Off-nominal event analysis in autonomous flights based on explainable artificial intelligence

Shivakumar I. Ranganathan*, Hari. S. Ilangovan†, Newton H. Campbell Jr.‡, Michael J. Acheson§, and Irene M. Gregory¶
NASA Langley Research Center, Hampton, VA, 23681, USA

A key objective in the Urban Air Mobility program at NASA is to intelligently perform an autonomous flight in a complex urban environment under all weather conditions with guaranteed levels of safety. To accomplish this, the mission manager (central decision-making module) of the vehicle needs to make informed decisions between various Courses of Action (CoA) based on its' interpretation of the inputs it receives. If an off-nominal event is detected either based on the amalgamation of sensor data or the use of machine learning models, the mission manager may greatly benefit from identification of the input features that most likely contributed to that specific event. Such an understanding is usually not possible to obtain from the classical machine learning models (deep learning) due to the inherent black box like structure. However, this understanding is achieved using eXplainable Artificial Intelligence (XAI) models that provide a human interpretable rationale for the predictions made. This work presents a game theory inspired XAI model for the off-nominal assessment of autonomous flights. The proposed approach based on Shapley values is model agnostic, provides local as well as global explanation and satisfies the four axioms (efficiency, symmetry, dummy, additivity) to achieve fair contribution. The versatility of the approach is first demonstrated on a simulated dataset in which the significance of each input to flight phase prediction is clearly identified. Subsequently, data from simulated flight trajectories are fed into the model which reveal the input features that most likely contributed to a rotor failure event thereby empowering the mission manager to take the appropriate CoA.

I. Acronyms

AI	=	Artificial Intelligence
$ARMA$	=	Autoregressive Moving Average
CoA	=	Courses of Action
$CVAE$	=	Conditional Variational AutoEncoder
$DARPA$	=	Defense Advanced Research Projects Agency
$DeepLIFT$	=	Deep Learning Important FeaTures
GVS	=	Generalized Vehicle Simulation
$HAVOK$	=	Hankel Alternative View of Koopman
ICE	=	Individual Conditional Expectation
ICM	=	Intelligent Contingency Management
$KernelSHAP$	=	Kernel Shapley Additive exPlanations
$LIME$	=	Local Interpretable Model-agnostic Explanations
LRP	=	Layer-wise Relevance Propagation
$NASA$	=	National Aeronautics and Space Administration
$OODA$	=	Observe-Orient-Decide-Act
PDP	=	Partial Dependence Plots
$SHAP$	=	SHapley Additive exPlanations

*Senior Principal Solutions Architect, Science Applications International Corporation, NASA Enterprise IT
†Senior Data Scientist, Science Applications International Corporation, NASA Enterprise IT
‡Director of Space Programs, Australian Remote Operations for Space and Earth (AROSE) Consortium
§Senior Researcher, Advanced Control Theory and Applications, NASA Langley Research Center
¶Senior Researcher, Advanced Control Theory and Applications, NASA Langley Research Center

TTT	=	Transformational Tools and Technologies
UAM	=	Urban Air Mobility
$VARMAX$	=	Vector Autoregression Moving-Average with Exogenous Regressors
XAI	=	eXplainable Artificial Intelligence

II. Introduction

The Transformational Tools and Technologies (TTT) project within the Aeronautics Research Mission Directorate at NASA strives to develop state-of-the-art computational and experimental tools to advance the prediction of aircraft performance in flight with varying levels of mission and environmental complexities. Among the various research initiatives undertaken in TTT, the objective of the Intelligent Contingency Management (ICM) team is to intelligently operate an autonomous flight in a complex urban environment under all weather conditions with guaranteed levels of safety. To accomplish this, the research team formulated the notion of a cognitive mission manager that develops and decides among Courses of Action (CoAs) to enable the vehicle to complete the mission safely. As the name indicates, the cognitive mission manager is a combination of two subsystems: i) Cognitive Architecture and ii) Mission Manager. From the perspective of the OODA (Observe-Orient-Decide-Act) loop concept, the Cognitive Architecture represents the Observe part while the Mission Manager represents the Orient-Decide-Act part. In essence, the Cognitive Architecture hosts a variety of algorithms whose primary purpose is to observe and assess the state of the vehicle and feed this information to the Mission Manager. In this research, the Cognitive Architecture hosts several algorithms including a deep learning based Conditional Variational Autoencoder (CVAE) that triggers a flag in case an assessment is made for an off-nominal event and/or loss of control in the autonomous flight. Although such a flag is a very useful indicator, its value to assist with the Orient-Decide-Act functions of the Mission Manager is limited. This is due to the complexity of the CVAE which accurately predicts the flag but is unable to assess the input features that likely triggered the event, which is often crucial to correctly determining a safe response. A possible approach to provide this insight (i.e. input features that most likely triggered the flag) would be to use eXplainable Artificial Intelligence (XAI) models that help understand the role of each input feature towards the predicted output.

The notion of eXplainable Artificial Intelligence (XAI) was pioneered by Defense Advanced Research Projects Agency (DARPA) who defined explainable AI as systems that could explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future [1]. This endeavor involved the development of more interpretable machine learning techniques, the creation of effective interfaces for explanations, and an exploration of the psychological aspects related to effective explanations. In a subsequent perspective, Rudin [2] examined the use of black box machine learning models in high-stakes decision-making across various domains, highlighting the problems they posed. The need to shift the focus towards designing inherently interpretable models rather than explaining black box models was emphasized due to the possibility of the latter to perpetuate bad practices or to potentially harm society.

