The impact of imbalanced datasets on Deep Neural Network predictions: A case study in scramjet performance

Anonymous authors
Paper under double-blind review

Abstract

Robust aerodynamic predictions for hypersonic vehicles increasingly rely on existing deep-learning tools. However, imbalanced datasets — often resulting from limited experimental data or insufficient coverage of operational conditions — can compromise model reliability and introduce bias into predictions. This work offers an application-centered account of how a feed-forward multilayer perceptron (PyTorch implementation) behaves when trained on (i) a data-rich yet operationally imbalanced set of scramjet simulations and (ii) a deliberately balanced counterpart generated with a conventional metaheuristic (MH) sampling scheme, but with a lower sample count. Without altering network architecture, loss function, or optimizer, we expose a clear trade-off: the imbalanced model achieves a 14% lower root mean square error (RMSE) but produces thrust predictions that violate first-principles trends, whereas the balanced model sacrifices a small amount of numerical accuracy to maintain physical coherence across Mach–altitude space. These results illuminate both the strength (high statistical accuracy) and the weakness (loss of physical fidelity under bias) of off-the-shelf deep neural networks (DNNs) when data coverage is uneven. The findings serve as a cautionary example for practitioners who might otherwise deploy such models uncritically, and underscore the methodological importance of rigorous dataset diagnostics — rather than chasing novel algorithms — for reliable AI adoption in aerospace design.

1 Introduction

The aerospace industry is increasingly leveraging Deep Learning (DL) techniques to enhance the design and performance predictions of hypersonic vehicles (Ai et al., 2021; Cui et al., 2018; McCall et al., 2018; Paulino & Passaro, 2023; Wang & Ma, 2024). These vehicles, capable of traveling at speeds exceeding Mach 5 — that is, five times higher than the speed of sound — present unique challenges due to their complex aerodynamic behaviors and operational conditions (Russo et al., 2021; Scarlatella et al., 2024; Speier et al., 2017). As the demand for efficient and reliable hypersonic travel grows, the need for accurate predictive models becomes imperative. DL models have shown promise in this area (Mao et al., 2021; Paulino & Passaro, 2024; Wang & Ma, 2024). By integrating domain-specific knowledge into the modeling process, researchers can improve the precision and reliability of the predictions made by DL models (Ren et al., 2023).

Despite the potential of DL in aerospace applications, a significant challenge arises from the lack of comprehensive training data covering all relevant operational conditions. Understanding the underlying physics of hypersonic flight, such as fluid dynamics, is essential in the development of predictive models, in a way that the interaction between vehicle geometry and atmospheric conditions significantly influences aerodynamic performance (Anderson, 2003). Therefore, it is required that the datasets used for training DL models reflect these complex relationships as faithfully as possible. Often, available datasets do not represent all critical variables related to the underlying physics, leading to biased predictions that do not align with physical reality. This imbalance may result from limited experimental data obtained from flight tests or simulations that fail to capture the full range of possible scenarios. Furthermore, the uncertainty present in the data can also be a problem (Paulino et al., 2020). Consequently, models trained on such datasets may perform well under specific conditions but struggle to generalize across a broader spectrum of scenarios (Dong et al., 2024; Johnson & Khoshgoftaar, 2019; Ling et al., 2016; Zhang et al., 2024). Thus, addressing data imbalance is of utmost importance for developing robust predictive models in aerospace engineering.

The issue of data imbalance can be addressed through techniques specifically designed for imbalanced datasets. Recent studies show different ways to address the problem, including the application of ensemble methods and algorithm-level adjustments that can enhance the performance of predictive models in scenarios with imbalanced data (Ghosh et al., 2024; He & Garcia, 2009; Johnson & Khoshgoftaar, 2019). This paper delves into one particular approach: the generation of additional samples in previously underrepresented regions to mitigate the effects of imbalance. New samples are generated using a metaheuristics-based approach, as shown in (Araújo et al., 2024). Metaheuristics (MHs) are particularly interesting in solving difficult problems whose complexity makes the search for exact optimal solutions unfeasible (Doğan & Ölmez, 2015). A case study is presented, which brings the impact of imbalanced data on the use of Deep Neural Networks (DNNs), i.e., a feed-forward multilayer perceptron, in predicting the aerodynamic performance of different compression section configurations of scramjets over a predefined operating range in terms of altitude and Mach number.

Considering the issues presented and the contemporary research landscape in Artificial Intelligence (AI) and Machine Learning (ML), there exists a potential for unconscious oversights. Researchers may, without due consideration, introduce AI applications into their methodologies, potentially overlooking the serious issue of data balance. This oversight can lead to an uncritical utilization of datasets for model training, which can risk the introduction of bias and skewed results Ghosh et al. (2024). Such practices are not uncommon; numerous scientific publications, despite their technical depth, often neglect to detail how data distribution is managed. As stated by Chen et al. (2024), many scientific articles implicitly assume a relatively equal number of samples in each studied class. This assumption could lead to a false sense of balanced data, masking potential issues that may significantly impact the performance of the model and the interpretation of the results Hellín et al. (2024).

