Toward Automated Instructor Pilots in Legacy Air Force Systems: Physiology-Based Flight Difficulty Classification Via Machine Learning

Article *in* SSRN Electronic Journal · January 2022

DOI: 10.2139/ssrn.4170114

4 authors, including:

William Caballero
United States Air Force Academy
21 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Nathan Gaw
Air Force Institute of Technology
20 PUBLICATIONS **782** CITATIONS
SEE PROFILE

Phillip R. Jenkins
Air Force Institute of Technology
26 PUBLICATIONS **162** CITATIONS
SEE PROFILE

Toward Automated Instructor Pilots in Legacy Air Force Systems: Physiology-based Flight Difficulty Classification via Machine Learning

William N. Caballero[c], Nathan Gaw[d], Phillip R. Jenkins[d], Chancellor Johnstone[d,*]

[a] *United State Air Force Academy, 2304 Cadet Drive. US Air Force Academy, CO, United States 80840*
[b] *Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH, United States 45433*

*Corresponding author for all stages of publication. Phone: +1 (509) 216-0574
Email addresses: william.caballero@us.af.mil (William N. Caballero), nathan.gaw@afit.edu (Nathan Gaw), phillip.jenkins@afit.edu (Phillip R. Jenkins), chancellor.johnstone@afit.edu (Chancellor Johnstone)
No data ethics considerations are foreseen related to this paper.

Toward Automated Instructor Pilots in Legacy Air Force Systems: Physiology-based Flight Difficulty Classification via Machine Learning

William N. Caballero[c], Nathan Gaw[d], Phillip R. Jenkins[d], Chancellor Johnstone[d,*]

[c] *United State Air Force Academy, 2304 Cadet Drive. US Air Force Academy, CO, United States 80840*
[d] *Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH, United States 45433*

Abstract

The United States Air Force (USAF) is struggling to train enough pilots to meet operational requirements. Technology has advanced rapidly over the last 70 years but USAF pilot training has not. Modern operational requirements demand a change and, for this reason, USAF senior leadership has advocated for innovation. The automation of instructor and evaluator pilots in select bottlenecks (e.g., simulators) is one such measure. However, to implement this vision, numerous technical issues must be mitigated. Accurate classification of flight difficulty is a foundational problem underpinning many of these technical issues, which requires either the acquisition of new systems or the development of new procedures. Therefore, given this need and the costly nature of purchasing new equipment, physiological-based classification of flight difficulty is our focus herein. Leveraging multimodal data from a designed experiment of pilots landing a simulated aircraft, we develop a high-quality machine learning pipeline for classifying flight difficulty, called the Multi-Modal Functional-based Decision Support System (MMF-DSS). MMF-DSS distills a tabular set of features from our multimodal and functional data through the use of functional principal component analysis, summary statistics, and BorutaSHAP. In this manner, information is derived from the time-series data via the generation of hundreds of features, of which a small subset having the most predictive capability is discerned. Four full factorial designs are used to perform hyperparameter tuning on a set of classifiers. In so doing, a superlative technique is identified. Impacts on executive decision making are examined as well as associated policymaking implications. Alternative classifiers are considered for use within our pipeline that trade predictive accuracy for cost efficiency, and recommendations for choosing among these alternatives is provided.

Keywords: Machine learning, Functional data analysis, Pilot training, BorutaSHAP

1. Introduction

Pilot training in the United States Air Force (USAF) remains largely unchanged from 1950s-era practices [1]. Although the service's current program, Specialized Undergraduate Pilot Training (SUPT), produces highly competent pilots, senior USAF leadership has determined this program is unable to meet modern operational requirements. More specifically, over the past five years, the USAF's pilot population has maintained a dangerous deficit. This shortfall is a grave national-security threat that cannot be ameliorated unless USAF pilot accession exceeds attrition; more pilots must be trained compared to those that leave the force. To do so, we posit that a more effective and efficient pilot training program can be developed through automated and personalized instruction.

*Corresponding author for all stages of publication. Phone: +1 (509) 216-0574
Email addresses: william.caballero@us.af.mil (William N. Caballero), nathan.gaw@afit.edu (Nathan Gaw), phillip.jenkins@afit.edu (Phillip R. Jenkins), chancellor.johnstone@afit.edu (Chancellor Johnstone)

No data ethics considerations are foreseen related to this paper.

Therefore, this research develops a machine learning approach to automate flight difficulty prediction as a first step toward modernizing the legacy SUPT program.

Military flight training, both SUPT and otherwise, places exceptional emphasis on feedback provided to students from experienced flight instructors during real-world and simulated sorties (i.e., flights of combat aircraft on missions). Such reliance corresponds with a bottleneck in pilot training pipelines; student throughput is limited by the number of qualified instructors. Although the USAF has experimented with novel training paradigms that place a greater emphasis on flight simulators (e.g., *Pilot Training Next* and *Accelerated Path to Wings*), these programs still rely heavily on human-instructor feedback, which implies the theoretical throughput of these new programs will exceed realized production.

To alleviate these traditional bottlenecks, both civilian academics and military officers have advocated for greater utilization of machine learning techniques in legacy pilot training programs [2, 3]. The efficacy of machine learning for flight-related classification and regression tasks is well-documented. Flight data has been utilized to build effective models of anomalous and unsafe behavior [e.g., see 4, 5, 6]; a variety of techniques, from Gaussian mixture models [7] to variational auto-encoders [8], have been utilized for this purpose. Machine learning methods have also been successfully adapted to predict flight trajectory [e.g., see 9, 10] and conduct air traffic management functions [e.g., see 11, 12]. However, the explicit application of machine learning for flight training purposes is relatively understudied. Therefore, building off the SUPT candidate selection work of [13], we develop the Multi-Modal Functional-based Decision Support System (MMF-DSS), a machine learning pipeline with the express purpose of providing support to improve USAF pilot training.

Any machine learning approach designed to support USAF operations should account for service-specific constraints. For example, many legacy SUPT training platforms are ill-equipped for the modern data collection required to systematically automate instructor feedback. Although the USAF is actively developing replacements [e.g., see 14], the acquisition process in the U.S. Department of Defense (DoD) is demanding and cumbersome [15], implying that the most pragmatic machine learning approach should be readily incorporated into existent technology and infrastructure.

Therefore, the machine learning methods developed herein do not rely on flight data generated by the training platform (i.e., aircraft or simulator). Instead, our models are exclusively based on physiological measurements of the student pilot. This characteristic allows our models to be more readily implemented within legacy training equipment. The methodology developed herein is such that an exogenous, physiological-monitoring and decision-support system [e.g., see 16], not dissimilar to modern driver state monitoring systems [17], can be appended to existent training platforms.

In particular, we consider if and how physiological measurements can be utilized to accurately predict flight difficulty levels. Our predictions are built off physiological data from seven qualitative streams: (1) electromyography of muscle activation corresponding to the flexion and extension of the pilot's arm, (2) acceleration of the pilot's forearm while operating the flight joystick, (3) electrodermal activity detecting changes in electrical conductance of the skin, (4) electrocardiography measuring electrical activity of the heart, (5) electrical signal of respiration derived using impedance pneumography, (6) eye tracking of gaze direction, position, and pupil diameter, and (7) photoplethysmography measuring changes in blood volume.

This data set was collected via human-subject testing conducted by the *USAF-MIT AI Accelerator* on behalf of the *USAF Chief Data Office* (CDO). Each qualitative stream consists of myriad time-series measurements collected on participants during a simulated landing of a Beechcraft T-6 Texan II; the data are both functional and multimodal. Depending on the measurement, the collection rate was on the order of milliseconds or seconds, thereby ensuring a rich compendium of information. To effectively utilize this deluge of data, we jointly leverage functional principal component analysis (FPCA), summary statistics, and BorutaSHAP, a wrapper feature selection technique, within our developed machine learning pipeline [18, 19]. After identifying a preferred set of features from those generated and tuning four distinct machine learning models, a superlative Adaptive Boosting classifer is identified to predict flight difficulty. The performance of our model

provides an *excellent* area under the receiver operating characteristic curve of ≈ 0.88 [20]. Moreover, we also consider how the inclusion of flight data from future training platforms can augment our model and improve classification accuracy. Finally, this collective analysis is utilized to derive three distinct policymaking actions, explore their potential implications, and provide implementation recommendations.

The remainder of this manuscript more thoroughly discusses our model development and results. More specifically, Section 2 details the relevant literature and techniques that may be leveraged to process and analyze the functional data utilized herein. Section 3 provides an in depth explanation of the data set, and describes the machine learning pipeline derived for automated flight difficulty classification. More specifically, Section 3 details the empirical data collection process, descriptively summarizes the raw feature measurements, and explains the methodology underpinning our proposed machine learning pipeline, i.e., the MMF-DSS. Section 4 sets forth the analytic basis of our pipeline, presents its performance, and furnishes analysis predicating the selection of our preferred classification algorithm. Section 5 translates this analysis into policymaking guidance, provides insight into executive decision making, and furnishes implementation recommendations. Finally, Section 6 provides concluding remarks and considers promising areas of future research.

2. Relevant Techniques for Multimodal and Functional Data

The data in this study present a unique challenge in that they are both functional and multimodal. Processing functional data and integrating information from multiple modalities requires innovation that can best harness the time-series data from multiple streams. Numerous functional data analysis and multimodal fusion techniques exist to separately preprocess our time-series data and to optimally combine this information for flight difficulty classification for which we provide a brief discussion below.

