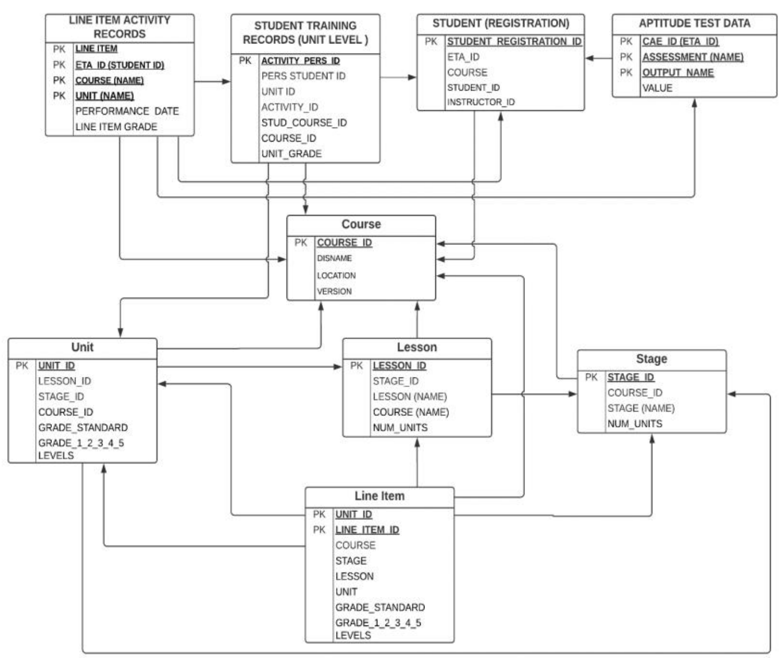


Figure 7.2 - Flight Performance Assessment Entity-Relationship Diagram

The distribution of the students who had taken psychometric tests spanning across the courses are depicted in the diagram below.

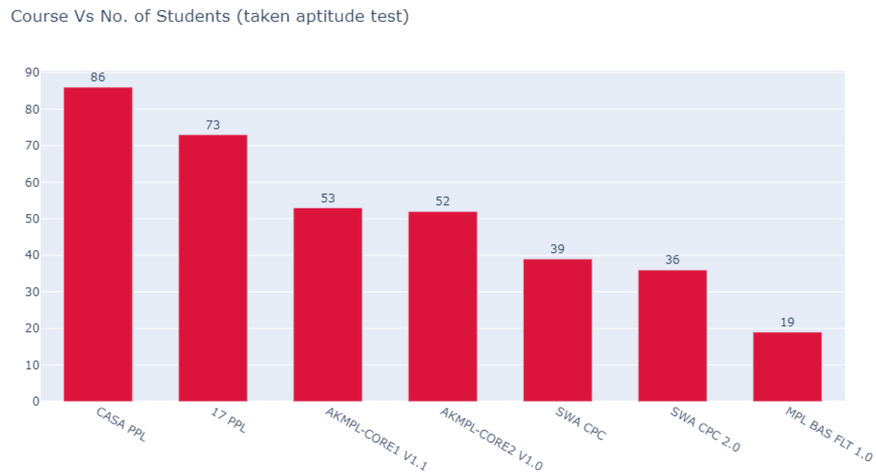


Figure 7.3 - Student count over courses

Definition of line-item grading: Line items are graded using absolute grading. Whether a student is under, on, or above performance is defined based on the line-item expected proficiency grade. Expected proficiency is defined for each line item in a lesson by a group of trainers who build the syllabus to meet the requirements of national authorities as well as the requirements of the customer airline.

7.3.4.2 Psychometric Data

Symbiotics Dataset consists of records of initial cadet applicant psychometric screening.

Collection of assessment data.

The assessment data is collected as part of the application process for potential cadets who wish to enrol in the CAE training programs. They are required to complete a battery of aptitude, preference and skills tests provided by Symbiotics Ltd. via their ADAPT platform. The actual tests included vary from one school to another, very often dependent on the objectives of any client airline associated with the program. However, most cadets complete the APQ (personality questionnaire) and the Cognitive Reasoning test. This is often supported by tests in Math, Physics and English Language. In addition, psychomotor skills are tested through FAST, Fixed Wing and Ball Game. Descriptors of these tests have been appended.

To give candidates access to the test platform, each one is sent a unique 20-digit licence code and login instructions for the ADAPT platform. All candidates create a user account, setting their own username and password, which enables them to log in to complete the tests in their own time. As part of the registration process, they double-accept the GDPR statements covering the data collection and use of stored data. Data are retained within the system for up to 24 months, before being fully anonymized, and raw scores are then archived in the data pool, where they are retained for future analysis purposes.

Once candidates have completed their account, they find that the tests for which they have been registered are available for completion. Some training centres require candidates to take all the tests in a supervised setting, in which case they usually issue the licence codes to the candidates once they are present on site, and they take the students through the tests under invigilation and in the same program order. Other centres wish to minimize the time for which students are present on site, and permit some of the tests, normally the self-reporting questionnaires, to be completed by the candidates at home prior to attending the centre. For these programs, the licence code only allows access to the permitted tests until the candidate is on site, and the supervisor unlocks the remaining tests. The psychomotor skills tests, Fixed Wing and Ball Game, can only be taken using additional hardware comprising a joystick and pedals, and therefore these tests have to be taken on site. FAST can be completed using only the keyboard, and many centres allow this to be taken at home.

Tests with correct/incorrect answers are always timed to ensure that the test is standardized, and results can be compared between candidates. Questionnaires about style or preference, such as personality questionnaires, are untimed and do not have right/wrong answers. Candidates are given clear guidance at the start of each test regarding the style of the test, the time allowed if limited, and the process of answering questions, in order to ensure as much as possible that candidates have fully understood test requirements before they start the test. Once started, any test in ADAPT must be completed within one sitting. However, candidates with multiple tests can take these individually if they wish, rather than sitting down and trying to complete all the tests “in one go”.

If candidates are taking the tests in a supervised environment, they should be given breaks between tests.

During the COVID-19 pandemic, many centres have been unable to have candidates attend on site for assessments, and ADAPT has been developed to offer the capacity for remote invigilation (remote proctoring) of multiple candidates simultaneously by one invigilator, and it is anticipated that this facility will continue to be used once normal processes have resumed, since it permits supervised assessment without requiring candidates to travel to a central point with all the associated expenses and time. Fixed Wing and Ball Game are then kept as a final-stage assessment, used only by candidates who are deemed to have met the required criteria through the other test elements.

Once all the candidates in a specified cohort have finished the required tests, either on site or remotely, the system collates their results into an integrated report for each individual. The school is also able to access a matrix of results showing the high-level scores of key attributes for all the candidates, allowing a quick comparison across the cohort. Results are reported as sten scores for each trait, with candidates compared with the norm group of pilot applicants. ADAPT also applies a colour banding overlay, with five colours—Red, Amber, Yellow, Blue, Green—showing the degree of fit for each candidate to the desired level of relevant traits. Centre administrators at each location have individual logins to a portal for retrieving reports and matrices at any time.

ADAPT tests typically used in cadet selection assessments

Personality Questionnaire

The Symbiotics' Personality Questionnaire (APQ) is widely used within the aviation sector, and includes traits that are specific to this sector, such as Hazardous Attitudes, in addition to the general traits of the Big Five Factor personality measures.

Derived from a theoretical basis to support the measurement of these traits, the APQ has proven predictive validity when used as part of a battery of appropriate tests.

The cultural measures within APQ are derived from Hofstede and can be adjusted to the needs of the particular country or region, to recognize necessary diversity.

Knowledge-based tests

Math Progressive: These measures pure mathematical capability, looking at the user's ability to apply mathematical knowledge, formulae and calculations. They are aligned to an 18+ level. Different levels of test complexities are available depending on the client's needs.

Physics Progressive: This test measures physics ability and increases in difficulty through 3 levels—Foundation, Intermediate and Advanced. The test is aligned to an 18+ level.

English Language: This test assesses English Language, Listening and Comprehension skills across six (6) key areas. When used for aviation candidates, it is designed to demonstrate a minimum ICAO level 4 proficiency.

FAST

These tests a candidate's multitasking, learning, cognitive, situational awareness and physical skills.

Designed by our psychologists, FAST takes around 15 minutes to complete and places candidates in an ever-increasingly stressful situation to determine how well they can function as the number and intensity of overlapping tasks increase.

The test determines how well a candidate can prioritize tasks, manage multiple activities and learn from each flight scenario to improve their overall performance.

Excellent at filtering a large number of candidates down to identify those with natural aptitude for ab-initio training.

Fixed Wing

The high cost of full-flight simulators makes it impracticable to use them in the assessment of large numbers of individuals who are applying to become pilots, many of whom do not have the personality traits, knowledge, skills or aptitudes to achieve their ambitions.

Our Situational Judgment tests use simulation software either fixed wing or rotary wing and take 30 minutes. They have proved to be a cost-effective environment for predicting the potential for individuals to succeed as pilots. The candidate is required to fly a route provided by a recorded mission, which is debriefed afterward. A joystick and throttle are required as well as the ability to listen to audio.

Aptitude tests

Symbiotics offer a range of aptitude tests that are designed to help measure an individual's natural talents and cognitive abilities for doing and learning to do. This sort of cognitive testing process can help a business to measure work-related cognitive capacity and a candidate's suitability for a role or training. Current tests include: cognitive reasoning, logical reasoning, mechanical reasoning, error checking.

It should be noted that all cases used in this study had completed the screening tests and met the minimum levels required for entry into the training programs.

7.3.5 Data Analysis Strategy

7.3.5.1 Performance Labelling

The line-item grades across the units are a way of bucketing the students into three performance categories based on their performance in each unit. Students are ranked on their overall performance in each course across the units; Top 10% (of the ranks) are taken as excellent, bottom 10% as poor, and the remaining as average performers.

This approach is used to map the knowledge of subject matter experts in aviation training pertaining to performance using statistical methods so as to scale this approach easily across courses without specific knowledge of them.

7.3.5.2 Clustering

The students are clustered into groups/profiles based on the aptitude test scores on various components (outputs) of assessments. Various clustering methods have been studied and assessed on the data, and lastly, agglomerative hierarchical clustering has been chosen. The primary reason for choosing this specific method was the ability to select a convincing distance threshold after observing a dendrogram resulting in four clusters and the 2D representation, as well as a heatmap of these clusters based on some patterns. The density-based clustering methods (DBSCAN, OPTICS, etc.) performed poorly on the data; though these algorithms performed on reduced dimensions, dimensionality reduction is not used since it takes the ability of cluster explainability. Methods like the Gaussian Mixture Model and K-Means were not adopted because of the requirement to input the number of clusters beforehand, prohibiting the natural formation of clusters.

Algorithm 1: Hierarchical Agglomerative Clustering

Input:	$\{X_n\}_{n=1}^{N}$ The distance threshold DT (After plotting the complete dendrogram, distance threshold has been chosen as 45 (in Ward's Linkage))	X_i is the psychometric test vector of ith student. (Students have taken different aptitude tests; Only the common test components are considered in the vector (142 components) for analysis.)
Output:		
Initialize:	Cluster Groups $\mathbf{\Omega}$ $\mathbf{\Omega} \leftarrow \emptyset$ DT ← 45 DMIN ← 0	 Initial cluster set starts as Null Based on dendrogram DMIN is the minimum distance between any two clusters

Step 1:	for n ← 1 to N: do
	$\Omega \leftarrow \Omega \cup \{\{X_n\}\}$
	end for
		Each data point is assigned as its own cluster

Step 2:	while DMIN <= DT do
	$C_1^*, C_2^* = \underset{C_1, C_2 \in \Omega}{\text{argmin}}\, DIST(C_1, C_2)$
	DMIN ← $DIST\,(C_1^*, C_2^*)$
	If DMIN > DT break, else
	$\Omega \leftarrow \Omega - \{C_1^*\} - \{C_2^*\}$
	$\Omega \leftarrow \Omega \cup \{C_1^* U C_2^*\}$
	end while
	Return $\mathbf{\Omega}$
		Choose the pair in $\mathbf{\Omega}$ with the closest distance
		Update DMIN as the current minimum distance among any of the clusters in the set
		Remove C_1^*, C_2^* from the cluster set
		Update the cluster set with newly formed cluster

7.3.5.3 Explainability & Correlations

Four clusters are formed based on the clustering method. A supervised algorithm is trained with the clustering labels, and SHAP is used to interpret the clusters resulting in four defined characteristic groups. Macro-level correlations are identified between the cluster groups and performance

7.4 Complete framework & experimental results

7.4.1 Performance Labelling Methodology

Given notions of performance being multi-dimensional and complex, a statistical approach has been adopted to scale the method to multiple courses without specific knowledge expertise. Some key notions of performance related to the grading being used are as follows:

1. The number of failed line items in a unit roughly measures the performance. However, it has to be noted that failing many line items in the initial units of a course, versus the later units, is expected as the grading of line items are absolute, and it is based on the instructor's judgment and likely to be biased.
2. Units are of varied difficulty. While failing 10 line-items in one unit is considered poor performance, the same could be considered expected in some other units. This is typically observed in some "staging units" where instructors have to suddenly check the overall skill level of a student to allow the student to pass that stage, because of some strict requirements on skill sets; many students are expected to fail relatively high number of line-items in these units.
3. Poor performance in a unit by failing many line items cannot be considered as overall poor performance in the course; this means the poor performance in that specific unit
4. Some students would not have taken all the units in a course, due to being transferred from one course to another.

We also identified other key performance metrics along with the line-item failure counts to measure the performance. They are as follows:

- Probability of passing a unit.
- The percentage of time students falls in top, average or bottom ranks in unit level performance across all the units in the course.
- Average number of times (per unit) the student is graded with excellent performance (Consistency Rate).
- The average unit takes (attempts to complete the unit) (the lower the better).

Based on the metrics, each student's performance in each unit for all the courses under consideration were gauged as "top", "average" and "bottom" based on the number of failed line items in those units. The top 20% (those with the least number of fail counts in the unit) would be tagged as "top" in that unit. Similarly, bottom 20% as "bottom" and the remaining as "average". The probability of being a "top", "average", "bottom" performer in a unit is calculated for each of the student in all the courses under consideration as:

$$P(student, tag, course) = \frac{Unit\ Count_{student,tag,course}}{Unit\ Count_{student,course}}$$

Here, $Unit\ Count_{student,course}$ is the total number of units student has taken in the course; while $Unit\ Count_{student,tag,course}$ is the total number of units student has taken in the course and belongs to that specific tag ("top", "average", "bottom")

Passing probability of a unit, Consistency Rate and Average attempts are calculated as follows

$$P_{pass}(student, course) = \frac{No.of\ Units\ passed_{student,course}}{No.of\ Units\ taken_{student,course}}$$

$$\text{Consistency Rate}\ (student, course) = \frac{No.of\ line\ times\ graded\ high\ _{student,course}}{No.of\ Units\ taken_{student,course}}$$

$$\text{Average Attempts}\ (student, course) = \frac{No.of\ unit\ takes/attempts\ _{student,course}}{Unique\ No.of\ Units_{student,course}}$$

Students are sorted (in descending order) based on the metrics below in decreasing order of priority

1. $P_{pass}(student, course)$ – Ascending order
2. $P(student, "top", course)$ – Ascending order
3. Consistency Rate $(student, course)$ – Ascending order
4. $P(student, "average", course)$ – Ascending order
5. Average Attempts $(student, course)$ – Descending order
6. $P(student, "bottom", course)$ – Descending order

If two students obtain all the same values for all metrics, they would obtain the same rank. After ranking the students in each course, the ones in the top 10% of the ranks (≤ 0.1 quantile of ranks) are marked as “Above average”, the ones in the bottom 10% of the ranks (> 0.9 quantile of ranks) are marked as “Below average”, and the remaining are marked as “Average”.

Given we are also interested in the temporal aspect of performance (How are students progressing over time?), we have also divided each course into two stages (sections); stage1 includes roughly first half of the course, and stage2 includes the second half and the student performance labels are drawn using the same method explained here, considering only the units that are part of the specific stage. It has to be noted that some students would not have attended any units in a particular stage; these instances are marked “NA” indicating the non-availability of performance measures for the students at that stage.

After performance labelling, the distribution of labels across the courses under consideration in overall and temporal view can be summarized as follows.

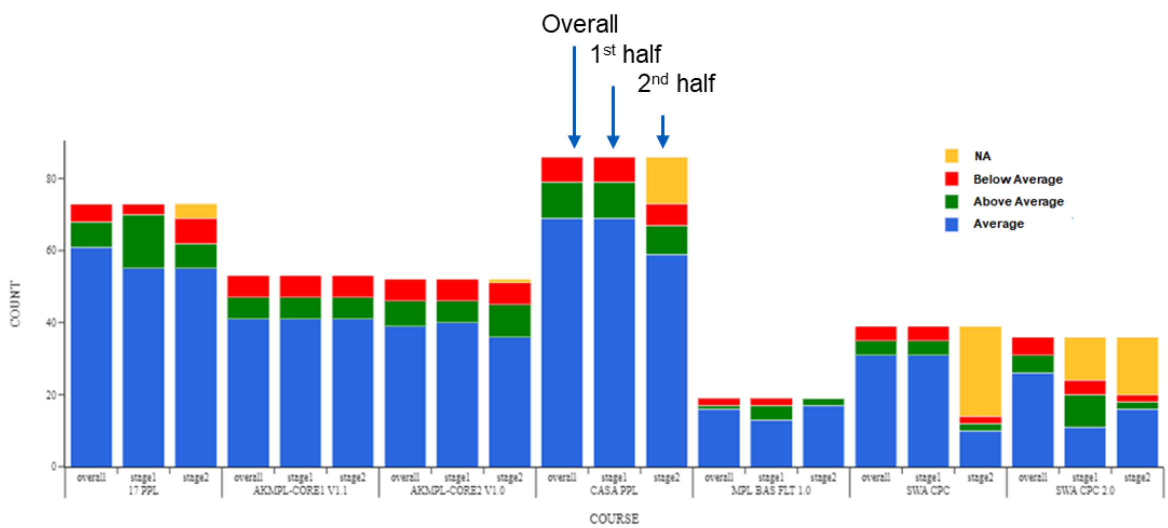


Figure 7.4 - Student performance label distribution across courses

7.4.2 Clustering

7.4.2.1 Clustering Visualization

The representation of psychometric test data after reducing the dimensions to 2D for visual representation is below. Uniform Manifold Approximation and Projection (UMAP) (Leland McInnes, 2018) for Dimension Reduction has been used to portray the data in 2D for its ability to capture non-linear relationship versus techniques like PCA. Components are values that correspond to a compression of 142 parameters in the aptitude test results. From the high-dimensional space, we are keeping data properties in order to evaluate distance between psychometric profiles.

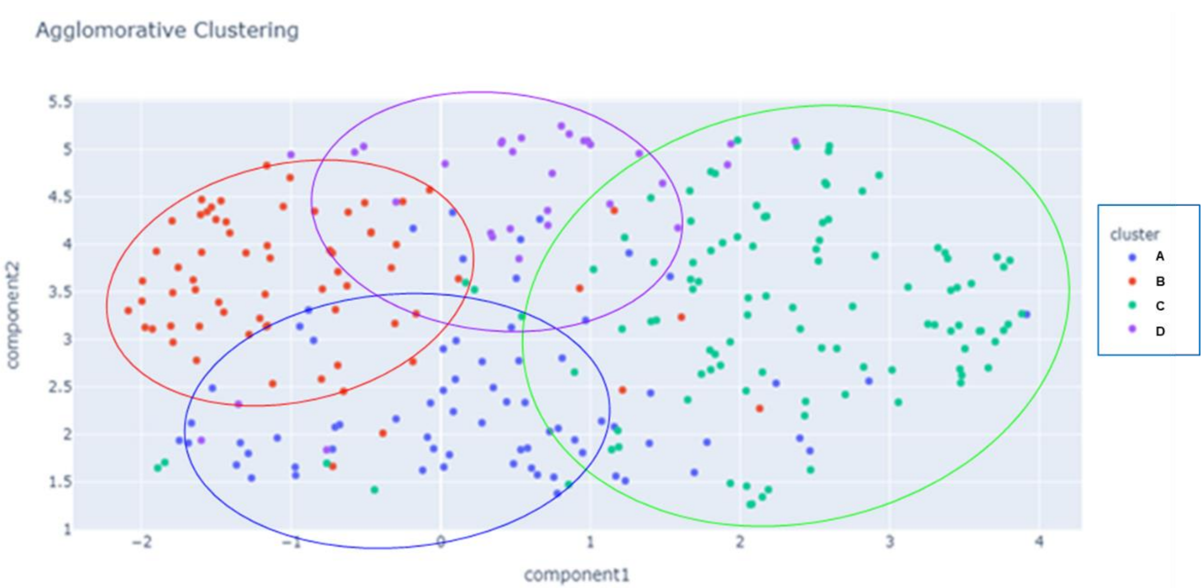


Figure 7.5 - Cluster formation using agglomerative hierarchical clustering

The pros and cons using different clustering methods are tabulated as follows. The choice of agglomerative hierarchical clustering is based on the cluster formation, interpretability and convenience.

Table 7.2:. Clustering Method Benchmarking

Clustering Method	Pros	Cons
Affinity Propagation	Less computational complexity Good for large datasets Not required to select the number of clusters upfront No iteration in matrix implementation	It tends to choose exemplars based on distance; might not be suitable in varying data densities. Though 6 clusters are formed, the separation in 2D space is not exclusive.
Agglomerative Clustering	***In each step, the most similar clusters are clubbed together until there is a single cluster.***	***Not suitable for a huge data set (Not our case). Dendrogram becomes unusable for huge data.***
BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)	Can cluster large datasets based on a smaller, meaningful summary. Useful for big data.	Can only process metric attributes While BIRCH is scalable as complexity increases only linearly, it is sensitive to the order of data insertion. The fixed size of leaf nodes results in unnatural clusters.
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	DBSCAN is computationally efficient and separates outliers. Can cluster arbitrary shapes.	The DBSCAN clustering results in many data points not part of the cluster. It is highly sensitive to parameters MinPts and Eps. The method failed to form any clusters in higher dimensions. DBSCAN performed on lower dimensions also has outliers (Not suitable for us as dimensionality reduction takes away explainability)
K-Means	K-Means divides the data into groups of equal variances, trying to minimize inertia or within-cluster sum-of-squares.	It depends largely on initial cluster centres; k-means++ leading to better results than random initialization. Inertia doesn’t work well with large dimensions. K-means after dimensionality reduction proved better in many cases. (Not suitable for us as dimensionality reduction takes away explainability.)