In response to these documented shortcomings, the present work formulates and tests a clear hypothesis: when the network architecture and training protocol are held constant, an operationally imbalanced training set can produce misleadingly favorable global error metrics while degrading the physical consistency of thrust predictions for underrepresented Mach-altitude regimes. To examine this proposition, we combine two established elements: a feed-forward DNN architecture previously validated for scramjet compression ramps (Paulino & Passaro, 2024) and a Sea Turtle Optimization (STO) routine that can populate the design space with synthetic samples (Araújo et al., 2024). Two complementary datasets are constructed: Dataset 1 (236,640 samples) exhibits over and underrepresentation of certain Mach–altitude pairs, while Dataset 2 (99,000 samples) contains a uniform number of samples for each pair. Using identical topology, loss function, and optimizer on each dataset, we will (i) compare conventional performance indicators such as RMSE, (ii) evaluate physical plausibility through first-principles checks, and (iii) analyze error distributions throughout the flight envelope. By isolating dataset balance as the sole independent variable, the study aims to provide quantitative evidence and practical guidance for researchers who rely on deep learning surrogates in hypersonic aerodynamics and other simulation-rich domains.

2 Related Work and Theoretical Framework

2.1 Training Deep Learning models and Deep Neural Networks with imbalanced data

According to Goodfellow et al. (2016), ML stands out as the only feasible approach to build AI systems capable of operating in complex environments. This methodology harnesses Neural Networks (NNs) — a fundamental concept within ML — which draw inspiration from the structure and function of biological brains (McCulloch & Pitts, 2021) and can address complex problems through a flexible framework (Paulino et al., 2019) that is a cost-effective alternative to expensive experiments and simulations while maintaining a reasonable level of accuracy (Gond & Sengupta, 2025). These networks are composed of interconnected units called artificial neurons that process information across layers (Schmidhuber, 2015). Artificially constructed neurons aim to mimic their biological counterparts in operation; they receive input from other neurons, apply activation functions, and transmit output signals to the subsequent layer (Haykin, 2009). DL, a specialized subfield of ML, utilizes NNs featuring numerous hidden layers stacked between the input and output layers, known as DNNs. These hidden layers facilitate the ability of the network to learn increasingly complex data representations, allowing it to address problems beyond the scope of shallower networks (Goodfellow et al., 2016; Haykin, 2009; LeCun et al., 2015; Mao et al., 2021). DL has brought about transformative changes across various sectors due to its ability to discern patterns within vast datasets and achieve high precision in tasks such as image identification, natural language understanding, and speech recognition (Brunton et al., 2021; Pontoh et al., 2024; Schmidhuber, 2015; Sharma et al., 2020).

In the field of aerospace engineering, particularly in hypersonics, DL applications have demonstrated advances and challenges in data-driven modeling (Mao et al., 2021; Wang & Ma, 2024). Recent studies have highlighted the effectiveness of DL techniques in optimizing various aspects of aircraft design and performance analysis. These applications range from flight control (Bu et al., 2023; Lv et al., 2023; Shou et al., 2023) to trajectory optimization (Shi et al., 2022; Wang et al., 2022), and even to the prediction of flowfield characteristics (Fujio & Ogawa, 2022; Ozbenli et al., 2020) and aerodynamic load estimation (Beachy et al., 2021). Paulino & Passaro (2024) proposed the use of DNNs trained with hundreds of thousands of samples derived from multi-objective optimizations of a scramjet compression section, aiming to replace expensive simulations and broaden the capacity for aerodynamic performance evaluation in previously unexplored flight conditions. The results showed that DNNs were able to satisfactorily represent the solution space, evidencing their potential as an efficient tool in the engineering of hypersonic systems.

However, training DNN models on imbalanced datasets can lead to significant performance biases (Chen et al., 2024; Johnson & Khoshgoftaar, 2019; Wang et al., 2016). These models tend to become overly optimized for the majority class, resulting in high accuracy under normal operating conditions but poor detection rates for the minority classes of interest, such as faults or anomalies (Chen et al., 2024; Dangut et al., 2020; Dong et al., 2024). In critical aerospace applications, the failure to accurately identify these rare events can have severe consequences for safety and operational efficiency (Shen & Zhao, 2023). Models may exhibit a strong bias towards predicting the most frequent class, potentially ignoring or misclassifying instances of the minority class altogether (Ghosh et al., 2024; Johnson & Khoshgoftaar, 2019; Wang et al., 2016). This issue is intensified by the fact that traditional evaluation metrics such as overall accuracy can be misleading in imbalanced scenarios, as a model that simply predicts the majority class for most instances can still achieve a high accuracy score despite its inability to identify the minority class (Chen et al., 2024; Hellín et al., 2024). Therefore, the use of imbalanced data in the training of DNNs for aerospace applications can lead to a misleading good overall performance metrics while masking a fundamental inability to detect critical abnormal conditions (Hellín et al., 2024; Johnson & Khoshgoftaar, 2019). Various techniques exist to address the problem of data imbalance, and can be broadly categorized into data-level approaches, algorithm-level approaches, and hybrid approaches (Dangut et al., 2020; Johnson & Khoshgoftaar, 2019; Kaur et al., 2020).

Data-level approaches focus on rebalancing the class distribution of the training data (Chen et al., 2024; Thabtah et al., 2020). These techniques include oversampling the minority class by randomly duplicating samples or by generating synthetic samples using methods like Synthetic Minority Over-sampling Technique (SMOTE) and its variants such as Borderline-SMOTE, Adaptive Synthetic Sampling (ADASYN), Safe-Level-SMOTE, k-means SMOTE, and Disjuncts-Robust Oversampling (DROS) (Chen et al., 2024; Johnson & Khoshgoftaar, 2019; Shen et al., 2023; Werner de Vargas et al., 2023). Although oversampling aims to increase the representation of the minority class, potentially improving classifier performance, it can also lead to overfitting and increased computational cost (Chen et al., 2024; Khan et al., 2018; Thabtah et al., 2020). Furthermore, generative methods, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), can be used to generate samples of synthetic minority class, effectively augmenting the dataset and addressing the imbalance, using real data and class labels to guide the generation of realistic and diverse synthetic samples, thus balancing the dataset for subsequent processing (Ghosh et al., 2024; Shen & Zhao, 2023). In contrast, undersampling reduces the number of samples in the majority class by random removal (Random Under-Sampling - RUS) or more sophisticated methods such as cluster-based selection or One Side Selection (OSS) (Johnson & Khoshgoftaar, 2019; Zhang et al., 2024). Although undersampling can balance the data and reduce training time, it can also result in the loss of potentially important information (Khan et al., 2018; Wang et al., 2016). Hybrid sampling methods combine oversampling and undersampling techniques, such as SMOTE with OSS, to leverage the benefits of both while mitigating their individual drawbacks (Dangut et al., 2020; Kaur et al., 2020).