Functional data analysis (FDA) encompasses a variety of techniques including forecasting [21, 22, 23], regression analysis [24, 25, 26], non-parametric modeling [27, 28], clustering [29, 30, 31], smoothing [32, 33, 34], and data reduction [24, 35, 36, 33]. In our particular application, FDA is used for data reduction, i.e., to diffuse the infinitely dimensional physiological functional data into features that can be used as input for machine learning models. The pre-eminent means to reduce dimensionality of functional data is functional principal component analysis (FPCA). In a functional data analysis review article, it was found that 60.7% (51 out of 84) of the reviewed studies utilized FPCA in at least some part of their analysis [37]. FPCA is a key technique in functional data analysis that extracts underlying patterns from temporal data having either sparsely or densely sampled time courses [38, 39]. The method was used first for growth curves by [40], for which various applications and theoretical properties were studied until the turn of the century [41, 42, 43, 44, 45]. One limitation of FPCA was that it relied on complete functional data collected at regular time points, which created problems with its use for many practical applications that could not meet these stringent requirements. To handle this problem, various papers made headway to propose practical solutions to handle problems with an irregular grid of timepoints [46, 47, 48]. However, in cases where the number of time points greatly differed across the functional curves (especially in sparse cases with only a few time points), the FPC scores could not be approximated well through the typical, numeric methods of integration. One solution that has been proposed is using B-spline basis functions to model the individual functional data curves via mixed effects models [49, 50, 51, 52]. However, with these methods, eigenfunctions of the principal components are not directly determined from the data. In contrast, [53] proposed a method called Principal Analysis by Conditional Estimation (PACE) which is capable of handling sparse and irregular functional data for which pooled time points across the functional curve instances are sufficiently dense. PACE can optimally pool trends observed across different time series of varying lengths. This work opened up possibilities to apply FPCA to a vast array of time series data and ushered in a new renaissance of dimensionality reduction in functional data analysis.

PACE has been applied to a broad array of applications, including: healthcare [54, 55, 56, 57], manufacturing [58, 59, 60], energy [61, 62, 63], remote sensing [64, 65], and agriculture [66, 67]. In healthcare, [54] utilized PACE to characterize longitudinal prostate-specific antigen across various levels of age. In manufacturing, [60] utilized multiple degradation signals to predict failure time of a partially degraded signal. In energy, [63] integrated PACE into a framework that generates a real-time prognostic policy for large-scale windmill farms. In remote sensing [64] performed FPCA from monsoon precipitation satellite images over Eastern India to build interpretable functions that characterize weather patterns over this region. Finally in agriculture, [66] used PACE to process daily minimum and maximum temperature trajectories for crop yield per acre prediction.

Many features can be generated from PACE across the multiple streams of physiological data, which necessitates additional methodological development to determine which features are most relevant to the classification task at hand. Feature selection is a broad field that encompasses (1) filter, (2) embedded, and (3) wrapper approaches. Filter methods are the quickest to generate since they do not require the labels of each instance/sample. However, because filter methods do not consider the predictive value of the features to the response variable, there is a high risk that the relevant features may be filtered out. Embedded methods incorporate feature selection into the objective function used to fit the predictive model [68, 69, 70, 71, 72]. These methods enjoy the benefit of integrating feature selection and predictive modeling fitting. The limitation, however, is that the integration/embedding mechanism must be designed specifically for each type of predictive model, and therefore is not universally applicable. Wrapper methods, for their part, leverage a specific, machine learning algorithm to guide the selection of the most important features; features are iteratively removed (added) until some termination criteria is met based upon the machine learning algorithm's output. Of particular importance to this research is the BorutaSHAP wrapper approach to feature selection. BorutaSHAP is an extension of the Boruta algorithm that leverages the SHAP value [73] as a measure of feature importance with a random forest classifier, i.e., via the TreeSHAP [74] routine.

Since we have yet to identify a method that flexibly and seamlessly integrates both functional data dimension reduction and multimodal fusion into a single framework, we adopt a pipeline approach herein. FPCA is first leveraged for dimensionality reduction, BorutaSHAP is utilized for feature selection, and a classifier is tuned to the resultant data. Section 3.3 discusses this methodology in further detail.

3. Data Set Overview and Machine Learning Pipeline

The machine learning techniques developed in this research rely on functional data analysis for dimensionality reduction and feature generation. Therefore, we begin this section by discussing the experimental setup and design conducted by the USAF-MIT AI Accelerator; subsequently, the raw data is summarized. We then present the architecture of the derived MMF-DSS machine learning pipeline that processes this data into classification predictions.

3.1. Experimental Setup and Design

Data leveraged within this study, obtained from the USAF-MIT AI Accelerator, were collected via a designed experiment wherein participants of varying aptitudes were required to land a T-6 Texan II in a simulated environment [75]. More specifically, utilizing a virtual reality implementation of the X-Plane 11 software, participants were tasked with landing an aircraft, originating approximately 19 miles from the runway, under one of four scenarios described as either low- or high-difficulty. Table 1 summarizes the conditions encountered by the pilots within each scenario.

Data was collected on 21 test subjects in total. Every participant performed a total of 12 landings, three under each scenario. Scenarios were presented in a non-sequential manner across each subject. While performing the experimental task, physiological data was collected via a suite

Table 1: Summary of Flight Scenarios by Difficulty Classification Level

Difficulty	Scenario	Wind**	Clouds/Visibility	Turbulence
Low	A	None	No clouds Visibility unlimited	None
Low	B	140° at 10 knots	Overcast at/above 2500′ Visibility of 5 statute miles	None
High	C***	1000′: 140° at 10 knots 3000′: 080° at 10 knots	Overcast at/above 1000′ Visibility of 3 statute miles	Mild at/above 1000′
High	D***	800′: 200° at 15 knots 2800′: 080° at 15 knots (20 knot gusts) 4800′: 250° at 10 knots	Overcast at/above 400′ Visibility of 1 statute mile	Mild

*Altitudes listed in terms of elevation above Mean Sea Level

** Wind above (below) the highest (lowest) altitude are constant to space (ground); wind direction indicates heading of origin.

*** Wind direction and velocity between altitudes are linearly interpolated.

of sensors attached to the participants' bodies. Data aggregation and synchronization of the time-series data provided by these sensors was accomplished with the Lab Streaming Layer software. For further information regarding the sensor configuration and experimental configuration we refer the interested reader to [75]. Flight data was also collected from the X-Plane 11 simulator; however, the use of such data is not the primary focus of this manuscript.

3.2. Summarizing the Raw Data Features

Seven different streams of physiological data and two different streams of non-physiological data were collected during experimentation. The physiological data included a variety of electromyography (EMG), forearm acceleration (ACC), electrodermal activity (EDA), electrocardiography (ECG), respiration (RES), eye tracking (ETK), and photoplethysmography (PPG) data. Among the non-physiological data were pilot (PX) and aircraft (AX) data streamed from the X-Plane 11 simulator. Each data stream is composed of multitudinous, time-series measurements summarized in Table 2. The sensor configuration for these data streams are as follows.

Table 2: Description of Raw Signals

Variable	Stream	Description	Unit	Variable	Stream	Description	Unit
$X_1(t)$	EMG	Electromyogram of wrist flexor muscles	mV	$X_{31}(t)$	ETK	Normalized vertical location of right pupil	-
$X_2(t)$	EMG	Electromyogram of wrist extensor muscles	mV	$X_{32}(t)$	ETK	Depth of binocular fixation	mm
$X_3(t)$	EDA	Electrodermal activity on left hand	$k\Omega$	$X_{33}(t)$	ETK	Boolean value of binocular convergence validity	-
$X_4(t)$	ECG	ECG signal measured from left leg to right arm	mV	$X_{34}(t)$	ETK	Fixation-event membership indicator	-
$X_5(t)$	ECG	ECG signal measured from left arm to right arm	mV	$X_{35}(t)$	ETK	Saccade-event membership indicator	-
$X_6(t)$	ECG	ECG signal measured from chest to right leg	mV	$X_{36}(t)$	PPG	Plethysmogram measured at middle-finger's tip	mV
$X_7(t)$	RES	Electrical measure of chest wall excursion	mV	$X_{37}(t)$	PX	Pilot's lateral head position	m
$X_8(t)$	ETK	Bits containing all left-eye validity	-	$X_{38}(t)$	PX	Pilot's longitundinal head position	m
$X_9(t)$	ETK	Bits containing all right-eye validity	-	$X_{39}(t)$	PX	Pilot-vertical-head position	m
$X_{10}(t)$	ETK	Lateral-axis origin of left-eye gaze	mm	$X_{40}(t)$	PX	Pilot-head-yaw angle	deg
$X_{11}(t)$	ETK	Longitudinal-axis origin of left-eye gaze	mm	$X_{41}(t)$	PX	Pilot-head-pitch angle	deg
$X_{12}(t)$	ETK	Vertical-axis origin of left-eye gaze	mm	$X_{42}(t)$	PX	Pilot-head-roll angle	deg
$X_{13}(t)$	ETK	Lateral-axis origin of right-eye gaze	mm	$X_{43}(t)$	AX	Aircraft latitude	deg
$X_{14}(t)$	ETK	Longitudinal-axis origin of right-eye gaze	mm	$X_{44}(t)$	AX	Aircraft longitude	deg
$X_{15}(t)$	ETK	Vertical-axis origin of right-eye gaze	mm	$X_{45}(t)$	AX	Aircraft height above WSG84 ellipsoid	m
$X_{16}(t)$	ETK	Normalized lateral direction of left-eye gaze	-	$X_{46}(t)$	AX	Aircraft above ground altitude	ft
$X_{17}(t)$	ETK	Normalized longitudinal direction of left-eye gaze	-	$X_{47}(t)$	AX	Aircraft indicated airspeed	kt
$X_{18}(t)$	ETK	Normalized vertical direction of left-eye gaze	-	$X_{48}(t)$	AX	Aircraft ground speed	m/s
$X_{19}(t)$	ETK	Normalized lateral direction of right-eye gaze	-	$X_{49}(t)$	AX	Aircraft speed in true east direction	m/s
$X_{20}(t)$	ETK	Normalized longitudinal direction of right-eye gaze	-	$X_{50}(t)$	AX	Aircraft inertial vertical speed	m/s
$X_{21}(t)$	ETK	Normalized vertical direction of right-eye gaze	-	$X_{51}(t)$	AX	Aircraft speed in true north direction	m/s
$X_{22}(t)$	ETK	Diameter of left pupil	mm	$X_{52}(t)$	AX	Aircraft climb rate	m/s
$X_{23}(t)$	ETK	Diameter of right pupil	mm	$X_{53}(t)$	AX	Aircraft landing gear deployment	%
$X_{24}(t)$	ETK	Value characterizing openness of left eye	-	$X_{54}(t)$	AX	Aircraft ILS lateral deflection	%
$X_{25}(t)$	ETK	Value characterizing openness of right eye	-	$X_{55}(t)$	AX	Aircraft ILS glide slope deflection	%
$X_{26}(t)$	ETK	Normalized lateral location of left pupil	-	$X_{56}(t)$	AX	Aircraft yaw angle	deg
$X_{27}(t)$	ETK	Normalized longitudinal location of left pupil	-	$X_{57}(t)$	AX	Aircraft pitch angle	deg
$X_{28}(t)$	ETK	Normalized vertical location of left pupil	-	$X_{58}(t)$	AX	Aicraft roll angle	deg
$X_{29}(t)$	ETK	Normalized lateral location of right pupil	-	$X_{59}(t)$	AX	Elevation trim status	-
$X_{30}(t)$	ETK	Normalized longitudinal location of right pupil	-	$X_{60}(t)$	AX	Aileron trim status	-

