AUTOMATED FLIGHT MANEUVER QUALITY ASSESSMENT FOR AIRCRAFT TESTING

Ozan Ege Camcı
Turkish Aerospace
Graduate Student, Hacettepe Univ.
Ankara, Turkey

Mehmet Önder Efe
Dept. of Computer Engineering,
Hacettepe Univ.
Ankara, Turkey

ABSTRACT

This paper presents a dual-method approach for the automated evaluation of flight maneuver quality, addressing limitations in current flight test practices that rely heavily on subjective visual assessments, manual data quantization, or static thresholds. For steady maneuvers, a statistical formulation is introduced in which flight parameter behavior is modeled as an empirical probability distribution. A new metric, termed *redundant cross-entropy*, quantifies trim quality by measuring the divergence of the observed data from ideal trimmed behavior which is modeled as a Gaussian distribution. This metric enables consistent comparison across parameters with differing physical units. For transient maneuvers, a fully convolutional network augmented with Squeeze-and-Excitation blocks is proposed to classify maneuver success using real and simulator data from the T625 Gökbey multirole utility helicopter developed by Turkish Aerospace. The architecture supports variable-length time series through masked global pooling and leverages downsampling to extend the model's receptive field without increasing complexity. Evaluation results demonstrate the model's capability to identify high-quality flight segments and reliably assess maneuver outcomes based on predefined success criteria. The proposed framework offers a scalable and interpretable solution to enhance the reliability and efficiency of maneuver evaluation in aircraft flight testing.

NOTATION

*

Symbols

ΔH Redundant cross-entropy (trim error), dimensionless

$\mathbb{E}[x]$ Expected value of random variable x, same units as x

$H(p,q)$ Cross-entropy between distributions p and q, nats

$H'(q)$ Entropy of the distribution q, nats

N Number of samples or parameters, unitless

$p(x)$ Observed (empirical) probability distribution, unitless

$q(x)$ Ideal probability distribution, unitless

σ Standard deviation of a distribution, same unit as variable

μ Mean of a distribution, same unit as variable

Ω Main rotor rotational speed, RPM

ψ, θ, ϕ Yaw, pitch, and roll angles, deg

u, v, w Translational velocities in body axes, m/s

p, q, r Angular velocity components in body axes, rad/s

$\delta_{\textbf{controls}}$ Control inputs (cyclic, collective, pedal), % or deg

Vario Vertical speed, ft/min

$t_{\textbf{rest}}$ Duration of resting phase after flare, s

Δh Altitude loss during flare, m

$v_{\max}$ Maximum allowable ground speed at touchdown, knots

*

Acronyms

ERF European Rotorcraft Forum

VFS Vertical Flight Society

FCN Fully Convolutional Network

SE Squeeze-and-Excitation

CNN Convolutional Neural Network

IAS Indicated Airspeed

RPM Revolutions Per Minute

TQ Main Rotor Torque

USNTPS United States Naval Test Pilot School

INTRODUCTION

Flight testing is one of the most crucial and critical component of lifecycle of a newly designed aircraft. It begins with the aircraft's initial flight and continues until certification is achieved. During this phase, engineers carefully observe the aircraft's performance to make necessary design adjustments. Following developmental testing, the aircraft undergoes certification tests to demonstrate compliance with international aviation standards and regulations. Aircraft testing is an expensive and risky operation usually subjected to time constraints due to project schedules. Hence, efficient and safe testing is important. Tests are conducted as follows. A team of design engineers, led by a flight test engineer, monitors critical parameters from a ground station known as telemetry to ensure safety and intervene if problems arise. Their primary objectives are to ensure the collection of high-quality data and the safe conduct of tests. A typical test involves executing a series of maneuvers specified by the design team. As each maneuver begins, the flight test engineer marks the time and informs the pilot; the end time is noted similarly. During testing, right after the maneuver, the team reviews the data to determine its usability for further analysis, marking whether the test point was executed correctly. However, these assessments are often performed through quick visual inspections, heavily relying on subjective judgments. Such evaluations can introduce significant challenges in interpreting results, particularly when certification flights must meet specific success criteria for maneuvers. Inaccurate evaluations of maneuver quality can render flight data unusable, underscoring the need for more quantitative and reliable evaluation methods in the evaluation of flight maneuver quality. On the other hand, even when flight test is executed perfectly, engineers usually use only the marked test points while rest of the flight still contains valuable data. Hence, an automated system that assesses maneuver quality can significantly enhance the efficiency of flight testing by providing immediate feedback during tests and extracting valuable segments from post-data.

On the other hand, there has not been significant attention focused on the area of flight maneuver quality assessment. Most of the works are centered around pilot performance evaluation. In some of these works (Ref. 1), (Ref. 2), pilots are asked to fly through a route precisely defined by GPS positions, and their performance is evaluated in terms of tracking errors. In the work of Zahed et al. (Ref. 2), flight data is labeled directly based on the experience level of the pilots, and classification with tracking error is performed using decision trees. In the work of Jong (Ref. 3), data is collected in a similar manner, but instead of calculating tracking error, deep neural networks are directly implemented to establish classification. Another method in the literature involves determining a set of pilot performance metrics, extracting these metrics from raw flight data, and classifying pilot performance based on these metrics. For example, Zhang et al. (Ref. 4) utilized a CNN architecture, while Li et al. (Ref. 5) used a fuzzy comprehensive evaluation method on pilot performance indicators. Although these works may provide valuable insights, they are not aimed at sensitive maneuver quality evaluation but rather focus on a general understanding of pilot performance.

One of the few approaches that directly address the challenge of maneuver quality assessment was proposed by Lerro et al. (Ref. 6), who developed a method to identify quasi-steady flight segments using a statistical scoring function. The algorithm evaluates multiple parameters (e.g., pitch rate, altitude variation) across fixed-length time windows and computes a heuristic score to identify suitable training data for synthetic sensor design. While the objective of their study differs from that of the present work, it demonstrates that steady flight validation can be effectively automated using parameter-level criteria. Another related study is the work of Tian et al. (Ref. 7), which trains a variational autoencoder and uses its encoder as a feature compression network for maneuver quality classification. Their approach relies on a standard maneuver library, which is a collection of typical maneuver executions used as a reference. The quality of a maneuver is estimated by comparing its latent space representation with those in the library using dynamic time warping. However, since the VAE is not trained to distinguish between good and poor executions, distances in the latent space may not reliably capture maneuver quality. Moreover, maneuver execution varies due to factors like weight, center of gravity, and altitude. Thus, instead of a single standard maneuver, an exemplary maneuver should be defined as a collection of well-executed maneuvers.

In this work, the problem of maneuver quality assessment is investigated by first categorizing maneuvers into two types: steady and transient. Steady maneuvers (e.g., level flight, turns, and climbs) involve maintaining a trimmed state and moving steadily along a trajectory. Success criteria are typically defined by maximum deviations from target values over a specific duration. For instance, during level flight testing, airspeed may be allowed a ±1-knot deviation within a 30-second data window. However, some criteria are operationally unrealistic, while increasing allowable deviations may incorrectly classify oscillatory data as satisfactory. To address this issue, a statistical metric is proposed to quantify trim quality. In this approach, time-series data are treated as an empirical probability distribution, and success is assessed based on proximity to ideal distributions of selected parameters. These ideal probability distributions, representing trimmed flight, are defined in a manner that preserves existing industry success criteria while more accurately reflecting the nature of deviations observed during flight testing.