In turn, algorithm-level approaches aim to improve the learning process of classifiers on imbalanced datasets without modifying the data distribution (Dangut et al., 2020; Johnson & Khoshgoftaar, 2019). Cost-sensitive learning is a key technique that assigns different misclassification costs to different classes, typically higher costs to misclassifications of the minority class, to encourage the model to learn the minority class better (Boughorbel et al., 2017; Chen et al., 2024; Dangut et al., 2020; Khan et al., 2018). This can be implemented by adjusting the loss functions or weighting the samples based on their class. Examples of modified loss functions include rescale weighted cross-entropy loss, focal loss, hinge loss, Kullback-Leibler divergence loss, sparse cross-entropy, Balanced Mean Squared Error (BMSE), and Label Distribution-Aware Margin (LDAM) loss (Chen et al., 2024; Dangut et al., 2020; Liu & Tian, 2024). Another algorithm-level strategy is threshold moving, where the classification threshold is adjusted to favor the predictions of the minority class (Dangut et al., 2020; Ghosh et al., 2024).

Hybrid approaches integrate techniques from both data and algorithm levels to address data imbalance more effectively (Dangut et al., 2020; Johnson & Khoshgoftaar, 2019; Kaur et al., 2020). A common strategy is to combine sampling methods with ensemble learning techniques such as Bagging and Boosting. For example, SMOTE with Boosting (SMOTEWB), WSMOTE-ensemble, and combinations of oversampling with bagging or boosting have shown promise (Chen et al., 2024; Werner de Vargas et al., 2023). These methods aim to create a more balanced dataset for training the individual members of the ensemble, leading to improved and more robust performance. Finally, while not direct solutions to data imbalance, Transfer Learning (TL) can leverage knowledge from pre-trained models on large, balanced datasets to improve performance on imbalanced target datasets, especially when data is scarce (Alzubaidi et al., 2023; Chen et al., 2024; Zhao et al., 2024a). Self-Supervised Learning (SSL) can utilize abundant unlabeled data to learn effective features that can then be used for downstream tasks with imbalanced data, potentially improving the representation of minority classes (Akrim et al., 2023; Alzubaidi et al., 2023). Regularization techniques such as Dropout and early stopping could indirectly help prevent overfitting, which can be exacerbated by oversampling (Hellín et al., 2024; Johnson & Khoshgoftaar, 2019). The choice of the most suitable technique depends on the specific characteristics of the dataset, the severity of the imbalance, and the objectives of the application (Mooijman et al., 2023; Shen & Zhao, 2023; Shen et al., 2023; Werner de Vargas et al., 2023; Zhao et al., 2024a). Often, a combination of different approaches can yield the best results (Kaur et al., 2020; Mooijman et al., 2023; Zhao et al., 2024b).

2.2 Metaheuristics-based sampling

This research explores a data-level approach by generating two distinct datasets. These datasets are produced using MH algorithms to derive optimized solutions from a mathematical model that accurately characterizes the underlying physical phenomena. Although identical MH algorithms and physical models are employed in their generation, the datasets remain independent due to the inherent stochasticity of the MH processes. Each dataset encompasses the entire set of solutions obtained from multiple MH executions. Comprehensive details regarding both the mathematical formulation of the physical phenomena and the corresponding optimization model are described in (Araújo et al., 2024). The model offers consistent solutions in a relatively short time, which is interesting for generating physically reliable datasets with the appropriate dimensions to structure the problems.

Metaheuristics are best described as high-level general-purpose algorithmic frameworks designed to guide the search for good solutions in computationally hard optimization problems. They are intelligent search strategies that are commonly employed when exact methods are too slow or fail to find a solution (Doering et al., 2019; Drake et al., 2020). Although MHs do not guarantee the finding of a globally optimal solution, they are widely used due to their ability to provide near-optimal high-quality solutions in a reasonable time in various problem domains (Houssein et al., 2024; Talbi, 2009). Unlike problem-specific heuristics, MHs aim to overcome the limitation of being tailored to a single problem or instance. As frameworks, MHs are domain-independent, but their implementation requires adaptation to the specific problem at hand (Doering et al., 2019; Talbi, 2009). The effectiveness of these methods relies on their ability to adapt to the specific instance, avoid getting stuck in local optima, and effectively exploit the mathematical structure of the problem. Metaheuristics can be broadly classified into population-based methods that work with a set of solutions and single-solution methods that iteratively improve a single candidate solution. They can also be categorized as nature-inspired (e.g., Genetic Algorithms, Particle Swarm Optimization) or non-nature-inspired (e.g., Tabu Search, Simulated Annealing) (Doğan & Ölmez, 2015; Du & Swamy, 2016; Oliveira et al., 2020; Sörensen, 2015). Other taxonomies may be applied as well, as seen in (Houssein et al., 2024).