EMG data was collected via a Shimmer Sensing device [76]. Muscle activation was measured from each subject's right forearm (i.e., the arm controlling the flight joystick). To measure flexion and extension[2], two EMG electrodes were placed close to the brachioradialis and extensor digitorum, respectively (i.e., near the elbow). A reference electrode was also placed on the ulna (i.e., the boney protrusion near the wrist). The nominal sampling rates for the EMG sensors were 128 Hz and 512 Hz. Whereas the Shimmer Sensing device also recorded accelerometry of the forearm, the quality of these measurements was insufficient for inclusion in this manuscript.

EDA and PPG data were collected using a Shimmer Sensing GSR+ and Optical Pulse device, respectively [77, 78]. Measurements were collected on the left hand (i.e., hand controlling the throttle). Electrodes were placed at the bases of the ring and index fingers, whereas a PPG sensor was placed on the tip of the middle finger. The nominal sampling rate for these sensors were 128 Hz and 1024 Hz.

ECG and RES data were recorded with electrodes placed at the torso using a Shimmer Sensing device [79]. ECG electrodes were placed at the sternum under the fourth rib, left mid-axillary line, right mid-axillary line, apex of the left hip, and the reference electrode was at the apex of the right hip. The respiration signal was measured via chest wall excursion. For all sensors, the nominal sampling rate was 128 Hz.

ETK data was recorded from the HTC Vive Pro Eye [80]. Measurements of the left and right eyes were recorded independently. Prior to experimentation and after headset adjustment, an eye-tracking calibration was performed to ensure accuracy. The nominal sampling rate for this data modality was 250 Hz. Finally, AX and PX data were obtained directly from the X-Plane 11 flight simulator [81]. The nominal sampling rate of this data was 4 Hz.

3.3. Pipeline Architecture

This section discusses the components of our proposed MMF-DSS pipeline, i.e., the feature extraction, feature selection, and machine learning techniques. The experiment described in Sections 3.1 and 3.2 results in multi-modal, functional data containing a substantial degree of information. However, without pre-processing, machine learning algorithms struggle to identify the underlying relationship between the input data and response labels. Therefore, we begin by describing how a countable set of features, i.e., $\mathcal{F}$, can be generated utilizing techniques from functional data analysis; the definition of each feature in $\mathcal{F}$ is described subsequently. Furthermore, we utilize a state-of-the-art feature selection technique to identify an $F^* \subset \mathcal{F}$ having superlative predictive value. This subset of features is utilized to build the predictive models discussed in Section 4 by identifying the superlative classifier from the subset described herein. Figure 1 depicts the overall flow of MMF-DSS, whereas the rest of this section describes its architecture.

The MMF-DSS pipeline begins by leveraging the raw, multimodal signals to generate a more manageable set of features (i.e., in tabular format). Smoothing and FDA methods were utilized. To reduce computational complexity, all time-series data are sampled at a frequency of 1 Hz for input before applying any subsequent analysis. Windowing at 15s, 30s, 45s and 60s non-overlapping time windows was also generated for each data stream. However, subsequent analysis did not find them to provide valuable information to improve predictive model performance. Thus, we do not report any results for the windowed features. Moreover, to build some elementary features, summary statistics were generated from each of the data streams. Table 3 summarizes the summary statistics (i.e., $f^S_{i,j}$) that were generated along with their corresponding descriptions. PACE is also utilized to generate functional principal components (FPCs) and FPC scores; the scores, denoted by $f^P_{i,k}$, are derived from the FPCs. Both the summary statistics and FPC scores are utilized as input features to the feature selection component of MMF-DSS. That is, $\mathcal{F} = \{f^S_{i,j} : i \in \mathcal{I}, j \in \mathcal{J}\} \cup \{f^P_{i,k} : i \in \mathcal{I}, k \in \mathcal{K}_i\}$

[2]Flexion refers to motion that decreases the angle at the joint, whereas extension refers to movement that increases this angle.

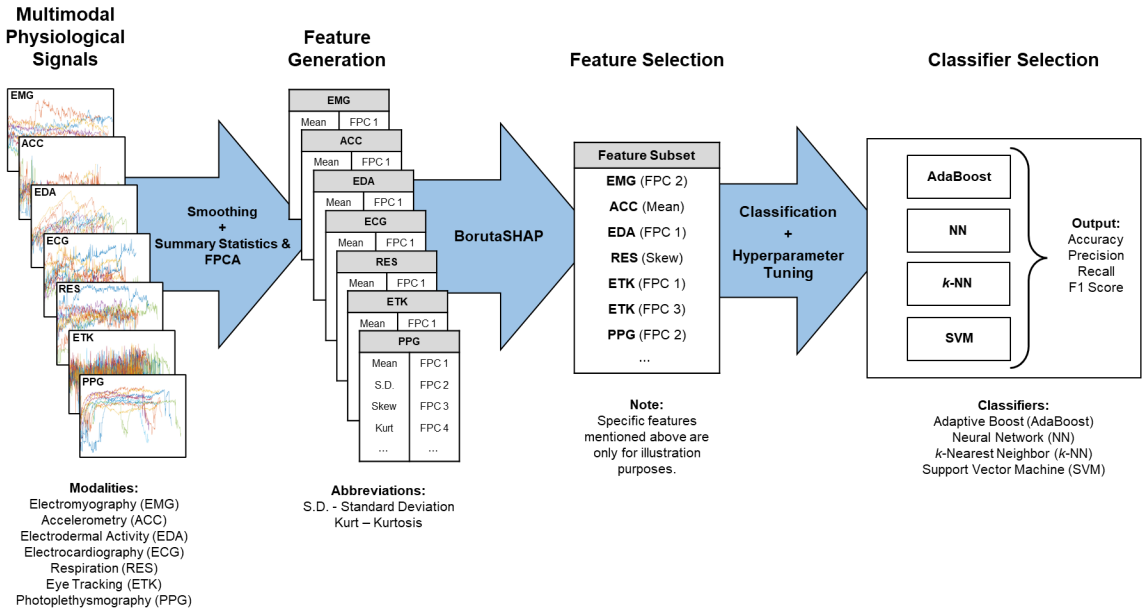

Figure 1: Overall model framework of the Multi-Modal Functional-based Decision Support System (MMF-DSS), consisting of (1) preprocessing of multimodal physiological signals, (2) generating features from the signals, (3) performing feature selection with BorutaSHAP, and (4) model selection to perform a classification task on flight difficulty.

where $\mathcal{I}$ is the set of raw feature indices, $\mathcal{J}$ is the set of observed sorties, and $\mathcal{K}_i$ is the set of FPC indices retained for $X_i(t)$. For additional information on our feature-generation approach and FPCA, we refer an interested reader to Appendix A and [82], respectively.

Table 3: Summary statistics generated for each data stream.

Variable	Summary Statistic Name	Summary Statistic Description
$f^S_{i,1}$	**Mean**	Mean of the data stream
$f^S_{i,2}$	**SD**	Standard deviation
$f^S_{i,3}$	**AvgAbsDiff**	Mean absolute deviation
$f^S_{i,4}$	**Min**	Minimum value of data stream
$f^S_{i,5}$	**Max**	Maximum value of data stream
$f^S_{i,6}$	**Range**	Difference between Min and Max
$f^S_{i,7}$	**Median**	Median value of data stream
$f^S_{i,8}$	**MedAbsDev**	Median absolute deviation
$f^S_{i,9}$	**IQR**	Interquartile range
$f^S_{i,10}$	**NegCount**	Count of negative values in the data stream
$f^S_{i,11}$	**PosCount**	Count of positive values in the data stream
$f^S_{i,12}$	**AboveMean**	Count of values above mean
$f^S_{i,13}$	**Skew**	Skew of the data stream
$f^S_{i,14}$	**Kurt**	Kurtosis of the data stream
$f^S_{i,15}$	**TimeMax**	Time at which data stream has maximum value
$f^S_{i,16}$	**TimeMin**	Time at which data stream has minimum value
$f^S_{i,17}$	**TimeAbsDiff**	Absolute value of the difference between TimeMax and TimeMin

The feature generation step results in a high number of features, and empirical testing reveals that their collective use produces inferior-performing, overfitted classifiers. Therefore, the next step in our pipeline utilizes the BorutaSHAP feature selection algorithm to determine which FPC scores and summary statistics should be retained. This algorithm is allocated 100 iterations in our pipeline. For further information on BorutaSHAP and our implementation, we refer an interested reader to [19] and Appendix B, respectively.