Transient maneuvers (e.g., landing, takeoff, flare) involve a sequence of actions performed for operational purposes. Unlike steady maneuvers, each transient maneuver has unique success criteria that require manual quantization by engineers based on predefined conditions. Consequently, a statistical approach cannot be directly applied. Instead, time-series classification is performed using a fully convolutional network (FCN) augmented with squeeze-and-excite blocks. The FCN architecture is selected primarily for its exceptional ability to handle data of varying durations. Furthermore, given the challenge of data scarcity in the aerospace domain, FCNs are preferred for their demonstrated sample efficiency and robust

generalization capabilities in time-series classification tasks (Ref. 8).

QUALITY ASSESSMENT OF STEADY MANEUVERS

In the context of flight mechanics, trim refers to a steady-state flight condition in which all the forces and moments acting on the aircraft are balanced (Ref. 9). Mathematically, this corresponds to a solution of the equations of motion where the time derivatives of translational and angular velocities are zero, and control inputs remain fixed

Trim validation in flight test practice is generally implemented by defining thresholds for key flight parameters (such as airspeed, pitch attitude, altitude, or heading), and requiring these parameters to stay within the given limits over a time window — often 30 seconds. For example, the US Naval Test Pilot School (USNTPS) notes suggest thresholds of ±1 knot in airspeed and ±20 feet in pressure altitude for assessing trimmed level flight in helicopters (Ref. 10). These values are intended to prevent false positives due to drift or sustained oscillations and to ensure the aircraft remains close to a single equilibrium point.

However, this threshold-based approach has inherent limitations:

1. Sensitivity to Momentary Disturbances: Even if the aircraft is mostly trimmed, a small disturbance (e.g., turbulence, pilot correction) might briefly push a parameter outside the threshold and cause the data segment to be rejected. As a result, a genuinely trimmed interval might be discarded due to single outlier events.

2. Threshold Tuning Dilemma: If thresholds are tightened, many intervals may be deemed invalid due to noise or small oscillations. If thresholds are loosened, an highly oscillatory data may be classified as satisfactory or the criteria may fail to distinguish between two different trimmed states in the same time window. For instance, allowing ±2 knots for a 70-knot trimmed flight might incorrectly group two distinct trimmed conditions at 68 and 72 knots as a single trim state.

3. No Measure of Trim "Quality": Threshold-based logic yields only binary results — either a condition is trimmed or it is not. It does not quantify how well a flight condition satisfies the desired trimmed state, leaving no room for graded assessments or nuanced interpretation of marginal cases.

To address these limitations, a distribution-based approach is proposed for trim validation. Rather than treating time-series data as sequences to be evaluated against static thresholds, they are modeled as empirical probability distributions. Each parameter's behavior over time (e.g., airspeed, altitude) is captured in its distributional shape. An ideal trimmed distribution is then defined for each parameter, serving as a boundary condition that reflects the expected behavior in a trimmed flight.

Cross-entropy is used as a statistical measure to quantify the proximity between the observed and ideal distributions. This enables the evaluation of not only whether the aircraft was in trim, but also how closely it matched the expected trimmed behavior.

Cross-Entropy for Evaluating Trimmed Flight Segments

Overview of Cross-Entropy Cross-entropy is a fundamental concept in information theory that measures the average number of bits (or nats) required to encode data from a true distribution $p(x)$ using a model distribution $q(x)$ (Ref. 11). It quantifies how well the model approximates the actual data.

The cross-entropy between $p(x)$ and $q(x)$ is defined as:

$$H(p,q) = -\int p(x)\log q(x)\,dx \tag{1}$$

In practice, when the true distribution $p(x)$ is only available through samples, the cross-entropy can be approximated using Monte Carlo integration:

$$H(p,q) \approx -\frac{1}{N}\sum_{i=1}^{N}\log q(x_i), \quad x_i \sim p(x) \tag{2}$$

Motivation for Use in Trim Evaluation In this study, cross-entropy is used to assess how closely a flight segment aligns with a predefined ideal distribution that characterizes trimmed flight. The ideal distribution $q(x)$ is modeled as a Gaussian representative of steady flight behavior.

By computing the cross-entropy between observed flight data $p(x)$ and the ideal model $q(x)$, a numerical score is obtained that reflects how "trimmed" the data is. Crucially, cross-entropy favors distributions that are more concentrated around the mode of $q(x)$, even if their shapes differ. This property aligns with the objective of trim assessment, as low-variance segments indicate steadier and more controlled flight behavior.

This behavior contrasts with other divergence measures such as Kullback–Leibler (KL) divergence, which penalize mismatches in distribution shape and might assign lower scores to higher-variance data that visually appears closer to the ideal.

Figure 1 illustrates this concept using three Gaussian distributions with equal means but different standard deviations. Although the distribution with $\sigma = 5$ appears visually closer to the ideal $\sigma = 4$, the narrower distribution with $\sigma = 1$ yields a lower cross-entropy score. This is because its values are more tightly clustered near the mean, making them easier to encode, which is a property aligned with trimmed flight.

Implementation of Cross-Entropy The probability density function (PDF) of a Gaussian (normal) distribution with mean μ and variance σ^2 is given by:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3}$$

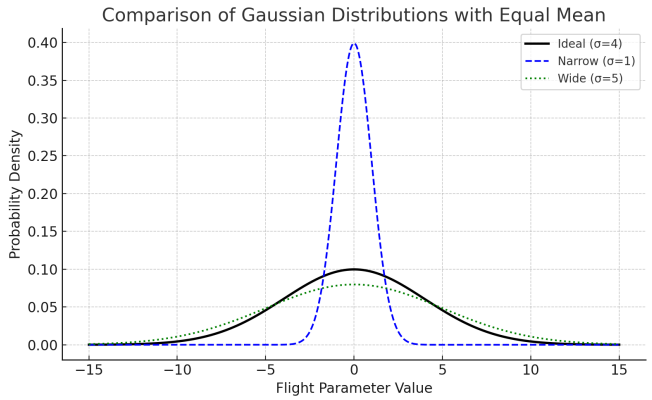

Figure 1. Comparison of Gaussian distributions with equal means and different standard deviations.

Let $p(x) = \mathcal{N}(x \mid \mu, \sigma_1^2)$ represent the empirical distribution of a flight parameter, and let $q(x) = \mathcal{N}(x \mid \mu, \sigma_2^2)$ be the predefined ideal trim distribution. Assuming both distributions share the same mean μ, the cross-entropy between them is expressed as:

$$H(p,q) = \frac{1}{2}\log(2\pi\sigma_2^2) + \frac{\sigma_1^2}{2\sigma_2^2} \tag{4}$$

In Equation 4, the second term accounts for how well the variance of the real distribution p is represented by the ideal model q. However, the first term depends solely on the variance of the model q. This is problematic, as different flight parameters may have ideal distributions of vastly different magnitudes. As a result, cross-entropy values may vary significantly across parameters, making them difficult to interpret or compare directly.