Metaheuristics inherently generate samples of the search space through their iterative optimization processes (Talbi, 2009). In population-based MHs, a set of candidate solutions is maintained and evolves over generations, effectively exploring different regions of the search space simultaneously. This diversification allows for a broad sampling of potentially good solutions. Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995) is a population-based MH that uses a swarm of particles, each representing a potentially good solution, to simultaneously explore the search space. The movement of these particles is guided by their individual best-found positions and the overall best position discovered by the swarm, facilitating information sharing and reducing the risk of becoming trapped in local optima. PSO has demonstrated effectiveness in tackling diverse optimization problems, including feature selection, portfolio optimization, and as a component of hybrid algorithms (Almaiah et al., 2024; Doering et al., 2019; Talbi, 2009). As for single-solution MHs, while focusing on a narrower region through intensification, they still explore the neighborhood of the current solution, generating a sequence of samples as they attempt to improve the objective function value. For example, the single solution Vortex Search (VS) algorithm (Doğan & Ölmez, 2015) begins by searching a wide area, thus considering a large portion of the solution space, and eventually reduces the size of the search to the region of the most promising solutions, resembling a vortex. The ability of some MHs to temporarily accept worse solutions allows them to escape local optima and explore other — potentially more promising — areas of the search space, directly leading to further sampling. Therefore, both population-based and single-solution MHs, through their mechanisms of exploration and exploitation, naturally sample the solution space as they search for optimal or near-optimal solutions (Doğan & Ölmez, 2015; Houssein et al., 2024; Talbi, 2009). This study employs STO, which is a MH that has not been published to date and has also been used in (Araújo et al., 2024).

Araújo et al. (2024) focused on the multi-objective optimization of a hypersonic air-breathing vehicle, specifically a scramjet engine, an enabling technology for sustained high-speed atmospheric flight. The authors employed multi-objective optimization techniques, utilizing MHs and an analytical model that was verified against computational fluid dynamics (CFD) data. The primary goal of optimization was to maximize thrust and minimize drag while simultaneously adhering to critical design constraints, such as preventing engine unstart by ensuring that the pressure ratio across shock waves remained below the Korkegi limit (Korkegi, 1975). To determine the most effective approach, the study tested three different formulations of the multi-objective function to identify the one that could achieve the highest net thrust while satisfying operational limits. The research findings indicate that incorporating Total Pressure Recovery (TPR) into the optimization process yielded superior results, concentrating the search in regions with greater uninstalled thrust and lower drag, and the numerical simulations corroborated the reliability of their analytical methodology for preliminary scramjet engine design, suggesting a way to efficiently design these engines in the early stages and reduce the need for extensive computational simulations. That study generated hundreds of thousands of samples that could be used to train DL models, as seen in (Paulino & Passaro, 2024).

This strategy is potentially capable of replacing computationally expensive simulation and optimization calculations in the early stages of vehicle design, given that enough testing is performed. However, the presence of data imbalance can severely degrade the quality of the predictions, as will be shown in the next sections. The present study highlights the effect of the data imbalance on the application of DNNs to predict the aerodynamic performance of scramjets.

3 Materials and methods

To systematically assess the impact of data imbalance on the predictive performance of DNNs, both quantitative and qualitative analyzes were performed using two datasets: an imbalanced dataset (Dataset 1) and a balanced dataset (Dataset 2). Both were created with the same optimization framework, and the imbalance is created by imposing limitations on the search space. In this sense, the imbalance is naturally created by these limitations, so that the two datasets are effectively distinct — not only in terms of data distribution, but also in terms of non-repeated data. In other words, the imbalanced dataset is not created simply by eliminating data from the balanced dataset. The following sections describe both datasets and the balancing procedure, incorporating quantitative summaries and visual analysis to support the discussion.

3.1 Datasets

The datasets analyzed in this study are derived from the optimization process described in Araújo et al. (2024), which generated candidate solutions relating net thrust to a set of input parameters representing both operational flight conditions and geometric characteristics of generic scramjet configurations. The input variables, detailed in Table 1, include geometric descriptors (such as the height of the combustion chamber and the intensities of the ramp shock) and freestream conditions (temperature, pressure, and Mach number). Thus, each sample represents a unique scramjet compression section under specific operational conditions, allowing the DNNs to learn and generalize aerodynamic responses throughout the hypersonic flight envelope.

Table 1: Description of input parameters. Target feature is marked with an asterisk (*).

FEATURE	DESCRIPTION	ROLE
h3	Combustion chamber height	Input
y0	1st ramp shock intensity	Input
y1	2nd ramp shock intensity	Input
y2	3rd ramp shock intensity	Input
T0	Freestream temperature	Input
P0	Freestream pressure	Input
M0	Freestream Mach number	Input
F_net	*Scramjet's net thrust**	**Output**

Dataset 1 covers a wide range of flight conditions, with Mach numbers ranging from 6 to 10 and altitudes from 25 to 35 km. The relationship between freestream pressure ("$P0$"), temperature ("$T0$"), and Mach number ("$M0$") results from their dependence on flight altitude ("Z"), although "Z" itself is not directly included as an input parameter. Table 2 summarizes the distribution of samples per flight condition in Dataset 1. In its turn, Dataset 2 was constructed to achieve a uniform representation across the parameter space, encompassing Mach numbers from 6 to 10 and altitudes from 20 to 40 km, which doubles the range in terms of total kilometers covered. Sampling was performed systematically, resulting in consistent coverage of all defined Mach-altitude combinations. Table 3 presents the sample distribution in Dataset 2. Both datasets serve as the basis for assessing the influence of data distribution on the training and performance of the DNN model.