Finally, after feature generation and feature selection, our MMF-DSS pipeline concludes with a classification step. This research explores and compares four distinct machine learning techniques:

(1) adaptive boosting (AdaBoost), (2) neural network (NN), (3) k-nearest neighbor (k-NN), and (4) support vector machine (SVM). The mathematical details associated with each of these techniques are discussed in Appendix C. As shown in the next section, AdaBoost is the superlative classifier of those considered and, as such, is leveraged within our machine learning pipeline.

4. Testing, Results, and Analysis

This section tests, analyzes, and calibrates our machine learning pipeline. Results are provided for each step and, as appropriate, hyperparameter tuning is conducted to optimize the pipeline's performance. The best-performing model is presented and its ability to correctly classify a sortie's difficulty is examined. Experiments and analyses are conducted utilizing BorutaShap (Version 1.0.16) within the Python modeling environment as well as the Statistics and Machine Learning Toolbox within MATLAB 2021B on a Lenova ThinkPad equipped with a 2.60 GHz Intel i7-9850H processor and 64GB of RAM.

4.1. Feature Generation: FPCA Results

To reduce the dimensionality of the resulting FPC scores but ensure they retain sufficient information, we include for each time-series signal i the minimum number of eigenfunctions (i.e., K_i) that describe at least 95% of the variation. Results are summarized for every raw feature in Table 4.

By inspecting Table 4, one can immediately draw conclusions regarding the degree of variability within each stream of functional data. For example, the non-physiological streams required, on average, more FPCs to describe 95% of the data's variability than the physiological streams. No physiological variable required greater than five FPCs to do so, but some non-physiological variables (e.g., AX-variables) are characterized by up to nine FPCs. Moreover, among the physiological variables, some data streams can be expressed more compactly than others. Whereas the EMG- and ECG-variables require at most two FPCs, select ETK-variables require four to five FPCs.

Additional analysis of the FPCs associated with Table 4 can identify each variable's primary modes of variation. However, we refrain from conducting such analysis until Section 4.2 after feature selection has been performed.

4.2. Feature Selection: BorutaSHAP Results

Through our application of FPCA, we generate a finite set of features that efficiently extracts the temporal information from the infinitely dimensional raw signals. The finite cardinality of the generated feature set is an improvement; however, as can be observed from Tables 3 and 4, the large set of generated features could benefit from further dimensionality reduction; of the 1111 features generated, we desire to identify a most-valuable subset of physiological features that will not overburden a classification algorithm

In our implementation, BorutaSHAP was allocated 100 iterations. In so doing, the importance of the following nine physiological features was confirmed: $f^S_{3,11}$, $f^S_{11,13}$, $f^S_{16,8}$, $f^S_{17,2}$, $f^S_{20,2}$, $f^P_{22,3}$, $f^P_{23,3}$, $f^S_{23,15}$, and $f^S_{36,11}$. With reference to Tables 3 and 4, it can be observed that seven of the nine features correspond to summary statistics, whereas two of the nine correspond to FPC scores. Likewise, seven features relate to the eye-tracking measurements, and the remaining two correspond to electrodermal and cardiac measurements, respectively.

The relative importance of eye-tracking, electrodermal, and cardiac measurements on a flight task's difficulty accords with historical, psychological research. Whereas [83] first showed a positive association between pupil dilation and cognitive load, these results were later extended to identify a relationship between skin resistance and heart rate as well [84]. Therefore, our BorutaSHAP results provide a defensible extrapolation of these foundational, psychological results to the present setting.

Given that many of the identified features correspond with summary statistics, their interpretation is rather straight-forward. For example, $f^S_{11,13}$, $f^S_{16,8}$, $f^S_{17,2}$, and $f^S_{20,2}$ all correspond, in some

Table 4: Summary of FPCA Results.

Variables	No. of Retained FPCs	% Explained Variation	Variable	No. of Retained FPCs	% Explained Variation
$f^P_{1,k}$	2	98.5 %	$f^P_{31,k}$	4	95.6 %
$f^P_{2,k}$	1	98.4 %	$f^P_{32,k}$	1	99.2 %
$f^P_{3,k}$	3	96.5 %	$f^P_{33,k}$	1	98.7 %
$f^P_{4,k}$	3	98.3 %	$f^P_{34,k}$	N/A	N/A
$f^P_{5,k}$	2	98.3 %	$f^P_{35,k}$	N/A	N/A
$f^P_{6,k}$	1	96.7 %	$f^P_{36,k}$	1	97.8 %
$f^P_{7,k}$	1	96.7 %	$f^P_{37,k}$	7	95.7 %
$f^P_{8,k}$	1	98.7 %	$f^P_{38,k}$	3	96.7 %
$f^P_{9,k}$	1	98.7 %	$f^P_{39,k}$	3	96.3 %
$f^P_{10,k}$	1	96.6 %	$f^P_{40,k}$	5	95.7 %
$f^P_{11,k}$	2	98.3 %	$f^P_{41,k}$	5	95.4 %
$f^P_{12,k}$	2	97.8 %	$f^P_{42,k}$	4	95.5 %
$f^P_{13,k}$	2	98.5 %	$f^P_{43,k}$	3	95.5 %
$f^P_{14,k}$	1	95.9 %	$f^P_{44,k}$	3	97.7 %
$f^P_{15,k}$	2	98.5 %	$f^P_{45,k}$	4	96.2 %
$f^P_{16,k}$	5	96.5 %	$f^P_{46,k}$	4	95.7 %
$f^P_{17,k}$	4	95.5%	$f^P_{47,k}$	8	95.0 %
$f^P_{18,k}$	5	96.5 %	$f^P_{48,k}$	8	95.2 %
$f^P_{19,k}$	4	95.9 %	$f^P_{49,k}$	7	96.5 %
$f^P_{20,k}$	5	96.2 %	$f^P_{50,k}$	9	95.8 %
$f^P_{21,k}$	2	97.3 %	$f^P_{51,k}$	8	95.6 %
$f^P_{22,k}$	3	95.9 %	$f^P_{52,k}$	9	96.0 %
$f^P_{23,k}$	3	95.7 %	$f^P_{53,k}$	1	99.0 %
$f^P_{24,k}$	1	95.8 %	$f^P_{54,k}$	9	95.0 %
$f^P_{25,k}$	2	95.8 %	$f^P_{55,k}$	9	95.5 %
$f^P_{26,k}$	1	99.2 %	$f^P_{56,k}$	3	95.4 %
$f^P_{27,k}$	3	96.3 %	$f^P_{57,k}$	7	95.4 %
$f^P_{28,k}$	3	95.4 %	$f^P_{58,k}$	4	96.2 %
$f^P_{29,k}$	2	99.9 %	$f^P_{59,k}$	5	95.9 %
$f^P_{30,k}$	1	95.7 %	$f^P_{60,k}$	7	95.8 %

way, to gaze variability. The implication being that there exists a relationship between gaze volatility and task difficulty. Alternatively, although $f^P_{22,3}$ and $f^P_{23,3}$ are clearly related to pupil dilation over time, given that these features were generated by FPCA, their interpretation is more nuanced.

Figure 2 plots the associated FPCs for $f^P_{22,3}$ and $f^P_{23,3}$. Recalling that $f^P_{22,3}$ and $f^P_{23,3}$ are the third FPC scores for the diameter of the left and right pupil, respectively, we may inspect Figure 2 to infer their meaning by interpreting the underlying FPCs (i.e., eigenfunctions). The left graph in Figure 2 displays the FPC associated with $f^P_{22,3}$ and quantifies the weighted difference of left pupil diameter between mid-flight (approximately 250-525 s) and the beginning/end of flight. Similarly, the right graph depicts the FPC associated with $f^P_{23,3}$, and quantifies the weighted difference of right pupil diameter between mid-flight (approximately 300-550 s) and the beginning/end of flight. Weights applied to each time instance are objectively determined by the PACE algorithm. Both of these FPCs show a relatively similar pattern of the difference in pupil diameter during the middle of the landing task versus the beginning and end. This pattern likely relates to the increased concentration of pilots at the beginning and end of the landing task (when pupil sizes are larger) versus the middle (when pupil sizes are smaller).

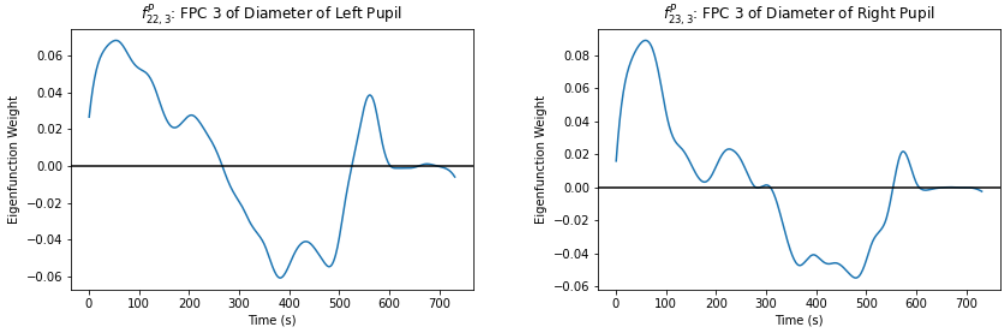

Figure 2: BorutaSHAP-identified Functional Principal Components

4.3. Classifier Development

A naive approach utilizing only the baseline settings of each technique yields adequate performance that we improve upon subsequently via hyperparameter tuning. However, inspection of Table 5 provides interesting insights in its own right. For example, the AdaBoost and SVM algorithms dominate the NN and k-NN approaches across each of four examined metrics, i.e., accuracy, precision, recall, and F1 score. Conversely, the baseline SVM and AdaBoost algorithms perform comparably across these measures. Subsequent analysis determines whether these results are anomalous or indicative of algorithmic performance at large.