To address this, the special case is considered where the real and ideal distributions are identical, i.e., $\sigma_1 = \sigma_2 = \sigma$. In this case, the cross-entropy reduces to the entropy of the Gaussian: H

$$H(q,q) = H'(q) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \tag{5}$$

This case is interpreted as the ideal, fully trimmed condition. To normalize and make the cross-entropy values comparable across different parameters, the entropy of the ideal distribution $H'(q)$ is subtracted from the cross-entropy $H(p,q)$, yielding what is referred to in this work as the *redundant cross-entropy* or trim cost when evaluating steadiness of flight segments:

$$\Delta H = H(p,q) - H'(q) \tag{6}$$

Placing 4 and 5 into 6, we obtain

$$\Delta H = \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2} \tag{7}$$

In practice, as $\sigma_1 \to 0$ (i.e., the real distribution becomes perfectly concentrated around the mean), the relative cross-entropy approaches:

$$\Delta H \to -\frac{1}{2} \tag{8}$$

This scaling naturally centers the scoring such that:

$$H(p,q) - H'(q) \leq 0 \qquad \text{if } p \text{ is better trimmed} \tag{9}$$
$$H(p,q) - H'(q) > 0 \qquad \text{if } p \text{ is less trimmed} \tag{10}$$
$$H(p,q) - H'(q) = -0.5 \qquad \text{if } p \text{ is perfectly trimmed} \tag{11}$$

In this work, the value ΔH is interpreted as a scalar measure of trim quality and referred to as the *trim error*. Lower trim cost values indicate closer adherence to the ideal trimmed condition, while higher values correspond to more transient or oscillatory behavior.

Although $\Delta H = H(p,q) - H'(q)$ is a straightforward difference of standard information-theoretic quantities, it is not commonly isolated or employed as a standalone metric in the literature. In this work, it is interpreted as a form of expected code length redundancy which is measuring the excess description cost incurred by encoding observed data using a reference model, normalized by the model's own entropy. To our knowledge, this specific formulation has not previously been formalized or applied in the context of trim quality assessment or elsewhere.

Determining Ideal Distributions

The empirical rule states that for a Gaussian distribution, approximately 68% of the data lies within $\mu \pm 1\sigma$, 95% within $\mu \pm 2\sigma$, and 99.7% within $\mu \pm 3\sigma$. Figure 2 demonstrates how the data is distributed for a Gaussian. Benefiting from empirical rule, ideal distributions can be designed such that encompassing traditional thresholds and allowing reasonable deviations. In this manner, parameters of interest and corresponding μ and σ values should be determined.

Determining Trim Parameters Trim is identified by the steadiness of the states. When modeling only the six rigid-body degrees of freedom, the state vector $\mathbf{x}$ includes the three translational velocity components (u, v, w), the three angular velocity components (p, q, r), and the Euler angles (ϕ, θ, ψ), which define the aircraft's orientation (Ref. 9).

For translational velocity and Euler angles, accelerometers and angular rate sensors can theoretically be used to measure the time derivatives directly. However, there are practical limitations when assessing steadiness through these derivatives. Specifically, the total change in a velocity component over a time window can be calculated by integrating the acceleration along the corresponding axis:

$$\Delta V = \sum_{n=1}^{N} a(n) \cdot \Delta t \tag{12}$$

Table 1. Trim parameter behavior for different flight maneuvers.

Parameter	Rectilinear Flight	Coordinated Turn	Autorotation	Autorotative Turn
IAS	Steady	Steady	Steady	Steady
SideSlip	Steady	Steady	Steady	Steady
Vario	Steady	Steady	Steady	Steady
p	0	Steady	0	Steady
q	0	Steady	0	Steady
r	0	Steady	0	Steady
ϕ	Steady	Steady	Steady	Steady
θ	Steady	Steady	Steady	Steady
ψ	Steady	N/A	Steady	N/A
Ω	N/A	N/A	Steady	Steady
TQ	Steady	Steady	0	0
δ_{controls}	Steady	Steady	Steady	Steady

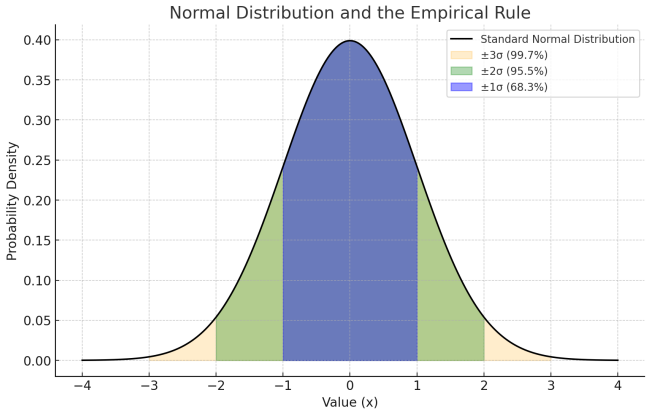

Figure 2. Demonstration of empirical rule

To ensure that the velocity remains steady, this integrated sum should be close to zero. Since the flight data is sampled at a fixed rate, this implies the conclusion in 13 :

$$\sum a(n) \approx 0 \tag{13}$$

From a probabilistic perspective, this condition corresponds to a histogram of acceleration values centered at zero. The discrete expected value of such a distribution is:

$$\mathbb{E}[x] = \sum_i P(x_i)x_i \approx 0 \tag{14}$$

where, x_i are the histogram bin centers and $P(x_i)$ represents the empirical probability mass at x_i.

This implies that the center of mass of the acceleration histogram should be close to zero. However, achieving a low cross-entropy value does not necessarily enforce this symmetry. Figure 7 illustrates an example, where a yaw rate segment yields low cross-entropy despite causing a continuous drift in yaw angle, resulting in a uniform distribution with high cross-entropy.

One possible solution is to reduce the ideal standard deviation for the yaw rate, forcing the distribution to be more centered. Yet, this approach often imposes an overly strict constraint that excludes segments with minor oscillations of yaw angle.

On the other hand, if a parameter has low cross-entropy (i.e., it follows the ideal distribution), its time derivative tends to exhibit a symmetric distribution around a zero mean — which also leads to low cross-entropy. However, this relationship breaks down when the time-derivative data contain relatively high-frequency oscillations.

Based on these observations, the steadiness of state variables is evaluated directly through their values rather than their derivatives. In accordance with flight test practice, indicated airspeed (IAS), vertical speed (vario), and sideslip angle are used as substitutes for the body-frame velocity components (u,v,w). Angular velocity and attitude sensors provide the remaining states (p,q,r,ϕ,θ,ψ). Assuming the helicopter employs a rotor speed governor, the main rotor speed Ω becomes relevant only in power-off conditions. Additionally, control inputs (collective, longitudinal and lateral cyclic, and pedal) and total torque are included to ensure force–moment balance.