In practice, there is a conflict between prediction accuracy and model interpretability which was clearly elucidated by Lundberg et al. [3], who introduced the SHAP (SHapley Additive exPlanations) framework for explainability. The approach introduced by the authors assigned importance values to features for individual predictions which was inspired by ideas from cooperative game theory. SHAP unified a class of additive feature importance methods, including six existing ones, and demonstrated a unique solution with desirable properties. This unity in SHAP's approach suggested a common thread in model interpretation principles for future methods. The paper also discussed estimation methods for SHAP values and suggested further research directions for faster, model-specific estimation methods, integration of game theory for interaction effects, and the development of new explanation model classes.

In a subsequent study, Schlegel et al. [4] explored the use of eXplainable Artificial Intelligence (XAI) methods, typically designed for black-box machine learning models, in the context of time series data. The authors introduced a methodology to adapt and evaluate XAI methods for time series, emphasizing the temporal dimension. Preliminary experiments indicated that SHAP was robust for all models, while DeepLIFT (Deep Learning Important FeaTures), LRP (Layer-wise Relevance Propagation), and Saliency Maps performed better for specific architectures. Similarly, LIME's (Local Interpretable Model-agnostic Explanations) performance was shown to be suboptimal due to high dimensionality when converting time to features. The results suggested the need for more tailored XAI methods for time series to enhance human understanding, especially in cases where visual saliency masks were challenging to interpret. In addition, Belle et al. [5] discussed the role of artificial intelligence (AI) in enhancing different facets of private and public life, with a specific emphasis on data-driven methods like machine learning. They underscored the significance of understanding AI decisions to establish trust and surveyed developments in explainable machine learning, addressing

both technical and non-technical challenges. The paper stressed the importance of effective communication and tailored explanations for various stakeholders, especially in financial applications. Regulatory integration of XAI was presented as an ongoing process, necessitating further interdisciplinary research. On the technical side, challenges included combining different explanation types, developing efficient implementations, and exploring stronger model-specific approaches to enhance fidelity. Trust in explanations was another area of concern, with a focus on robustness against adversarial attacks and the suitability of proposed techniques. Hybrid models that combined opaque and transparent models were identified as a promising area of research, and causal analysis was anticipated to play a pivotal role in the future of XAI literature.

More recently, Villani et al. [6] analyzed the significance of feature importance techniques in XAI for understanding machine learning models. It focused on Shapley value-based approaches for time series data, providing closed form solutions for SHAP values in various time series models, including VARMAX (Vector Autoregression Moving-Average with Exogenous Regressors). The study introduced KernelSHAP (Kernel Shapley Additive exPlanations) for time series and demonstrated its use in "event detection." The paper also explored Time Consistent Shapley values. The authors noted the need for further research on decomposing SHAP values in ARMA (Autoregressive Moving Average) models and the comparison of different feature importance algorithms' strengths, focusing on stability, faithfulness, and robustness. They mentioned the challenge of autocorrelations in time series data and the potential of Time Consistent SHAP values to address it. The paper highlighted the importance of tailoring explanations to problem-specific assumptions, such as the presence of a time component, and underscored the need for adapting XAI techniques to complex model architectures.

The primary goal of this manuscript is to identify and rank the input features that most likely had a significant impact during an off-nominal event. Subsequently, this information is furnished to the Mission Manager, allowing it to make informed decisions regarding the most appropriate Courses of Action (Orient-Decide-Act) for successfully completing the autonomous flight mission. The forthcoming sections will provide more in-depth explanations of the methodology employed and discuss the results obtained from simulated trajectories.

III. Background

Consider a scenario in which an autonomous flight encounters a Loss of Control or an off-nominal event. In this situation, the flow of information between the Cognitive Architecture and the Mission Manager is depicted in Figure 1. The Cognitive Architecture is responsible for the Observe part of the OODA loop and makes an assessment of the vehicle's current and future capabilities based on the environmental and mission complexities. Future capabilities assessment performed by the Cognitive Architecture may include a revision to the model(s) of the current system based on sensor observation triggered by an off nominal event occurrence. The Cognitive Architecture in our research hosts a variety of algorithms such as Hankel Alternative View of Koopman (HAVOK) Analysis [7] and a Conditional Variational Autoencoder (CVAE) [8] for Loss of Control detection among other algorithms. Based on the probability generated by these algorithms, the Cognitive Architecture makes an assessment for Loss of Control or an off-nominal event and passes on this information to the Mission Manager. In addition to receiving the off nominal notification and any associated learned system model changes, the Mission Manager also reads data from the sensors which enables it to make a determination if the signal is a true positive. Once an off-nominal determination is made, the Mission Manager uses any system model updates and XAI to identify the input features that most likely triggered the event as decision aids in executing the Orient-Decide-Act part of the OODA loop.