Table 5: Results of Baseline Models

Technique	Accuracy (%)	Avg Precision (%)	Avg Recall (%)	Avg F1 Score (%)
Adaptive Boost	75.71	75.71	75.74	75.71
Neural Network	62.86	62.86	63.25	62.58
k-Nearest Neighbor	70.00	70.00	70.02	69.99
Support Vector Machine	75.71	75.71	76.25	75.59

We design, develop, and conduct a full factorial computational experiment consisting of a variety of hyperparameter settings for each machine learning technique utilizing the selected features from Section 4.2. Table 6 reports the hyperparameters, factor levels, baseline hyperparameter values, and best-tuned hyperparameter values for each of the four machine learning techniques. A 70-30 train-test split was utilized as well as a five-fold cross-validation.

Table 6: Hyperparameter Experimental Design Factor Levels

Technique	Hyperparameter	Factor Levels	Baseline Value	Best-Tuned Value
Adaptive Boosting	Learn Rate	{0.01, 0.02, . . . , 1}	1	0.09
	Num. Learning Cycles	{10, 20, . . . , 200}	100	35
Neural Network	Layer Sizes	{5, 10, . . . , 50}	10	5
	Num. Layers	{1, 2, 3}	1	1
	Lambda (i.e., Regularization)	{0, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3}	0	0.01
	Activations	{Identity, Relu, Sigmoid, tanh}	Relu	tanh
k-Nearest Neighbor	Num. Neighbors	{1, 2, . . . , 70}	1	31
	Distance	{Chebychev, Euclidean, Minkowski, Hamming}	Hamming	Euclidean
	Distance Weight	{Equal, Inverse, SquaredInverse}	Equal	Equal
Support Vector Machine	Kernel Function	{Linear, Polynomial, RBF}	Linear	Linear
	Box Constraint (i.e., C)	{0.0001, 0.0003, 0.001, 0.003, . . . , 1000, 3000}	1	0.0001
	Polynomial Order*	{2, 3, 4}	3	N/A

* Polynomial Order is only necessary if a polynomial kernel function is utilized

Table 7 presents each machine learning technique's best performance on the test set, and Figure 3 depicts their variability in accuracy across the examined hyperparameter settings. Noteworthy is that, when we compare the performance of each algorithm's best-tuned settings, AdaBoost not only dominates the NN and k-NN methods, but SVM as well. Alternatively, the SVM dominance over NN and k-NN vanishes under these conditions, implying that the results from Table 5 are not indicative of a global relationship between algorithms on this data set. Inspection of Figure 3 provides insight in this regard. Relative to the other three algorithms, AdaBoost is associated with a diminutive degree of variability; however, the same cannot be said about the NN and k-NN methodologies. Their performance is highly variable with substantial dependence on the selected hyperparameters, implying that the the baseline results derived from naturally occurring variation rather than structural disparities.

Table 7: Results of Best-Tuned Models

Technique	Accuracy (%)	Avg Precision (%)	Avg Recall (%)	Avg F1 Score (%)
Adaptive Boost	87.14	87.14	87.42	87.12
Neural Network	82.86	82.86	83.29	82.80
k-Nearest Neighbor	82.86	82.86	83.29	82.80
Support Vector Machine	81.43	81.43	83.65	81.12

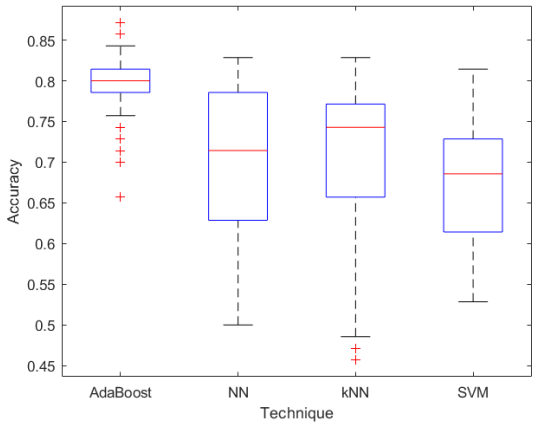

Figure 3: Boxplot of Technique and Accuracy (%)

4.3.1. Confirming the Best-tuned AdaBoost Model

Additional testing is also conducted to further investigate the relative variability within the AdaBoost classifer. More specifically, multiple linear regression analysis is conducted to investigate

the relationship between the hyperparameter settings (i.e., the learning rate and the number of learners) and the dependent variable (i.e., accuracy). Results from this analysis are depicted in Tables 8 and 9. Although not provided herein, analogous results to those in the aforementioned tables were observed with the remaining performance measures as well. Inspection of Table 8 reveals that a substantial degree of the variability observed within our AdaBoost tuning is associated with the first-order effects of our hyperparameter levels, i.e., approximately one third of the model error derives from these effects. Moreover, of the these two independent variables, the learn rate appears to affect the classifier's accuracy more than the number of learning cycles. Such results confirm the importance of hyperparameter tuning in our setting as well as the relative importance of the learning rate and learning cycle parameters in our application of AdaBoost, but they also imply that higher-order effects between these hyperparameters non-trivially affect the algorithm's output.

Table 8: Multiple Linear Regression Analysis of Variance

Source of Variation	SS	DF	MS	F_0	P
Model	0.4143	2	0.2072	421.9216	< 0.0001
Error	0.9805	1997	0.0005		
Total	1.3948	1999			

Table 9: Multiple Linear Regression Analysis Effects Test

Source of Variation	SS	DF	MS	F_0	P
Learn Rate	0.4129	1	0.4129	840.8910	< 0.0001
Num. Learning Cycles	0.0014	1	0.0014	2.9414	0.0865

Furthermore, the collective insights of Table 7 and Figure 3 confirm that AdaBoost is the preferable approach among those considered. Not only is its performance superlative across each of the examined metrics, but its relatively low degree of variability across varying hyperparameter settings suggests that it should yield comparable performance on unseen data. As a matter of course, we confirm this conjecture by additional validation presented in Figure 4. This analysis fixes the learning rate at 0.09 and varies the number of learning cycles. These hyperparameter pairs are tested upon the training set, testing set, and a five-fold cross-validation set. The classifier's respective accuracy is estimated as a function of the number of learning cycles for each data set. Intuitively, accuracy tends to increase with the number of learners and, whereas the graph indicates a minor degree of overfitting, the classifier still achieves high-quality performance on the test set with 87% accuracy. It is also noteworthy that the validation curves maximum accuracy corresponds to that of the test set, thereby lending further confidence in the selected hyperparameter values. Confirmatory testing was also conducted on the learning rate with similar results.

4.3.2. Implementing the Best-tuned AdaBoost Classifier

Having confirmed the superlative performance of the the best-tuned AdaBoost model, we turn our attention to implementation concerns. Whereas analysis in the previous section highlights the holistic efficacy of the AdaBoost classifer, from an implementation perspective, its efficacy when conditioned upon a sortie's true label is of utmost importance.

Such is the focus of Figure 5 wherein the output probabilities of a sortie being classified as high-difficulty are presented conditioned upon their true label. The histogram on the left of Figure 5 aggregates the AdaBoost high-difficulty probabilities for true low-difficulty sorties. Its right-skew indicates that the majority of low-difficulty sorties are associated with a low high-difficulty score. Similarly, the right histogram depicts the same response for true high-difficulty sorties. Its left skew indicates that a majority high-difficulty sorties are associated with a greater high-difficulty score.

Such analysis can be extended by focusing on the classifier's susceptibility to varying thresholds. The best-tuned AdaBoost algorithm was used to construct the ROC curve presented in Figure 6.

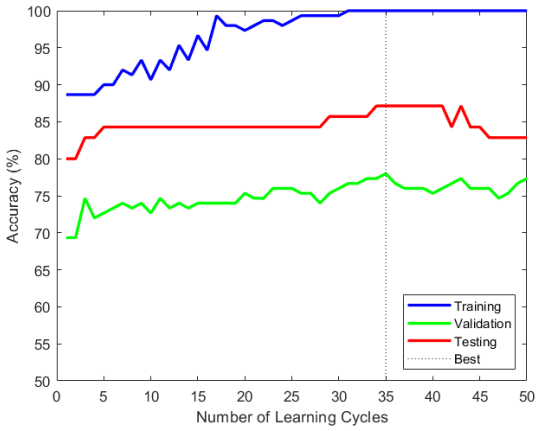

Figure 4: Accuracy Curves For Training, Validation, and Testing Sets Obtained From AdaBoost Model With Learn Rate = 0.09

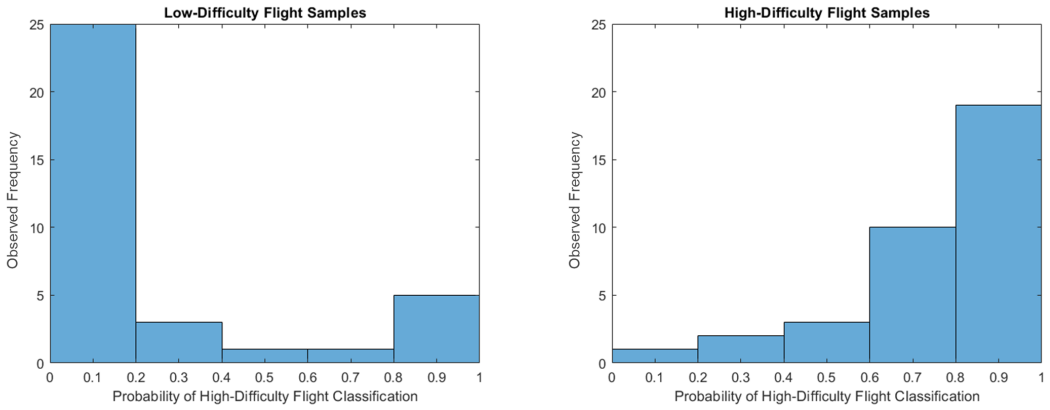

Figure 5: Best-tuned AdaBoost Classifier Output by True Difficulty Level

It can be observed that, as the false-positive-rate threshold is allowed to increase, the true positive rate increases drastically (i.e., where a positive response corresponds to a high-difficultly sortie). For a minor decrease in specificity, the classifier makes substantial gains in sensitivity. The area under the curve (AUC) associated with this classifier is ≈ 0.88, which is *excellent* and bordering on the 0.9 *outstanding* benchmark, following the standards set forth by [20].