Table 7.3:. Clustering Method Benchmarking (cont'd)

Clustering Method	**Pros**	**Cons**
Mini-Batch K-Means	This method has shown similar results to the standard K-Means with the performance advantage on working with a smaller data set.	Similar to K-Means
Mean Shift	Mean shift can capture complex shaped clusters	Convergence in high-dimensional space is arguable. Bandwidth is a parameter of kernel function depicting its standard deviation. This influences the modes and the cluster. It resulted in many outliers along with a single cluster.
OPTICS (Ordering Points To Identify the Clustering Structure)	Good for automatic and interactive alternative to DBSCAN cluster analysis	Just like DBSCAN, OPTICS also do not keep many data points within any cluster, in higher dimensions. OPTICS performed on lower dimensions also has outliers (Not suitable for us)
Spectral Clustering	Makes no assumption in the shape of the data. Not being an iterative method is computationally viable. Spectral clustering with KNN (neighbours=4) seemed to produce comparable results with Agglomerative clustering.	Internally, after converting to lower dimensions, the specific implementation of Spectral clustering uses KNN; Hence need to specify the no. Of clusters
Gaussian Mixture Model & EM	GMM clustering assumes data is from different gaussian mixtures. It is a very useful for non-circular shapes (on contrary to k-means). GMM (gaussian functions=4) also seemed to produce comparable results with Agglomerative clustering.	Needs to mention the no. Of clusters (Gaussian Mixtures)

1.1.1. Dendrogram

Number of clusters are not arbitrary, but rather are calculated using a dendrogram. The dendrogram shows the hierarchical relationship and the order of cluster formation

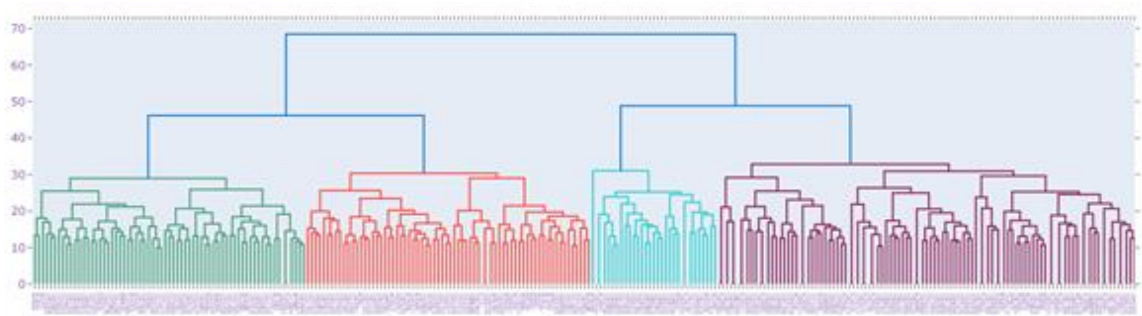


Figure 7.6 - Dendrogram of agglomerative hierarchical clustering

1.2. Explainability

A classifier has been trained on psychometric test scores across 142 test components as the predictor variables and the cluster labels as the response variable. The outcome of the classification is as follows:

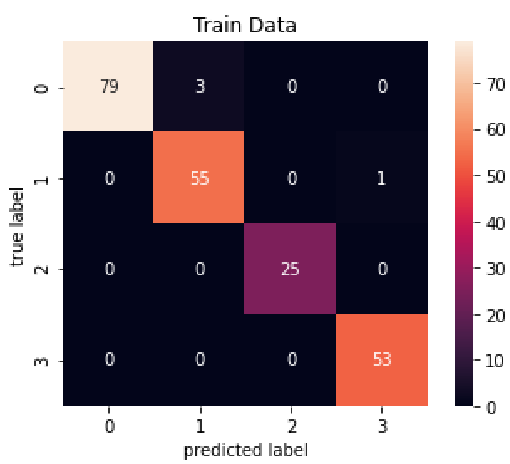


Figure 7.7 - Train Data

Figure 7.8 - Test Data

Accuracy on training set: 0.981 whereas the accuracy on test set: 0.836

The top 15 defining features of individual clusters are as follows. The colour indicates the feature value, and positive value in x axis indicates the positive association with the cluster. To interpret the cluster, the psychometric test components are analyzed with psychologists.

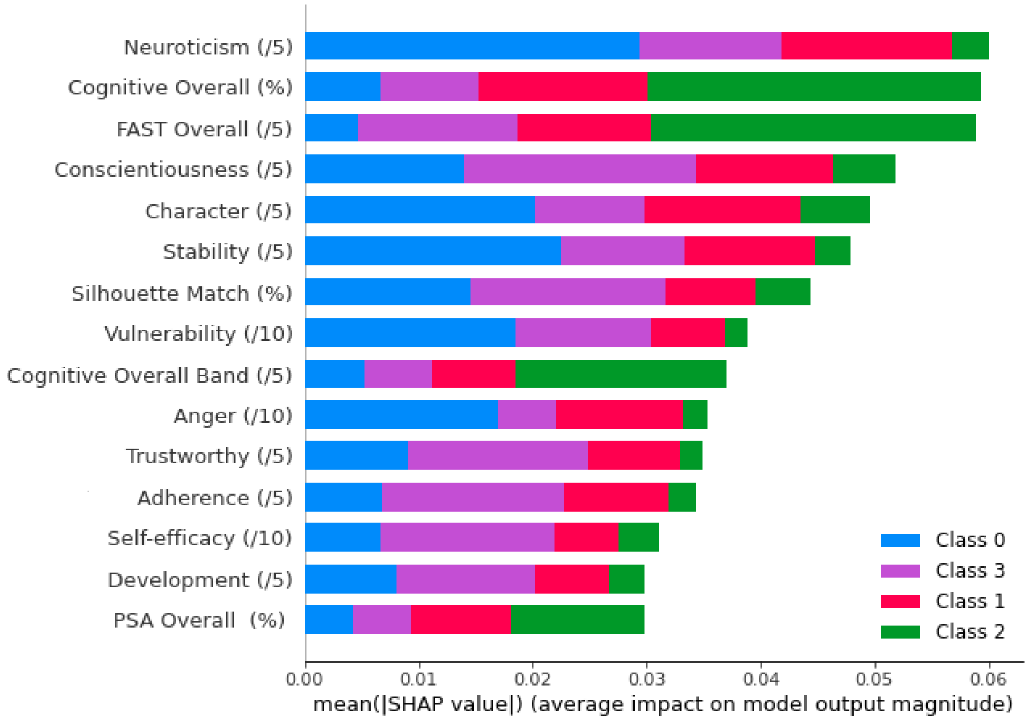


Figure 7.9 Top 15 Defining Features of clustering

7.4.2.2 Applying SHAP methodology on the clusters

Explainability in intelligence artificial is very important to gain thrust with the user. **SHAP (Shapley Additive exPlanations)** is a unified approach to interpreting model predictions based on game theory to interpret the output of any machine-learning model. SHAP identifies the class of additive feature importance methods and help users interpret the output predictions of a classifier and allows for the cluster interpretation (Scott M. Lundberg, 2017).

Cluster A can be roughly interpreted as ***neurotic, unstable, vulnerable, angry, abnormal, non-trustworthy & not a stereotypical pilot***

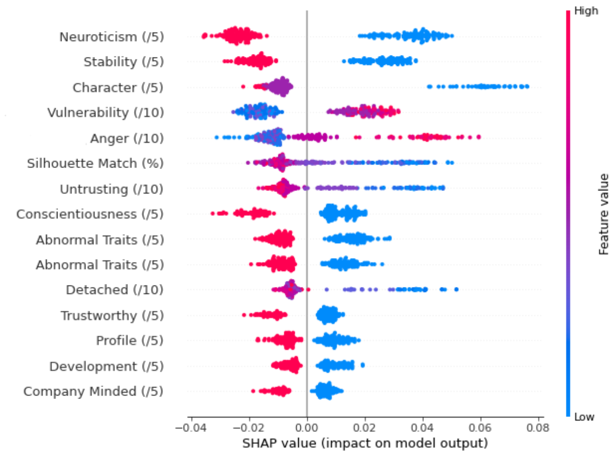


Figure 7.10 - Top 15 features of Cluster A

Cluster B can be roughly interpreted as ***neither adherent nor a stereotypical pilot; intelligent, calm, stable while neither company-minded nor trustworthy.***

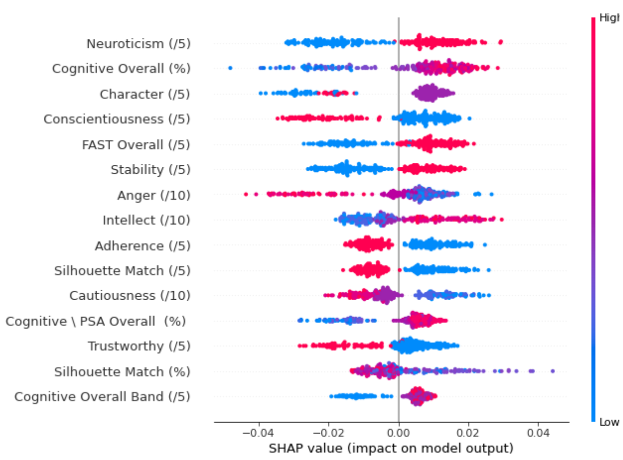


Figure 7.11 Top 15 features of Cluster B

Cluster C can be roughly interpreted as ***challenged in cognitive, speed and physical skills, while accommodating and trustworthy.***

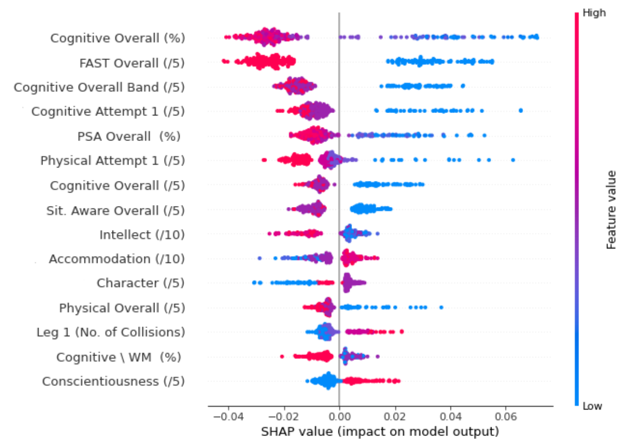


Figure 7.12 - Top 15 features of Cluster C

Cluster D can be roughly interpreted as ***self-aware, stereotypical pilot, stable, intelligent, trustworthy with scope to develop and adhering to the company values.***

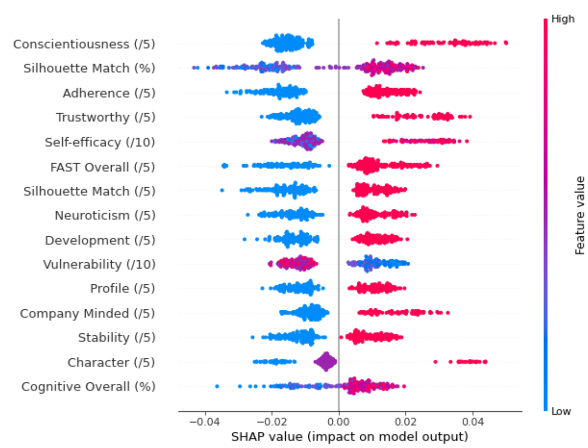


Figure 7.13 - Top 15 features of Cluster D

The detailed explanation of top 15 defining features of **cluster D** are as follows. The same explanations of cluster A, B and C have been extracted and can be provided on demand.

- ***Conscientiousness (/5)*** – Band indicating suitability across the constructs – Self-efficacy, Orderliness, Dutifulness, Achievement-Striving, Self-Discipline, Cautiousness (High values are associated; higher scores better)
- ***Silhouette Match (%)*** – Silhouette is a personality "type" profile looking at combinations of preferences and how they influence each other. This is matched against a target silhouette, which is the ideal for the role/organization (High values are associated; higher scores better)
- ***Adherence (/5)*** – Candidate's ability to commit to the organization, take on board the organization's ethos and values (High values are associated; high scores better)
- ***Trustworthy (/5)*** – The level of reliability of an individual and worthiness of +trust from others (High values are associated; high scores better)
- ***Self-efficacy (/10)*** – Understanding of own strengths and weaknesses (Comparatively High values are associated; non-linear sten scale; moderately high scores better)
- ***FAST Overall (/5)*** – Overall performance band (looks at performance across physical, cognitive, situation awareness) – (High values are associated; high scores better)
- ***Silhouette Match (/5)*** – Silhouette is a personality 'type' profile looking at combinations of preferences and how they influence each other. This is matched against a target silhouette, which is the ideal for the role/organization (High values are associated; higher scores better)
- ***Neuroticism (/5)*** – Stress coping; (High values are associated; high scores better)
- ***Development (/5)*** – The ability of the individual to develop and to "bridge the gap" between current and required skill levels (High values are associated; high scores better)
- ***Vulnerability (/10)*** – Tendency to lack resilience, become overwhelmed (Low values are associated; lower scores better)

- ***Profile (/5)*** – How close candidate silhouette profile aligns with target profile (High values are associated; high scores better)
- ***Company Minded (/5)*** – A positive attitude and willingness to align themselves with company values (High values are associated; high scores better)
- ***Stability (/5)*** – Capacity of an individual to maintain an emotional balance under stressful circumstances – Stress coping (High values are associated; high scores better)
- ***Character (/5)*** – Band indicating suitability across the six (6) constructs – Stability, Confidence, Abnormal Traits, Trustworthy, Sociability & Adherence – (High values are associated; high scores better)
- ***Cognitive Overall (%)*** – Overall cognitive performance band (high values are associated; high scores better)

The Figure 7.14 below summarizes the mean and median statistics of top 15 features of each cluster, totaling 32 distinct features.

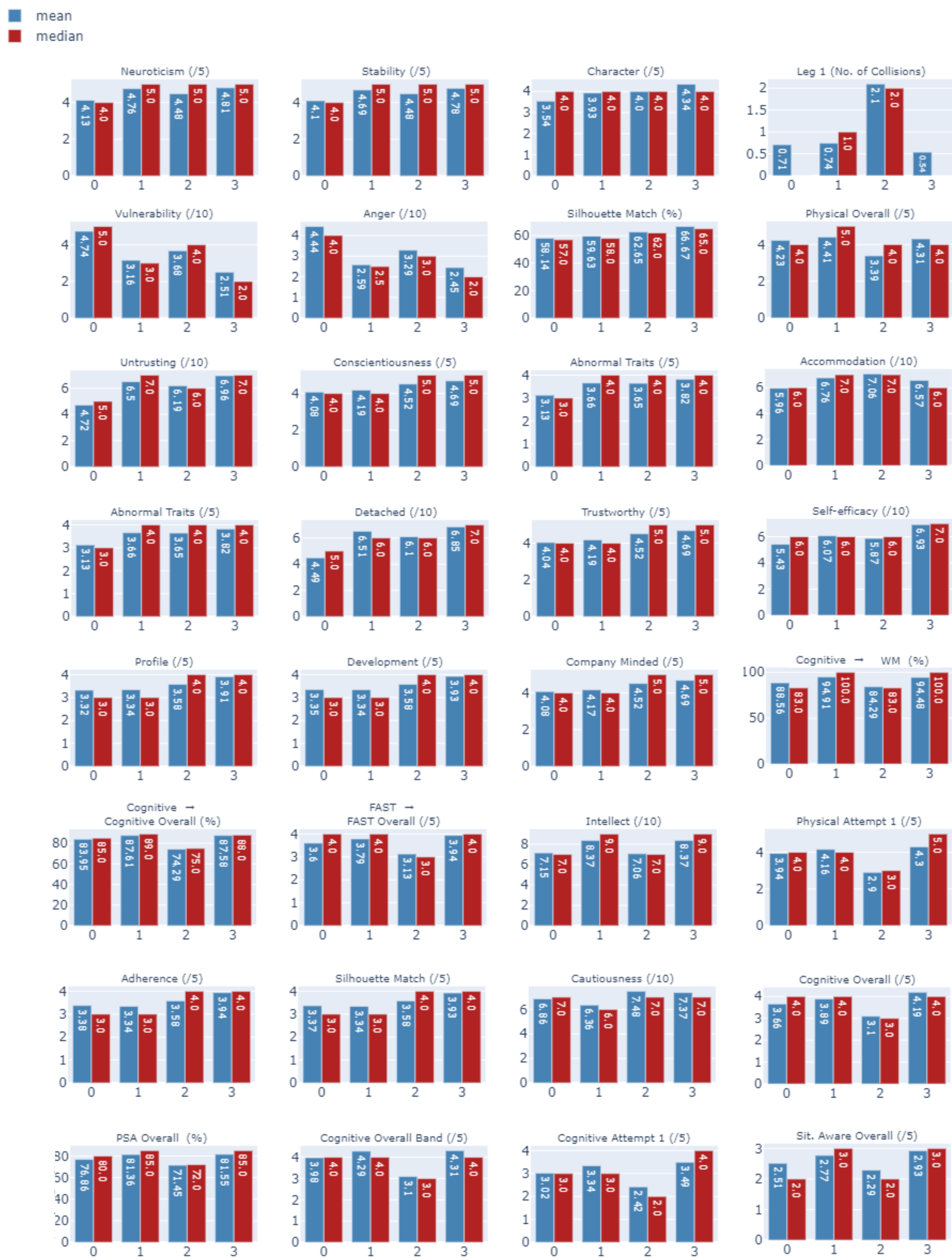


Figure 7.14 - Mean and median statistics of top features in four clusters

7.4.3 **Learner Profile**

7.4.3.1 Overall performance

The psychometric clusters are correlated with overall performance.

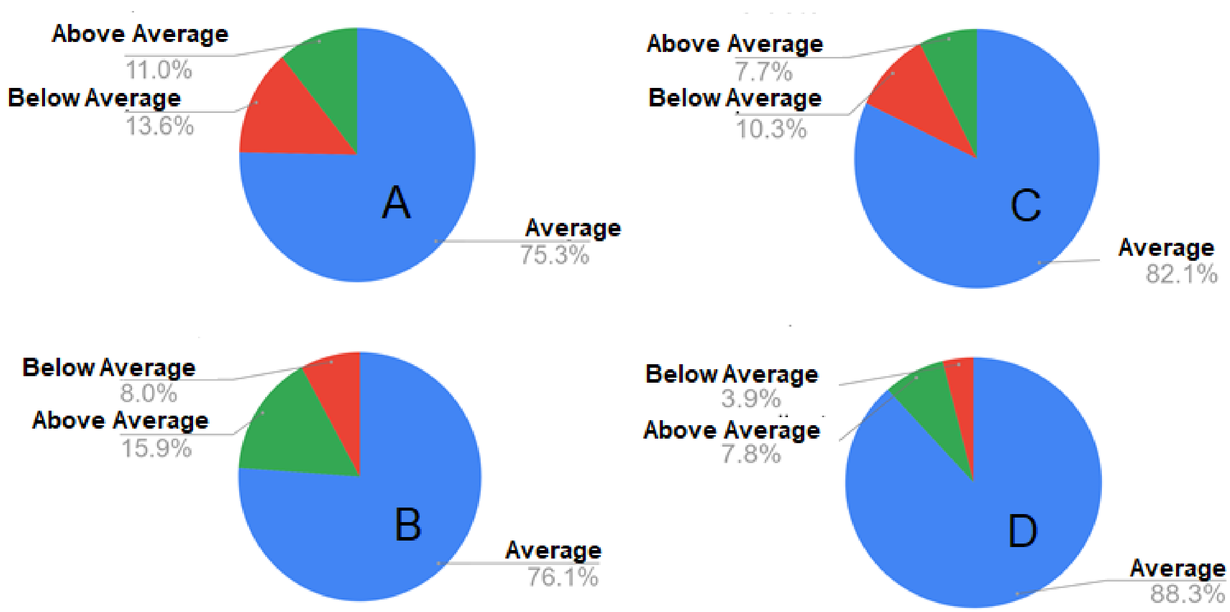


Figure 7.15 - Psychometric Clusters versus Overall Performance

- Cluster A: Neurotic, unstable, vulnerable, angry, abnormal, non-trustworthy & not a stereotypical pilot
- Cluster B: Neither adherent nor a stereotypical pilot; Intelligent, calm, stable, while neither company minded nor trustworthy.
- Cluster C: Challenged in cognitive, speed and physical skills, while accommodating and trustworthy.
- Cluster D: Self-aware, stereotypical pilot, stable, intelligent, trustworthy with scope to develop and adhering to the company values.

Cluster B has the highest percentage of Above Average performance, whereas Cluster D has the lowest percentage of Below Average performance, leading to a belief that intelligence (a characteristic for both Cluster B and Cluster D) has a role in performance, be it excellence or avoiding failure. The difference between Cluster B and Cluster D is that the former being introverted and not stereotypical and latter being extroverted and stereotypical pilot.

Cluster A has the highest percentage of poor performance even more than Cluster C, while Cluster C struggle cognitively. Cluster A has the most significant behavioural problems compared with all other clusters, leading to a belief that poor performance is correlated with attitude more than intelligence.

7.4.3.2 Temporal performance

The individual student performances in the first and second stages of the courses are compared and correlated to understand the relationship between the psychometric and performance progression.

Performance improvement has been encoded as an integer number [-2, 2], which represents the level of improvement from the former to the latter stage of the course. (The cases of students who had taken only one of the stages are ignored.)