Determining Ideal μ Trim is the condition where states converges to specific values that makes state derivatives zero. In our approach, instead of calculating these specific values, we assume the states have the trimmed value if they are steady. Hence, ideal μ value is directly calculated as the mean of time window being investigated. However, special considerations are needed for different maneuvers. In rectilinear flight the angular rates p,q,r are expected to be zero. On the contrast, for the coordinated turn manuever, angular rates are expected to be steady with a mean value. In this case, the yaw angle continuously change. In Table 1, details of utilized trim parameters and mean value approaches specific to flight maneuvers are presented.

In addition, our method does not restrict the user from selecting the target mean. For example, ideal IAS μ can be set to 80 in order to calculate how well the 80 kts level flight trim is established. Morever, for autorotation, torque should be setted to zero indicating maneuver is in real power-off condition.

Table 2. Standard deviation bounds for each parameter.

Parameter	$\pm1\sigma$ (68%)	$\pm2\sigma$ (27%)	$\pm3\sigma$ (4.7%)
IAS	0.8	1.6	2.4
SideSlip	1	2	3
Vario	60	120	180
p	0.5	1.5	2
q	0.5	1.5	2
r	0.5	1.5	2
ϕ	0.8	1.6	2.4
θ	0.8	1.6	2.4
ψ	1	2	3
Ω	1	2	3
TQ	1	2	3
δ_{controls}	1	2	3

Determining Ideal σ Ideal standard deviation values are selected based on the empirical rule, as illustrated in Figure 2. For a Gaussian distribution, approximately 68% of the data lies within the interval $\mu \pm 1\sigma$. Therefore, the standard deviation σ should be chosen small enough so that this level of deviation can be considered negligible under trimmed flight conditions.

Following this, about 27% of the data falls between $\mu \pm 1\sigma$ and $\mu \pm 2\sigma$, which represents the maximum allowable deviation range. The remaining 4.7% of the data lies between $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$, and is interpreted as representing brief, momentary disturbances that can be tolerated.

Based on these considerations, the ideal standard deviation values for each flight parameter are defined as shown in Table 2. Although a single set of values is proposed here, it is entirely possible—and often necessary—to tailor the ideal distributions for specific maneuvers. Defining different success thresholds is common in flight testing, as the difficulty of maintaining control over certain parameters can vary significantly depending on the nature of the maneuver.

QUALITY ASSESSMENT OF TRANSIENT MANEUVERS

Transient maneuvers are an essential part of rotorcraft testing. These maneuvers are typically mission-oriented and consist of a sequence of actions and defined checkpoints. Analyzing such maneuvers is more challenging than evaluating steady-state conditions, as the time-series data must be segmented, and success must be assessed based on the aircraft's state at specific checkpoints as well as the quality of transitions between them. This segmentation and evaluation process can be difficult to encode using rule-based logic. As a result, a data-driven approach based on neural networks is proposed to facilitate robust assessment.

In this work, the autorotative flare maneuver is selected as a representative transient case. The success of the maneuver is evaluated based on the following criteria:

- The helicopter must reach a near-zero vertical velocity and maintain this "resting" state for a duration of time indicated by t_{rest}, without exhibiting a ballooning effect (i.e., an unintentional climb following the flare).
- The total altitude loss from initiation of flare to the resting state must remain below a predefined value indicated by Δh .
- The ground speed must be reduced to a v_{max} value or less by the end of the resting period.
- The rotational speed of the main rotor must remain within transient limits.

This section details the data collection, preprocessing, and model architecture. Training procedures and performance evaluation are discussed in the *Results* section.

Data Collection and Preprocessing

This study utilizes real flight and simulator test data from the T625 Gökbey helicopter, with sensor selection tailored to the specific maneuver being assessed. For the flare maneuver, ground speed, rate of descent, altitude, torque, and main rotor rotational speed are collected.

The sensors record data at a sampling rate of 128 Hz while simulator runs at 100 Hz, which are higher than necessary for assessing maneuver quality. This high sampling rate increases the required receptive field to capture sufficient duration, complicating data processing. To manage the resulting data volume and ensure feasibility, the signals are downsampled to a common frequency. The choice of downsampling factor and its effect on the receptive field are discussed in the section on *Receptive Field.*

Given the significant variability in the duration of flight test points, the model architecture must accommodate variable-length input. However, for efficient batch training, uniform sample lengths are required. This is achieved through zero-padding to standardize data lengths. However, during training, it is essential to prevent the model from learning from the non-informative padded regions. To this end, masking is incorporated within the global pooling layer to ensure that the model focuses solely on the meaningful portions of the data. The implementation details of this masking technique are further discussed in the section on *Masked Global Pooling*.

Model Architecture

The proposed model architecture is designed to classify transient maneuvers from variable-length time series data. It combines a fully convolutional backbone with attention-enhancing Squeeze-and-Excitation (SE) blocks and integrates masking strategies to handle zero-padded inputs during batch training. This section outlines the structure and rationale behind each architectural component. The base model structure is illustrated in Figure 3.

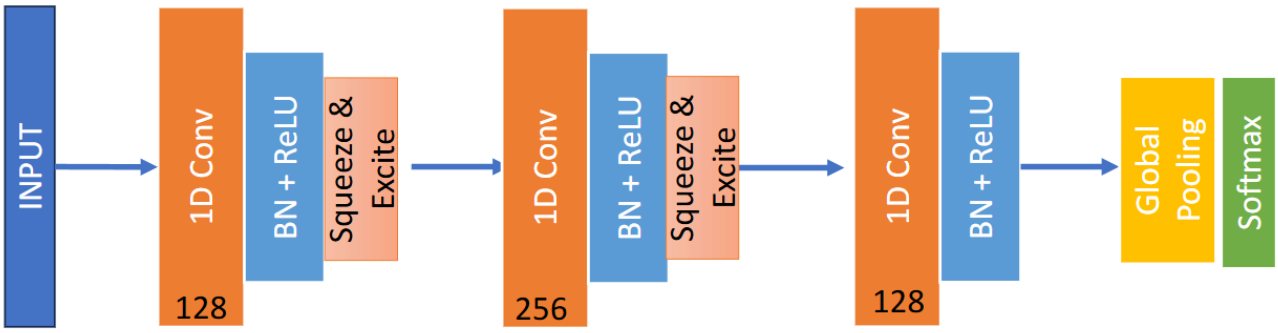

Figure 3. Fully convolutional network with squeeze and excite blocks

Fully Convolutional Networks FCNs are distinguished by their architecture, which is composed entirely of convolutional layers. Their implementation on time series classification is first proposed by Wang et al. (Ref. 8). Their design enables FCNs to effectively process time-series data. Typically, an FCN includes several convolutional layers, each followed by batch normalization and a ReLU activation function, which introduce non-linearity and help stabilize the training process.

A key feature of FCNs is their ability to handle inputs of varying lengths, thanks to the inclusion of global average pooling (GAP) at the end of the convolutional stack. This pooling layer condenses the feature maps into a single value per map, significantly reducing the number of parameters in the model. This reduction not only helps in preventing overfitting but also simplifies the network, making it more computationally efficient.