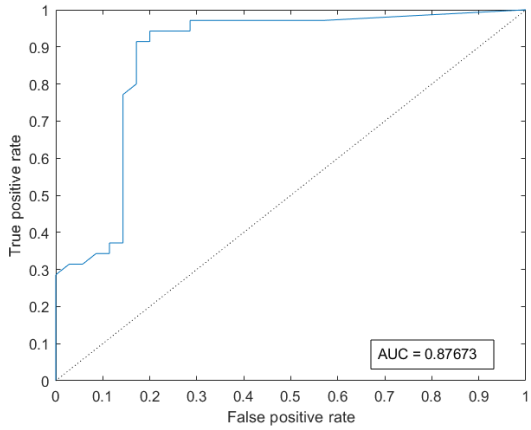

Figure 6: ROC Curve and AUC for Best-Tuned AdaBoost Model

5. Executive Decision making and Policymaking Implications

Having developed an effective machine learning pipeline for flight difficulty classification, we turn our attention to the practical matters of executive decision making and policymaking. As noted by [85], to enable executive decision making, a pipeline's performance and function must be clearly communicated to all relevant stakeholders. The USAF in particularly has placed a great emphasis on explainability for this exact reason [86]. As such, we present herein a simple means to explain the rationale behind our pipeline's predictions via an analysis of each sorties' SHAP value. Furthermore, although we contend that the classifier developed in Section 4 is preferable, executive decision making demands it be compared to alternatives. Therefore, in this section, we present two alternative classifiers that rely on less and additional data streams, respectively. The modification of the utilized data streams affects not only the classifier's accuracy but also its implementation cost. The balance of these two competing objectives must be weighed by the executive decision maker, and we discuss the corresponding implications of each decision.

5.1. Model Explainability and Feature Importance

SHAP values extend the canonical Shapley value from coalitional game theory to a machine learning setting. SHAP values are calculated on a sortie-by-sortie bases such that, for each observation, one may quantify how each feature affected its classification. Figure 7 plots each feature's SHAP values across every sortie; its mean SHAP value is depicted as well (i.e., via the red diamond). Through inspection of this plot, insight can immediately be drawn with regards to the relative importance of each factor.

The most influential factors correspond to the FPCA-generated features, i.e., $f^P_{22,3}$ and $f^P_{22,3}$. Not only do their mean SHAP values have the greatest magnitude, but they are also associated with the greatest variability. More specifically, the differences between their minimum and maximum SHAP values are approximately 1.1 and 0.9, respectively. In juxtaposition, $f^S_{20,2}$ has a range of approximately 0.2. Whereas the relative importance of these two features reinforces our utilization

of FPCA, the dominance of the ETK features in Figure 7 emphasizes the utility of the Vive Pro Eye headset in our application. The top five most influential features correspond to ETK measurements; the ECG and PPG data (i.e., $f^S_{3,11}$ and $f^S_{36,11}$) are the sixth and eighth most-influential factor as measured by their mean SHAP values.

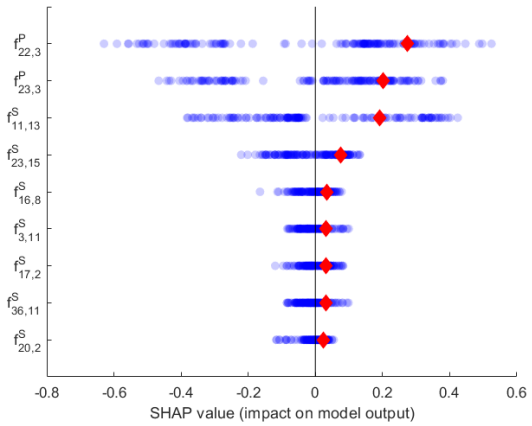

Figure 7: SHAP Values using Best-Tuned AdaBoost Model

From a policycmaking perspective, this SHAP-value approach to feature importance is of significant utility. Given that SHAP values are calculated on a sortie-by-sortie basis, model diagnostic checking is greatly facilitated. This is particularly important in flight training in the case of accidents. For example, if a training accident occurs, the USAF assigns an accident investigation review board to examine it. SHAP values may be directly leveraged in this setting to ascertain why the on-board decision support system behaved as it did.

5.2. Reducing Implementation Cost via an Eye-tracking-focused Model

The relative importance of the ETK features in comparison to the other modality naturally suggests a policymaking alternative. Namely, if the ETK features are the most important, it may be possible to construct an adequately performing classifier that does not require ECG or PPG data. Whereas such a classifer may result in degraded performance, its implementation also requires less equipment and is therefore less expensive.

To examine this tradeoff, we perform a second factorial experiment based on the levels outlined in Table 6; however, in this case, only the ETK features derived from Section 4.2 are included in the model. The performance of the best-tuned model across each algorithm class is presented in Table 10. The results illustrate that AdaBoost remains the superlative algorithm, though its performance is slightly degraded from the classifier developed with multi-modal data in Section 4.3. Of note, is that, even though AdaBoost performs well across all performance measures, it no longer dominates the other algorithms. The best-tuned SVM and k-NN algorithms have slightly better recall than AdaBoost. Furthermore, it is also interesting to note that, whereas as the best-tuned AdaBoost, NN and SVM classifiers are degraded when trained on ETK data alone, the k-NN classifier improves on each metric.

Table 10: Results of Best-Tuned Models using only ETK Features

Technique	Accuracy (%)	Avg Precision (%)	Avg Recall (%)	Avg F1 Score (%)
Adaptive Boost	85.71	85.71	85.71	85.71
Neural Network	78.57	78.57	80.59	78.21
k-Nearest Neighbor	84.29	84.29	86.71	84.02
Support Vector Machine	82.86	82.86	85.78	82.50

Nevertheless, the performance of the best-performing model (i.e., AdaBoost) is degraded when utilizing ETK data alone. Ultimately, this results in a policymaking tradeoff between implementation cost and classifier performance that must be resolved by executive-level decision makers. We discuss this matter further in Section 5.4.

5.3. Increasing Predictive Accuracy with Aircraft and Pilot Data

The focus of this research is the rapid introduction of an automated, flight difficulty classifier via the exclusive utilization of physiological data. From a policymaking perspective, such a classifier may be viewed as an intermediate step due to the costly (and cumbersome) nature of acquiring new, modern aircraft and simulators. However, in truth, the purchase of new aircraft and simulators with subsystems optimized for machine learning applications is a third course of action that may be pursued by executive decision makers. In fact, depending upon how the inclusion of additional aircraft-and-pilot data modalities affects the classifier's performance, executives may decide that such subsystems are not necessary.

Therefore, in this section, we perform a third full factorial design on based on the levels outlined in Table 6, but now allow every feature outlined in Table 2 to flow through the pipeline depicted in Figure 1. Results from this experiment are provided in Table 11 wherein the performance of the best-tuned models are presented across each of our four performance metrics. As in Sections 4.3 and 5.2, AdaBoost remains the superlative methodology for use in our pipeline.

Table 11: Results of Best-Tuned Models using every Data Modality

Technique	Accuracy (%)	Avg Precision (%)	Avg Recall (%)	Avg F1 Score (%)
Adaptive Boost	97.14	97.14	97.30	97.14
Neural Network	94.29	94.29	94.87	94.27
k-Nearest Neighbor	97.14	97.14	97.14	97.14
Support Vector Machine	94.29	94.29	94.87	94.27

When utilizing the full data set, the performance of every classifier improves over every performance measure. This implies that the AX and PX data modalities contain valuable information for flight difficulty classification. Moreover, as in Section 5.2, the performance of the best-performing k-NN model is noteworthy; its performance mirrors AdaBoost across each measure. The best-performing NN and SVM models also perform comparable but at a slightly diminished level compared to the AdaBoost and k-NN models. Finally, it is also interesting to point out how the feature selection step performed when utilizing all data models as opposed to the physiological streams alone. The BorutaSHAP iterations remained constant at 100 and, when utilizing the the full data set, no physiological features were selected, implying that, when given the option, a high-quality classifier does not require physiological data. Although it is costly to acquire the necessary equipment, the USAF could construct a better-performing, automated classifier by acquiring new aircraft and simulators.