- 2 » (Below Average performance in stage 1, Above Average performance in stage 2)
- 1 » (Below Average performance in stage 1, Average performance in stage 2), (Average performance in stage 1, Above Average performance in stage 2)
- 0 » (Below Average performance in stage 1 & in stage 2), (Average performance in stage 1 & in stage 2), (Above Average performance in stage 1 & in stage 2)
- -1 » (Average performance in stage 1, Below Average performance in stage 2), (Above Average performance in stage 1, Average performance in stage 2)
- -2 » (Above Average performance in stage 1, Below Average performance in stage 2)

Psychometric cluster	improvement	student count	perc count
A	0	84	61.3%
	-1	29	21.2%
	1	20	14.6%
	-2	2	1.5%
	2	2	1.5%
B	0	35	59.3%
	-1	16	27.1%
	1	5	8.5%
	2	2	3.4%
	-2	1	1.7%
C	0	22	75.9%
	-1	4	13.8%
	1	3	10.3%
D	0	47	75.8%
	-1	9	14.5%
	1	6	9.7%

Figure 7.16 - Psychometric Clusters versus Temporal Performance Change

It must be noted that Cluster C and D have the most consistent performance between stages 1 and 2, resulting in significantly higher percentages for cases related to 0 improvement score. Though the samples are not enough to show that it is not a chance that the biggest jumps and falls of performance are associated to Cluster A and B, the same has been observed with very small supporting numbers. The key features common to both Cluster A & B differing from Cluster C & D are Silhouette match, Conscientiousness, Trustworthy, Profile, Development and Company minded. These attributes together seem to capture consistency.

7.5 Conclusion

Artificial intelligence and a clustering approach bring difference regarding the traditional view by providing insights on pilots as well as different levels of support of training to adapt training methodology. AI techniques can increase access to pilot training for cadets, communicate success potential to stakeholders and improve overall success of cadets undergoing training

Clusters are created based on psychometric test scores from various test components. The cluster labels are used for building a supervised classifier model, which will later be used in SHAP framework for cluster interpretation. This results in student demarcation into four interpretable clusters. These clusters are correlated with overall performance of students observed during training, and the statistics suggests that intelligence has a role in performance, be it excellent performance or avoiding poor performance. Similarly, poor performance seems to relate to bad attitude rather than cognitive challenges. The learning consistency has a correlation with either or a combination of any of the Silhouette match, Conscientiousness, Trustworthy, Profile, Development and Company minded components of the psychometric test.

This study contributes in a good part to advance the notion of adaptive training and learning. One of the foundational elements of the adaptation process is the ability to segment the learners and this lays the groundwork that we have different learner profiles with the relevant insights. It helps identifying the individualized support required by learners for the successful completion of identified specific courses based on the psychometric test results.

This study shows that psychometric test results and performance are correlated. Psychometric test results alone will not be sufficient to predict the performance, however, this study convinces to use psychometric results in individualized learning methods like adaptive learning; for instance, to initialize adaptive learning-based recommendations based on the psychometric test scores.

By using historic flight performance data and psychometric test results, we can have better insights on the eventual flight performance of new cadet before they start the flight training program. The following potential actions have been identified, which could support the above goals:

- Predict and explain a cadet’s probability of success based on admission screening
- Provide individualized support based on results (more granular segmentation)
- Predict performance on remaining training sessions
- Identify training pathways that lead to success
- Adapt instructor briefing to improve probability of success
- Identify pilots who need more support and take appropriate action based on the cluster mapping

7.5.1 **Future Work**

Provide efficient pilot pairing & assignment. Are there some psychological profiles that boost each other's performance? Pair Students together during flight training and classroom. Pair cadets with the most suitable instructor.

Drop-out. How many terminations should each course include? Can we see any correlations between negative performance and drop-outs?

Career path. Can we redirect students to a specialized curriculum that corresponds to their profile? Aircraft type/long distance, fire-fighting mission, military assignment or any specific flight operation that requires a particular path?

Learning content. Explore the “difficulty” level of content based on pilot profile. Adapt learning content based on the student performance/profile? Add additional learning content to students who struggle. Help the content creator to understand how the various profiles use their content.

7.6 Acknowledgments

Acknowledgments to Andrea Lodi, Research Director, Polytechnique de Montréal and CAE contributors: Marc St-Hilaire, Vice-president Technology Innovation; Arnaud Van de Voorde, Vice-president Global Digital Strategy; Kevin Graham, Director Artificial Intelligence; Anthoine Dufour, Data Scientist; Janet Parham Digital Product Journey Owner; Seth Beckhardt, Training Solution Specialist Technical Services, Courseware Center of Excellence; Stephen Harry, Manager Product Management; Timothy Schoenauer, Director Global Training Solution; Veronique Gosselin-Bouchard, Digital Business Analyst

7.1 Conference

We presented the results at the European aviation training summit (EATS) 2021 to support the paper publication and share research results with commercial partner from Symbiotics Inc.

Figure 7.17 - EATS 2021 Conference

CHAPTER 8 ARTICLE 5: USING AI AND NEUROSCIENCE FOR ADAPTIVE LEARNING TO ACCELERATE PILOT TRAINING

Using AI and neuroscience for adaptive learning to accelerate pilot training, Jean-François Delisle, Theophile Demazure, Hamza Nabil, Pierre-Majorique Léger, Submitted in *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2022 proceedings,* June 2022

8.1 Abstract

To improve and accelerate pilot training, this paper explores the capture of cognitive/psychophysiological states using biometric sensors and flight telemetry to drive an intelligent human performance assessment system for adaptive learning.

Assessing pilot performance during a training session is a capability that can partially be performed by an AI-based algorithm. With technical data gathered during a flight maneuver, such assessment can provide objectivity during flight training, can be a predictor of future pilot performance, and adapt simulation training using a combination of flight telemetry (technical skills) and biometric/behavioral data (non-technical skills).

Evaluation of non-technical skills remains difficult without the support of data analytics and proper visualization tools. Additionally, soft skills are inherently more difficult to grade compared to technical performance. An AI engine can provide cues on behaviors and cognitive/psychophysiological states that cannot be easily observed by the instructor.

With a cohort of 16 novices, we explored neuroscience capabilities that could enable real-time adaptive flight training using a variety of data collected from a flight training session. By using electroencephalogram (EEG), eye tracking device and flight telemetry data with N-Back, BART & IGT cognitive baseline methodologies in a fast-jet flight simulator with an e-Series Medallion visual, we intend to provide a training scenario & maneuver analysis during initial training for both technical and non-technical flight performance.

8.2 Introduction

The objective of this paper is to explore artificial intelligence capability & human factors in the context of pilot training in a flight simulation environment by using Electroencephalography (EEG), Eye trackers, Facial Emotion recognition, and Flight Telemetry data with N-Back, BART & IGT paradigms. Biometrics analytics and data science methodologies are used to determine pilot performance & human factors associated with cognitive workload and decisions involving risk in a mission rehearsal using a 3D immersive simulation environment. In other words, our objective is to explore autonomous artificial intelligence capability that could enable adaptive flight training using a variety of data collected from a full-flight simulator and biometric sensors data. An intelligent adaptive flight-training system shall consider the human mental state in the learning process of the pilot. A better understanding of cognitive workload, risk-taking behaviour, and immersion levels are essential aspect to succeed in the real-time adaptation.

The risk-taking behaviour of the pilot is a flight safety issue considering that the reward effect can incite pilots to make riskier decision. This behaviour is very hard to reproduce during flight simulation and pilots know that they are in a simulation, so their risk tolerance is way higher. By increasing the immersion level of a training device, we hypothesis that we can improve the normalization of the risk-taking behaviour and better train on this aspect during mission rehearsal. The visual system of a training device is essential for the immersion level. Consequently, the technology level in the human-machine interface can influence the learning effectiveness. Introduction of a new immersive device brings the need for understanding key factors that impact human-technology interaction such as decision-making, reasoning, memorization, and perception.

One of the research goals is to identify a method that can optimize the cognitive load calculation using non-intrusive biometric sensors. Can machine learning be leveraged for model transferability from engineering where intrusive sensors can be used, to live operation where intrusive sensors cannot be used because of privacy & logistics reasons? Another goal is to explore if flight performance & pilot behaviour are correlated during initial training. In a sequence of initial training manoeuvres, we aimed to answer questions such as: What is the variability of cognitive load between manoeuvres and pilot profiles?

To ensure that the specific areas of focus and the experiment protocol is influenced by previous research studies, a focused literature review on section 8.3 is first carried out. In section 8.4, we present the concept using artificial intelligence to improve the cognitive load computing with the constraint of non-intrusiveness of biometric sensors. We also present the design of the experiment protocol as a training scenario during initial training. With an increasing difficulty scenario, we assessed performance based on criteria and thresholds. This is followed by low flight altitude manoeuvres involving risk-taking decisions as part of a contest between participants. The experiment protocol ends in a 2D/3D AB/Testing during an Air-to-air refuelling manoeuvre to observe if there are significant differences in performance with this new technology addition. In section 4, we provide results with a performance analysis. In conclusion, we identified potential future research using biometric sensors in flight training operations. More in-depth analysis of the risk-taking behaviour and a comparative evaluation of the perceive experience of the 3D visual are left for another publication.

8.3 Literature Review

EEG and Eye tracking data are tools used to study and understand the mental state of pilots in aviation as well as the processing of visual information. The applicability to the aviation domain as well as the methodology of data collection and analysis will be the primary focus of this literature review.

8.3.1 Eye Tracking for Cognitive Workload Estimation

Pupil size and eye blinking can be used as an index of cognitive workload where a lower eye blink rate is thought to indicate a higher workload, a higher eye blink rate may indicate fatigue, and larger pupils may also indicate greater cognitive effort or more pleasurable stimuli. Cognitive workload evaluation based on eye tracker data is a well-known method where (Marshall, 2002), a well-recognized author and a patented method and apparatus (Marshall, 2000) are certainly considered in the evaluation of cognitive activity in aviation. (Cabestrero et al., 2009) also analyzed how pupil diameter can be used to reflect mental effort and processing resource allocation when performing a recall task under multiple cognitive load conditions. The pupillary diameter increased systematically until the appearance of the small plateau. No reduction in pupil diameter was observed when exceeding processing resource limits besides the appearance of the plateau during the last tasks. This indicates that participants can continue to actively process even if resources are exceeded.

8.3.2 Scan Pattern with Eye-Tracking Data

Other important eye tracking metrics include blinks, fixation duration and location, and saccades, the rapid eye movements occurring between fixations. A higher number of saccades indicate seeking behaviour. (Škvareková, Iveta 2020) uses an eye tracker to record eye movements. The article confirms that experienced pilots were able to receive information in less time and had higher saccades per minute than inexperienced pilots. (Stephanie Brams et al. 2018) examined differences in gaze and visual scanning behaviour between high-performing and low-performing pilots. They also provide insight into the underlying processes that may explain perceptual-cognitive expertise under the theory of long-term working memory, the information reduction hypothesis, and the holistic model of perception of knowledge. The number of downtime and the number of transitions between AOIs differed between high-performing and low-performing pilots. Poorly performing pilots perform a more exhaustive search and make more transitions between extreme areas of interest. Pilots are better able to shift their attention between global and local information processing.

8.3.3 Understanding pilot's reaction in flight operation using neuroscience

(Lan et al., 2019) recorded EEG data with 15 subjects using a 32 EEG channels. In their paper, the authors adopted Differential Entropy (DE) as features for emotion recognition. DE features have been extensively used in cited literature studying the application of transfer learning techniques in EEG-based emotion recognition. Extending our data sample with emotion will be an addition that will complement well in an intelligent adaptive flight training system.

(Binias et al., 2020) attempted to predict the reaction time to an unexpected event based on the brain activity recorded before the event using EEG data. They measured the time lag in the participant's reaction time to a visual cue using regression in a flight simulation experiment with autopilot enabled. The prediction algorithms used are the least absolute shrinkage operator, Kernel Ridge regression and Radial Basis Support Vector Machine. A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. Automated systems placed the pilot in a passive role which introduced an additional challenge should any issues arise as the pilot must move into an active role and resolve complex issues. (Binias et al., 2018) dealt with the problem of discrimination between brain activity states related to anticipation and reaction to a visual signal and the selection of an appropriate classification algorithm for such problems. In this work, an EEG signal processing and classifier tuning methodology was proposed with the aim of analyzing data containing brain activity states related to an inactive but focused anticipation of a visual signal and a reaction to this signal. Experimental electroencephalographic data were acquired using an Emotiv EPOC+ headset. The methodology involved the use of different classification algorithms, such as Linear Discriminant Analysis, *k*-nearest neighbours, Support Vector Machines, Random Forests, and Artificial Neural Networks. The results suggested that the performance of a neural network could be significantly better than that of other algorithms and validated by an analysis of variance (ANOVA).

In an investigative article by (Monteiro et al., 2019), an in-depth review of techniques for using EEG to assess MF mental fatigue was performed and supported by an overview of the principles of acquisition, interpretation, algorithms, and trends. There are subjective ratings based on self-report to assess MF states and include the NASA Task Load Index, Karolinska Sleepiness Scale, Epworth Sleepiness Scale, and Chalder Fatigue Scale, but they are subject to bias. When evaluating MF sensing, EEG signals are composed of five main frequency bands: delta, theta, alpha, beta, and gamma. The theta (θ) band can be found during drowsiness and sleep in adults. The alpha (α) band can be found in adults who are awake, relaxed, or mentally inactive. Frontal θ and occipital α and parietal activity are likely to increase as a person becomes fatigued. The beta (β) band signifies tension and anticipation and can be found in alert and anxious subjects. The most used preprocessing methods for MF detection using EEG include digital filtering, independent component analysis (ICA), and discrete wavelet transform (DWT). Commonly used feature extraction methods include power spectral density (PSD), statistics, and entropy measurements. When a person is fatigued, a decrease in the level of entropy of their EEG signals can be expected, indicating a decrease and weakening of brain synapses. The most used measures of entropy are Sample Entropy (SampEn), Fuzzy Entropy (FuzEn), Approximate Entropy (AppEn), and Spectral

Entropy (SpecEn). Since the MF state is a constructed process where fatigue accumulates over time, a dynamic approach considering the temporal aspect becomes possible with the development of models such as LSTM. The article concluded by suggesting a model based on kernel partial least squares discrete output linear regression as a good overall option for an FM evaluation system.

8.3.4 **Neuroscience for Pilot Workload**

With ECG, Eye Tracker, and EEG complemented by NASA-TLX questionnaires (Thomas C. Hankins, 1998) measured the mental workload of pilots collected during flight scenarios. Combining multiple measures such as psychophysiological states and subjective measures can provide a broader picture of the mental state of the pilot. Heart rate is useful to measure the flight demand but not on the mental workload. Eye tracker was more powerful for the diagnostic task while EEG theta band increased during mental calculation.

Augmented cognition is a form of human-computer interaction in which sensing a user's cognitive state is used to invoke system automation on demand. The study by (Nicholas Wilson, 2021) monitored the pilot's in-flight physiological state to determine the optimal combination of EEG cues to predict changes in cognitive workload. Data collection was executed in a real-world flight environment with scenarios that varied in workload with a group of undergraduate aviation students using a single-engine trainer equipped with Garmin G1000 avionics. Some of the higher workload flight manoeuvres executed a missed approach to minima and perform consecutive steep turns. While manoeuvres categorized as low workload included straight and level flight and taxiing in a towerless airport. Power spectral density values were calculated and subjected to machine learning methods to distinguish periods of high and low workload. The feature extraction step was performed using power spectral analysis. Fast Fourier transform (FFT) was used to transform EEG into power spectral density (PSD). The Lasso cross-validation algorithm was used to select the most important features. The support vector machine (SVM) algorithm was used as a binary classifier for its robust approach to complex pattern recognition, good generalization performance, and efficient computational cost. The results show that the selected characteristic can be successfully used as an indicator of the level of cognitive workload of pilots.

8.3.5 **Neuroscience in flight training**

(Zhang et al., 2020) compared the pilot's EEG signal at different phases of flight, different weather conditions, and different levels of training. The results showed that EEG entropy could be used to assess the pilot training effect. The entropy value in windy and rainy conditions was more dispersed, which means that the frontal workload is greater than in sunny conditions. According to the cerebral plane, the load on the occipital lobe, part of the parietal lobe and the right temporal lobe increased. The increase in the occipital load during the take-off phase comes from the change in the exterior view of the cabin. This change will bring higher mental load to the pilots and the difficulty in processing information has caused the load on the frontal lobe as the whole processing centre of the brain to fluctuate considerably. The student will adapt to this environmental change in the later stages of flight, reducing the load on the temporal lobe. Trained pilots demonstrate that regular training increases excitation of the frontal and occipital lobes and as training time increases; the average level of entropy approach a fixed value.

The ability to identify the learner's workload is crucial for their implementation of an adaptive training system. The study by (Baldwin & Penaranda, 2012) used an artificial neural network (ANN)-based classification algorithm using neurophysiological measures requiring effective real-time mental workload classification based on electroencephalographic (EEG) activity during the performance of short-period tasks. Classifiers determined the workload and the cognitive-emotional response of a learner during training were essential for the implementation of adaptive training. With fifteen participants, signals from the EEG and EOG electrooculogram were recorded using a 40-channel NeuroScan NuAmps amplifier, and a 40-channel QuikCap Ag/AgCl electrode cap. Three different working memory tasks, the Reading Span task, the Visuospatial n-back (n-back) task, and the Sternberg Memory Scanning task were used. The ANN could distinguish between low and high difficulty levels quite reliably.

In their article, (Liu et al., 2019) proposed the use of emotion, workload, and stress recognition algorithms based on the use of an EEG, in addition to questionnaires and traditional feedback to study the optimal duration of the training of air traffic control officers (ATCOs). A 14-channel Emotiv EEG device was used to monitor the brain states of ATCOs as they learn to use a new 3D interface in performing aircraft trajectory operations in different weather and terrain conditions in addition to the traditional 2D display. Emotion and workload recognition algorithms from EEG signals and a stress recognition algorithm are proposed. It was observed that while emotions did not have a definite impact on training duration, workload and stress levels were significantly different between training duration and optimal training duration. Correlation analysis indicates that if ATOs have more confidence in a new system, their emotion is more positive; stress and workload are lower when they learn to use this new interface.

8.4 Methodology

8.4.1 Problem Statement

Observing and operating the simulator can be complex for the instructor; they may miss some behaviour in fast pilot operations during the evaluation of complex manoeuvres. Moreover, trust in the instructors' evaluation could be challenged. The use of an expert is not scalable or cost-effective when assessing non-technical skills. Soft skill cannot be easily assessed by another human comparatively to technical skills. Without the support of data analytics and visualization tools, it will be impossible to identify levels of correlation across the entire rich collection of data and parameters available. The understanding of the cognitive load at the microlevel task requires high temporal resolution data that electroencephalogram can provide. However, the intrusiveness of the EEG in flight operation is preventing the collection of this data.

8.4.2 AI Solution Concept

Successful training of supervised machine-learning approaches for classification required objective ground truth to provide annotated examples of the classification target. In the case of the research, we used two deep learning models, a fully convolutional neural network (FCN and a residual network (ResNet) to estimate mental workload. As presented by the Figure 8.1, the usage of artificial intelligence can provide an optimization of the cognitive load index measured by eye tracker and pupillometry. By using supervised machine-learning with pupillometry as a feature and EEG cognitive load index as the target label, we can provide a machine-learning model that is deployable without the intrusion of an EEG in flight operation.

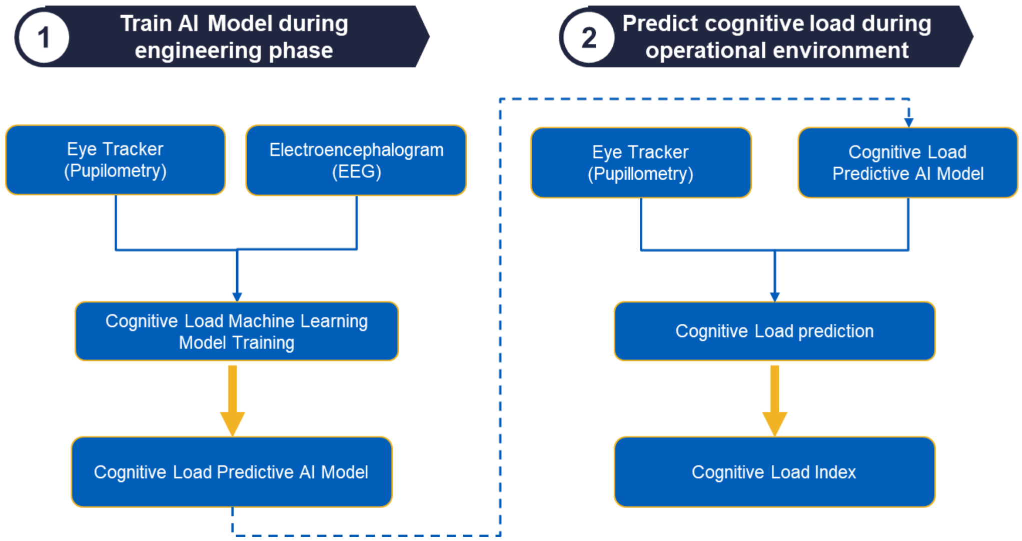


Figure 8.1 - Cognitive Load Optimization with EEG label in machine learning process

8.4.3 **Experimentation Device**

In our research, we used a fastjet flight simulation device with an immersive visual and flight controls. The experimentation occurred in an F16 - Flight Simulator with CAE Medallion MR e-Series visual system that provided natural hi-fidelity visual immersion with the objective of reducing eyestrain and fatigue. The prototype simulation of a consisted of a 200° partial sphere screen with a radius of 1m and a height of 1.5m giving a 9,42m viewable surface area as shown at Figure 8.2. The aim of such prototype was to provide Smearing/Motion-Blur reduction from unequalled dynamic 120Hz resolution; head movement compensation to virtually eliminates parallax error and 3D depth perception via active eyewear as shown at Figure 8.3. The flight simulator will not have motion enable during the experimentation.