XAI models provide the following benefits in addition to making model predictions human interpretable:

- *Transparency*: Autonomous systems that can explain their decision-making processes provide greater transparency, which can help build trust between users and the system.
- *Accountability*: In cases where an autonomous system causes harm, XAI may aid in understanding why the system made the decision it did.
- *Regulation*: Transparency in the decision-making process can help autonomous systems to comply with legal and ethical standards.
- *Human-AI collaboration*: XAI can facilitate collaboration between humans and AI by providing a shared understanding of how the system arrived at a decision (would help improve the system over time).

The XAI approach utilized in this research is game theory-inspired and provides post-hoc explainability (Figure 2). SHapley Additive exPlanations (SHAP), rooted in cooperative game theory, has gained prominence in machine learning for insights into feature importance. SHAP extends the concept of Shapley values, initially designed for fair distribution in cooperative games, to understand the contribution of individual features in predictions. It systematically

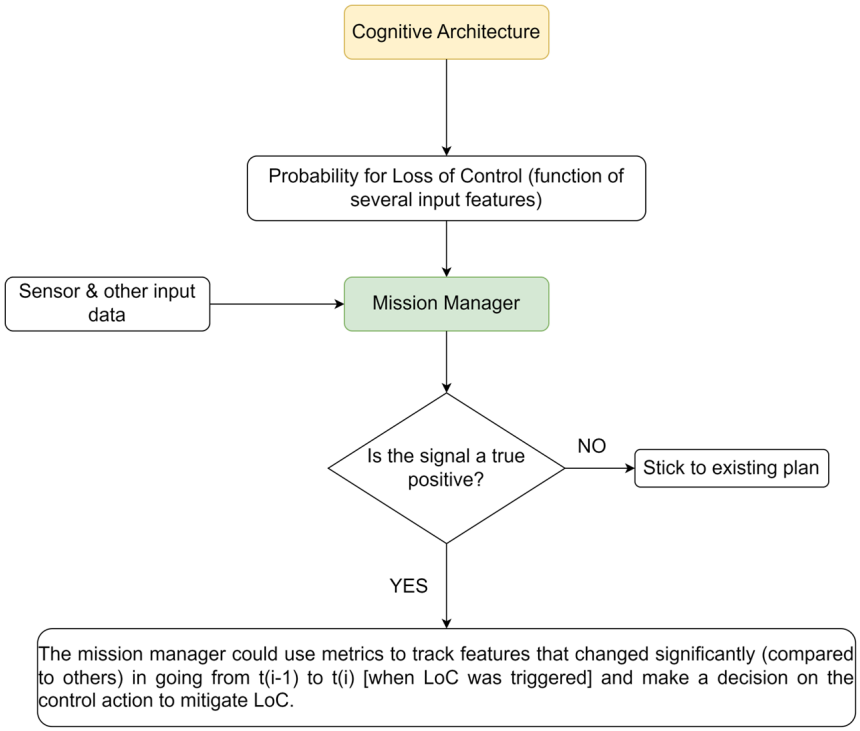

Fig. 1 Information flow within the cognitive mission manager in a Loss of Control scenario.

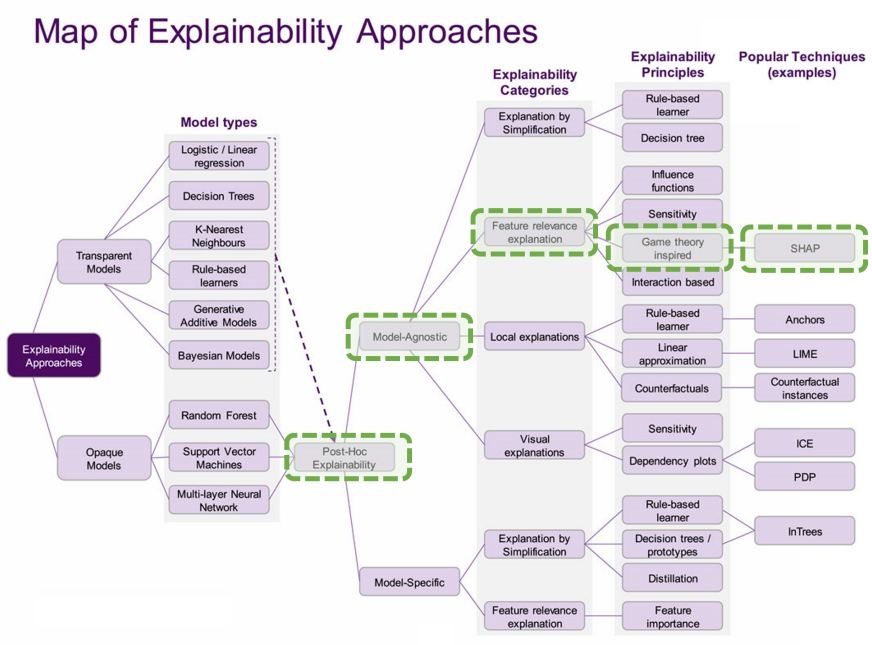

Fig. 2 XAI approach. This Figure was first presented in [5].

evaluates feature impact by considering all possible feature combinations, measuring output changes with or without a specific feature. This process calculates Shapley values, quantifying each feature's contribution and ensuring a fair distribution of influence. SHAP's nuanced analysis unveils feature interactions, fostering a model-agnostic, transparent interpretation of complex machine learning models. This unique approach enhances transparency and trust in AI

systems (see TABLE 1). SHAP applications include feature selection, model debugging, and fairness assessments, making it valuable for understanding black-box models and improving interpretability. Furthermore, this approach generates both local and global explanations (provides insight into the behavior of the model both at a specific instance and across the entire dataset) and satisfies all the four Axioms required to achieve a fair contribution, namely:

- *Efficiency*: Efficiency is the principle that a fair distribution should make the best use of the available resources. In the context of feature selection, this means that the selected features should collectively contribute to the best possible performance or utility. In other words, the selected features should be the most informative and relevant ones for the task at hand.
- *Symmetry*: Symmetry implies that the distribution of resources or the selection of features should not favor one individual or feature over another, all else being equal. It ensures that each feature or individual is treated equally in the distribution process. In feature selection, symmetry means that no feature should have an unfair advantage or disadvantage in the selection process unless there is a legitimate reason for differentiation.
- *Dummy*: The dummy axiom requires that if a feature does not contribute to the overall utility or performance, it should not be selected. In the context of feature selection, this means that irrelevant or redundant features should not be included in the final set of selected features. Dummies, in this case, are features that do not add value to the task at hand and should be excluded.
- *Additive*: The additive axiom suggests that the overall utility or performance of a set of features should be the sum of the utilities or performances of the individual features. This means that the value of the selected feature set should be a linear combination of the values of the individual features, without any complex interactions or dependencies that could lead to non-additive effects.

Table 1 Comparison of XAI techniques.

XAI Technique	Model Specific/Agnostic	Local/Global/Both	Efficiency	Symmetry	Dummy	Additivity
Shapley	Model Agnostic	Global & Local	Yes	Yes	Yes	Yes
Integrated Gradients	Model Specific	Global	Yes	Yes	No	Yes
Local Interpretable Model-agnostic Explanations (LIME)	Model Agnostic	Local	No	No	No	No
Anchors	Model Agnostic	Local	No	No	No	No
Counterfactual instances	Model Agnostic	Local	No	No	No	No
Individual Conditional Expectation (ICE)	Model Agnostic	Global & Local	Yes	No	No	Yes
Partial Dependence Plots (PDP)	Model Agnostic	Global & Local	Yes	No	No	Yes
InTrees	Model Specific	Global	Yes	No	No	Yes
Feature Importance	Model Specific	Global	Yes	No	No	Yes

IV. Results

To benchmark the algorithm, the XAI technique is first applied on a dataset based on a known function and then extended to nominal flight data, as well as flight data with partial rotor failure, as discussed in the sections below.

1) *Benchmarking with known dataset*: Consider a synthetic time series dataset with five features $(x_1, x_2, x_3, x_4, x_5)$ and a target variable y defined as follows:
   - x_1 is a linearly increasing function over time, $x_1 = 3t$
   - x_2 is a quadratic decreasing function over time, $x_2 = 10 - t^2$
   - x_3 is a constant value, and its value remains ten for all time points, $x_3 = 10$
   - x_4 is an exponentially increasing function over time, $x_4 = e^{t/4}$
   - x_5 represents Gaussian white noise generated with a mean of 2 and a standard deviation of 0.7.
   - y is the target variable, defined as $y = 1 + 0.5x_1 + 0.5x_2 + 0.5x_3 + 0.5x_4 + x_5$

   After generating the dataset, SHAP values are computed using the KernelExplainer from the SHAP library in Python. The explainer is configured with the target variable as the prediction function and a baseline dataset (typically the mean). This baseline establishes the model's prediction if no specific feature had an influence. The disparity between the actual model prediction and the baseline prediction is attributed to the features, with SHAP values indicating the magnitude and direction of their impact. This comparative analysis aids in understanding individual feature contributions to the model's predictions. SHAP values are then computed for a specified time interval (t_{start} to t_{end}), introducing perturbations to each feature and observing corresponding model responses. This iterative process provides reliable estimates of SHAP values, offering insights into each feature's importance at specific time points. In the plotted example (Figure 3), at 27 seconds, the predicted output is $y = 116.11$, with

the hierarchy of input features contributing the most: $x_4 > x_2 > x_1 > x_5 > x_3$. Notably, x_2 and x_4 are important, but x_2 decreases the output, while x_4 increases it. Furthermore, Figure 4 illustrates the specific effect of x_4 on y between 25 and 30 seconds. As illustrated in Figure 4, the input features $(x_1, x_2, x_3, x_4, x_5)$ vary identically in both the top and bottom plots. However, the predicted output y (depicted by the black curve) significantly differs between the two plots. This discrepancy arises from the fact that, unlike the top plot, the predicted output y in the bottom plot is independent of x_4 based on the data generation process. The contrast in the average absolute SHAP values effectively captures this distinction between the top and bottom plots, underscoring the adaptability of the methodology employed in this paper.