The global pooling layer ensures that the network can accommodate different input sizes, which is particularly useful for tasks involving variable data dimensions, without requiring any adjustments to the input size or network architecture.

Squeeze and Excite Blocks Squeeze-and-Excitation (SE) blocks are a type of enhancement for convolutional neural networks that improve how the network focuses and uses its features (Ref. 12). Each SE block starts by compressing spatial information from the input into a smaller form through a process called 'squeeze'. This is done using global average pooling, which reduces the amount of data by summarizing each channel of features into a single value.

After squeezing, the 'excitation' step uses two simple layers connected in a network and a sigmoid function to adjust the emphasis on certain features over others. This allows the network to better focus on important features while ignoring less useful ones.

Researchers have found (Refs. 12, 13) that while the inclusion of SE blocks introduces some additional parameters, the resulting increase in computational cost is relatively small compared to the performance gains achieved. This makes SE blocks a cost-effective mechanism for enhancing neural network performance by improving the network's focus on relevant features.

Receptive Field The receptive field (RF) in convolutional neural networks (CNNs) is a fundamental concept that refers to the portion of the input data that influences a specific response at a particular layer. In practice, it represents the section of the input that a given feature in the CNN can "see" and process.

The formula to compute the RF for a network of depth d with each layer having a filter length equal to k_i with $i \in [1, d]$ and a stride equal to one is:

$$RF = 1 + \sum_{i=1}^{d} (k_i - 1) \tag{15}$$

Based on the model introduced by Wang et al. (Ref. 8), which employs kernel sizes of 8, 5, and 3, the resulting receptive field spans 14 time steps. Given a sensor sampling rate of 128 Hz, this corresponds to a temporal window of approximately 0.11 seconds.

However, prior work by Fawaz et al. (Ref. 14) demonstrated that longer temporal patterns require larger receptive fields to be effectively captured. Thus, a receptive field of 0.11 seconds is considered insufficient for reliably capturing the temporal characteristics relevant to maneuver quality.

Although extending the receptive field by increasing kernel sizes or network depth is a viable option, doing so significantly enlarges the model by introducing additional parameters. This increase in complexity can lead to overfitting, especially when data availability is limited.

To overcome this limitation, downsampling is applied during preprocessing, as described in the section on *Data Collection and Preprocessing*. Time data is downsampled to 16 Hz resulting in effective receptive field duration of 0.875 seconds.

This adjustment enables the model to capture longer-term patterns in the data without the need for excessive network depth or parameter count, thereby maintaining a balance between performance and generalizability.

Handling Zero-Padded Input with Masked Global Pooling Masked global pooling is a technique developed to handle variable-length data by ignoring padded regions during global pooling operations. Hertel et al. (Ref. 15) appear to be among the first to explicitly formalize and implement this method. This method is particularly useful in neural networks, where it is necessary to pad the input data to uniform lengths for batch training.

To ensure that the zero-padded regions do not influence the learning outcomes of the model, we calculate the effective length of these regions in the global pooling layer using the following formula:

$$N_{\text{padded,eff}} = N_{padded} \prod_{i=1}^{d} \frac{1}{\text{stride}_i} \tag{16}$$

In this equation, d represents the number of layers in the network, and stride_i refers to the stride of each corresponding layer. This calculation is crucial as it helps to effectively exclude the padded areas during the pooling process, thus allowing the network to focus solely on the meaningful, actual data.

RESULTS

Results for steady and transient maneuver quality assessment are presented under separate headings, as the methods used for each are fundamentally different.

Steady Maneuver

In Figure 4, Figure 5, Figure 6, and Figure 7, the application of the trim quality assessment metric—referred to as *redundant cross-entropy*—is illustrated for individual parameters: indicated airspeed (IAS), vertical velocity, yaw rate, and yaw angle, respectively. This metric is defined in Equations 9, 10, and 11. Redundant cross-entropy is computed over the flight data using a sliding window approach, where segments yielding the lowest scores are identified as the most steady intervals. These steady segments are highlighted in blue on the upper time-series plots.

The corresponding steady segments are also shown in the form of histograms in the lower plots. Each histogram is normalized to a probability density function, such that the total area under the curve equals one, as described by Equation 17:

$$\text{Height of each bin} = \frac{\text{Count in bin}}{\text{Total count} \times \text{Bin width}} \tag{17}$$

The bin width is set to one-tenth of the ideal distribution's standard deviation, i.e., $\sigma_{\text{ideal}}/10$. Overlaid on each histogram is the corresponding ideal probability density function, which serves as a reference, enabling visual comparison between the empirical and ideal distributions. The resulting redundant cross-entropy values for each trimmed segment are provided as legends in the top-right corner of the corresponding histograms.

Figure 4 presents three test points, each containing a trimmed segment for IAS. The columns are arranged from left to right in increasing steadiness:

- The first column depicts a nearly trimmed condition.
- The second column shows a well-trimmed test point.
- The third column illustrates an almost perfectly steady segment, with a redundant cross-entropy score of approximately -0.486.

Conversely, Figure 5 displays test points that do not contain any steady segments for IAS. These plots are similarly organized from left to right in increasing trim cost:

- The first column represents a borderline trim condition.
- The second column represents a transitional condition.
- The final column corresponds to a highly transient flight segment.

Additionally, in Figure 6, the trim error is calculated for the vertical velocity parameter, whose ideal distribution is defined with a standard deviation of 60 feet per minute (as listed in Table 2). As seen in the figure, the trim error function penalizes non-ideal distributions with magnitudes comparable to those observed for IAS, whose ideal standard deviation was 0.8 knot. This example demonstrates the inherent normalization property of the metric, which enables consistent comparison across parameters with different physical scales. This property is essential for multivariate optimization tasks aimed at identifying the best-trimmed segments.

A composite trim cost can be defined by summing the trim error values across all relevant parameters:

$$\text{Trim Cost} = \sum_{i=1}^{N} \Delta H(p_i, q_i) \tag{18}$$

Here, i indexes the parameters listed in Table 2, and N denotes the total number of parameters considered in the trim evaluation.

Finally, Figure 7 presents a special case in which the yaw rate segment yields a low trim error, yet results in a continuous drift in yaw angle. This occurs because maintaining steadiness in yaw angle requires the yaw rate to exhibit a symmetric distribution around a zero mean, as discussed in the section on *Determining Trim Parameters*. However, symmetry is not explicitly enforced by the cross-entropy formulation. Consequently, steadiness should be evaluated directly on the state variable of interest—such as yaw angle—rather than on its derivative, such as yaw rate.

These results demonstrate the behavior of the proposed metric across diverse flight segments and parameter distributions. The trim error changes consistently in response to variation in empirical distributions and is effectively shaped by the corresponding ideal distribution. The redundant cross-entropy thus penalizes deviations in a parameter-specific yet normalized manner, supporting its utility as a scalable and interpretable scoring function for steady maneuver quality assessment.