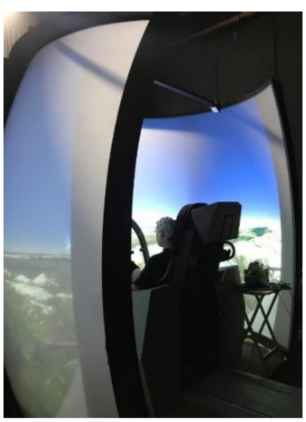


Figure 8.2 - Flight Simulation Visual System

Figure 8.3 - Head tracker and 3D glasses

The flight simulator is not using motion system during the experimentation and the key recorded flight parameters were: Latitude (LAT), Longitude (LONG), Mean Sea Level (MSL), Altitude Above Ground Level (AGL), Calibrated Airspeed (CAS), Ground Speed (GSPD), G-Force (GTRK), Heading/Yaw (HDG), Pitch (PITCH), Roll/Banking (ROLL), Angle of Attack (AOA), Engine Thrust (ENG_THRUST)

8.4.4 **Sensor's selection**

The first step of the research was to select the biometric sensors to be used during the execution of the experiment. This selection must be able to consider the specific nature of a fastjet flight simulator offering the pilot a 200-degree field of view while maintaining data quality for the purpose of calculating cognitive and emotional state, as well as the path of gaze on the instruments or Area of Interest (AOI)

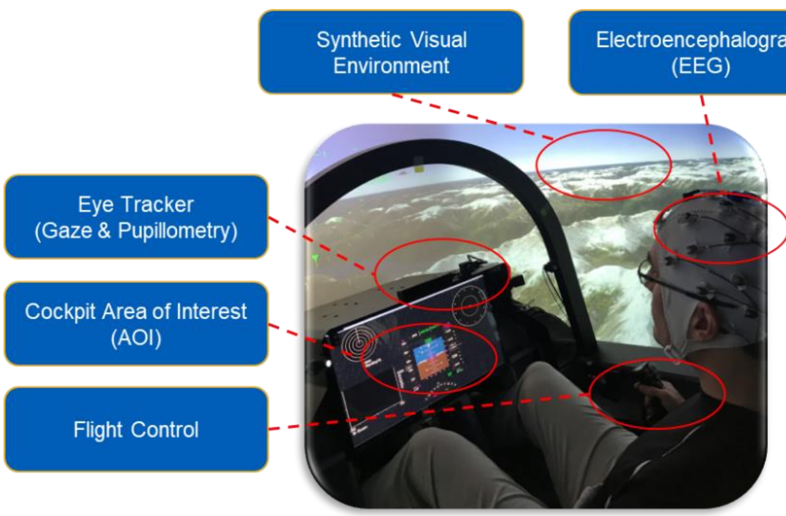


Figure 8.4 - Flight Simulation Cockpit and Biometric Sensors

The EEG device is a tool for measuring brain electrical activity and infer mental states such as mood, anxiety, or stress as well as cognitive state. We recorded EEG with 32 electrodes at 1000 Hz with BrainAmp amplifiers from Brain Product. The EEG signal was recorded using the standard 10–20 montage. The signal was first referenced into Fz and then filtered using a 2nd order Butterworth 1-40 Hz bandpass IIR filter and a 60 Hz notch filter. Muscle and eye movement related artifacts were removed using blind source separation by independent component analysis. Before analysis, EEG signal was downsampled to 500 Hz.

The Eye Tracker system is using 5-cameras and a Bar Tracker model 5-CAM DX 2.0 MP Smart Eye Pro with IR mini flashes 60Hz. Data Collected are Fixation/Dwell time, Blinks, Pupil diameter, Saccades and Intersection names with a 3D world model composed of Area of Interest (AOI) of the instrument panel.

Figure 8.5 - Smart Eye Pro dx camera + IR

Figure 8.6 - Smart Eye Aurora system (Bar tracker)

The bar tracker has 2 built-in cameras and 2 infrared illuminators. The system can obtain more precise metrics for instrument checks with the 5-camera system and was suitable for the context of a wide range FOV. The definition of the world is composed of Area of Interests (AOIs) corresponding to the specific cockpit instrument area, which is done using a laser and a laser chessboard to calculate the world coordinates using various geometric shape to design the 3D world. The device is able to obtain the metric to about 150 cm around the subject. However, if the subject is seeing a point beyond, the system does not detect properly the world intersection, or it might have interference. The gaze calibration is performed with every subject using the device. If a gaze calibration is not performed, the results may have a variation between 3 and 7 degrees with respect to the point to which the subject is looking at. The **Error! Reference source not found.** presents the result of tested accuracy and device specifications using a variety of candidates prior to the experimentation.

Table 8.1:. Eye Tracker Accuracy Assessment and Specifications of Smart Eye's eye tracker

	Gaze Tracking (Able to track gaze) (1)		**Gaze tracking - Calibration results (Deviation/Accuracy)**		**Head tracking, (Able to track head) (3)(4)(5)**		**Head Tracking, Field of vision (Range) (2)**	
	5-Cam	Bar Tracker	5-Cam	Bar Tracker	5-Cam	Bar Tracker	5-Cam	Bar Tracker
No glasses	99%	90%	1.5° deviation, 0.8° accuracy	0.2° deviation, 0.6° accuracy	> 97%	> 97%	> 180°	90° – 130°
Wearing glasses	80%	70%	3.5° deviation, 2.5° accuracy	2.7° deviation, 3.2° accuracy	> 97%	> 97%	> 180°	90° – 110°
Long hair	99%	90%	1.3° deviation, 0.9° accuracy	0.7° deviation, 1.3° accuracy	> 90%	> 90%	> 180°	90° – 130°
Helmet	99%	90%	1.1° deviation, 0.2° accuracy	0.6° deviation, 0.2° accuracy	> 90%	> 90%	90° – 150°	90° – 110°
Communication device	90%	90%	1.9° deviation, 2.7° accuracy	0.5° deviation, 0.8° accuracy	> 90%	> 90%	90° – 150°	90° – 110°
Covering ears	90%	90%	1.5° deviation, 2.4° accuracy	0.3° deviation, 0.8° accuracy	> 90%	> 90%	90° – 110°	90° – 110°
Facial hair	90%	90%	1.8° deviation, 2.6° accuracy	0.4° deviation, 0.7° accuracy	> 97%	> 97%	90° – 150°	90° – 130°
Make-up	99%	90%	N/A	N/A	> 97%	> 97%	> 180°	90° – 130°
Covering mouth	90%	90%	N/A	N/A	> 90%	> 90%	> 180°	90° – 130°

(1) The use of multiple cameras compensates both eyes and a unified gaze direction is streamed, (2) The metric of the 5-Cam according to the position of the cameras, (3) Affected by the field of vision and the angle between the subject and the camera, (4) Affected by covering facial features, (5) This result can improve creating a manual profile for the subject

8.4.5 Experimentation Protocol

The human testing activities will consist of simulated real-world scenarios in various controlled setting environments.

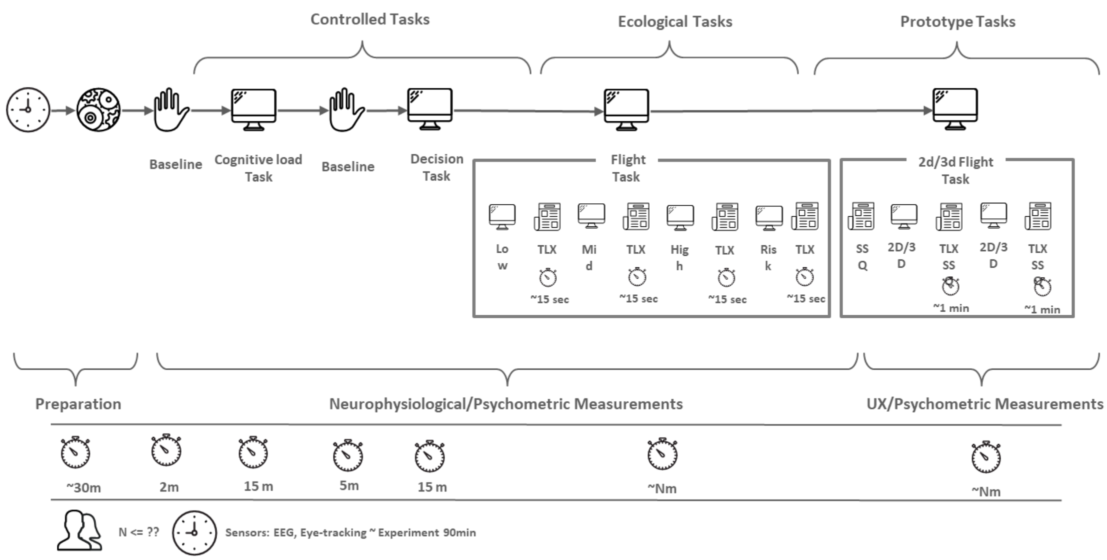


Figure 8.7 - Experimentation Protocol Diagram

The experimental protocol was designed in a mirror fashion composed of two artificial and controlled tasks and then, ecologically flying tasks valid. The first experimental task was a synthetic n-back task (Susanne M Jaeggi, 2010), which is known to gradually manipulate mental workload. The corresponding simulator task was sequence of maneuvers designed to incrementally increase the mental workload of the pilot through maneuver difficulty. The second experimental task was a BART, created to manipulate risk taking behaviors. The corresponding task was a risk-taking free flying task during which the participant had to fly near mountains and valleys. The artificial tasks were implemented using the software package E-Prime 3.0 from Psychology Software Tools and performed before the two ecologically valid flying task.

Sixteen participants recruited from a Flight Training Company with a beginner profile or first step in piloting were favoured in order to promote the collection of data in cognitive overload. All human subjects have signed a consent form to participant in this experiment.

8.4.5.1 Simulation Task

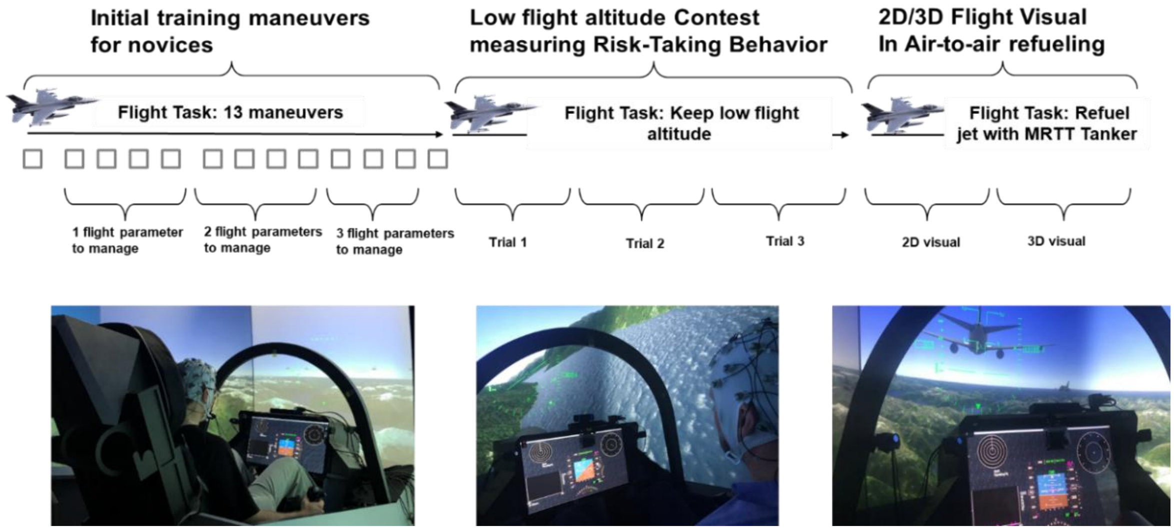


Figure 8.8 - Flight Simulation Tasks Sequence

As show at Figure 8.8, the simulation tasks are divided in three categories, the initial training manoeuvres, the low flight altitude task and the Air-to-Air refueling with a MRTT Tanker.

8.4.5.2 Initial manoeuvre with increasing difficulty

The first phase of the experiment consists of flying "twisler". It consists of a few little simple manoeuvres, which requires looking at various flight parameters such as banking, pitch, altitude, and speed. This phase is inline with the technical test done on the A310 simulator where a pilot was coached by an instructor to execute a few simple manoeuvres in an initial training context. Manoeuvres corresponded to a variation of 4 parameters: speed, altitude, heading, and banking.

After a free flight to familiarize with flight control and the aircraft reaction, we started the exercise with an aircraft stabilization manoeuvre. The following 4 manoeuvres consisted of changing one of the flight parameters in isolation, one at a time. The following 4 manoeuvres consisted of the same principle but varying two parameters simultaneously. The following two manoeuvres varied 3 and then 4 parameters simultaneously to complete this exercise phase with a vertical loop before stabilizing the aircraft.

There are three blocks of flight manoeuvres shown as low, moderate, and high mental load. Each manoeuvre was associated with appropriate flight actions and led by a flight instructor, who recorded each participant's performance and signalled the end of each manoeuvre as a synchronization marker between the data.

Assessment was made by an instructor that evaluated the participants after the execution of each maneuver. The factor for assessment was time, fluid stabilization, slope of the change, and compliancy to the threshold. We used a tolerance for the various flight parameters as followed: Altitude: 100 ft, Knots: 5 ft, Deg. Heading: 5 deg, Deg. and Banking: 5 deg

8.4.5.3 Low flight altitude for Risk-Taking Behaviour

To assess risky behaviours, we performed a computerized decision-making and risk-taking activity followed by a simulated flight activity involving a risk-taking task. The first task, two experimental computerized risk-reward risk-taking tasks were performed, an Iowa game task (Antoine Bechara, 1994) and a Balloon Analog Risk Task (BART) (Lejuez, 2002). The second phase involved risk taking. Where a participant was asked to sustain an altitude above ground at various threshold, where each threshold paid points to a leaderboard. The objective was to incite pilots to manage risk and their strategy in order to gain points. In a simulation through the British Columbia Rockies Mountains, we asked the participate to choose a mountain or a valley that may create a multiplier factor of: Valley = x1, Small Mountain = 2x, Big Mountain= x3 with the above point distribution per altitude level 0-250 ft = 100 pts, 251-500 ft = 50 pts, 501 – 1000ft = 25 pts, Crash or >1000 ft = 0 pts. Participants got 3 trials to accumulate points and we gave a TLX questionnaire after the third trial.

8.4.5.4 2D/3D AB Testing during Air-to-Air Refuelling Manoeuvre

In the third phase, half of the participants started with 2D visual display then switch to 3D, another half started with 3D then finish with 2D. We were inspired by (Wen-Chin Li, 2014) where the scenario was an air-to-air task in a jet fighter simulator studying eye movement. We asked participant to execute a Mission rehearsal of an air-to-air refuelling. We assessed the performance based on time to reach the in-flight refueling pole (boom) of the MRTT Tanker and the stability of the flight while keeping the boom in range. We gave a TLX questionnaire after each trial.

8.4.6 Data Analysis Methodology

In this section, we will present the methodology of the analysis of the performance data for this experiment. Our goal was to develop an autonomous method to score or grade each pilot's performance through various flight manoeuvres based on their telemetry data. Our prediction would be as the manoeuvre's difficulty increases; the performance of the pilots would decrease. In relationship to the telemetry data, we also want to analyze and correlate the results of the eye-tracking data using an ANOVA test. It is important to note that all participants were anonymous for the experiment and analysis to be objective.

8.4.6.1 Data Description

There were three data sources taken from this experiment, Smarteye's eye Tracker, Brain Vision's EEG, and Objective Assessment. Smarteye log files are generated from the videos captured by the Smarteye camera system. There is one log file per participant. The log files contain tab-separated entries. The number of rows in each log file is the frame numbers captured by the Smarteye camera. Each log file has 485 columns. The log files can be loaded as a data frame in pandas, which could be very useful for various analytics tasks. Each time stamp is 16.7 ms with a frequency of 60 HZ. Objective Assessment was a measure of the telemetry data from the flight manoeuvres performed. The measurement of the objective assessment was on different parameters that include Altitude, Banking, Heading, Speed, and Pitch. The objective assessment of the experiment was based on the exceedance occurrences, standard deviation, and time. Each time stamp is 240ms. We used two deep learning models, a fully convolutional neural network (FCN and a residual network (ResNet) to estimate mental workload. EEG data was exported as CSV files that include FCN & RESNET methods.

8.4.6.2 Algorithms

Performance assessment method

Objective assessment was based on the telemetry data of each pilot and the following factors: Exceedance occurrences, Standard Deviation and Time of flight. The objective score was scored from 0-4 and was averaged based on the 3 factors above.

The Standard Deviation factor is scored based on the standard deviation of each parameter: Altitude, Banking, Heading, Speed and Pitch. Each manoeuvre differed from one another, since some of them had one, two, or all parameters. The pilot's standard deviation for a given manoeuvre was in comparison to the preset tolerance each manoeuvre had for each parameter (Altitude, Banking, Heading, Speed & Pitch).

$$std = \sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Time was to be scored based on the difference the pilot took on each manoeuvre in comparisons to the time tolerance for the specific manoeuvre. The time tolerance was preset based on the mean time it took for all participants to finish the given manoeuvre.

Time score was also scored from 0-4.

The Exceedance occurrence factor is based on how consistent the pilot would stay within the tolerance of each manoeuvre and its parameters. The factor was measured on how much time of the entire manoeuvre did the pilot spend outside of the tolerance. It was also scored from 0-4.

Hypothesis test method

One-way ANOVA test was performed to compare the means of multiple grouped sets of data.

Null hypothesis: Groups means are equal (no variation in means of groups)

H_0: $\mu_1=\mu_2=\ldots=\mu_p$

p-value< 0.05

Alternative hypothesis: At least, one group mean is different from other groups

H_1: All μ are not equal

p-value> 0.05

After finding the p-value, a post hoc comparison was made using a Tukey honestly significantly differenced (HSD) test to know which group was significantly different from each other.

The ANOVA test assumption was primarily checked by using the Shapiro-Wilk test that analyzed the normal distribution of the residuals.

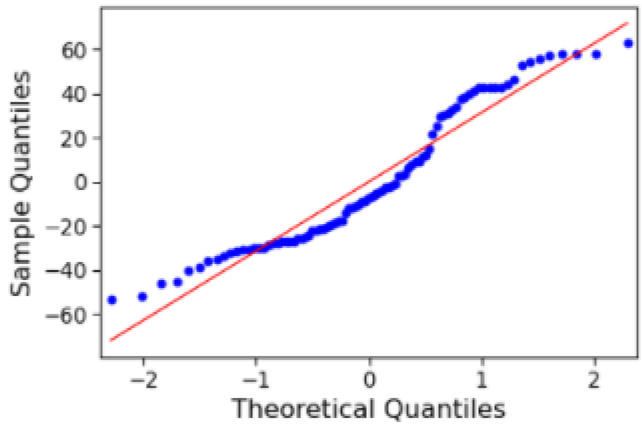


Figure 8.9 - ANOVA test

Depending on whether the results were drawn from normal distribution or not, a Bartlett's test was used to check the Homogeneity of the variances to see that it was normally distributed and Levene's test when not normally distributed.

Null hypothesis: Group variances are equal (no difference in variance of groups)

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

p-value< 0.05

Alternative hypothesis: At least, one group variance is different from other groups

H_1: All σ are not equal

p-value> 0.05

8.4.6.3 AOI Correlation method

A one-way ANOVA test was performed to compare the means of 3 grouped sets of data in comparison to each flight parameter (Altitude, Banking, Heading, Speed & Pitch).

Each AOI data set was split in 3 groups:

1. Not looking
2. Slightly looking
3. Looking a good amount

ANOVA test was done 5 times in total, split up by manoeuvres that flight parameter mattered.

8.4.6.4 Workload Correlation Using the Gaze Tracking method

A one-way ANOVA test was to be performed twice for the workload (fixation & saccade). The means of 3 different groups for the objective and subjective scores were compared to the fixation and saccade.

The objective and subjective scores were split into:

1. Bad (Score under 2),
2. Average (Score between 2-3),
3. Good (Score between 3-4)

8.5 Results

8.5.1 Flight Profiles

With the 5 parameters studied as time series per task (Altitude, Speed, Heading, Pitch, Banking). Figure 8.10 is an example of a flight profile for the altitude change task. We can see different profiles with good/bad stabilization, some short/long manoeuvre time, and multiple peaks prior to the targeted altitude prescribed by the training scenario.

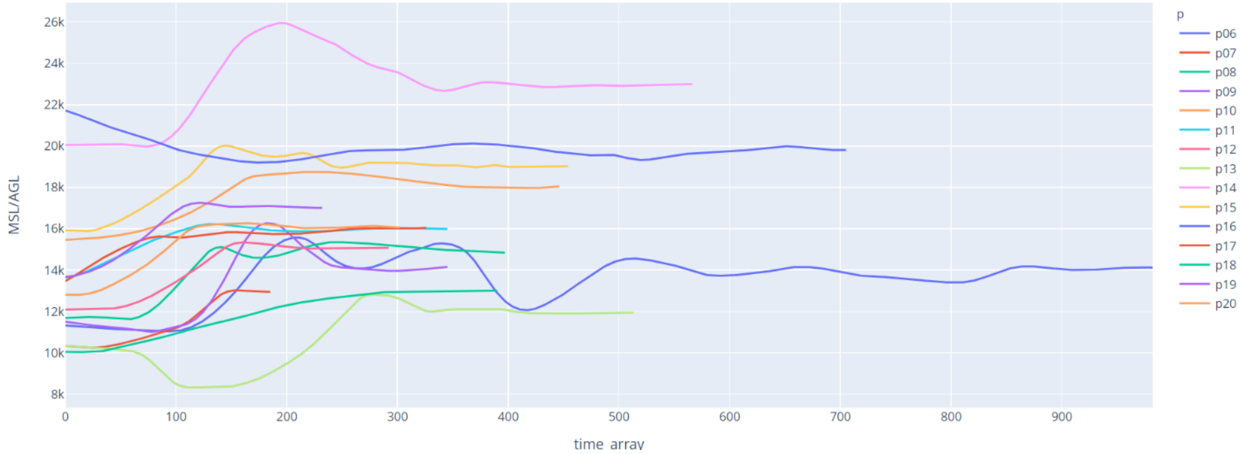
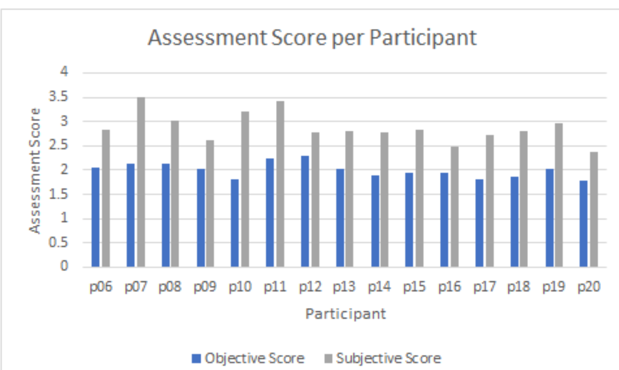


Figure 8.10 - Flight profile - Altitude change task

8.5.2 **Performance Assessment Analysis**

The following Figure 8.11 represents the average of objective and subjective scores for the entire experiment for each participant. As shown in the above figure, there is a discrepancy between the objective and subjective scoring. The subjective scoring can be influenced by the emotional aspect of the person scoring the participants. However, with the objective scoring, everything is rational, and feelings are not a factor when giving a score.