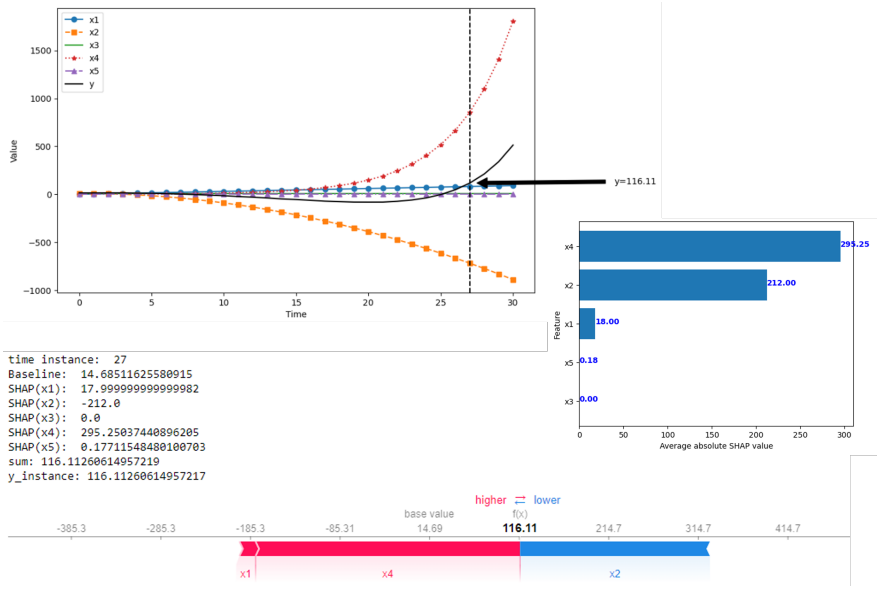

Fig. 3 Benchmarking the algorithm using a simulated test function.

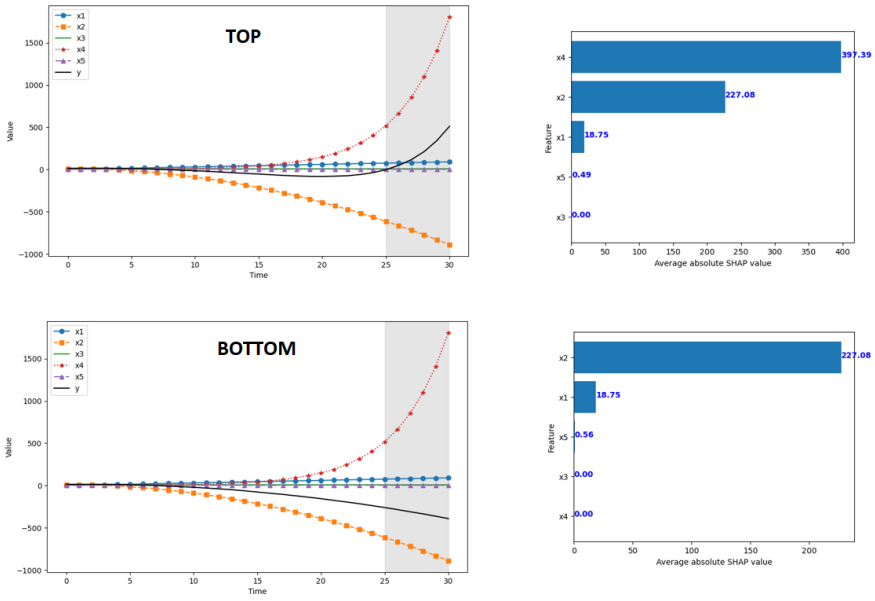

Fig. 4 Effect of specific input feature (x_4) on the target variable in the time interval $25 \leq t \leq 30$. Consistent with the nature of data, top plot captures dependence of target on (x_4), where as the bottom plot does not.

2) *Analysis of nominal flight data*: Having tested the XAI methodology on a simulated test function, the next step is to analyze the flight data. This dataset consisted of over 150 features drawn from flight dynamics, sensors and environmental variables. By specifying the waypoints for a nominal flight path, the Generalized Vehicle Simulation (GVS) generates a time series data that tracks the flight phases ('climb,' 'cruise,' 'turn,' 'descent') as a function of various input parameters (see Figure 5). Once the GVS generates the flight data, a random forest classifier (decision tree-based ensemble model) is trained to model the flight phases in combination with SHAP for interpretability. A critical aspect of this methodology is the time series data split. TimeSeriesSplit from scikit-learn is employed to partition the dataset into training, validation, and test sets while preserving the chronological order of observations. The maintenance of temporal order is essential for accurate evaluation of the model's predictive capabilities in a real-world context. Within each training-validation split (following a 5-fold cross-validation strategy), a Random Forest classifier is instantiated. This ensemble model is configured with 300 decision trees, each constrained to a maximum depth of 30. A Random Forest classifier was trained to predict the flight phase labels as a function of the input features. The accuracy metric computed in the validation set indicates the model's prediction performance. Additionally, the F1 score is computed since it provides a balanced assessment of precision and recall. SHAP values are then harnessed to enhance model interpretability. These Shapley values elucidate the contributions of individual features to the model's predictions, rendering it more comprehensible. To assess the model's performance, a confusion matrix is presented as a heatmap for ease