Transient Maneuver

Before implementing the proposed model, datasets were prepared by identifying and labeling flare test points. Each test point was extracted with a duration ranging from 13 to 20 seconds and downsampled to 16 Hz. The segments were manually analyzed to quantify performance indicators related to predefined success criteria: resting duration, altitude loss, minimum ground speed, and main rotor speed range. These metrics were then used for labeling.

The experiments aimed to achieve the simplest model configuration without compromising performance. During training, substantial fluctuations were observed in both training and validation loss, even after extended epochs. To ensure reliable model performance, early stopping was employed with a patience of 100 epochs and a minimum delta of 0.005, using the Adam optimizer with a learning rate of 0.0005. Due to data scarcity for flare maneuvers, the batch size was set to 16, and the reduction ratio of the squeeze-and-excitation (SE) block was fixed at 16. Additionally, 5-fold cross-validation was used to improve robustness given the limited sample size.

Some of the training results are presented in Table 3. The candidate configuration set includes models with three and four convolutional layers. Channel sizes are systematically varied. Accuracy results are reported as the mean and standard deviation across five folds. The best performing model achieved a mean accuracy of 0.874. However, the relatively high standard deviation indicates that model performance varies significantly across folds. This suggests that the dataset may contain diverse conditions, and the number of samples may be insufficient for the model to consistently capture those variations across different train-test splits.

Furthermore, no significant degradation in accuracy is observed for the smaller configurations, indicating that model complexity can be reduced without substantial loss in performance.

In addition, model performance is evaluated specifically with respect to altitude loss. This parameter is of particular interest because detecting the initiation of the flare maneuver would be challenging using a rule-based approach. For this evaluation, the altitude loss of each flare sample is manually measured. The dataset is then relabeled based solely on altitude loss, and only altitude and vertical velocity sensor readings are included in the input data. A low threshold for altitude loss is selected to ensure a balanced number of samples in each class.

Figure 8 presents the classification results for samples with varying altitude losses. The x-axis shows normalized altitude loss, and the y-axis indicates the predicted probability of the sample being classified as satisfactory. The model's predictions exhibit a clear trend: the probability of a sample being satisfactory increases as altitude loss decreases. Notably, most misclassified samples are concentrated around intersection of the chosen threshold and the 0.5 probability line, suggesting that the model is effectively learning to associate altitude loss with maneuver quality.

CONCLUSIONS

This study addresses a fundamental limitation in current flight test practices: the reliance on subjective, visual, and manually quantified assessments for evaluating maneuver quality. Such approaches hinder consistency, traceability, and scalability—particularly as datasets grow and test complexity increases.

To overcome this, a dual-method framework was developed. For steady maneuvers, a distribution-based metric—termed redundant cross-entropy—was introduced to quantify trim quality. Unlike classical approaches that define trim based on fixed parameter thresholds, this method redefines trim as a statistical distribution benefiting from Gaussian. For transient maneuvers, a fully convolutional network (FCN) was proposed to classify success based on mission-relevant criteria.

The redundant cross-entropy metric offers interpretable, unit-normalized, and threshold-centered scoring. It enables consistent comparison across parameters and reveals subtle deviations from trimmed behavior. The FCN architecture supports variable-length inputs through masked global pooling, is lightweight, and performs reliably even under data scarcity. With cross-validation, the model achieved a mean classification accuracy of up to 87.4% on real and simulated autorotative flare maneuvers.

These tools automate maneuver evaluation, reduce analyst workload , and enable the extraction of previously unmarked high-quality segments. The approach has strong potential for real-time test support, reducing unnecessary retests and improving flight test coverage across platforms and mission scenarios.

Future work includes extending this framework to other certification maneuvers, refining output to not only classify but also quantify success metrics. Overall, this study proposes a practical and scalable methodology suitable for adoption as a flight maneuver analysis tool in modern flight test programs.

Author contact: Ozan Ege Camcı ozanege.camci1@tai.com.tr
Mehmet Önder Efe onderefe@gmail.com

ACKNOWLEDGMENTS

The author would like to thank Uğurcan Özalp for his mentorship in the field of artificial intelligence and his valuable insights throughout the development of this work. Special thanks are also extended to former chief engineer Ilgaz Doğa Okçu for initially encouraging the author to pursue this topic,

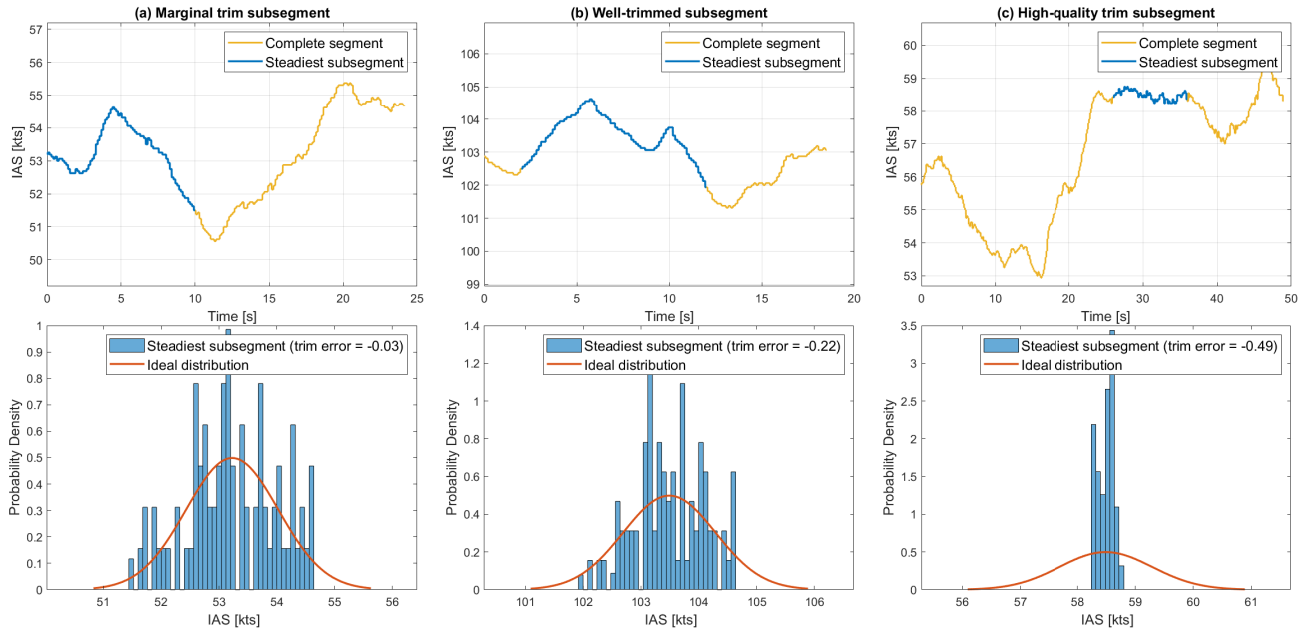

Figure 4. Demonstration of trim error on the IAS across three steady flight segments, sorted by increasing quality. All segments meet the steady condition but differ in how closely they match the ideal trim condition.