The differences in the sample distributions between Dataset 1 and Dataset 2 are visually illustrated in Figures 1 to 4. Figure 1 shows the histogram of Mach numbers in Dataset 1, highlighting the pronounced disparities in sample frequency, while Figure 2 presents a scatterplot of normalized net thrust ("F_net") as a function of Mach number, color-coded by atmospheric pressure ("$P0$"). For Dataset 2, Figure 3 shows the histogram of Mach numbers with uniform bar heights, and Figure 4 provides a scatterplot of normalized "F_net" as a function of Mach number, also color-coded by altitude.

3.2 Characterization of data imbalance and generation of balanced dataset

Dataset 1 reveals clear asymmetries in sample representation across Mach numbers and altitudes. Mach numbers 6, 8, and 9 are each associated with three altitude conditions, whereas Mach 7 and 10 are represented by only two and one altitude condition(s), respectively. The sample counts per condition also vary, resulting in aggregate disparities — such as 59,400 samples for Mach 6 and 9, compared to only 19,800 for Mach 10 — leading to a 3:1 ratio between the most and least sampled regimes. These imbalances, summarized in Table 2, are visually confirmed by the histogram in Figure 1, which shows that Mach 6, 8, and 9 each contain

Table 2: Number of samples per flight conditions – Dataset 1

MACH "M0"	ALTITUDE "Z" (Km)	SAMPLES	SAMPLES/MACH	TOTAL SAMPLES
6	25	19,800		
6	30	19,800	59,400	
6	35	19,800		
7	25	19,800		
7	30	19,800	39,600	
8	25	19,800		
8	30	19,800	58,440	236,640
8	35	18,840		
9	25	19,800		
9	30	19,800	59,400	
9	35	19,800		
10	30	19,800	19,800	

Table 3: Number of samples per flight conditions – Dataset 2

MACH "M0"	ALTITUDE "Z" (Km)	SAMPLES	SAMPLES/MACH	TOTAL SAMPLES
6	20	3,960		
6	25	3,960		
6	30	3,960	19,800	
6	35	3,960		
6	40	3,960		
7	20	3,960		
7	25	3,960		
7	30	3,960	19,800	
7	35	3,960		
7	40	3,960		
8	20	3,960		
8	25	3,960		
8	30	3,960	19,800	99,000
8	35	3,960		
8	40	3,960		
9	20	3,960		
9	25	3,960		
9	30	3,960	19,800	
9	35	3,960		
9	40	3,960		
10	20	3,960		
10	25	3,960		
10	30	3,960	19,800	
10	35	3,960		
10	40	3,960		

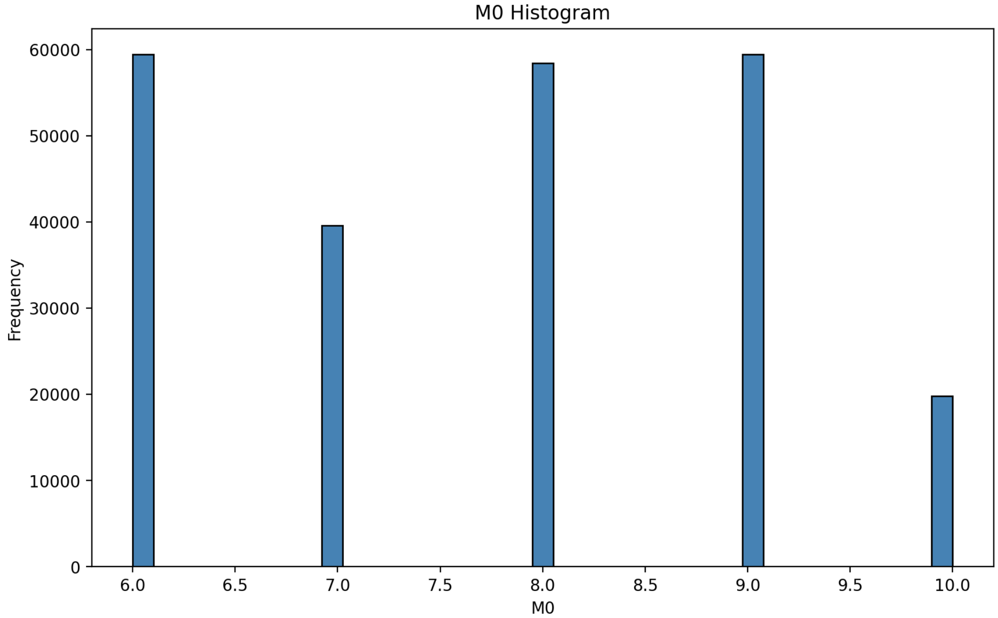

Figure 1: Histogram of "$M0$" values for Dataset 1.