5.4. Policymaking Tradespace

Whereas the SHAP value analysis provided herein ensures the explainability of our pipeline, the additional analysis based upon alternative feature selection assumptions presents a policymaking tradespace that must be decided at the executive level. The classifier developed in Section 4.3 can be viewed as a compromise between two competing objectives: implementation cost and classifier accuracy. However, it represents but a single point on the Pareto front associated with these objectives. The classifier presented in Section 5.2 can be implemented with less cost but with degraded performance, whereas the classifier presented in Section 5.3 is most costly to implement but yields improved performance. A proper decision among these alternatives, depends upon the relative worth executive-level decision makers place on each objective. There is no right answer. However, canonical techniques from decision analysis may be leveraged to distinguish among them [87]. Doing so requires the specification of a multi-attribute utility function as well as the elicitation

of related uncertainties and outcomes (e.g., by treating implementation cost as a random variable); [88] and [89] provide a thorough explanation of the quantitative techniques necessary to do so. Such an undertaking is a promising avenue of future inquiry in its own right, but the results provided herein are foundational to its conduct.

6. Conclusions

The USAF must change with the times. Its pilot shortfall is a grave national-security threat that will not be solved by the status quo. The automation of instructor and evaluator pilots is an innovative idea meant to alleviate the shortfall by reducing bottlenecks in the training pipeline. However, numerous technical complications surround its implementation. The automated classification of flight difficulty is one such issue and is the focus of the machine learning pipeline developed herein.

Utilizing multimodal, functional data from a designed experiment of pilots landing a simulated aircraft, our MMF-DSS pipeline leverages functional principal component analysis to distill a large set of tabular features that are then reduced to a most-valuable subset via the BorutaSHAP feature selection algorithm. Multiple full factorial designs illustrated that the AdaBoost classifier is then best able to utilize these features to predict a flight's difficulty. SHAP values are then utilized to ensure the explainability of the classifier and inform executive decision making. Bearing in mind the trade-off between classifier accuracy and cost effectiveness, alternative data subsets and classifiers are considered that represent distinct, approximately Pareto-efficient solutions. The policymaking tradespace between these classifiers is explored and techniques are discussed through which a resolution may be ascertained.

This research has therefore laid a foundation upon which future research can build. However, much like the automation of self-driving cars, the problem of automating an instructor pilot is far from solved. Myriad avenues of future research exist in such diverse areas as maneuver grading to performance-improvement recommender systems. The study of this problem is in its infancy, and it remains a research field of significant inquiry.

Disclaimer

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, United States Department of Defense, or United States Government.

Acknowledgments

This research is partially supported by the Air Force Office of Scientific Research (AFOSR) under the Dynamic Data and Information Processing (DDIP) portfolio. CogPilot Dataset provided by the United States Air Force pursuant to Cooperative Agreement Number FA8750-19-2-1000

Author Contributions

W.N.C performed conceptualization, methodology, and writing - review and editing. N.G. and P.R.J both performed conceptualization, methodology, formal analysis, and were involved in all phases of writing. C.J. performed conceptualization, methodology and writing - review and editing.

References

[1] J. Hunter, The Truth About The Air Force's Biggest Changes To Pilot Training Since The Dawn Of The Jet Age, The Drive, accessed 29 November 2021 (2021).

[2] S. Yang, K. Yu, T. Lammers, F. Chen, Artificial Intelligence in Pilot Training and Education–Towards a Machine Learning Aided Instructor Assistant for Flight Simulators, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 581–587.

[3] P. Tucker, An Army Pilot Just Re-Invented Flight Training for the Digital Era, https://www.defenseone.com/technology/2021/09/army-pilot-just-re-invented-flight-training-digital-era/185690/, accessed 29 November 2021 (2021).

[4] L. Li, Anomaly detection in airline routine operations using flight data recorder data, Ph.D. thesis, Massachusetts Institute of Technology (2013).

[5] L. Li, S. Das, R. John Hansman, R. Palacios, A. N. Srivastava, Analysis of Flight Data using Clustering Techniques for Detecting Abnormal Operations, Journal of Aerospace Information Systems 12 (9) (2015) 587–598.

[6] W. Zhao, L. Li, S. Alam, Y. Wang, An Incremental Clustering Method for Anomaly Detection in Flight Data, Transportation Research Part C: Emerging Technologies 132 (2021) 103406.

[7] L. Li, R. J. Hansman, R. Palacios, R. Welsch, Anomaly Detection via a Gaussian Mixture Model for Flight Operation and Safety Monitoring, Transportation Research Part C: Emerging Technologies 64 (2016) 45–57.

[8] M. Memarzadeh, B. Matthews, I. Avrekh, Unsupervised anomaly detection in flight data using convolutional variational auto-encoder, Aerospace 7 (8) (2020) 115.

[9] S. T. Barratt, M. J. Kochenderfer, S. P. Boyd, Learning probabilistic trajectory models of aircraft in terminal airspace from position data, IEEE Transactions on Intelligent Transportation Systems 20 (9) (2018) 3536–3545.

[10] S. M. Katz, A.-C. Le Bihan, M. J. Kochenderfer, Learning an urban air mobility encounter model from expert preferences, in: 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), IEEE, 2019, pp. 1–8.

[11] S. Choi, Y. J. Kim, S. Briceno, D. Mavris, Prediction of weather-induced airline delays based on machine learning algorithms, in: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), IEEE, 2016, pp. 1–6.

[12] G. Gui, Z. Zhou, J. Wang, F. Liu, J. Sun, Machine learning aided air traffic flow analysis based on aviation big data, IEEE Transactions on Vehicular Technology 69 (5) (2020) 4817–4826.

[13] P. R. Jenkins, W. N. Caballero, R. R. Hill, Predicting success in united states air force pilot training using machine learning techniques, Socio-Economic Planning Sciences (2021) 101121.

[14] O. Pawlyk, New T-7 Red Hawk Trainer Faces Delays over Parts Shortages, Testing, https://www.military.com/daily-news/2021/06/18/new-t-7-red-hawk-trainer-faces-delays-over-parts-shortages-testing.html, accessed 30 November 2021 (2021).

[15] S. S. Oakley, GAO-19-439:DOD Acquisition Reform: Leadership Attention Needed to Effectively Implement Changes to Acquisition Oversight , Tech. rep., U.S. Government Accountability Office (2019).

[16] T. Heldt, B. Long, G. C. Verghese, P. Szolovits, R. G. Mark, Integrating data, models, and reasoning in critical care, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2006, pp. 350–353.

[17] Y. Dong, Z. Hu, K. Uchimura, N. Murayama, Driver inattention monitoring system for intelligent vehicles: A review, IEEE transactions on intelligent transportation systems 12 (2) (2010) 596–614.

[18] J.-L. Wang, J.-M. Chiou, H.-G. Müller, Functional data analysis, Annual Review of Statistics and Its Application 3 (2016) 257–295.

[19] E. Keany, BorutaShap: A Wrapper Feature Selection Method which Combines the Boruta Feature Selection Algorithm with Shapley Values (2020).

[20] J. N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (9) (2010) 1315–1316.

[21] B. Erbas, M. Akram, D. M. Gertig, D. English, J. L. Hopper, A. M. Kavanagh, R. Hyndman, Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the united states and england–wales, Journal of epidemiology 20 (2) (2010) 159–165.

[22] R. J. Hyndman, H. Booth, Stochastic population forecasts using functional data models for mortality, fertility and migration, International Journal of Forecasting 24 (3) (2008) 323–342.

[23] A. Laukaitis, Functional data analysis for cash flow and transactions intensity continuous-time prediction using hilbert-valued autoregressive processes, European Journal of Operational Research 185 (3) (2008) 1607–1614.

[24] M. Febrero-Bande, P. Galeano, W. González-Manteiga, Functional principal component regression and functional partial least-squares regression: An overview and a comparative study, International Statistical Review 85 (1) (2017) 61–83.

[25] S. Greven, F. Scheipl, A general framework for functional regression modelling, Statistical Modelling 17 (1-2) (2017) 1–35.

[26] P. T. Reiss, J. Goldsmith, H. L. Shang, R. T. Ogden, Methods for scalar-on-function regression, International Statistical Review 85 (2) (2017) 228–249.

[27] N. Ling, P. Vieu, Nonparametric modelling for functional data: selected survey and tracks for future, Statistics 52 (4) (2018) 934–949.

[28] G. Geenens, Curse of dimensionality and related issues in nonparametric functional regression, Statistics Surveys 5 (2011) 30–43.

[29] J. Jacques, C. Preda, Functional data clustering: a survey, Advances in Data Analysis and Classification 8 (3) (2014) 231–255.

[30] P.-S. Wu, H.-G. Müller, Functional embedding for the classification of gene expression profiles, Bioinformatics 26 (4) (2010) 509–517.

[31] S. B. Kim, P. Rattakorn, Y. B. Peng, An effective clustering procedure of neuronal response profiles in graded thermal stimulation, Expert Systems with Applications 37 (8) (2010) 5818–5826.

[32] B. J. Parker, J. Wen, Predicting microrna targets in time-series microarray experiments via functional data analysis, BMC bioinformatics 10 (1) (2009) 1–10.

[33] C.-R. Jiang, J. A. Aston, J.-L. Wang, Smoothing dynamic positron emission tomography time courses using functional principal components, NeuroImage 47 (1) (2009) 184–193.

[34] V. Baladandayuthapani, B. K. Mallick, M. Young Hong, J. R. Lupton, N. D. Turner, R. J. Carroll, Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis, Biometrics 64 (1) (2008) 64–73.

[35] E. A. Crane, R. B. Cassidy, E. D. Rothman, G. E. Gerstner, Effect of registration on cyclical kinematic data, Journal of biomechanics 43 (12) (2010) 2444–2447.

[36] S. Ullah, C. F. Finch, Functional data modelling approach for analysing and predicting trends in incidence rates—an application to falls injury, Osteoporosis international 21 (12) (2010) 2125–2134.