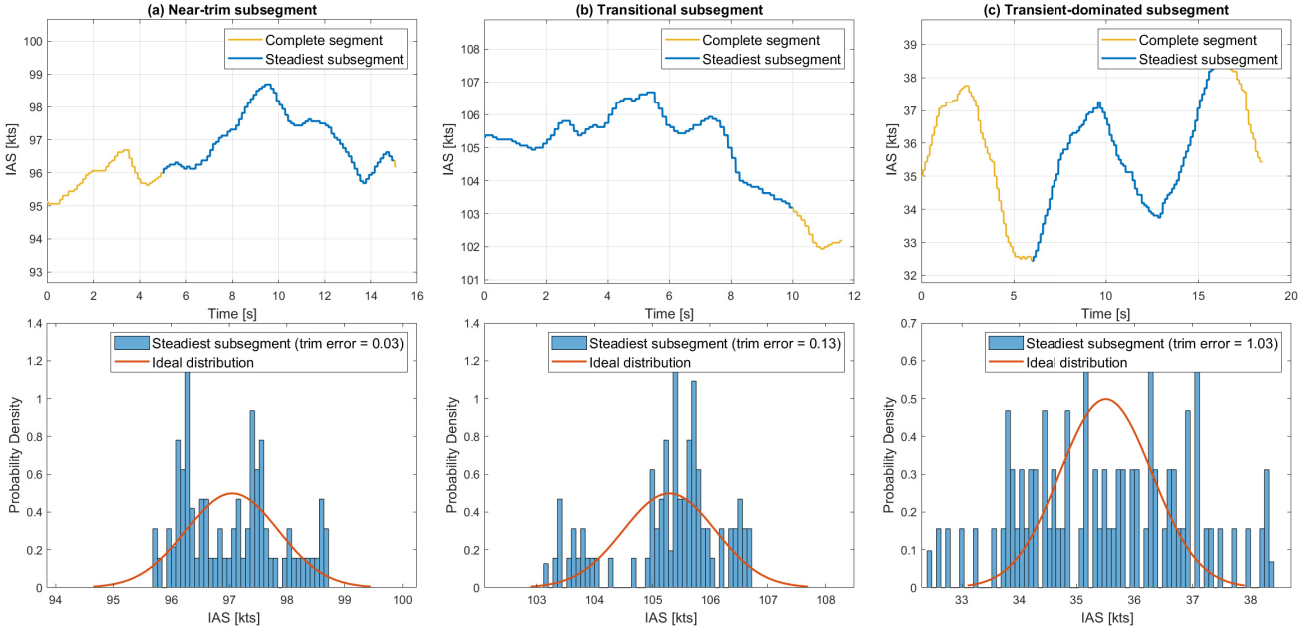

Figure 5. Demonstration of trim error on the IAS parameter across three transient flight segments exhibiting increasing deviation from ideal steady-state behavior. Segment (a) represents a borderline case close to trim, while (b) and (c) show progressively stronger transient characteristics.

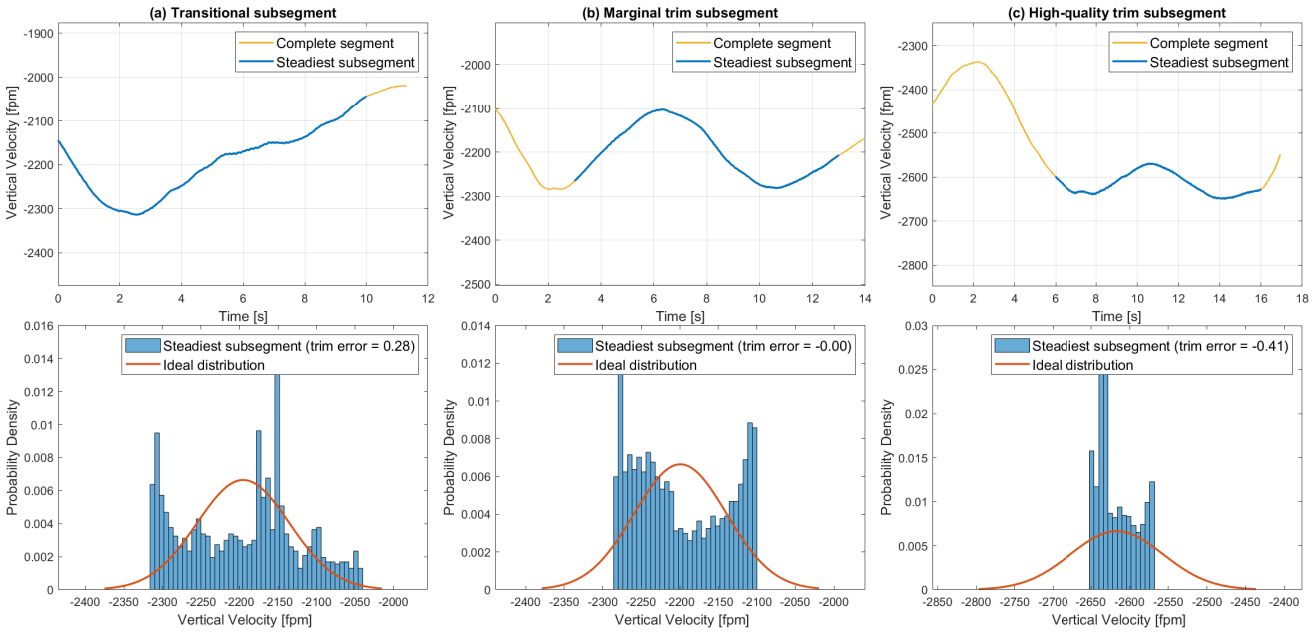

Figure 6. Demonstration of trim error on the vertical velocity parameter across three flight segments. Subsegment (a) represents a transitional case, (b) shows marginal trim alignment, and (c) reflects a high-quality trim condition with minimal divergence.