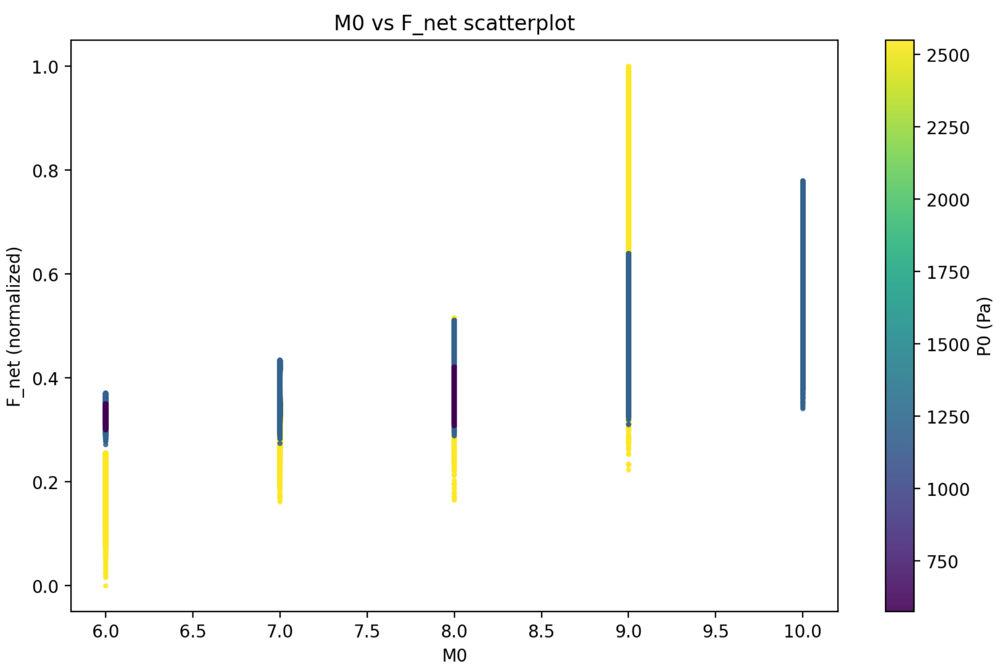

Figure 2: Scatterplot of the normalized values of "F_net" as a function of "$M0$" for Dataset 1. Each point indicates a sample. The color-coded scale on the right shows atmospheric pressure in pascals (Pa), which inversely correlates to the flight altitude.

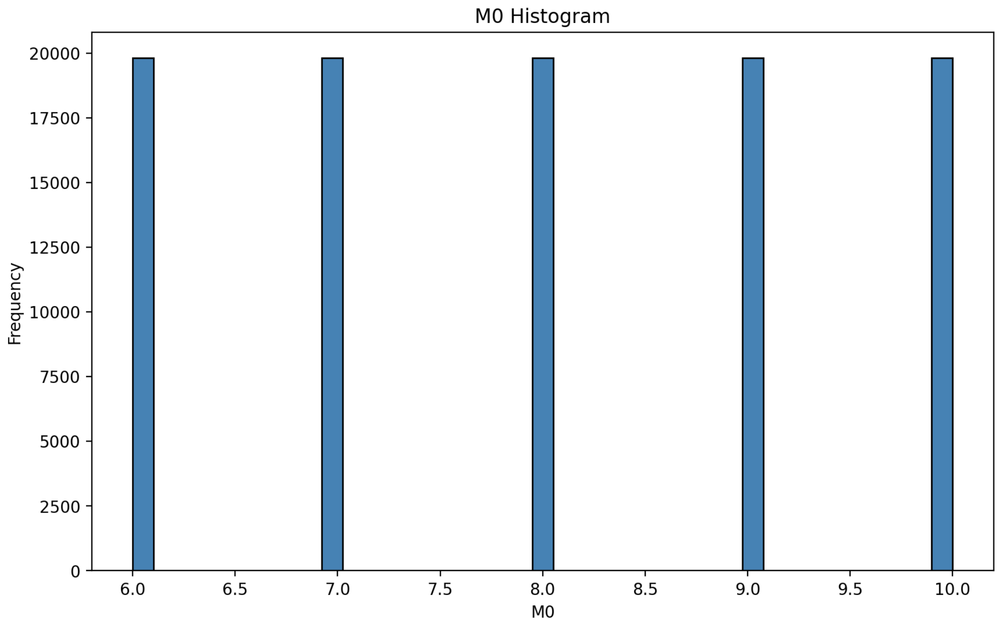

Figure 3: Histogram of "$M0$" values for Dataset 2.

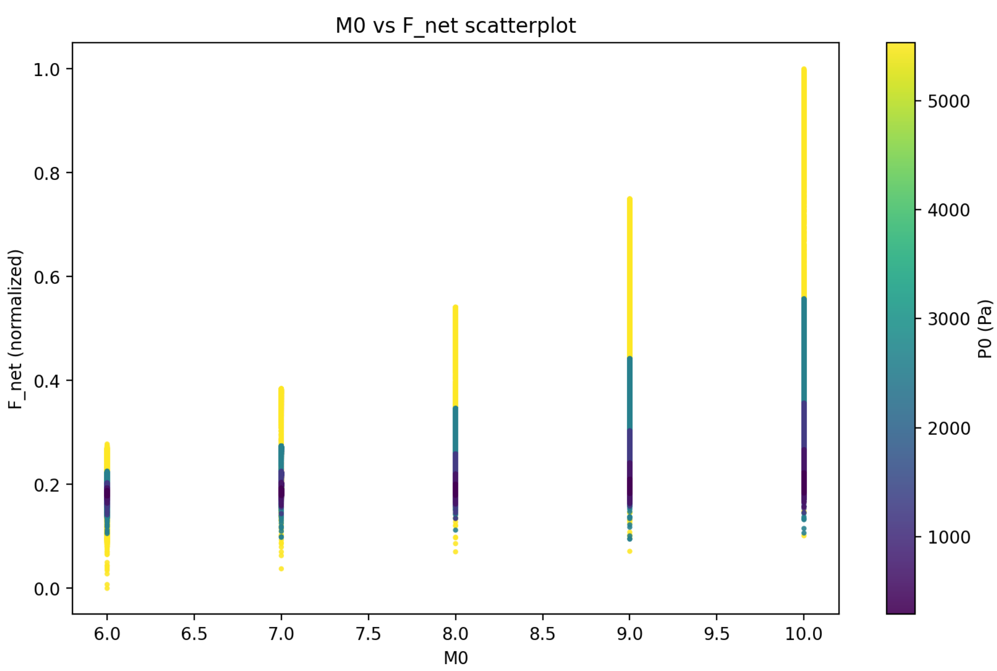

Figure 4: Scatterplot of the normalized values of "F_net" as a function of "$M0$" for Dataset 2. Each point indicates a sample. The color-coded scale on the right shows atmospheric pressure in pascals (Pa), which inversely correlates to the flight altitude.

approximately 59,000 samples, while Mach 7 and 10 are substantially underrepresented. Figure 2 further demonstrates coverage gaps in the parameter space: certain Mach-altitude combinations — particularly at high altitudes for Mach 7, 9, and 10 — are insufficiently sampled or entirely absent. Furthermore, Mach 10 lacks samples at low altitudes (high pressures). Such sampling irregularities have important implications for statistical analysis and predictive modeling.