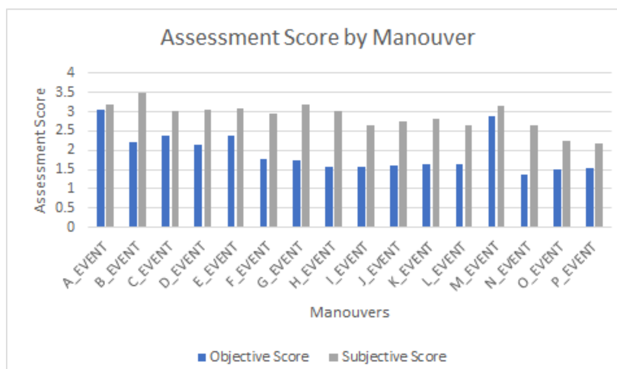


Figure 8.11- Grades per participants

Figure 8.12- Grades per manoeuvres

The Figure 8.12 above presents each manoeuvre graphed with the comparison of the objective and subjective scores. As the manoeuvres increase with difficulty, the objective scores start to decrease which goes along with the initial hypothesis.

8.5.3 **AOI Correlation results**

An ANOVA was done to analyze the correlation between the eye tracking of the participants and their performance with the objective score. From the results below, it was concluded that the hypothesis of manoeuvres that altitude, pitch, and speed factored in, the more time the participants/pilot looked at the gauges for those parameters, the better their objective score would be. The results, however, differed for the speed and heading parameter as a p-value above 0.05 was achieved.

Test	P-Value	Result
Altitude	0.003	Reject, means are different
Pitch	0.046	Reject, means are different
Banking	0.049	Reject, means are different
Speed	0.56	Accept, means are same
Heading	0.22	Accept, means are same

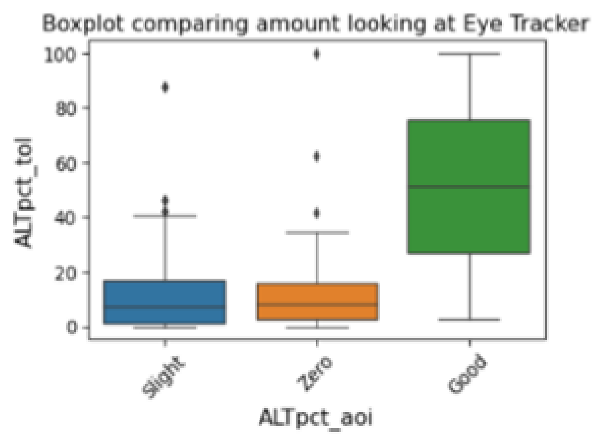


Figure 8.13 - Boxplot comparing amount looking at Eye Tracker

8.5.4 **Workload analysis**

Two end-to-end deep learning models were trained to learn features and mental workload levels based on the EEG signal recorded during the n-back tasks. Two models, an FCN (91% accuracy) and a ResNet (92% accuracy) algorithms trained on the task data provided us a good mental workload level estimation. The models were trained to discriminate mental workload level from the cleaned signal data. Sanitary checks of the models were performed for there physiological plausibility. The results show expected results, an increase of manoeuvre complexity led to an increase of high mental workload classifications, and a decrease of maneuver performance.

There is also a close link between the pilot's mental workload and eye movements, if a pilot does not have optimum conditions, eye movement changes. We define the two parameters as saccades and fixation. The saccade is defined as the rapid movement of the eyes between two fixations. Fixation is defined as a condition in which an individual visually collects and interprets information available in the range of the eye over a period of time.

In an article from the 9th International Conference on Air Transport "Number of Saccade & Fixation Durations as Indicators of Pilot Workload". They measured the effects of mental stress caused by a lack of training and flight experience. They’ve concluded that with a higher Saccade per minute, experienced pilots were able to receive information in less time. A shorter stay on flight instruments allowed more experienced pilots to scan other areas of interest. They also had more time to detect any errors and then start the correction.

To compare our results with the research paper, we’ve used an ANOVA test to validate the performance assessment in relationship to the Saccade & Fixation of each pilot. The hypothesis was done four times, twice each for saccade and fixation using both objective and subjective scores. The scores were divided into 3 categories of bad, average, and good.

	Bad	Average	Good
count	179.000000	38.000000	23.000000
mean	17.110838	30.928947	141.886087
std	19.348152	30.881550	434.929001

<table>
<tr><th>Test</th><th>Objective Score P-Value</th><th>Subjective Score P-Value</th><th>Result</th></tr>
<tr><td>Saccade</td><td>0.0002</td><td>0.48</td><td rowspan="2">Reject, means are different.
Major Correlation between "Good" and other scores.</td></tr>
<tr><td>Fixation</td><td>4.9E-9</td><td>0.1</td></tr>
</table>

As seen in the results above, the results of the ANOVA test further validate the research previously done. Both the Saccade and Fixation p-values were under 0.05 by a significant amount showing that the means of the different groups of the objective score have a major correlation between each other. With a higher saccade per time frame, the pilots achieved a higher objective score. However, the same cannot be said using the subjective score, further proving the objective scoring scheme is more accurate and should be utilized more than the subjective score.

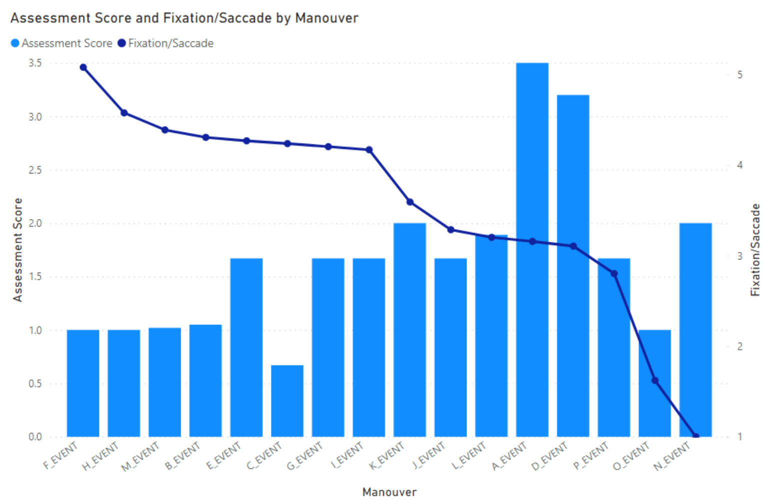


Figure 8.14 - Ratio of fixation over saccades per manoeuvre’s assessment score

For further validation, we have also graphed the ratio of fixation over saccades in comparison to the assessment score per manoeuvre to give a more robust level of attention to the pilots. As shown in Figure 8.14, pilots who achieve a higher score tend to have a lower ratio of fixation per saccade. Incorrect scanning patterns from the pilots could lead to information overload and staying longer on a flight instrument. The difference between these numbers of fixation per saccade for pilots is mainly because successful pilots were able to receive information in a shorter time and continued with the scanning technique of instruments.

8.5.5 Performance of initial manoeuvre tasks, Low Flight Altitude tasks and Air-to-Air Refuelling tasks

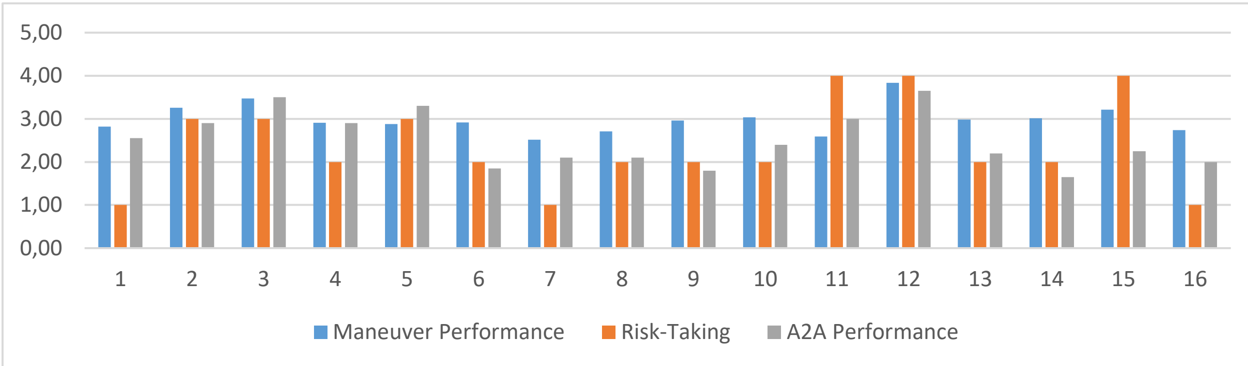


Figure 8.15 - Average Performance per participants with simulation tasks

We see in Figure 8.15 that in general, the performance correlate with risk-taking behaviour that can indicate an engagement, however, many factors can influence the results and the presence of outliers, such as the usage of a new training device, the impact of 3D, etc. Also, their results showed differences in pilots' fixations among five different areas of interest (AOIs) of the cockpit using head-mounted eye tracker. Further results will be presented into another publication.

8.6 Conclusion

The objective of this study was to explore artificial intelligence capability & human factors during initial flight training session. Those insights contribute to explore how we can enhance the instructor's awareness of the cognitive workload, and scan student's pattern. As an eventual capacity of an intelligent adaptive flight training system, the session can be tailored to maximize the students in real time. We also aimed to develop a conceptual method by using artificial intelligence to keep the strength of an EEG device in the engineering phase and prevent the addition of intrusive sensors during real flight training operation.

For this we used biometric sensors in a simulator training session and explored how we could use data science on a number of training manoeuvres that could be used to assess pilot performance during initial training tasks, risk-taking behaviour tasks and air-to-air refueling tasks within a new human interfaces machine technology to improve training programs and immersive systems.

As expected, we found that as manoeuvre complexity increased, workload and perceived mental workload measured by NASA-TLX also increases. We also found that correlation exists between the scanning pattern and the flight performance. This indicates to us that biometry sensors can bring a new kind of insight that can bring objective measure in the assessment of the human performance in the flight operation.

For future work, we consider going further in the cognitive workload estimation methodology. We would like to identify and compare multiple algorithms that are able to classify and predict cognitive load, flight performance, and risk-taking behaviour. We would like to address questions such as: How can we predict the outcome (technical/non-technical) of a manoeuvre on biometric & flight telemetry data? Does cognitive load risk-taking behavior is affected by cognitive load? What is the correlation between flight performance with telemetry and psychophysiological state? How can we detect and predict flight performance based on the cognitive load index?

The dataset we accumulate can be used in future analysis around risk taking behaviour in low altitude manoeuvre. The results of formal data analysis using anticipated statistical methods would provide insight into participants' risk behaviours and the level of cognitive workload required when taking risks. This study will aim to analyze the human factors associated with risky behaviours using the characteristics of central and autonomic nervous system activity and answer questions such as: What is the neurophysiological state involves risk-taking behaviour. What is the performance impact for the various risk behaviour? We will also propose to use machine-learning methods to build a risk classification model using other flight simulation models to valid portability of machine learning algorithms for pilot performance & behaviour. Analysis using biometric sensors to assess initial training and risk-taking behaviour of novice pilots. Which machine learning algorithms are more suitable to apply risk-taking behaviour classification and prediction?

To contribute to flight safety, those capabilities can be matured up with an R&D project into a business-centric initiative. Using a real Crew Resource Management (CRM) training session on a large number of experienced pilots, we will explore how we can augment the technical readiness level of neuroscience capability. We will consider operating emergency manoeuvres in a flight-training session commercial aircraft simulator in a flight-training centre using non-intrusive biometric sensors and certified flight instructor.

Also in future publication, we will show a method to evaluate perceived experience using self-reported data with the 2D vs. 3D visual system as a A/B testing results.

8.7 Acknowledgements

Acknowledgements to Andrea Lodi, Research Director from Polytechnique de Montréal, Patricia Gilbert, Marc St-Hilaire, Andrew Fernie, Ginet Andreina Calderon Perez Philippe Perey from CAE Inc. and Alexander Karan Tech3lab, Nicolay Nonchev from HEC-Tech3lab

8.8 Conference.

The research methodology was presented at the « Institut de valorisation des données IVADO 2021 - Zoom sur l’intelligence numérique collaborative » conference.

Figure 8.16 - IVADO 2021 Conference

CHAPTER 9 PATENTS - IMMERSIVE FLIGHT TRAINING SYSTEM

This chapter presents the technological capability that was invented to build an intelligent adaptive flight training system with systems/software components such as intelligent flight simulation control and monitoring system, automatic objective assessment & e-Grading system, record and playback with just-in-time visualization of insights, adaptive lesson plan system, virtual instructor, and cognitive assistant artificial intelligent solutions in immersive flight training devices. From data collection to dynamic visualization, the enablement of analytic & software capability is represented in the thesis with multiple granted patents published to the following patent offices: Canadian Intellectual Property Office (CIPO), United States Patent and Trademark Office (USPTO), China National Intellectual Property Administration (CNIPA), European Patent Office (EPO), and the World Intellectual Property Organization (WIPO). **Annexe A** presents the patent portfolio

In the articles of the previous chapters, the scientific hypotheses are presented as well as the experiments to validate them. The hypotheses and experiments are new, so the instrumentation and the software used to perform the experiments are also new. Patents therefore describe the new capabilities of flight training systems built in the context of this research.

An article and a patent are technical documents that both describe a new concept. They both have the objective of making new scientific discoveries available to the public. They both promote the same principles of innovation and creativity and are an indicator of the industrial contribution of the research project. Our patents make possible to publish this source of creation having an industrial impact and describe the instrumentation in detail to enable researchers to easily duplicate the experiments to verify the scientific hypothesis.

With twelve granted patents, we contribute to the flight training community where their claims prove the additions to flight training systems for the purpose of solving a specific problem that meet the criteria of novelty and of non-obviousness. They are verified by one or more expert patent examiners.

Figure 9.1 presents the mapping between the patents and the publications and the supporting instrumentation. The scientific experimentation is using an innovative instrumentation composed of the training support systems and the flight simulation systems as a platform supporting the execution of the flight scenarios in the experimentation protocol and the analysis of the results.

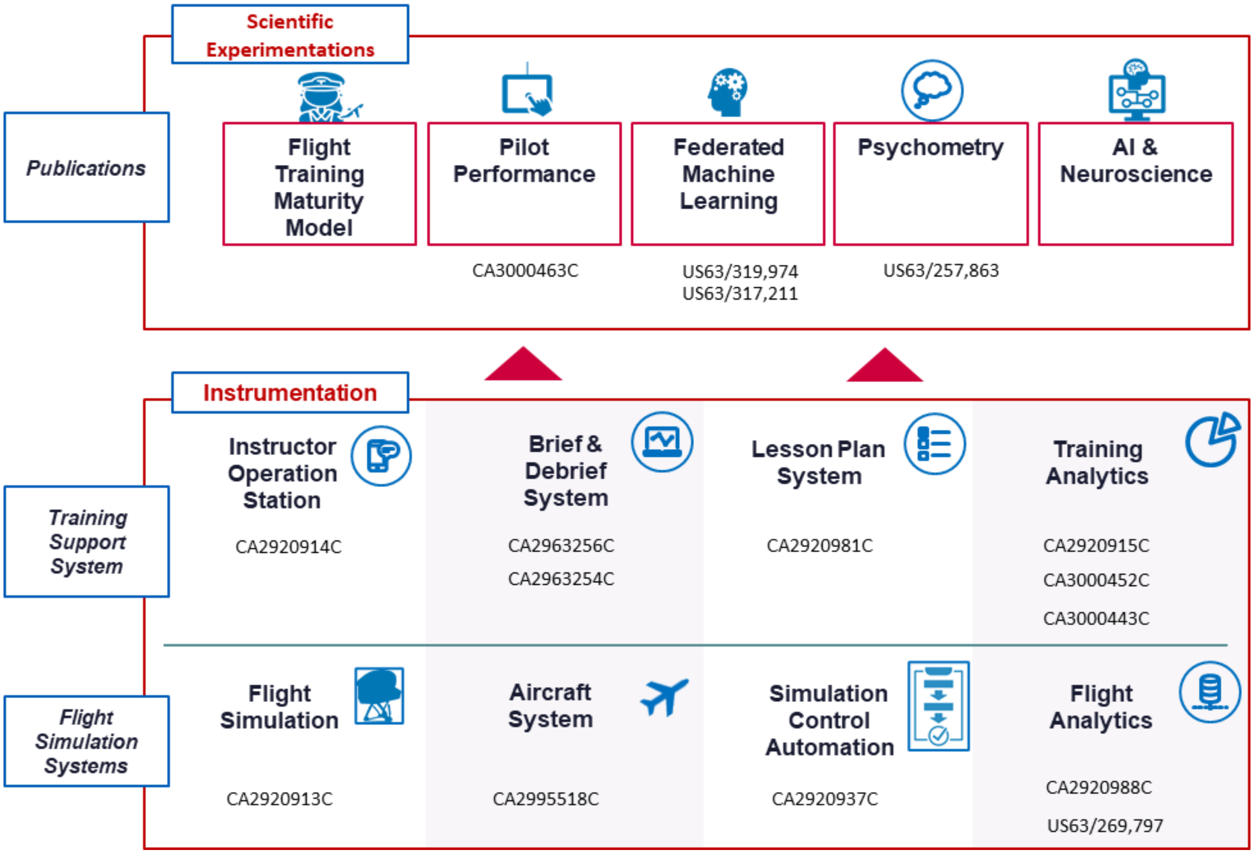


Figure 9.1 - Mapping of the patents with the scientific experimentation

9.1 Just-in time visualization in immersive flight training device patents.

A simulation server capable of transmitting a visual alarm representative of a simulation event discrepancy to a computing device (Delisle & Malo, 2017). The simulation server stores lesson plans comprising at least one event; each event comprising at least one rule. The simulation server executes a simulation according to a particular lesson plan and transmits a visual representation of the executed simulation to a computing device.

A simulation server capable of transmitting a visual prediction indicator representative of a predicted simulation event discrepancy (Delisle, 2017b). The simulation server collects simulation data representative of the executed simulation, processes the simulation data, and compares a simulation value of the particular event with the corresponding prediction metric. When the prediction metric is met, the simulation server transmits information for displaying in the visual representation of the executed simulation a visual prediction indicator representative of the prediction metric being met to the computing device.

Portable computing device and method for transmitting instructor operating station (IOS) filtered information (Delisle, 2017a). A portable computing device receives IOS control and monitoring data from a simulation server, displays the IOS control and monitoring data on the portable computing device, and receives a selection by a user of at least one component of the displayed IOS control and monitoring data. The selection is performed by an interaction of the user with the displayed IOS control and monitoring data.

Contextual monitoring perspective selection during training session (Delisle, 2018a). Monitoring a training session from a trainee in an interactive computer simulation system. During the training session, while the trainee performs actions in an interactive computer simulation station on one or more tangible instruments thereof for controlling a virtual simulated element, dynamic data is logged related to the actions of the trainee. At a monitoring station of the interactive computer simulation system and during the training session, a graphical user interface is displayed depicting a contextual scene related to the interactive computer simulation from a first point of view and detecting a predetermined event in the dynamic data during the training session.

Perspective selection for a debriefing scene (Delisle, 2018b). Debriefing a session from a user in a system. A graphical user interface depicting a debriefing scene, related to the session, is displayed from a first point of view starting at a first time within the session timeline.

Visualizing sub-systems of a virtual simulated element in an interactive computer simulation system (Delisle, 2020). Method and system for visualizing dynamic virtual sub-systems of a virtual simulated element in an interactive computer simulation system comprising a computer generated environment.

Simulation server capable of interacting with a plurality of simulators to perform a plurality of simulations (Delisle, 2018d). The simulation server comprises a communication interface for exchanging data with other entities. The processing unit also generates simulator simulation data and transmits the simulator simulation data to at least one simulator via the communication interface. The processing unit further processes the simulator interaction data and controls the execution of the at least one simulation based on the processed simulator interaction data.

9.2 Training Performance Patents

A simulation server capable of configuring events of a lesson plan through interactions with a computing device (Delisle et al., 2018). The simulation server receives a lesson plan selection from the computing device. The simulation server further receives from the computing device, a selection of at least one event to be used for the selected lesson plan with a configuration of the at least one rule for each selected event.

A simulation server capable of creating events of a lesson plan based on simulation data statistics (Delisle, 2018c). The processing unit creates at least one event having at least one rule based on the simulation data statistics. The system provides at least one rule consisting in at least one measurable value to be measured by at least one of the simulation functionalities.

Assessing a training activity performed by a user in an interactive computer simulation (Delisle et al., 2019a). An interactive computer-based training system, station and method for assessing a training activity performed by a user interacting with tangible instruments for controlling the virtual element in an interactive computer simulation. The processor module computes the plurality of performance metric datasets to identify actual maneuvers of the virtual element during the training activity, identifies one or more failed actual maneuvers of the virtual element during the training activity against corresponding ones of the expected maneuvers and performs computational regression on the actual maneuvers of the virtual element compared to the expected maneuvers of the virtual element to identify one or more root causes of the failed actual maneuvers.