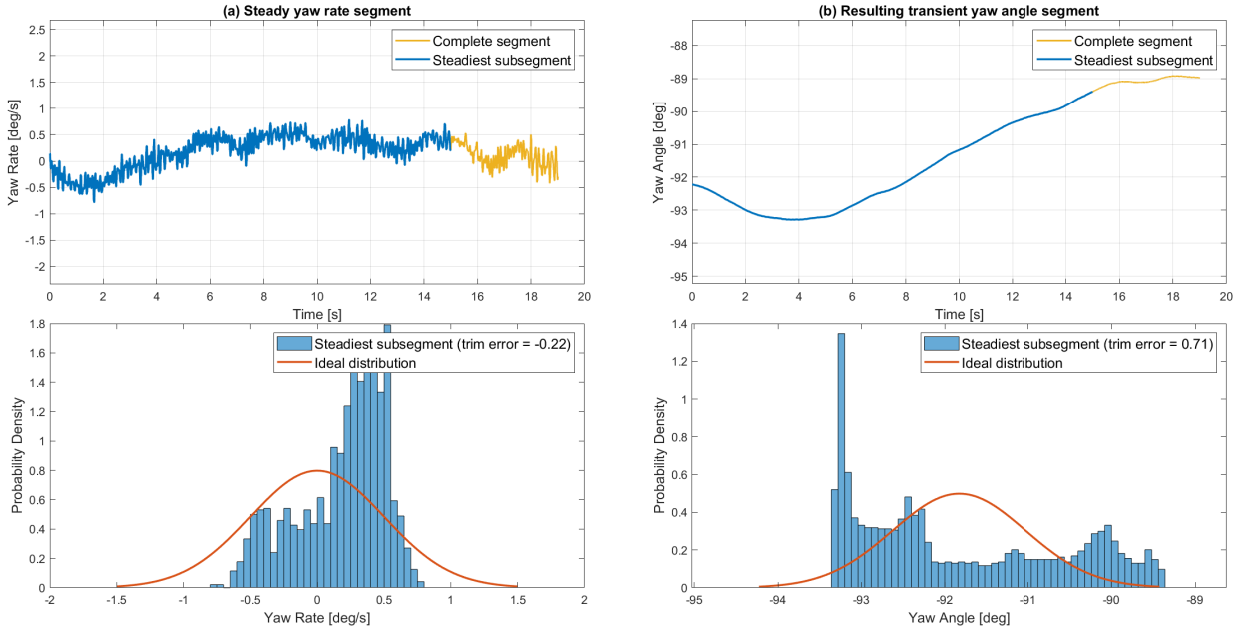

Figure 7. Yaw rate segment with low trim error despite causing yaw drift, highlighting the need to assess steadiness on the parameter itself (e.g., yaw angle) rather than its time derivative.

Table 3. Validation accuracy for different convolution configurations.

Convolution Channel Sizes	Convolution Kernel Sizes	Mean Validation Accuracy	Std. Dev. of Val. Accuracy
128-256-128	8-5-3	0.840	0.063
128-256-128-64	8-5-3-3	0.846	0.052
128-256-128-128	8-5-3-3	0.853	0.050
32-64-32	8-5-3	0.826	0.042
32-64-32-16	8-5-3-3	**0.874**	0.067
32-64-32-32	8-5-3-3	0.833	0.055
16-32-16	8-5-3	0.860	0.049
16-32-16-8	8-5-3-3	0.832	0.045
16-32-16-16	8-5-3-3	0.819	0.044

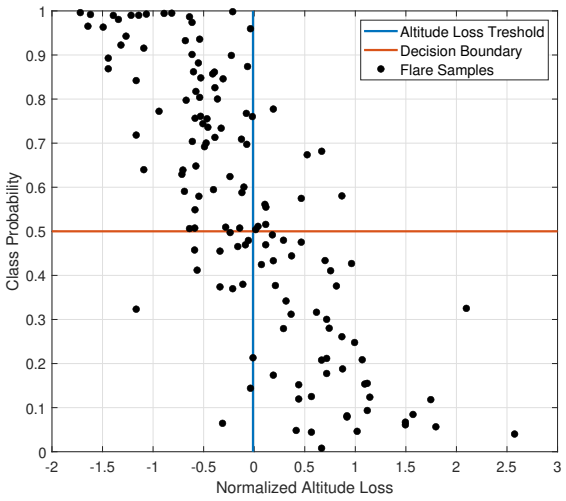

Figure 8. Classification probabilities versus normalized altitude loss. The vertical and horizontal lines represent the altitude loss threshold and the model's decision boundary, respectively.

as well as to current chief engineer Yusuf Onur Arslan and manager Ahmet Alper Ezertaş for their continued support of this study. The author also gratefully acknowledges Turkish Aerospace for fostering a research-supportive environment and encouraging academic contributions alongside engineering practice.

REFERENCES

1. Paces, P., and Insaurralde, C. C., "Artificially Intelligent Assistance for Pilot Performance Assessment," *Proceedings of the 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, San Antonio, TX, Oct. 2021, pp. 1–7. DOI: 10.1109/DASC52595.2021.9594487.

2. Zahed, M. J. H., and Fields, T., "Evaluation of Pilot and Quadcopter Performance from Open-Loop Mission-Oriented Flight Testing," *Proceedings of the Institution of Mechanical Engineers, Part G*, Vol. 235, No. 13, 2021, pp. 1817–1830. DOI: 10.1177/0954410020985987.

3. De Jong, M. J. L., "Classifying Human Pilot Skill Level Using Deep Artificial Neural Networks," M.Sc. Thesis, Delft University of Technology, Delft, Netherlands, 2021.

4. Zhang, S., Huo, Z., Sun, Y., Li, F., and Jia, B., "Pilot Maneuvering Performance Analysis and Evaluation with Deep Learning," *International Journal of Aerospace Engineering*, Article ID 6452129, March 2023. DOI: 10.1155/2023/6452129.

5. Li, G., Wang, H., Pan, T., Liu, H., and Si, H., "Fuzzy Comprehensive Evaluation of Pilot Cadets' Flight Performance Based on G1 Method," *Applied Sciences*, Vol. 13, No. 21, 2023, Article ID 12058. DOI: 10.3390/app132112058.

6. Lerro, A., Brandl, A., Battipede, M., and Gili, P., "A Data-Driven Approach to Identify Flight Test Data Suitable to Design Angle of Attack Synthetic Sensor for Flight Control Systems," *Aerospace*, Vol. 7, No. 5, 2020, Article ID 63. DOI: 10.3390/aerospace7050063.

7. Tian, W., Zhang, H., Li, H., and Xiong, Y., "Flight Maneuver Intelligent Recognition Based on Deep Variational Autoencoder Network," *EURASIP Journal on Advances in Signal Processing*, Vol. 2022, No. 1, 2022, Article ID 21. DOI: 10.1186/s13634-022-00850-x.

8. Wang, Z., Yan, W., and Oates, T., "Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline," *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, May 2017, pp. 1578–1585. DOI: 10.1109/IJCNN.2017.7966039.

9. Padfield, G. D., "Helicopter Flight Dynamics: The Theory and Application of Flying Qualities and Simulation Modelling," 2nd ed., Blackwell Publishing, Oxford, UK, 2007, p. 27–28.

10. U.S. Naval Test Pilot School, *Flight Test Manual – Rotary Wing Performance, USNTPS-FTM-No. 106*, Patuxent River, MD, Revised 31 December 1996.

11. Murphy, K. P., *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA, 2012, pp. 57–58.

12. Hu, J., Shen, L., and Sun, G., "Squeeze-and-Excitation Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, June 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

13. F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019. doi: 10.1016/j.neunet.2019.04.014

14. H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "InceptionTime: Finding AlexNet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1936–1962, 2020. doi: 10.1007/s10618-020-00710-y

15. Hertel, L., Phan, H., and Mertins, A., "Classifying Variable-Length Audio Files with All-Convolutional Networks and Masked Global Pooling," *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, September 2016, pp. 1–5.

CLASSIFICATION & COPYRIGHT

The paper must be unclassified for release to the public and cleared by the appropriate company and/or government agency if necessary. CEAS and its national member societies shall be allowed to publish the paper. The copyright information is stat-ed on the Forum website https://www.rotorcraft-forum.eu/ and on the front page of this template. All papers presented at the ERF will be published as ERF proceedings. In addition, they will be available in a freely accessible web-based repository 2 years after the respective conference.