Systematic biases may arise as a consequence of these irregularities, as models trained on the imbalanced Dataset 1 are likely to overfit to overrepresented regimes, reducing predictive confidence and accuracy in sparsely sampled conditions. Statistical inferences will also exhibit varying confidence levels depending on the Mach regime, with the least represented conditions (e.g., Mach 7 and 10) particularly susceptible to reduced reliability. To address these limitations, Dataset 2 was generated using a MH-based targeted strategy aimed at achieving uniform coverage across Mach number and altitude. The balancing procedure involved the definition of a systematic sampling grid, spanning Mach numbers 6 to 10 and altitudes from 20 to 40 km in regular increments, resulting in 25 unique combinations of operating conditions. Each combination was populated with 3,960 samples, for a total of 99,000 data points — ensuring an equitable distribution of 19,800 samples per Mach number (Table 3). Synthetic samples were generated using the STO MH, implemented in an optimization package named LOF-SYSTEM (Saba, 2017), following the methodology presented in (Araújo et al., 2024). LOF-SYSTEM offers a collection of established, single-solution and population-based MH algorithms, which have been applied in various academic research (Araújo et al., 2024; De Lima Filho et al., 2022; Santana et al., 2022; Silva et al., 2021; Soares et al., 2022).

The uniformity of sampling in Dataset 2 is confirmed by the histogram in Figure 3, where all Mach numbers are equally represented. Figure 4 shows that the normalized net thrust increases with the Mach number and decreases with altitude, reflecting expected physical trends such as higher air mass flow rates at lower altitudes (higher static pressures), thus validating the physical consistency of the synthetic samples. The homogeneous sample quantities across all experimental conditions eliminate potential weighting biases in subsequent analyses, providing a robust basis for assessing DNN generalization performance, which is further discussed in Section 4. By characterizing both datasets and detailing the methodology for constructing a balanced dataset, this section establishes a comprehensive framework for evaluating the effects of data distribution on deep learning models in the context of scramjet aerodynamic modeling.

3.3 DNNs training

This study adopts, without modification, the optimal DNN architecture proposed by Paulino & Passaro (2024). In that work, the architecture was determined through a systematic process that began with simple models available in the PyTorch library (Paszke et al., 2019) and progressively increased in complexity by adding hidden layers and tuning the number of neurons, until satisfactory performance was achieved. The training procedure used the mean squared error (MSE) loss function (Goodfellow et al., 2016). The accuracy of the model was evaluated using the root mean square error (RMSE) (Chai & Draxler, 2014) and the coefficient of determination (R^2), which is a statistical parameter commonly used to assess the quality of the estimates made (Brar & Singh, 2024; Staerk et al., 2024), both of which are widely used in regression tasks (Ispir et al., 2023; Staerk et al., 2024). In this work, the previously defined architecture (illustrated in Figure 5) is used to train both Datasets. The first layer (Input) receives a matrix in which the number of rows corresponds to the number of samples presented to the network and the number of columns matches the selected input parameters, which also define the number of input neurons. The last layer (Output) consists of a tensor that represents the predicted value of the target variable, "F_net". The number of neurons in each hidden layer was originally determined empirically, and the rectified linear unit (ReLU) activation, whose use is recommended in the literature (Goodfellow et al., 2016), is applied to each hidden layer.

3.4 Computational implementation

The development environment for this study was based on Python, leveraging the PyTorch library (Paszke et al., 2019) to construct, train, and evaluate the DNN models. PyTorch was selected for its intuitive debugging, computational efficiency, and native support for GPU acceleration. Additional libraries were used to support various tasks: NumPy (Harris et al., 2020) and Pandas (McKinney, 2010) for data manipulation

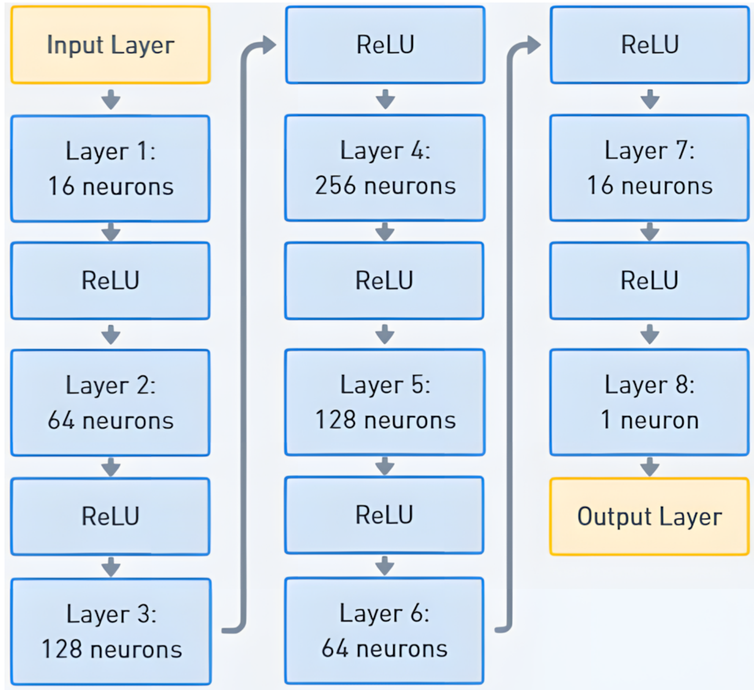

Figure 5: Architecture of the networks used for the experiments.