Standard operating procedures feedback during an interactive computer simulation (Delisle et al., 2019b). An interactive computer simulation system, station and method for training a user in the performance of a task through a training activity. During execution of the interactive computer simulation, in the plurality of performance metric datasets, a plurality of actual maneuvers of the virtual element are detected during the training activity, one or more standard operating procedures (SOPs) are identified for the training activity from a plurality of the individually detected actual maneuvers.

Performance metrics in an interactive computer simulation (Delisle et al., 2021). A simulation mapping system and method for determining a plurality of performance metric values in relation to a training activity performed by a user in an interactive computer simulation. The processor module constructs a dataset corresponding to the plurality of performance metric values from the dynamic data having a target time step by synchronizing dynamic data and by inferring, for at least one missing dynamic subsystems of the plurality of dynamic subsystems missing from the dynamic data, a new set of data into the dataset from dynamic data associated to one or more co-related dynamic subsystems.

Method and system for generating vehicle parameters for training a user to score a vehicle maneuver (Chaouachi & Delisle, 2022). There is described a method for training a user to score a vehicle maneuver using one of a machine learning models and a rule-based model, vehicle parameters based on a target grade and a vehicle maneuver, the target grade being indicative of a given performance level of a pilot performing the vehicle maneuver.

System and method for predicting performance by clustering psychometric data using artificial intelligence (Jean-Francois Delisle et al., 2022). A system for predicting performance of a student based on psychometric data includes data storage devices for storing psychometric data for students obtained via psychometric tests.

9.3 Adaptive Flight Training System patents

Federated machine learning in adaptive training systems (Jean-François Delisle et al., 2022). A federated machine learning system for training students comprises a first adaptive training system having a first artificial intelligence module for adapting individualized training to a first group of students and for developing a first learning model based on a first set of learning performance metrics for the first group of students.

Adaptive learning in a diverse learning ecosystem (Delisle & Qi, 2022). An adaptive learning artificial intelligence (ALAI) module receives student performance data to adapt training of the student. The ALAI module comprises a learner profile module that profiles the student, a training task recommendation module that generates AI-generated recommendations, and an explainability and pedagogical intervention module.

CHAPTER 10 CONCLUSION & RECOMMENDATIONS

Given the depth of human complexity, there is more than one area of science, which in combination with data science, can be used to explain human behavior in complex flight operations requiring a high degree of safety and security.

The objective of this research project was to discover the capabilities required to support adaptive flight training, to assess where artificial intelligence can be used within a flight training process and to assess the availability and quality of the data produced during flight training operations.

The flight training industry is embarking on a major shift towards CBTA for pilot training, where artificial intelligence and data science are at the heart of this new trend. According to our research and analysis conducted so far, artificial intelligence plays a key role in analyzing performance and adapting training, whether to promote excellent performance or simply to avoid poor performance affecting flight safety.

The capability model approach could provide a framework for integrating AI models into a businesslike process for pilot training programs. With the ADDIE and CFITES models present in training centers for the creation and delivery of training content, we can enable automation of the flight training process subject to a competency framework of aviation training.

One uncertainty we sought to eliminate was the current subjectivity of the pilot's flight performance using a wide variety of data available to us. Since the notion of performance is complex and multidimensional, different statistical approaches have been used to adapt the methods to several training centers at different levels of data security constraints. Our working hypothesis was that flight performance data combined with a variety of the pilot's profile data can improve the predictability of success based on multiple facets of training programs and recommend learning content. We also hypothesized that machine learning will improve the quality of the performance assessment, enable both real-time and non real-time adaptation and provide flight training operators explainable, robust and certifiable results while considering instructor biases in a human-in-the-loop system.

We started by collecting training center data on historical flight performance and we analyzed the data based on their quality, quantity and accessibility while preserving data privacy and General Data Protection Regulation (GDPR) compliancy during the data collection process. We then identified the additional data sets needed in a data acquisition strategy focused on pilot profile data that was the key data entity of our experimentation protocols executed in class D full flight simulation devices.

We have identified a supervised learning approach to classify the technical performance of pilots and creates a standard performance index, portable from one evaluation scheme to another depending on the nature of the training programs. To address the challenges of performance labeling, we have sometimes used instructor evaluations and automated approaches using rules created with SMEs. We also experimented an approach in which grades are used to segment students into sub-categories based on their historical performance. This approach is used to map the knowledge of subject matter experts about flight performance using statistical methods to easily generalize the method for multiple training programs.

We then made progress on the architecture of the global federated solution and analyzed the results of distributed machine learning model using data from multiple training centers. We were able to demonstrate that machine learning model scalability can be achieved with federated machine learning model portability between multiple training programs.

During this federation, we focused on the study of the use of AI components in an adaptive flight training platform and we formulated a solution hypothesis of a recommender systems. Several advancements were made from this experimental work including a platform architecture specific to virtual reality flight training and the identification of a list of AI capabilities that enable an adaptive flight training solution. We had to thoroughly analyze the initial pilot performance predictor to ensure that it can be used as input into recommender systems that can recommend additional maneuvers to add to the training plan, automatically adapt a lesson plan in real-time, recommend personalized training content to increase the pilot's engagement and recommend eLearning of computer-based training courses to maximize the learning experience.

We then experimented with an approach in which students are segmented into groups/profiles based on historical performance and aptitude test results. This pilot segmentation process used a data-driven clustering algorithm approach to create student profiles and identify each profile's pattern that can potentially provide actionable recommendations to instructors and training managers. We were able to test different clustering methods, including Agglomerative Clustering, that provided us good results. We used the SHAP methodology for cluster interpretation to provide us a high level of explainability, an important aspect of this research project.

An intelligent adaptive flight training system can provide autonomy to pilots in a self-learning paradigm by giving them tools to access their profile, performance metrics and recommendations. By examining the research questions, we claim that we improved their self-understanding of training performance with psychophysiological inferences. We have determined that statistical approaches are indicative of cognitive performance in initial learning and that subjective NASA-TLX self-report approaches perform significantly well to assess flight training experience executed in an immersive flight simulation device by a cohort of novice participants.

Overall, we testify that to get results in an industrial context as this research was carried out, an industrial engineering approach was necessary. By approach, we mean using a complete suite of processes of quality management, configuration management, data management, and software and system engineering. In the worldwide large-scale context where the AI capability has been deployed and used by certified pilots and instructors, most of the effort lies in the data management and software solutions implementation, including significant testing and deployment phase.

10.1 **Future work**

In future research, the student's journey within all learning environments will be analyzed and we will explore a personalized student micro-learning approach based on the learner profile and preferences. With a complete learning workflow optimization that include virtual classroom, Virtual/Mixed-Reality (VR/MR) devices, Computer-Based Training (CBT) application, Integrated Procedure Trainer (IPT) and Full Flight Simulators (FFS), we expect being able to provide new analytics capability such as Root Cause Analysis (RCA) of flight performance in a closed-loop process between the training and the flight operations.

Affective and cognitive computing will be considered with the goal of augmenting user’s decision-making with the presence of cognitive assistance in real-time operations. The modelization of the flight training actors will continue with the creation of a digital twin of the pilot or the instructor to either create a virtual instructor, a virtual co-pilot, a virtual air traffic controller or a virtual and synthetic pilot, as shown at Figure 10.1. With a robust mental modeling and behavior analysis using biometrics sensors, we can push further our research and provide a framework that will map the

Observeable Behaviors (OBs) of ICAO's standardized competencies with biometric derived key performance indicators (KPIs) that will be a serious enabler of EBT and CBTA.

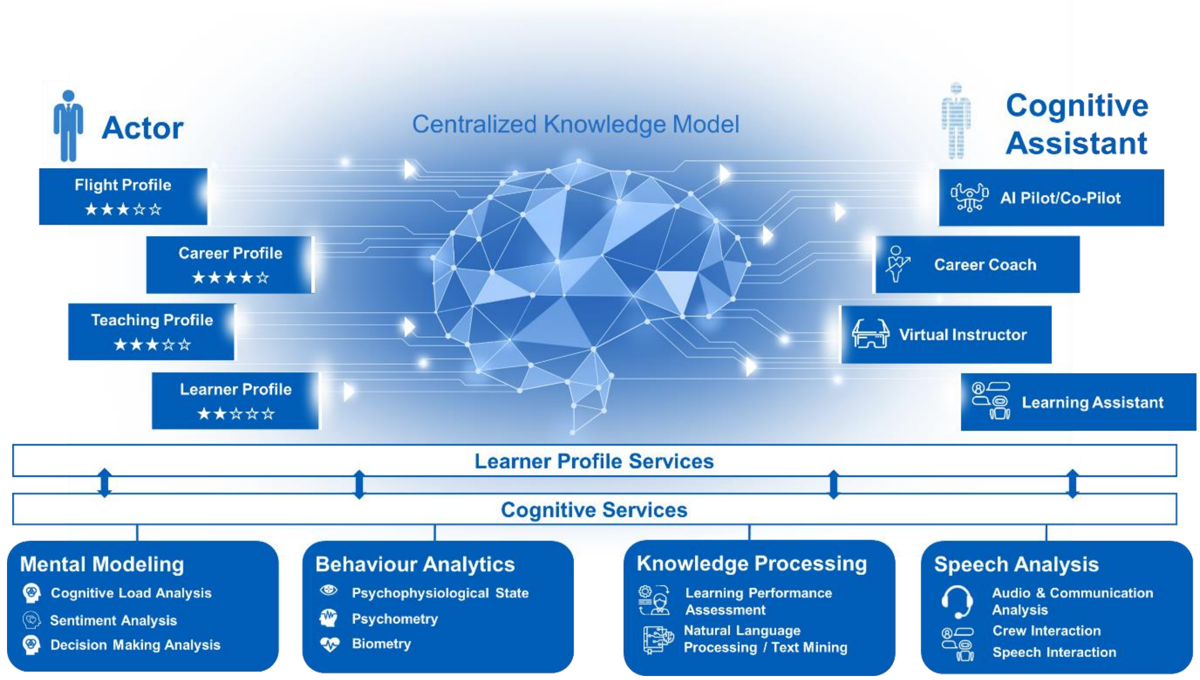


Figure 10.1 - Actor's modelisation of a central knowledge model supporting cognitive assistance

As a new AI capability that will augment the IAFTS, a cognitive assistant with voice interaction and a Natural Language Processing (NLP) layer can provide an advanced immersive Human-Machine Interface (HMI) for flight training device. This kind of HMI can provide natural interaction considering the Man-Machine-Teaming (MMT) aspects in distributed situation awareness and decision-making context of future aircrafts cockpit.

With the upcoming arrival of electric Vertical Takeoff and Landing (eVTOL) vehicle, new challenges are expected. While operating a single-seat vehicle through a stimulating urban environment, the presence of new system automations may require different skills to master for the pilots. Urban environment offers a multitude of new elements to consider such as pedestrian, ground vehicle, buildings and infrastructure in a new regulated airspace. By exploring the use of a Multi-Agent with Reinforcement Learning (MARL) AI capability in immersive simulation, we may be able to provide a rich and realistic synthetic urban environment as a training platform for eVTOL flight training. This innovation in the aerospace industry will represent a new technological transformation itself, on the road of the democratisation of flight training, in a new Open Flight paradigm.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*. Paper presented at the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Antoine Bechara, A. R. D., Hanna Damasio, Steven W.Anderson. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition,, 50*(1-3).

Aydogan Savran, R. T., Yasar Becerikli. (2006). Intelligent adaptive nonlinear flight control for a high performance aircraft with neural networks. *The Instrumentation, Systems and Automation Society Transaction, 45*(2).

Balakrishna, P., Ganesan, R., & Sherry, L. (2008). *Estimating Taxi-out Times with a Reinforcement Learning Algorithm*. Paper presented at the Digital Avionics Systems.

Baldwin, C. L., & Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *Neuroimage, 59*(1), 48-56. doi:10.1016/j.neuroimage.2011.07.047

Barron, L. G., Carretta, T. R., & Rose, M. R. (2017). Aptitude and Trait Predictors of Manned and Unmanned Aircraft Pilot Job Performance. *Military Psychology, 28*(2), 65-77. doi:10.1037/mil0000109

Bass, E. J. (1998). *Architecture for an Intelligent Instructor Pilot Decision Support System*. Paper presented at the EEE International Conference on Systems, Man, and Cybernetics.

Bekolay, T., Kolbeck, C., & Eliasmith, C. (2013). *Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks*. Paper presented at the 35th Annual Conference of the Cognitive Science Society.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*.

Binias, B., Myszor, D., & Cyran, K. A. (2018). A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. *Comput Intell Neurosci, 2018*, 2703513. doi:10.1155/2018/2703513

Binias, B., Myszor, D., Palus, H., & Cyran, K. A. (2020). Prediction of Pilot's Reaction Time Based on EEG Signals. *Front Neuroinform, 14*, 6. doi:10.3389/fninf.2020.00006

Blake C.W. Martin, C. H., Bernadette Dececchi, James Kerry, Stuart C. Grant. (2016). The Implementation of Canadian Forces Individual Training and Education System and Mission Essential Competency Training Analysis Methods, A Case Study.

Bor, R., Eriksen, C., Hubbard, T. P., & King, R. (2020). *Pilot Selection. Psychological Principles and Practive* (C. Press Ed.): CRC Press.

Bosoni, P., Tucker, A., Bellazzi, R., Nihtyanova, S. I., & Denton, C. P. (2016). *Combining Unsupervised and Supervised Learning for Discovering Disease Subclasses*. Paper presented at the 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS).

Branaghan, R. J., Covas-Smith, C. M., Jackson, K. D., & Eidman, C. (2011). Using knowledge structures to redesign an instructor-operator station. *Appl Ergon, 42*(6), 934-940. doi:10.1016/j.apergo.2011.03.002

Brown, J. E., Waag, W. L., & Eddowes, E. E. (1975). *USAF Evaluation of an Automated Adaptive Flight Training System*. Retrieved from

Bryan Matthews, S. D., Kanishka Bhaduri, Kamalika Das, Rodney Martin, Nikunj Oza, Ashok N. Srivastava, John Stutz. (2013). Discovering Anomalous Aviation Safety Events using Scalable Data Mining Algorithms. *Journal of Aerospace Information Systems, 10*(10).

Butcher, J. N. (1994). Psychological assessment of airline pilot applicants with the MMPI-2. *Journal of Personality Assessment, 62*(1), 31-44. doi:10.1207/s15327752jpa6201_4

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI-2. An administrative and interpretive guide*. Minneapolis: University of Minnesota Press.

Cabestrero, R., Crespo, A., & Quiros, P. (2009). Pupillary dilation as an index of task demands. *Percept Mot Skills, 109*(3), 664-678. doi:10.2466/pms.109.3.664-678

CAE. (2017). *Airline Pilot Demand Outlook 10-year view*. Retrieved from Canada:

Causse, M., Dehais, F., & Pastor, J. (2011). Executive Functions and Pilot Characteristics Predict Flight Simulator Performance in General Aviation Pilots. *The International Journal of Aviation Psychology, 21*(3), 217-234. doi:10.1080/10508414.2011.582441

Cercone, N., An, A., & Chan, C. (1999). Rule-Induction and Case-Based Reasoning Hybrid Architectures Appear Advantageous. *IEEE Transactions on Knowledge and Data Engineering, 11*(1).

Chang, M., Hwang, W.-Y., Chen, M.-P., & Muller, W. (2011). Edutainment technologies Educational Games and Virtual Reality/Augmented Reality Applications. In *6th International Conference on E-learning and Games, Edutainment 2011*. Taipei, Tawan.

Chaouachi, M., & Delisle, J.-F. (2022). United States Patent and Trademark Office.

Chen, F., Luo, M., Dong, Z., Li, Z., & He, X. (2018). Federated Meta-Learning with Fast Convergence and Efficient Communication. *arXiv:1802.07876*.

Chen, S., Xu, D.-L., & Jiang, W. (2020). *High value passenger identification research based on Federated Learning*. Paper presented at the 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC).

Chidester, T. R., Helmreich, R. L., Gregorich, S. E., & Geis, C. E. (1991). Pilot personality and crew coordination: implications for training and selection. *The International Journal of Aviation Psychology, 1*(1), 25-44. doi:10.1207/s15327108ijap0101_3

Chieu, V. M., & Herbst, P. (2010). Designing an intelligent teaching simulator for learning to teach by practicing. *ZDM Mathematics Education*.

Chrissis, M. B., Konrad, M., & Shrum, S. (2011). *CMMI for Development: Guidelines for Process Integration and Product Improvement*: Addison-Wesley Professional.

Chu, E., Gorinevsky, D., & Boyd, S. P. (2010). Detecting Aircraft Performance Anomalies from Cruise Flight Data. *Proceedings AIAA Infotech@Aerospace*.

Cox, C., Lewis, C., & Suchomel, C. (2000). *A Neural Network Based, Real-Time Algorithm for Detection and Mitigation of Pilot Induced Oscillations*. Paper presented at the IEEE International Conference on System, Man, Cybernetics, Nashville, USA.

D'Angelo, G., Tipaldi, M., Glielmo, L., & Rampone, S. (2017). Spacecraft Autonomy modeled via Markov Decision Process and Associative Rule-based Machine Learning. *IEEE International Workshop on Metrology for AeroSpace*.

Delisle, J.-F. (2017a). Canada Patent No. CA2920914C, US10395550B2, CN108701426A, WO2017139879A1. Canadian Intelectual Property Office: C. Inc.

Delisle, J.-F. (2017b). Canada Patent No. CA2920915C, WO2017139878A1. Canadian Intelectual Property Office: C. Inc.

Delisle, J.-F. (2018a). Canada Patent No. CA2963254C. Canadian Intelectual Property Office.

Delisle, J.-F. (2018b). Canada Patent No. CA2963256C / US20180232133A1. Canadian Intelectual Property Office.

Delisle, J.-F. (2018c). United States Patent No. CA2920981C, US10679513B2, WO2017139876A1, CN108701424B. United States Patent and Trademark Office.

Delisle, J.-F. (2018d). Canada Patent No. CA2920913C, US20200357294A1, WO2017139880A1. Canadian Intelectual Property Office.

Delisle, J.-F. (2020). CA2995518C, CN110462709A, WO2018148818A1, EP3574489A4. World Intellectual Property Organization.

Delisle, J.-F., Dufour, A., Proulx, M.-A., & Ho, D. T. (2019a). United States Patent No. US10957216B2, CA3000452C. United States Patent and Trademark Office.

Delisle, J.-F., Dufour, A., Proulx, M.-A., & Ho, D. T. (2019b). Canada Patent No. CA3000443C, US20190304325A1, EP3547289A1, CN110322098A. Canadian Intelectual Property Office.

Delisle, J.-F., Dufour, A., Proulx, M.-A., & Ho, D. T. (2021). United States Patent No. CA3000463C, US10991262B2, EP3547290A1, CN110322104A. United States Patent and Trademark Office.

Delisle, J.-F., Gingras, L., Bowness, D., & Ho, D. T. (2018). Canada Patent No. CA2920937, WO2017139875A1, US20200050720A1. Canadian Intelectual Property Office.

Delisle, J.-F., & Malo, S. (2017). Canada Patent No. CA2920988C, WO2017139877A1, US20200066178A1. Canadian Intelectual Property Office.

Delisle, J.-F., Naranth, B. B., & Dufour, A. (2022). United States Patent and Trademark Office.

Delisle, J.-F., & Qi, J. (2022). United States Patent and Trademark Office.

Delisle, J.-F., Winokur, B., & Singh, N. (2022). United States Patent and Trademark Office.

Dennis Drinka, M. Y.-M. Y. (2008). *Controlling Curriculum Redesign with a Process Improvement Model*.

Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., . . . Weber, B. (2012). On the Use of Neurophysiological Tools in IS Research: Developing a Research Agenda for NeuroIS. *MIS Quarterly, 36*(3).

Dong, B., Xie, S., Gao, J., Fan, W., & Yu, P. S. (2015). *OnlineCM Real-time Consensus Classifcation with Missing Values*. Paper presented at the Proceedings of the 2015 SIAM International Conference on Data Mining.

Earl, M. G., & D'Andrea, R. (2002). *Modeling and control of a multi-agent system using mixed integer linear programming.pdf>*. Paper presented at the proceedings of the 41st conference on decision and control, Las Vegas.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science, 338*(6111), 1202-1205. doi:10.1126/science.1225266

Eric Chu, D. G. (2010a). Detecting Aircraft Performance Anomalies from Cruise Flight Data. *Proceedings AIAA*.

Eric Chu, D. G. (2010b). Detecting Aircraft Performance Anomalies from Cruise Flight Data. *Proceedings AIAA Infotech@Aerospace*.

FAA. (2016). *Pilot's Handbook of Aeronautical Knowledge*. (FAA-H-8083-258). Federal Aviation Administration

FAA. (2018). FAA Training & Testing. Retrieved from https://www.faa.gov/training_testing/

Fabrizio Carcillo, Y.-A. L. B., Olivier Caelen, Gianluca Bontempi. (2017). *An Assessment of Streaming Active Learning Strategies for Real-Life Credit Card Fraud Detection*. Paper presented at the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Fairclough, S. H. (2008). Fundamentals of physiological computing. *Interacting with Computers, 21*(1-2).

Fischetti, M., & Jo, J. (2017). Deep Neural Networks as 0-1 Mixed Integer Linear Programs: A Feasibility Study. In.

Gartner Group, K. H. (2004). *Use a Maturity Model to Make the Most of E-Learning*. Retrieved from

Gibson, D., & de Freitas, S. (2015). Exploratory Analysis in Learning Analytics. *Technology, Knowledge and Learning, 21*(1), 5-19. doi:10.1007/s10758-015-9249-5

Glorot, X., & Bengio, Y. (2010). *Understanding the difficulty of training deep feedforward neural networks*. Paper presented at the Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.

GOV.UK. (2017). *Defence direction and guidance for training and education(JSP 822)*. (JSP 822).

Guan, X., Lv, R., Sun, L., & Liu, Y. (2016). *A study of 4D trajectory prediction based on machine deep learning*. Paper presented at the 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China.