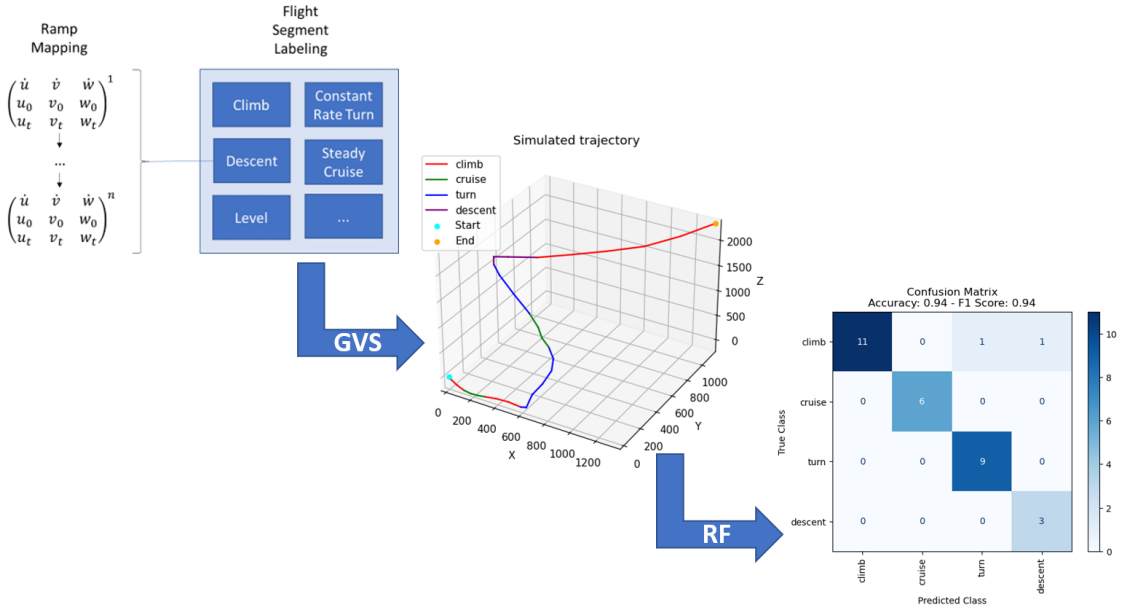

Fig. 5 Trajectory generation in Generalized Vehicle Simulation (GVS) and analysis using Random Forest (RF) classifier.

of interpretation. It can be seen from the confusion matrix that the model perfoms exceedingly well and the accuracy is (≈ 0.94). Using the SHAP (SHapley Additive exPlanations) for interpretability, the overall feature importance for the entirety of the flight plot is highlighted in Figure 6. The feature space is truncated so that the plot is easier to read. It can be noted from this plot that some of the features (for instance, y-component of the velocity vector Vel_bli_2) contribute significantly to all phases on flight while other features do not significantly contribute to all phases (for instance, engine command EngineCmd_7). In addition to making an assessment of overall feature importance, the model also provides insights on specific time points as seen in Figure 7. This plot displays several useful information including the time instant, true label and the model predicted label, the probability associated with the prediction and the contributing features ordered by importance. As an illustrative example, consider the top plot in Figure 7. This plot highlights that both the predicted and true labels correspond to the climb phase of the flight. The probability associated with this prediction is 0.97 (high confidence) and some of the important features at this time instant are Asensed_bIb_1, Q_i2b_1 and so on. Finally, Figure 8 ranks every feature based on the order of importance and clearly identifies the non-contributing features for the

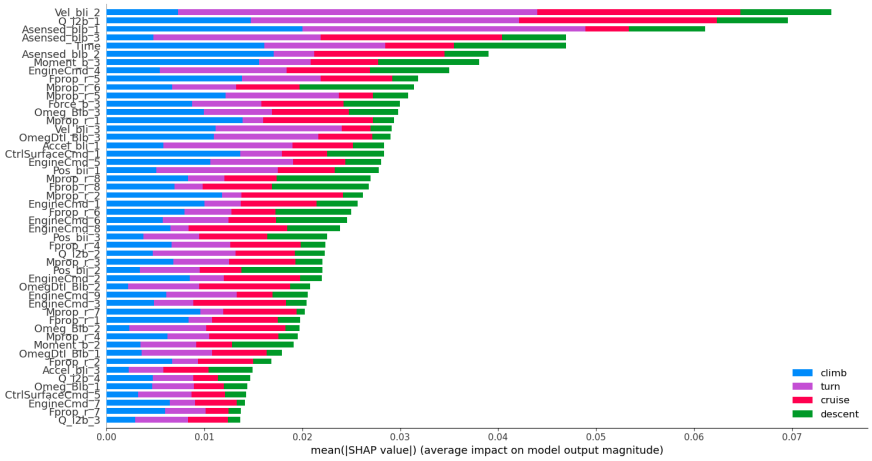

Fig. 6 Feature importance for the entirety of the flight (all phases) based on mean absolute SHAP values.

entire duration of the flight. In summary, it is possible to identify specific input features that contribute towards the flight path at any time instant using the Random Forest Classifier along with the explainability using SHapley Additive exPlanations for nominal flight. The next section will focus on ideas pertaining to the application of this approach on a flight with fault/off-nominal event.