[37] S. Ullah, C. F. Finch, Applications of functional data analysis: A systematic review, BMC medical research methodology 13 (1) (2013) 1–12.

[38] C. Croux, A. Ruiz-Gazen, High breakdown estimators for principal components: the projection-pursuit approach revisited, Journal of multivariate analysis 95 (1) (2005) 206–226.

[39] F. Ferraty, P. Vieu, Nonparametric functional data analysis: theory and practice, Springer Science & Business Media, 2006.

[40] C. R. Rao, Some statistical methods for comparison of growth curves, Biometrics 14 (1) (1958) 1–17.

[41] P. Besse, J. O. Ramsay, Principal components analysis of sampled functions, Psychometrika 51 (2) (1986) 285–311.

[42] P. E. Castro, W. H. Lawton, E. Sylvestre, Principal modes of variation for processes with continuous sample curves, Technometrics 28 (4) (1986) 329–337.

[43] C. S. Berkey, N. M. Laird, I. Valadian, J. Gardner, Modelling adolescent blood pressure patterns and their prediction of adult pressures, Biometrics (1991) 1005–1018.

[44] J. A. Rice, B. W. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves, Journal of the Royal Statistical Society: Series B (Methodological) 53 (1) (1991) 233–243.

[45] M. Jones, J. A. Rice, Displaying the important features of large collections of similar curves, The American Statistician 46 (2) (1992) 140–145.

[46] J. G. Staniswalis, J. J. Lee, Nonparametric regression analysis of longitudinal data, Journal of the American Statistical Association 93 (444) (1998) 1403–1418.

[47] P. C. Besse, H. Cardot, F. Ferraty, Simultaneous non-parametric regressions of unbalanced longitudinal data, Computational Statistics & Data Analysis 24 (3) (1997) 255–270.

[48] J. Boularan, L. Ferré, P. Vieu, Growth curves: a two-stage nonparametric approach, Journal of statistical planning and inference 38 (3) (1994) 327–350.

[49] M. Shi, R. E. Weiss, J. M. Taylor, An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves, Journal of the Royal Statistical Society: Series C (Applied Statistics) 45 (2) (1996) 151–163.

[50] J. A. Rice, C. O. Wu, Nonparametric mixed effects models for unequally sampled noisy curves, Biometrics 57 (1) (2001) 253–259.

[51] G. M. James, T. J. Hastie, C. A. Sugar, Principal component models for sparse functional data, Biometrika 87 (3) (2000) 587–602.

[52] G. M. James, C. A. Sugar, Clustering for sparsely sampled functional data, Journal of the American Statistical Association 98 (462) (2003) 397–408.

[53] F. Yao, H.-G. Müller, J.-L. Wang, Functional data analysis for sparse longitudinal data, Journal of the American statistical association 100 (470) (2005) 577–590.

[54] A. J. Simpkin, C. Metcalfe, R. M. Martin, J. A. Lane, J. L. Donovan, F. C. Hamdy, D. E. Neal, K. Tilling, Longitudinal prostate-specific antigen reference ranges: choosing the underlying model of age-related changes, Statistical methods in medical research 25 (5) (2016) 1875–1891.

[55] B. A. Goldstein, G. M. Pomann, W. C. Winkelmayer, M. J. Pencina, A comparison of risk prediction methods using repeated observations: an application to electronic health records for hemodialysis, Statistics in medicine 36 (17) (2017) 2750–2763.

[56] Q. Wang, S. Zheng, A. Farahat, S. Serita, C. Gupta, Remaining useful life estimation using functional data analysis, in: 2019 ieee international conference on prognostics and health management (icphm), IEEE, 2019, pp. 1–8.

[57] J.-M. Kim, N. Wang, Y. Liu, Multi-stage change point detection with copula conditional distribution with pca and functional pca, Mathematics 8 (10) (2020) 1777.

[58] Y. Dong, T. Xia, X. Fang, Z. Zhang, L. Xi, Prognostic and health management for adaptive manufacturing systems with online sensors and flexible structures, Computers & Industrial Engineering 133 (2019) 57–68.

[59] X. Fang, K. Paynabar, N. Gebraeel, Multistream sensor fusion-based prognostics model for systems with single failure modes, Reliability Engineering & System Safety 159 (2017) 322–331.

[60] X. Li, X. Fang, Multistream sensor fusion-based prognostics model for systems under multiple operational conditions, in: International Manufacturing Science and Engineering Conference, Vol. 85079, American Society of Mechanical Engineers, 2021, p. V002T09A003.

[61] Y. Cheng, C. Lu, T. Li, L. Tao, Residual lifetime prediction for lithium-ion battery based on functional principal component analysis and bayesian approach, Energy 90 (2015) 1983–1993.

[62] J. Guo, Z. Li, Prognostics of lithium ion battery using functional principal component analysis, in: 2017 IEEE International Conference on Prognostics and Health Management (ICPHM), IEEE, 2017, pp. 14–17.

[63] T. Xia, Y. Dong, E. Pan, M. Zheng, H. Wang, L. Xi, Fleet-level opportunistic maintenance for large-scale wind farms integrating real-time prognostic updating, Renewable Energy 163 (2021) 1444–1454.

[64] K. Jana, D. Sengupta, S. Kundu, A. Chakraborty, P. Shaw, The statistical face of a region under monsoon rainfall in eastern india, Journal of the American Statistical Association 115 (532) (2020) 1559–1573.

[65] X. Chang, Z. Zhu, X. Dai, J. Hobbs, A geospatial functional model for oco-2 data with application on imputation and land fraction estimation, arXiv preprint arXiv:2101.09418.

[66] N. Kantanantha, N. Serban, P. Griffin, Yield and price forecasting for stochastic crop decision planning, Journal of agricultural, biological, and environmental statistics 15 (3) (2010) 362–380.

[67] R. K. Wong, Y. Li, Z. Zhu, Partially linear functional additive models for multivariate functional data, Journal of the American Statistical Association 114 (525) (2019) 406–418.

[68] H. Liu, M. Zhou, Q. Liu, An embedded feature selection method for imbalanced data classification, IEEE/CAA Journal of Automatica Sinica 6 (3) (2019) 703–715.

[69] S. Maldonado, J. López, Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for svm classification, Applied Soft Computing 67 (2018) 94–105.

[70] C.-W. Chen, Y.-H. Tsai, F.-R. Chang, W.-C. Lin, Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results, Expert Systems 37 (5) (2020) e12553.

[71] R. Sheikhpour, M. A. Sarram, S. Gharaghani, M. A. Z. Chahooki, A robust graph-based semi-supervised sparse feature selection method, Information Sciences 531 (2020) 13–30.

[72] C. Shi, C. Duan, Z. Gu, Q. Tian, G. An, R. Zhao, Semi-supervised feature selection analysis with structured multi-view sparse regularization, Neurocomputing 330 (2019) 412–424.

[73] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.

[74] S. M. Lundberg, S.-I. Lee, Consistent feature attribution for tree ensembles, arXiv preprint arXiv:1706.06060.

[75] L. Brattain, S. Yuditskaya, J. Koerner, G. Ciccarelli, R. Carter, E. Cowen, T. Broderick, V. Sze, H. Reynolds, K. McAlpin, H. Rao, T. Heldt, CogPilot Challenge Problem Datasheet, Tech. rep., USAF-MIT AI Accelerator (2021).

[76] Shimmer Sensing, EMG User Guide, https://shimmersensing.com/wp-content/docs/support/documentation/EMG_User_Guide_Rev1.12.pdf, accessed 6 December 2021 (2021).

[77] Shimmer Sensing, GSR+ User Guide, https://shimmersensing.com/wp-content/docs/support/documentation/GSR_User_Guide_rev1.13.pdf, accessed 6 December 2021 (2021).

[78] Shimmer Sensing, Optical Pulse Sensor User Guide, https://shimmersensing.com/wp-content/docs/support/documentation/Optical_Pulse_Sensor_User_Guide_rev1.6.pdf, accessed 6 December 2021 (2021).

[79] Shimmer Sensing, ECG User Guide, https://shimmersensing.com/wp-content/docs/support/documentation/ECG_User_Guide_Rev1.12.pdf, accessed 6 December 2021 (2021).

[80] VIVE, VIVE Pro Eye, https://www.vive.com/us/support/vive-pro-eye/, accessed 6 December 2021 (2021).

[81] XPlane, X-Plane Datarefs, https://developer.x-plane.com/datarefs/, accessed 6 December 2021 (2021).

[82] H. L. Shang, A survey of functional principal component analysis, AStA Advances in Statistical Analysis 98 (2) (2014) 121–142.

[83] D. Kahneman, J. Beatty, Pupil diameter and load on memory, Science 154 (3756) (1966) 1583–1585.

[84] D. Kahneman, B. Tursky, D. Shapiro, A. Crider, Pupillary, heart rate, and skin resistance changes during a mental task., Journal of experimental psychology 79 (1p1) (1969) 164.

[85] A. Dhinakaran, Survey Of AI Teams Points To Promise And Peril Ahead, https://www.forbes.com/sites/aparnadhinakaran/2022/02/07/survey-of-ai-teams-points-to-promise-and-peril-ahead/?sh=709ef0d649fe (2022).

[86] G. Zacharias, Autonomous Horizons: The Way Forward, Air University Press; Curtis E. LeMay Center for Doctrine Development and . . . , 2019.

[87] R. L. Keeney, Decision analysis: an overview, Operations research 30 (5) (1982) 803–838.

[88] R. L. Keeney, H. Raiffa, R. F. Meyer, Decisions with multiple objectives: preferences and value trade-offs, Cambridge university press, 1993.

[89] P. P. Wakker, Prospect theory: For risk and ambiguity, Cambridge university press, 2010.