Harper, K. C., Kuanchin & Yen, David. (2004). Distance learning, virtual classrooms, and teaching pedagogy in the Internet environment. *Technology in Society, 26*(26), 585-598.

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction* (Second edition ed.): Springer.

Hatami, N., Gavet, Y., & Debayle, J. (2017). *Classification of Time-Series Images Using Deep Convolutional Neural Networks*. Paper presented at the The 10th International Conference on Machine Vision.

Hedge, J. W., Bruskiewicz, K. T., Borman, W. C., Hanson, M. A., Logan, K. K., & Siem, F. M. (2000). Selecting pilots with crew resource management skills. *The International Journal of Aviation Psychology, 10*(4), 377-392. doi:10.1207/S15327108IJAP1004_5

Husken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing, 50*.

IATA. (2013). *Evidence-Based Training Implementation Guide*. In IATA (Ed.).

Ibrahim, L. (2000). *Using an Integrated Capability Maturity Model®–The FAA Experience.* Paper presented at the Proceedings of the Tenth Annual International Symposium of the International Council on Systems Engineering (INCOSE).

ICAO. (2013). *Manual of Evidence-based Training* (Vol. Doc 9995 AN/497): International Civil Aviation Organization.

Ippolito, C., Yeh, Y.-H., & Kaneshige, J. (2007). *Neural Adaptive Flight Control Testing on an Unmanned Experimental Aerial Vehicle*. Paper presented at the AIAA 2007 Conference and Exhibit.

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). Machine Learning Based Student Grade Prediction (A Case Study). *arXiv*(arXiv:1708.08744).

Jacobs, R. S. (2014). *Optimization of daily flight training schedules.* Naval Postgraduate School,

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*: Springer.

Jens Pöppelbuß, M. R. (2011). *What Makes a Useful Maturity Model? A Framework of General Design Principles for Maturity Models and its Demonstration in Business Process Management.* Paper presented at the European Conference on Information Systems.

Jensen, R. S. (1997). The Boundaries of Aviation Psychology, Human Factors, Aeronautical Decision Making, Situation Awareness, and Crew Resource Management. *The International Journal of Aviation Psychology, 7*(4), 259-267. doi:10.1207/s15327108ijap0704_1

Jiefeng Xu, M. S., Mike McCleery, T. Glenn Bailey. (2006). A dynamic neighborhood based tabu search algorithm for real-world flight instructor scheduling problems. *European Journal of Operational Research, 169*(3), 978-993. doi:10.1016/j.ejor.2004.08.023

Jing, G., Feng, L., Wei, F., Yizhou, S., & Jiawei, H. (2013). A Graph-Based Consensus Maximization Approach for Combining Multiple Supervised and Unsupervised Models. *IEEE Transactions on Knowledge and Data Engineering, 25*(1), 15-28. doi:10.1109/tkde.2011.206

Jr., F. J. I., & Rubin, P. A. (2003). Feature Selection for Multiclass Discrimination via Mixed-Integer Linear Programming. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 25*(6).

Kelley, C. R. (1969). What Is Adaptive Training. *Human Factors, 11*(6).

Khalil, E. B. (2016). Machine Learning for Integer Programming. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

King, R. E. (2014). Personality (and Psychopathology) Assessment in the Selection of Pilots. *The International Journal of Aviation Psychology, 24*(1), 61-73. doi:10.1080/10508414.2014.860844

Kulkarni, V. B. (2015). *Intelligent air traffic controller simulation using artificial neural networks*. Paper presented at the International Conference on Industrial Instrumentation and Control (ICIC).

Lahmeri, M.-A., A.Kishk, M., & Alouini, M.-S. (2021). Artificial Intelligence for UAV-enabled Wireless Networks A Survey. *arXiv:2009.11522*.

Lan, Z., Sourina, O., Wang, L., Scherer, R., & Muller-Putz, G. R. (2019). Domain Adaptation Techniques for EEG-Based Emotion Recognition: A Comparative Study on Two Public Datasets. *IEEE Transactions on Cognitive and Developmental Systems, 11*(1), 85-94. doi:10.1109/tcds.2018.2826840

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*(2).

Leland McInnes, J. H., James Melville. (2018). UMAP Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*(e-prints 1802.03426).

LeVie, L. R. (2016). *Survey of Quantitative Research Metrics to Assess Pilot Performance in Upset Recovery*. Retrieved from

Liu, Y., Lan, Z., Traspsilawati, F., Sourina, O., Chen, C.-H., & Muller-Wittig, W. (2019). *EEG-Based Human Factors Evaluation of Air Traffic Control Operators (ATCOs) for Optimal Training*. Paper presented at the 2019 International Conference on Cyberworlds (CW).

Lodi, A. (2017). *Lecture Notes in Integer Programming* [Version 3.5].

Lodi, A., Belotti, P., Bonami, P., Fischetti, M., Monaci, M., Nogales-Gómez, A., & Salvagnin, D. (2016). On handling indicator constraints in mixed integer programming. *Computational Optimization and Applications, 65*(3), 545-566. doi:10.1007/s10589-016-9847-8

Lodi, A., Dell'Amico, M., & Maffioli, F. (1999). Solution of the Cumulative Assignment Problem With a Well-Structured Tabu Search Method. *Journal of Heuristics*.

Loh, C. S., & Li, I. H. (2016). *Using Players' Gameplay Action-Decision Profiles to Prescribe Training: Reducing Training Costs with Serious Games Analytics*. Paper presented at the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA).

Ludwig, J., & Ramachandran, S. (2005). *Developing an Adaptive Intelligent Flight Trainer*. Retrieved from

Lutteroth, C., Luxton-Reilly, A., Dobbie, G., & Hamer, J. (2007). *A maturity model for computing education*. Paper presented at the Proceedings of the ninth Australasian conference on Computing education - Volume 66, Ballarat, Victoria, Australia.

Luxhøj, J. T. (2013). Predictive Safety Analytics for Complex Aerospace Systems. *Procedia Computer Science, 20*, 331-336. doi:10.1016/j.procs.2013.09.281

Maiga Chang, W.-Y. H., Ming-Puu Chen, Wolfgang Muller. (2011). Edutainment technologies Educational Games and Virtual Reality/Augmented Reality Applications. In *6th International Conference on E-learning and Games, Edutainment 2011*. Taipei, Tawan.

Maldonado, S., Pérez, J., Weber, R., & Labbé, M. (2014). Feature selection for Support Vector Machines via Mixed Integer Linear Programming. *Information Sciences, 279*, 163-175. doi:10.1016/j.ins.2014.03.110

Maria E. Chaparro, M. C., Shem Malmquist. (2020). Personality Trends in the Pilot Population. *Collegiate Aviation Review International, 38*(2).

Marshall S., M. G. (2018). *An e-learning maturity model*.

Marshall, S. P. (2000). 6090051. U. S. Patent.

Marshall, S. P. (2002). *The Index of Cognitive Activity Measuring Cognitive Workload*. Paper presented at the IEEE 7' Human Factors Meeting, Scottsdale Arizona.

Matthews, B., Das, S., Bhaduri, K., Das, K., Martin, R., Oza, N., . . . Stutz, J. (2013). Discovering Anomalous Aviation Safety Events using Scalable Data Mining Algorithms. *Journal of Aerospace Information Systems, 10*(10).

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. Paper presented at the 20th International Conference on Artificial Intelligence and Statistics.

Meng, Y. (2021). *Machine Learning for Aviation Data.* (Master of Science 9). Trent University,

Ming Hou, C. M. F. (2014). Untangling Operator Monitoring Appreaches When designing intellignet adaptive systems p804910_A1b.pdf>. *Defence Research and Development Canada*, 147.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology, 41*(1), 49-100. doi:https://doi.org/10.1006/cogp.1999.0734

Miyashita, S., Lian, X., Zeng, X., Matsubara, T., & Uehara, K. (2017). *Developing Game AI Agent Behaving Like Human by Mixing Reinforcement Learning and Supervised Learning*. Paper presented at the International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Kanazawa, Japan.

Monteiro, T. G., Skourup, C., & Zhang, H. (2019). Using EEG for Mental Fatigue Assessment: A Comprehensive Look Into the Current State of the Art. *IEEE Transactions on Human-Machine Systems, 49*(6), 599-610. doi:10.1109/thms.2019.2938156

Neill, C. (2003). *The Vision For RAAF Learning In The Year 2020*. Retrieved from

Neuhauser, C. (2004). A maturity model: does it provide a path for online course design? *The Journal of Interactive Online Learning, 3*(1).

Nicholas Wilson, H. T. G., Jessica VanBree, Bradley Hoffman, Kouhyar Tavakolian. (2021). *Identifying Opportunities for Augmented Cognition During Live Flight Scenario An*

Analysis of Pilot Mental Workload Using EEG. Paper presented at the International Symposium on Aviation Psychology.

Oladimeji, M. O., Turkey, M., Ghavami, M., & Dudley, S. (2015). *A New Approach for Event Detection using k-means Clustering and Neural Networks*. Paper presented at the International Joint Conference on Neural Networks (IJCNN).

Ortiz de Guinea, A., Titah, R., & Léger, P.-M. (2013). Measure for Measure: A two study multi-trait multi-method investigation of construct validity in IS research. *Computers in Human Behavior, 29*(3), 833-844. doi:10.1016/j.chb.2012.12.009

Oza, N. C., Tumer, K., Tumer, I. Y., & Huff, E. M. (2003). *Classification of Aircraft Maneuvers for Fault Detection*. Paper presented at the International Workshop on Multiple Classifier Systems.

Papamitsiou, Z., & Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*.

Parasuraman, R. (2003). Neuroergonomics: Research and practice. *Theoretical Issues in Ergonomics Science, 4*(1-2).

Patrick Corsi, E. N. (2015). *Innovation Capability Maturity Model*: John Wiley & Sons.

Paul W. Caro, J. (1969). Adaptive Training-An Application to Flight Simulation. *Human Factors, 11*(6).

Picard, R. W. (2010). Affective computing: from laughter to IEEE. *IEEE Transactions on Affective Computing, 1*(1), 11-17.

Pino, F. J., Baldassarre, M. T., Piattini, M., & Visaggio, G. (2009). Harmonizing maturity levels from CMMI-DEV and ISO/IEC 15504. *Journal of Software Maintenance and Evolution: Research and Practice*.

Poos, A. M., Maicher, A., Dieckmann, A. K., Oswald, M., Eils, R., Kupiec, M., . . . Konig, R. (2016). Mixed Integer Linear Programming based machine learning approach identifies regulators of telomerase in yeast. *Nucleic Acids Res, 44*(10), e93. doi:10.1093/nar/gkw111

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications, 97*, 205-227. doi:10.1016/j.eswa.2017.12.020

R. L. Helmreich, J. T. S. (1978). The work and family orientation questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. *JSAS Catalog of Selected Documents in Psychology,, 8*(35).

Rantanen, E. M., Talleur, D. A., Taylor, H. L., Bradshaw, G. L., Tom W. Emanuel, J., Lendrum, L., & Hulin, C. L. (2007). Derivation of pilot performance measures from flight data recorder information. *IEEE Aerospace and Electronic Systems Magazine*.

Rechkoski, L., Ajanovski, V. V., & Mihova, M. (2018). *Evaluation of Grade Prediction using Model-Based Collaborative Filtering methods*. Paper presented at the IEEE Global Engineering Education Conference (EDUCON).

Remolina, E., Ramachandran, S., Fu, D., Stottler, R., & Howse, W. R. (2004). *Intelligent Simulation-Based Tutor for Flight Training*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender Systems Handbook*: Springer.

Richards, A., & How, J. P. (2002). *Aircraft Trajectory Planning With Collision Avoidance Using Mixed Integer Linear Programming*. Paper presented at the Proceedings of the American Control Conference, Anchorage, AK.

Riedl, R., Davis, F. D., & Hevner, A. R. (2014). Towards a NeuroIS research methodology: intensifying the discussion on methods, tools, and measurement. *Journal of the Association for Information Systems, 15*(10).

Riedl, R., & Léger, P.-M. (2016). Fundamentals of NeuroIS: Information Systems and the Brain. In: Springer.

Schnell, T., Keller, M., & Poolman, P. (2008). Quality of training effectiveness assessment (QTEA); a neurophysiologically based method to enhance flight training. *IEEE/AIAA 27th Digital Avionics Systems Conference*.

Scott M. Lundberg, S.-I. L. (2017). *A Unified Approach to Interpreting Model Predictions*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Settles, B. (2012). *Active Learning*.

Sharma, P. K., Park, J. H., & Cho, K. (2020). Blockchain and federated learning-based distributed computing defence framework for sustainable society. *Sustainable Cities and Society, 59*. doi:10.1016/j.scs.2020.102220

Slavin, R. E. (2006). Research on Cooperative Learning: an international perspective. *Scandinavian Journal of Educational Research, 33*(4), 231-243.

Smith-Jentsch, K. A., Jentsch, F. G., Payne, S. C., & Salas, E. (1996). Can pretraining experiences explain individual differences in learning? *Journal of Applied Psychology, 81*(1), 110-116. doi:10.1037/0021-9010.81.1.110

Speciale, P., Paudel, D. P., Oswald, M. R., Kroeger, T., Gool, L. V., & Pollefeys, M. (2017). *Consensus Maximization with Linear Matrix Inequality Constraints*. Paper presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA.

Spence, J. T., Helmreich, R. L., & Holahan, C. K. (1979). Negative and positive components of psychological masculinity and femininity and their relationships to self-reports of neurotic and acting out behaviors. *Journal of personality and social psychology, 37*(10), 1673-1682. doi:10.1037/0022-3514.37.10.1673

Stein, E. S. (1984). *The Measurement of Pilot Performance. A Master Journeyman Approach*. Retrieved from

Stevens-Adams, S. M., Basilico, J. D., Abbott, R. G., Gieseler, C. J., & Forsythe, C. (2010). *Performance Assessment to Enhance Training Effectiveness*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)

Strait, M., & Scheutz, M. (2014). What we can and cannot (yet) do with functional near infrared spectroscopy. *Frontiers in Neuroscience, 8*(8 MAY), 1-12. doi:10.3389/fnins.2014.00117

Strother, J. B. (2002). An Assessment of the Effectiveness of e-learning in Corporate Training Programs. *The International Review of Research in Open and Distributed Learning, 3*(1).

Sun, J., Liu, F., Zhou, Y., Gui, G., Ohtsuki, T., Guo, S., & Adachi, F. (2020). Surveillance Plane Aided Air-Ground Integrated Vehicular Networks: Architectures, Applications, and Potential. *IEEE Wireless Communications*, 1-7. doi:10.1109/mwc.001.2000079

Susanne M Jaeggi, M. B., Walter J Perrig, Beat Meier. (2010). The concurrent validity of the N-back task as a working memory measure. *PubMed, 18*(4).

Suzanne K. Kearns, T. J. M., Steven Hodge. (2016). *CompetencyBased Education in Aviation: Exploring Alternate Training Pathways*: Ashgate.

Swigger, K. M. (1988). *Intelligent Tutoring Systems for Air Force Applications*. Paper presented at the Aerospace and Electronics Conference, Dayton, USA.

Thomas C. Hankins, G. F. W. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space and Environmental Medicine*.

Thong Chee Ling, Y. Y. J., Rusli Abdullah, Nor Hayati Alwi. (2012). A Review Study Applying Capability Maturity Model in Curriculum Design Process for Higher Education. *Journal for the Advancement of Science and Arts, 3*(1).

Tupes, E. C., & Christal, R. E. (1961). *Recurrent Personality Factors based on Trait Ratings*. Retrieved from Lackland Airforce Base:

Volkovs, M., Yu, G., & Poutanen, T. (2017). *DropoutNet: Addressing Cold Start in Recommender Systems*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Wagenstein, H. N. (2006). *A capability maturity model for training & education*. Paper presented at the PMI® Global Congress 2006, Madrid, Spain.

Wang, Y., Dong, J., Liu, X., & Zhang, L. (2015). Identification and standardization of maneuvers based upon operational flight data. *Chinese Journal of Aeronautics, 28*(1), 133-140. doi:10.1016/j.cja.2014.12.026

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442. doi:10.1038/30918

Wen-Chin Li, C.-s. Y., Lon-Wen Li, Matthew Greaves. (2014). *Pilots Eye Movement Patterns during Performing Air-to-Air Mission*. Paper presented at the Proceedings of 31st European Association for Aviation Psychology Conference.

Wendler, R. (2012). The maturity of maturity model research: A systematic mapping study. *Information and Software Technology, 54*(12), 1317-1339. doi:10.1016/j.infsof.2012.07.007

Wu, X., Liang, Z., & Wang, J. (2020). FedMed: A Federated Learning Framework for Language Modeling. *Sensors (Basel), 20*(14). doi:10.3390/s20144048

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1-19. doi:10.1145/3298981

Yang, S., Wang, H., Zhang, C., & Gao, Y. (2020). Contextual Bandits With Hidden Features to Online Recommendation via Sparse Interactions. *IEEE Intelligent Systems, 35*(5), 62-72. doi:10.1109/mis.2020.3010298

Ye, C., Yung, N. C., & Wang, D. (2003). A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance. *IEEE Trans Syst Man Cybern B Cybern, 33*(1), 17-27. doi:10.1109/TSMCB.2003.808179

Zhang, F., Chen, D., & Wu, D. (2020). *Analysis of Pilot's Training Effect Based on EEG Signal*. Paper presented at the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT.

Zheng, H., Hu, H., & Han, Z. (2020). Preserving User Privacy for Machine Learning: Local Differential Privacy or Federated Machine Learning? *IEEE Intelligent Systems, 35*(4), 5-14. doi:10.1109/mis.2020.3010335

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). *Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions*. Paper presented at the Proceedings ofthe ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data.

APPENDIX A PATENT PUBLICATIONS

Published at the Canadian Intellectual Property Office (CIPO), United States Patent and Trademark Office (USPTO), China National Intellectual Property Administration(CNIPA), European Patent Office (EPO), World Intellectual Property Organization (WIPO)

Perspective selection for a debriefing scene

Abstract

> Debriefing a session from a user in a system. During the session, while the user performs actions on one or more tangible instruments of the system, dynamic data is logged in relation to the system along a session timeline. The dynamic data covers the actions of the user on tangible instrument(s). A graphical user interface depicting a debriefing scene, related to the session, is displayed from a first point of view starting at a first time within the session timeline. The debriefing scene is generated starting at the first time from at least a first image feed. Upon detection of a predetermined event in the dynamic data at a second time along the session timeline, a second point of view different from the first point of view is defined and the debriefing scene is generated therefrom after the second time using at least a second image feed.

Inventor: Jean-Francois Delisle
Status: Granted 2018-05-01
Patent ID: CA2963256C / US20180232133A1

A simulation server capable of transmitting a visual prediction indicator representative of a predicted simulation event discrepancy(Delisle, 2017b)(Delisle, 2017b)(Delisle, 2017b)(Delisle, 2017b)(Delisle, 2017)(Delisle, 2017b)

Abstract:

> A simulation server capable of transmitting a visual prediction indicator representative of a predicted simulation event discrepancy. The simulation server stores a lesson plan comprising one or more events. A particular event has a corresponding objective consisting in a value, and a corresponding prediction metric consisting in another value. The objective is not met when the prediction metric is met. The simulation server executes a simulation according to the lesson plan, and transmits a visual representation of the executed simulation to a computing device. The simulation server collects simulation data representative of the executed simulation, processes the simulation data, and compares a simulation value of the particular event with the corresponding prediction metric. When the prediction metric is met, the simulation server transmits information for displaying in the visual representation of the executed simulation a visual prediction indicator representative of the prediction metric being met to the computing device.

Inventors: Jean-Francois Delisle
Status: Granted 2017-09-12,
Patent ID : CA2920915C / WO2017139878A1

A simulation server capable of transmitting a visual alarm representative of a simulation event discrepancy to a computing device

Abstract:

> A simulation server capable of transmitting a visual alarm representative of a simulation event discrepancy to a computing device. The simulation server stores lesson plans comprising at least one event; each event comprising at least one rule. The simulation server executes a simulation according to a particular lesson plan, and transmits a visual representation of the executed simulation to a computing device. The simulation server collects simulation data representative of the executed simulation, processes the simulation data, compares the simulation data with the at least one rule of the at least one event of the lesson plan, and determines each rule is met based on the comparison. When a rule is not met, the simulation server transmits information for displaying on a timeline in the visual representation of the executed simulation a visual alarm representative of the event corresponding to the rule not being met to the computing device.

Inventor: Jean-Francois Delisle, Sebastien Malo
Status: Granted 2017-09-12
Patent ID: CA2920988C / WO2017139877A1 / US20200066178A1

Portable computing device and method for transmitting instructor operating station (IOS) filtered information

Abstract

A portable computing device and method for transmitting Instructor Operating Station (IOS) filtered information. A portable computing device receives IOS control and monitoring data from a simulation server, displays the IOS control and monitoring data on the portable computing device, and receives a selection by a user of at least one component of the displayed IOS control and monitoring data. The selection is performed by an interaction of the user with the displayed IOS control and monitoring data. The portable computing device determines IOS filtered information related to the selected at least one component, and transmits the IOS filtered information to a destination computing device. The determination of the IOS filtered information takes into consideration destination user access rights of a destination user. The destination device may be a simulator or a portable computing device, where the destination user performs a simulation session by interacting with the simulation server.

Inventor: Jean-Francois Delisle
Status: Granted 2017-07-18
Patent ID: CA2920914C / US10395550B2 / CN108701426A / WO2017139879A1

Contextual monitoring perspective selection during training session

Abstract

Monitoring a training session from a trainee in an interactive computer simulation system. During the training session, while the trainee performs actions in an interactive computer simulation station on one or more tangible instruments thereof for controlling a virtual simulated element, dynamic data is logged related to the actions of the trainee. At a monitoring station of the interactive computer simulation system and during the training session, a graphical user interface is displayed depicting a contextual scene related to the interactive computer simulation from a first point of view and detecting a predetermined event in the dynamic data during the training session. At the monitoring station, a second point of view is defined different from the first point of view and the contextual scene is generated in the graphical user interface after the predetermined event detection from the second point of view.