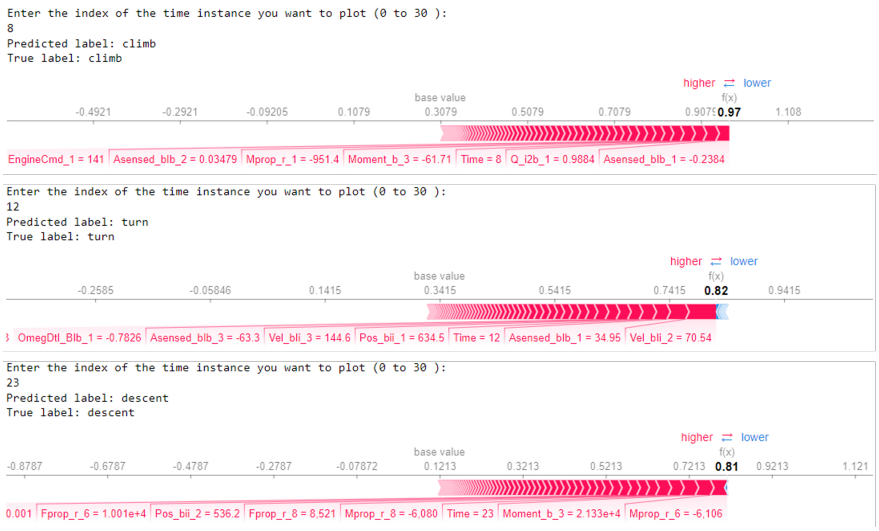

Fig. 7 Feature importance at specific instants of time.

3) *Analysis of partial rotor failure*: Consider a flight that encounters reduced rotor functionality for a certain time interval during the flight. To simulate this, a fault was introduced in one of the nine propellers driving the plane. Between time 5 s and 25 s, the functionality of Propeller 5 was degraded by 90 %. The simulated failure is highlighted in Figure 9, where it can be clearly seen that the actual engine speed of propeller 5 past the 5 s time is closer to 0 corresponding to the simulated degraded performance. The results of the XAI model for the cruise phase of the flight prior and post failure is shown in Figure 10. At time t=4 s (pre failure), Engine 5 (operating at $\approx$ 265 rpm) is trying to align with the command (350 rpm). However, at time, t=18 s (post failure), the engine (operating at 2 rpm) is unable to keep up with the command (350 rpm) indicating to the Cognitive Mission Manager the possibility of an off-nominal event for Engine 5. In addition, a subsystem within the Mission Manager named the Mission Monitor hosts anomaly detection algorithms which would raise a flag corresponding to this failure scenario.

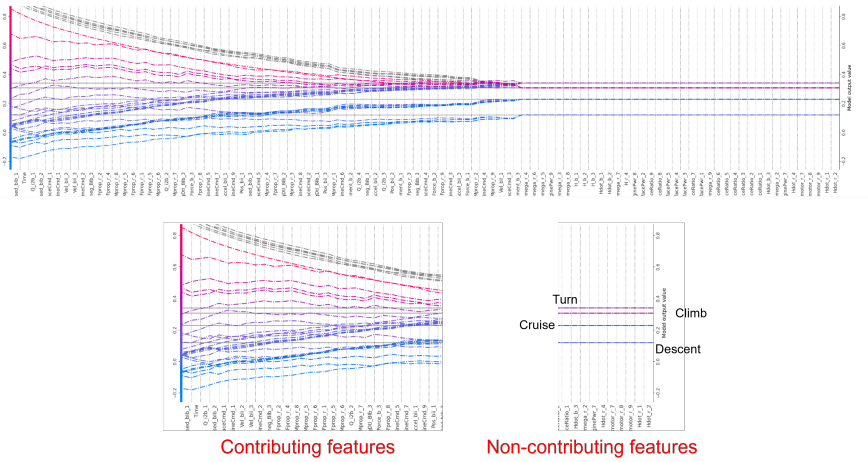

Fig. 8 Contributing and Non-contributing features.

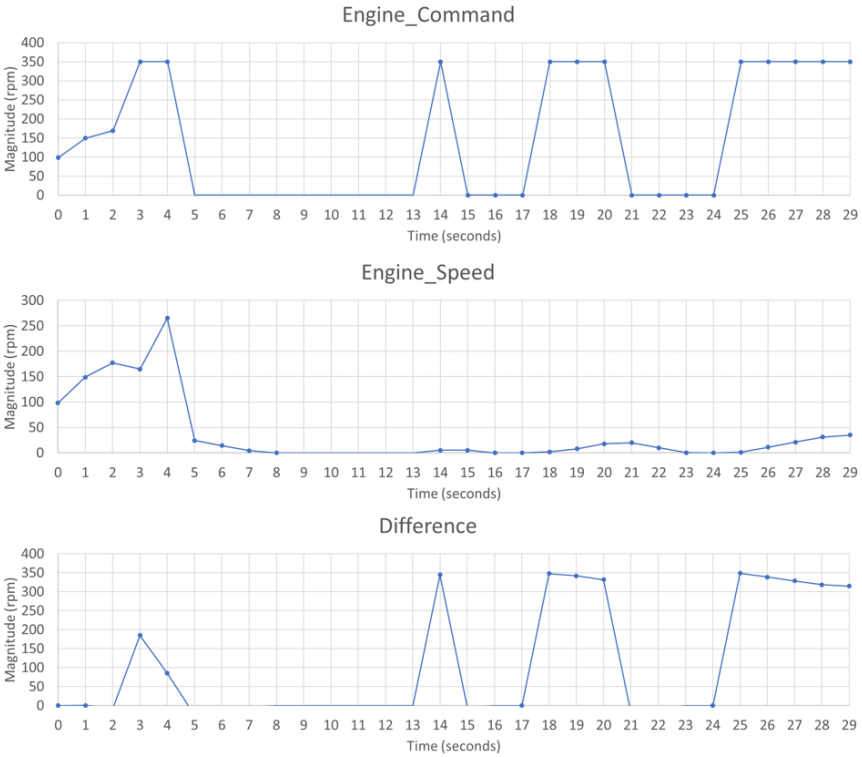

Fig. 9 Simulating a partial failure of a rotor. Engine Command vs. Time (top); Engine Speed vs. Time (mid); Difference vs. Time (bottom)