Inventor: Jean-Francois Delisle
Status: Granted 2018-05-01
Patent ID: CA2963254C

Visualizing sub-systems of a virtual simulated element in an interactive computer simulation system

Abstract

Method and system for visualizing dynamic virtual sub-systems of a virtual simulated element in an interactive computer simulation system comprising a computer generated environment. One or more tangible instruments control the virtual simulated element in the computer generated environment. A graphical user interface comprising an interactive display portion depicting a rendered view of the virtual simulated element. While an interactive computer simulation of the virtual simulated element is performed in the interactive computer simulation system, a storage system logs dynamic data in relation to the dynamic virtual sub-systems. At least one of the dynamic virtual sub-systems of the virtual simulated element is selected and a subset of dynamic data related to the selected virtual sub-system is loaded from the storage system. The selected virtual sub-system is displayed together with the related dynamic data on the graphical user interface.

Inventor: Jean-Francois Delisle
Status: Granted 2020-12-22
Patent ID: CA2995518C / CN110462709A / WO2018148818A1 / EP3574489A4

A simulation server capable of configuring events of a lesson plan through interactions with a computing device

Abstract

A simulation server capable of configuring events of a lesson plan through interactions with a computing device. The simulation server stores in a memory at least one lesson plan. Each lesson plan comprises at least one event, and each event comprises at least one rule. The simulation server receives a lesson plan selection from the computing device. The simulation server extracts from the memory the at least one event corresponding to the lesson plan selection and the corresponding at least one rule and transmits the extracted at least one event and at least one rule corresponding to the selected lesson plan to the computing device. The simulation server further receives from the computing device, a selection of at least one event to be used for the selected lesson plan with a configuration of the at least one rule for each selected event.

Inventor: Jean-Francois Delisle, David Bowness, Dac Toan Ho, Luc Gingras

Status: Granted 2018-03-06

Patent ID: CA2920937C / WO2017139875A1 / US20200050720A

A simulation server capable of creating events of a lesson plan based on simulation data statistics

Abstract

A simulation server capable of creating events of a lesson plan based on simulation data statistics. The simulation server comprises memory for storing simulation data, and a processing unit. The processing unit executes a plurality of simulations functionalities according to a lesson plan. The processing unit collects simulation data representative of the execution of the plurality of simulations functionalities according to the lesson plan. The processing unit stores the simulation data in the memory. The processing unit processes the simulation data to generate simulation data statistics. The processing unit creates at least one event having at least one rule based on the simulation data statistics. The at least one rule consists in at least one measurable value to be measured by at least one of the simulation functionalities.

Inventor: Jean-François Delisle

Status: Granted 2018-05-01

Patent ID: CA2920981C / US10679513B2 / WO2017139876A1 / CN108701424B

Assessing a training activity performed by a user in an interactive computer simulation

Abstract

An interactive computer-based training system, station and method for assessing a training activity performed by a user interacting with tangible instruments for controlling the virtual element in an interactive computer simulation. A processor module obtains a plurality of performance metric datasets related to the virtual element and obtains a plurality of expected maneuvers of the virtual element during the training activity. The processor module computes the plurality of performance metric datasets to identify actual maneuvers of the virtual element during the training activity, identifies one or more failed actual maneuvers of the virtual element during the training activity against corresponding ones of the expected maneuvers and performs computational regression on the actual maneuvers of the virtual element compared to the expected maneuvers of the virtual element to identify one or more root causes of the failed actual maneuvers.

Inventors : Jean-Francois Delisle, Anthoine Dufour, Marc-Andre Proulx, Dac Toan Ho

Status: Granted 2019-05-07

Patent ID : US10957216B2/ CA3000452C

Standard operating procedures feedback during an interactive computer simulation

Abstract

An interactive computer simulation system, station and method for training a user in the performance of a task through a training activity. A tangible instrument module allows a user to interact with the tangible instrument module for controlling a virtual element. A plurality of performance metric datasets representing results of the interactions between the user and the tangible instrument module is obtained. During execution of the interactive computer simulation, in the plurality of performance metric datasets, a plurality of actual maneuvers of the virtual element are detected during the training activity, one or more standard operating procedures (SOPs) are identified for the training activity from a plurality of the individually detected actual maneuvers. In real-time upon detection of the SOPs, information for display in the interactive computer simulation related the SOPs.

Inventors: Jean-Francois Delisle, Anthoine Dufour, Marc-Andre Proulx, Dac Toan Ho

Status: Granted 2019-05-07

Patent ID: CA3000443C / US20190304325A1 / EP3547289A1 / CN110322098A

Performance metrics in an interactive computer simulation

Abstract

A simulation mapping system and method for determining a plurality of performance metric values in relation to a training activity performed by a user in an interactive computer simulation, the interactive computer simulation simulating a virtual element comprising a plurality of dynamic subsystems. A processor module obtains dynamic data related to the virtual element being simulated in an interactive computer simulation station comprising a tangible instrument module. The dynamic data captures actions performed by the user on tangible instruments. The processor module constructs a dataset corresponding to the plurality of performance metric values from the dynamic data having a target time step by synchronizing dynamic data and by inferring, for at least one missing dynamic subsystems of the plurality of dynamic subsystems missing from the dynamic data, a new set of data into the dataset from dynamic data associated to one or more co-related dynamic subsystems

Inventor: Jean-Francois Delisle, Anthoine Dufour, Marc-Andre Proulx, Dac Toan Ho

Status: Granted 2021-04-27

Patent ID: CA3000463C / US10991262B2 / EP3547290A1 / CN110322104A

Simulation server capable of interacting with a plurality of simulators to perform a plurality of simulations

Abstract

A simulation server capable of interacting with a plurality of simulators to perform a plurality of simulations. The simulation server comprises a communication interface for exchanging data with other entities. The processing server also comprises a processing unit for executing at least one simulation. The processing unit also generates simulator simulation data and transmits the simulator simulation data to at least one simulator via the communication interface. The simulator simulation data are representative of the execution of the at least one simulation. The processing unit also receives simulator interaction data from the at least one simulator via the communication interface. The processing unit further processes the simulator interaction data and controls the execution of the at least one simulation based on the processed simulator interaction data. The simulation server may also interact with one or more of portable computing devices to perform the plurality of simulations.

Inventor: Jean-Francois Delisle

Status: Granted 2018-05-01

Patent ID: US20200357294A1 / WO2017139880A1 / CA2920913C

Method and system for generating vehicle parameters for training a user to score a vehicle maneuver

Abstract

There is described a method for training a user to score a vehicle maneuver, the method being executable by a processor, the method comprising: generating, using one of a machine learning model and a rule-based model, vehicle parameters based on a target grade and a vehicle maneuver, the target grade being indicative of a given performance level of a pilot performing the vehicle maneuver, the vehicle parameters representing parameters of the vehicle during the vehicle maneuver, the vehicle parameters enabling evaluating a performance of the pilot during the vehicle maneuver; providing the vehicle parameters for display on a display device; receiving, from a user device, a subjective grade indicative of an estimated performance level of a vehicle during the given vehicle maneuver; and providing the target grade for display to the user on the display device, thereby allowing the user to compare the subjective grade to the target grade.

Inventor: Maher Chaouachi, Jean-Francois Delisle

Status: Submitted 2022-03-23

Patent Application ID: US63/269,797

Federated machine learning in adaptive training systems

Abstract

A federated machine learning system for training students comprises a first adaptive training system having a first artificial intelligence module for adapting individualized training to a first group of students and for developing a first learning model based on a first set of learning performance metrics for the first group of students. A second adaptive training system provides individualized training to a second group of students and has a data property extraction module for extracting statistical properties from training data for the second group of students. A data simulator module generates simulated training data using extracted statistical properties from the training data for the second group of students to thereby generate a second learning model. A federation computing device receives first and second model weights for the first and second learning models and generate or refines a federated model based on the first and second model weights.

Inventor: Jean-François Delisle; Ben Winokur; Navpreet Singh

Status: Submitted 2022-03-15

Patent Application ID: US63/319,974

Adaptive learning in a diverse learning ecosystem

Abstract

A system for training a student to operate an actual machine includes an electronic learning module and a simulation system for simulating operation of the actual machine. An adaptive learning artificial intelligence (ALAI) module receives student performance data to adapt training of the student. The student performance data includes instructor-graded performance results of the student based on the student operating the actual machine, simulation performance results for the student operating a simulated vehicle in a simulation system that simulates operation of an actual machine and electronic learning content results from an electronic learning module that delivers electronic learning content to a student computing device used by the student. The ALAI module comprises a learner profile module that profiles the student, a training task recommendation module that generates AI-generated recommendations, and an explainability and pedagogical intervention module for displaying on the instructor computing device explanations for the AI-generated recommendations.

Inventor: Jean-François Delisle; Jian Qi

Status: Submitted 2022-03-07

Patent Application ID: US63/317,211

System and method for predicting performance by clustering psychometric data using artificial intelligence

Abstract

A system for predicting performance of a student based on psychometric data includes data storage devices for storing psychometric data for students obtained via psychometric tests, the psychometric data being indicative of a plurality of psychological traits of the students. A training management system having one or more simulation stations collects performance data for the students. One or more processors executing an artificial intelligence module clusters the psychological traits define aptitude clusters and correlates the aptitude clusters with the performance data to thereby associate different levels of performance with each of the plurality of aptitude clusters. A new student performance prediction module receives a set of new psychometric data for a new student and associates the set of new psychometric data for the new student with one of the plurality of aptitude clusters to thereby predict the performance of the new student.

Inventor: Jean-Francois Delisle; Anthoine Dufour; Bincy Baburaj Naranth

Status: Submitted 2021-27-10

Patent Application ID: US63/257,863

APPENDIX B SUPPORTED PUBLICATIONS

Publication directly supported during the realisation of the current research project:

Christophe Lazure, Laurence Dumont, Sofia El Mouderrib, Jean-François Delisle, Sylvain Sénécal, Pierre-Majorique Léger, « Certified Flight Instructors Performance Review of the Literature and Exploration of Future Steps", The International Journal of Aerospace Psychology, 2020

Abstract: Objective: We conducted a systematic review of peer-reviewed articles aimed at the evaluation of certified flight instructors' (CFI) performance in a training context and a scoping review of potential research avenues given the previously identified gaps. Background: As the demand for pilots will continue to grow significantly in the coming decades, so will the demand for CFIs, and for ways to improve their existing performance. Understanding performance factors of CFIs could benefit their training and help meet the increasing demand for pilots. Method: State-of-the-art research on the subject was surveyed via a systematic review of performance factors of CFIs and a scoping review to identify areas where other fields of research could inform CFI performance evaluation. Result: Only 20 articles since 1965 have directly assessed performance factors of CFIs. Their focus has mostly been on communication and educational processes. The scoping review brings forward concepts from cognitive psychology and psychophysiology as means of improving the current understanding of CFI situation awareness and task management. Conclusion: Very little work has been done on CFI situation awareness and task management. These are the two main domains in which psychophysiological tools could provide a clear understanding of the attentional and decisional processes at play while developing situation awareness in a dynamic environment and quantify the task load affecting it.

Yang Meng, "Machine Learning for Aviation Data", A Thesis Submitted to the Committee on Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science in the Faculty of Arts and Science, Trent University, Peterborough, Ontario, Canada, 2021 (Meng, 2021)

Abstract: This thesis is part of an industry project which collaborates with an aviation technology company on pilot performance assessment. In this project, we propose utilizing the pilots' training data to develop a model that can recognize the pilots' activity patterns for evaluation. The data will present as a time series, representing a pilot's actions during maneuvers. In this thesis, the main contribution is focusing on a multivariate time series dataset, including preprocessing and transforming. The main difficulties in time series classification is the data sequence of time dimension. In this thesis, I developed an algorithm which formats time series data into equal length.

Three classification and two transformation methods were used. In total, there are six models for comparison. The initial accuracy was 40%. By optimization through resampling. We increased the accuracy to 60%.

APPENDIX C PAPER PUBLICATIONS

Article 1 - *A Capability Maturity Model for Flight Training*, Jean-François Delisle, Stéphane Ouellet, Derek Linders, Published in *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2018 proceedings*, November 2018

Article 2 - *Pilot performance assessment using a hybrid expert system and machine learning for an automatic objective assessment in flight simulation*, Jean-François Delisle, Andrea Lodi, Maher Chaouachi, Melvyn Tan, Laurent Desmet, Submitted in *PLOS One*, August 2022

Article 3 - *Federated Machine Learning in Adaptive Flight Training*, Jean-Francois Delisle, Navpreet Singh, Jian Qi, submitted *in AIAA Journal*, `March 2022

Article 4 - *Predicting Pilot Training Performance Using Psychometry and Flight Performance Data*, Jean-François Delisle, Bincy Baburaj Narath, Karen Moore, Submitted in *The International Journal of Aerospace Psychology*, July 2022

Article 5 - *Using AI and neuroscience for adaptive learning to accelerate pilot training*, Jean-François Delisle, Theophile Demazure, Hamza Nabil, Pierre-Majorique Léger, Submitted in *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2022 proceedings*, February 2022

Jean-François Delisle, Manager AI Innovation, Digital Accelerator, CAE Inc.

Jean-François has over 20 years of experience in software engineering, artificial intelligence, and data solution. Part of the CAE's digital accelerator aims to digitize its training landscape and its training solution ecosystem for pilots, crews, and customers. Focusing his research on AI solution that helps optimizing and adapting our training solution based on human behavior and performance analysis. He joined CAE in 2010 and his current mandate is to define flight training strategies using data analysis and artificial intelligence capabilities in advanced air mobility and eVTOL engineering solutions. He is a PhD candidate in cognitive science for adaptive flight training, under the supervision of Professor Andrea Lodi, Canada Excellence Research Chair in Data Science for Real-Time Decision Making at the Polytechnique de Montréal

Andrea Lodi, Andrew H. and Ann R. Tisch Professor at the Jacobs Technion-Cornell Institute at Cornell Tech and the Technion

Andrea Lodi is the Andrew H. and Ann R. Tisch Professor at the Jacobs Technion-Cornell Institute, Cornell Tech and Technion–Israel Institute of Technology. His research interests include mixed-integer linear and nonlinear programming and data science. His work has been recognized by IBM and Google faculty awards and the 2021 Farkas Prize by the INFORMS Optimization Society. He has been Canada Excellence Research Chair in Data Science for Real-Time Decision Making at Polytechnique Montréal, network co-ordinator and principal investigator of EU and Canadian projects and consultant of the IBM CPLEX research and development team.

Stéphane Ouellet, Solution Manager, Business Development—Product Management and Solutions, CAE

Stéphane has been with CAE for 28 years. With CF-18 systems engineering department for 10 years where he was exposed to the development of mission-critical applications and involved in the implementation of CMMI processes. Later, at CAE, he occupied various positions in project management and technical leadership, enabling him to acquire an understanding of the simulation industry and processes, from management, technical, and business development standpoints. Since 2009, Stéphane has focused on Training System Integration (TSI), and has been at the forefront of

CAE's efforts to develop a unique offering and expertise in training solutions and integration. Stephane has worked at defining, designing, delivering, and leading major TSI programs, such as the Operational Training System Provider program, where CAE delivered integrated training solutions to the Royal Canadian Air Force, including the CC-130J Training Center and the Medium Heavy Lift Helicopter Training Center. As CAE evolved by delivering TSI programs, Stéphane played a key role in developing the CAE training solutions groups. Stéphane is now in the Business Development team, as a solution manager responsible for defining the training solutions CAE will offer in the future. Stéphane has a bachelor`s degree in engineering from Polytechnique de Montréal.

Derek Linders, Group Lead, Integrated Learning Environments, CAE

Derek Linders has nearly 30 years of experience in IT services and process design and implementation, with 15 years serving the aerospace, health care, and defense industries. He has worked in all phases of training systems design, delivery, and operations, including learning systems and operational process design for Canada's CH148 helicopter maintenance training and deployment services, technology refreshes for the NATO Flight Training Center in Canada and C-130J Hercules Operational Training Services Provider program, as well as design for the Canadian Air Force's AFIILE e-learning platform.

Derek has been with CAE since 2008 and currently heads a team dedicated to the design, implementation, and operational processes for integrated learning environments. He is currently leading a transformation of CAE's military aviation training centers, aimed at making training more efficient and engaging, and at moving the training centers forward in their own delivery maturity. Derek has a bachelor's degree in science from Dalhousie University.

Laurent Desmet, Data Scientist, Digital Accelerator, CAE Inc.

Laurent Desmet has a master's degree in data science from Polytechnique Montreal (M.Ing) and was hired by CAE three years ago. He integrated the rotation program of the company and had the opportunity to work in several teams where he filled several patents. He started in the sound department where he developed an algorithm to harmonize acoustic energy, and then went to the

Healthcare division of the company where he used his knowledge in several applications, from time-series forecasting to the analysis of images. Finally, he officially joined the data science team where he worked on a predictive engine to assess the performance of pilots during their training phase.

Maher Chaouachi AI Strategist, Digital Accelerator, CAE Inc.

Dr. Maher Chaouachi holds a PhD in computer science specialized in the field of Artificial Intelligence obtained from the University of Montreal. He worked as a postdoctoral researcher at McGill University. Dr. Chaouachi is AI strategist at CAE working in various programs involving machine learning, process and data mining, optimization and predictive modelling.

Melvyn Tan, Data Analyst, Global Engineering, CAE Inc.

Melvyn Tan joined CAE in 2019 and is part of the company's rotational development program. He has completed past stints at various sectors within aviation, including Scoot Airways in Singapore and Bombardier Aerospace in Canada. A believer in lifelong learning, he is currently completing his professional certificates in data science and artificial intelligence at the University of Toronto's School of Continuing Studies. He also holds a bachelor's degree in aerospace engineering from the University of Toronto.

Navpreet Singh, McGill University

Navpreet Singh received his B.E. (2014) in electrical engineering from Thapar University in Patiala, India, and a M.S. in electrical and computer engineering from Carnegie Mellon University (2016) in Pittsburgh, USA. He is currently working on his Ph.D. in electrical and computer engineering at McGill University in Montreal, Canada. His research interests include smart physical sensors, microelectromechanical systems (MEMS), machine learning, and circuit design. He is working on designing miniaturized air quality sensors coupled with machine learning. He is a recipient of Natural Sciences and Engineering Research Council of Canada Doctoral Scholarship and Mitacs Accelerate scholarship.

Jian Qi, Senior Data Scientist, CAE

Jian Qi received a Ph.D. degree in Telecommunications from the Institut national de la recherche scientifique (INRS), Montreal, QC, Canada, in 2011, with distinction (Dean's Honor List) and is a recipient of Doctor Research Scholarship from the Quebec Government Fonds Québécois de la Recherche dur la Nature et les Technologies (FQRNT) – Merit Scholarship Program. He has more than 10 years of experience as a professor, data scientist, and entrepreneur in artificial intelligence for telecommunications, eCommerce, aviation, and finance; with a variety of institutions and industries, namely, King Abdullah University of Science and Technology (KAUST), Saudi Arabia, University of Reading, UK, Institut national de la recherche scientifique (INRS), SSENSE, and CAE, Canada. From 2013 to 2015, he was an Assistant Professor in Electrical Engineering at the University of Reading, UK. Since October 2018, he has been with CAE as a Senior Data Scientist. His research interests include artificial intelligence, big data, adaptive flight training federated learning, etc.

Bincy Baburaj Narath, Data Scientist, Lixar I.T

Bincy Baburaj Narath is a professional data scientist, experienced in the Educational Industry, Telecommunications, and Networking sectors. Her research interests span from algorithm development, to using machine learning and statistical techniques in various industries to uncover hidden patterns, automate decision-making and effective recommendations. She has a Master's degree in Electrical Engineering, specialized in wireless communication with a research focus on contention resolution, and various industry-level certifications in Data Science.

Karen Moore, CPsychol, CSci, AFBPsS, EuroPsy, MRAeS, Symbiotics

Karen Moore is a Chartered Occupational Psychologist with over 30 years' experience assessing individuals at all levels, from graduate to board directors, and in a diverse range of industries from nuclear, through utilities, to banking and aviation. She is interested in ensuring that assessment decisions are of value, not only to the hiring company in terms of effective performance, but also to the individual, and that outputs can assist them with their personal development either in training or on the job.

Karen joined Symbiotics in 2017 as Principal Occupational Psychologist to further develop their assessment processes for high consequence industries. Symbiotics specialize in the assessment of aviation roles and particularly of pilots from cadets through to instructor/examiner. Karen published Moore, K. (2020). Ability, Aptitude and Performance Assessment. In: R.Bor, C.Eriksen, T.Hubbard, & R.King (Eds), *Pilot Selection; Psychological Principles and Practice* pp161-168. London, UK: Routledge.

Hamza Nabil, Data Analyst at CAE

Hamza Nabil is a Data Analyst professional, with a Bachelor's degree specialized in aerospace engineering. He joined CAE in 2019 within a leadership rotation program to pursue a career in Data. His mandate was to research and analyze specific patterns between the correlation of biometric data and flight telemetry of pilots using various statistical models and to also automate the performance of each pilot using an objective score. He has a Master's degree in Mechanical Engineering specialized in Computational Fluid Dynamics.

Theophile Demazure

Théophile Demazure is a Ph.D. candidate in Information Technology at HEC Montréal under the NSERC-Prompt Industrial Research Chair in User Experience. Théophile published in multidisciplinary outlets such as Frontiers in Human Neuroscience, Business & Information Systems Engineering, or NeuroIS. His research explores the human factor and neurophysiological side of human-computer interaction in performance-oriented environments.

Pierre-Majorique Léger

Professor Pierre-Majorique Léger is the senior chairholder of the NSERC-Prompt Industrial Research Chair in User Experience (UX). He is a full professor of IT at HEC Montreal. He holds a Ph.D. in industrial engineering from École Polytechnique de Montréal and has done post-doctoral studies in information technologies at HEC Montréal and NYU Stern School of Business. He is the co-director of the Tech3Lab, a research laboratory and devoted to the investigation of human factors in information technologies.