V. Conclusion

From a regulatory perspective, it is imperative to ensure that intricate AI models generate decisions comprehensible to humans. In the event of an adverse incident, interpretable models play a crucial role in identifying the root cause by providing a lucid explanation of the most probable sequence of events leading to the disaster. This paper delves into the utilization of Shapley Additive Explanations, drawing inspiration from game theory, to enhance explainability. Initial findings indicate that this approach shows promise when applied to a range of test functions, as well as nominal and off-nominal flight data.

In the case of the simulated test function, the employed methodology successfully eliminated undesirable features

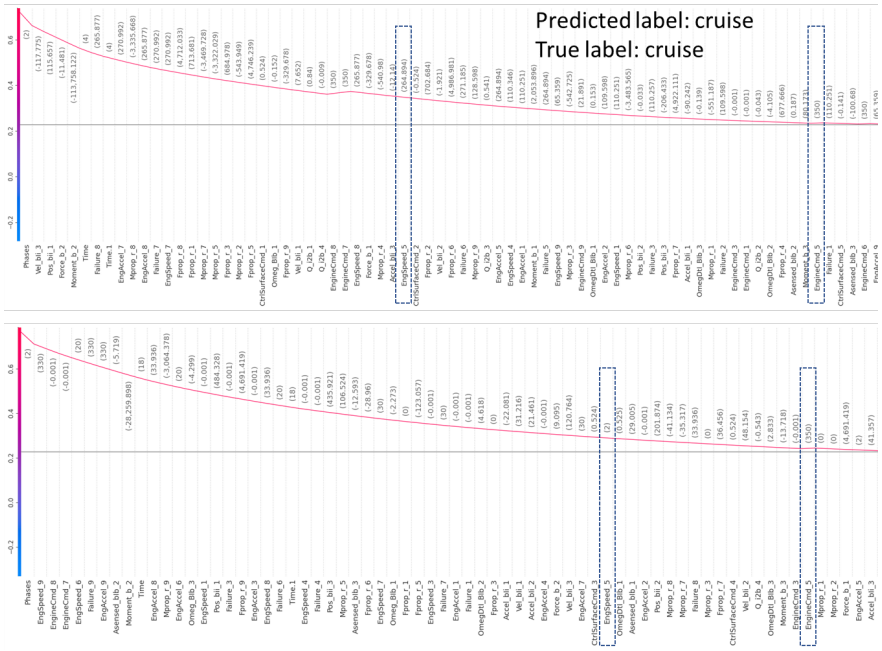

Fig. 10 Shapely analysis pre and post failure of the rotor.

that did not contribute to the output. For nominal flight data with labels, the random forest classifier housed within the cognitive mission manager accurately predicted flight phases, and the eXplainable Artificial Intelligence (XAI) model distinctly identified both contributing and non-contributing features during that phase of flight. Lastly, the XAI model proficiently identified the engine with partial rotor failure based on pre and post-failure Shapley analysis.

In conclusion, it is essential to develop a fully functional, explainable cognitive mission manager to realize the potential of utilizing autonomous flights in complex urban settings under various weather conditions while ensuring safety.

Acknowledgments

"Authors thank the Transformational Tools and Technologies (TTT) project within the Aeronautics Research Mission Directorate at NASA for supporting this research".

References

[1] Gunning, D., and Aha, D., "DARPA's Explainable Artificial Intelligence (XAI) Program," *AI Magazine*, Vol. 40, No. 2, 2019, pp. 44–58. https://doi.org/10.1609/aimag.v40i2.2850, URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850.

[2] Rudin, C., "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, Vol. 1, No. 5, 2019, pp. 206–215.

[3] Lundberg, S. M., and Lee, S.-I., "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 4768–4777.

[4] Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., and Keim, D. A., "Towards A Rigorous Evaluation Of XAI Methods On Time Series," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 4197–4201. https://doi.org/10.1109/ICCVW.2019.00516.

[5] Belle, V., and Papantonis, I., "Principles and Practice of Explainable Machine Learning," *Frontiers in Big Data*, Vol. 4, 2021. https://doi.org/10.3389/fdata.2021.688969, URL https://www.frontiersin.org/articles/10.3389/fdata.2021.688969.

[6] Villani, M., Lockhart, J., and Magazzeni, D., "Feature Importance for Time Series Data: Improving KernelSHAP," , 2022.

[7] Brunton, S. L., Brunton, B. W., Proctor, J. L., Kaiser, E., and Kutz, J. N., "Chaos as an intermittently forced linear system," *Nature communications*, Vol. 8, No. 1, 2017, p. 19.

[8] Campbell, N. H., Grauer, J. A., and Gregory, I. M., "Loss of control detection for commercial transport aircraft using conditional variational autoencoders," *AIAA Scitech 2021 Forum*, 2021, p. 0778.