and numerical operations; Scikit-learn (Pedregosa et al., 2011) for statistical tools and machine learning utilities; and Matplotlib (Hunter, 2007) and Plotly (Inc., 2015) for data visualization. A complete Python-based pipeline was developed to preprocess the datasets, define and train neural network architectures, and carry out performance evaluation in various flight conditions. It is important to note that, as previously described, the synthetic samples used to construct the Datasets were generated separately using the STO MH implemented in the LOF-SYSTEM software (Saba, 2017), outside of the Python environment.

3.5 Experiments

Following the computational implementation described in the previous section, the performance of DNNs in predicting normalized "F_net" from input parameters was evaluated using a simplified version of the Seven-Fold Cross-Validation (7FCV) strategy (Paulino et al., 2019; 2020). Instead of considering all 42 possible combinations of training (5 folds), validation (1 fold), and generalization (1 fold) sets among the seven partitions, this study adopts a simplified approach, hence called "Simplified 7FCV" (S7FCV). Before partitioning, all samples were randomly shuffled to ensure statistical independence across Folds. One of the seven folds is fixed as the generalization set for all cross-validation iterations, while the remaining six folds are used to generate all combinations of five folds for training and one for validation, resulting in six iterations ($C_5^6 \cdot C_1^1 = 6$). This configuration allows for model selection and evaluation under a consistent generalization scenario while reducing computational cost compared to the full combinatorial approach. A depiction of the S7FCV scheme is shown in Table 4. For Dataset 1, each fold contained 168,994 training samples and 33,799 validation samples. For Dataset 2, which is smaller but uniformly balanced, each fold had 70,715 training samples and 14,143 validation samples. Prior to training, both datasets were normalized to the range [0, 1]. A grid search (Bergstra et al., 2012) was applied to tune the learning rate ("lr"), selected for its simplicity and suitability to explore a small, manually defined hyperparameter space. The hyperparameter "lr" was varied in the discrete set 0.0002, 0.0005, 0.001, 0.002. This evaluation setup enables comparison of models trained on diverse data, as detailed next.

Table 4: Cross-validation folds and data partitioning

ITERATION	FOLD A	FOLD B	FOLD C	FOLD D	FOLD E	FOLD F	FOLD G
1	Validation	Training	Training	Training	Training	Training	Generalization
2	Training	Validation	Training	Training	Training	Training	Generalization
3	Training	Training	Validation	Training	Training	Training	Generalization
4	Training	Training	Training	Validation	Training	Training	Generalization
5	Training	Training	Training	Training	Validation	Training	Generalization
6	Training	Training	Training	Training	Training	Validation	Generalization

3.6 Prediction Comparison

This methodological framework was designed to address the data imbalance in predicting aerodynamic performance based on deep learning for hypersonic vehicles. Using Dataset 2 — a uniformly sampled dataset — and using the established neural architecture, the study aims to assess the extent to which balanced sampling improves model reliability and generalization. To this end, two models were trained: Model 1, using the imbalanced Dataset 1, and Model 2, using the balanced Dataset 2. Both models were evaluated under identical flight condition scenarios to ensure comparability. The results and comparative analysis of their predictive performance are presented and discussed in the next Section.

4 Results and discussions

The neural architecture adopted in this study was trained and evaluated using multiple input configurations derived from the optimized datasets. The assessment focused on two complementary aspects: (i) the statistical predictive performance of the models, measured through cross-validation using the S7FCV algorithm, and (ii) the physical coherence of the predictions under a wide range of flight conditions. These dimensions are important to validate the applicability of the models to realistic aerodynamic scenarios.

4.1 Generalization performance

To evaluate generalizability, the predictive performance was compared between the two models. Although both models achieved low error levels, with accuracies superior to 99%, Model 1 exhibited slightly better RMSE (0.0122) compared to Model 2 (0.0142), representing a relative difference of approximately 14%. This marginal advantage is most likely due to the larger sample size in Dataset 1, which contains roughly 2.4 times more training data than Dataset 2 — a factor that leads to a proportionally greater number of weight updates during training. However, statistical analysis indicates that this difference is not significant in practical terms. The upper and lower bounds of the prediction errors are also comparable: ± 0.1819 for Model 1 and ± 0.1882 for Model 2 (Figures 6 and 7). In addition, both models show a high concentration of pointwise differences near zero, suggesting similar predictive accuracy in aggregate. However, this similarity in the numerical error metrics may create a misleading impression of interchangeability between the two datasets. As detailed in Subsection 4.2, the limitations of Dataset 1 emerge under flight conditions deviating from the original optimization parameters.

4.2 Evaluation of aerodynamic performance under different flight conditions

Figures 8 and 9 illustrate the predicted values of normalized "F_net" across a range of Mach number and altitude combinations, as produced by Model 2 and Model 1, respectively, for a given scramjet geometry. The analysis begins with Model 2 (Figure 9), which was trained on the balanced Dataset 2 and demonstrates behavior that is consistently aligned with physical expectations. Specifically, the predicted thrust increases monotonically with Mach number and decreases with altitude, clearly reflecting the physical principle that higher air mass flow ("$\dot{m}$") at lower altitudes enhances thrust generation. Consequently, the resulting dis-

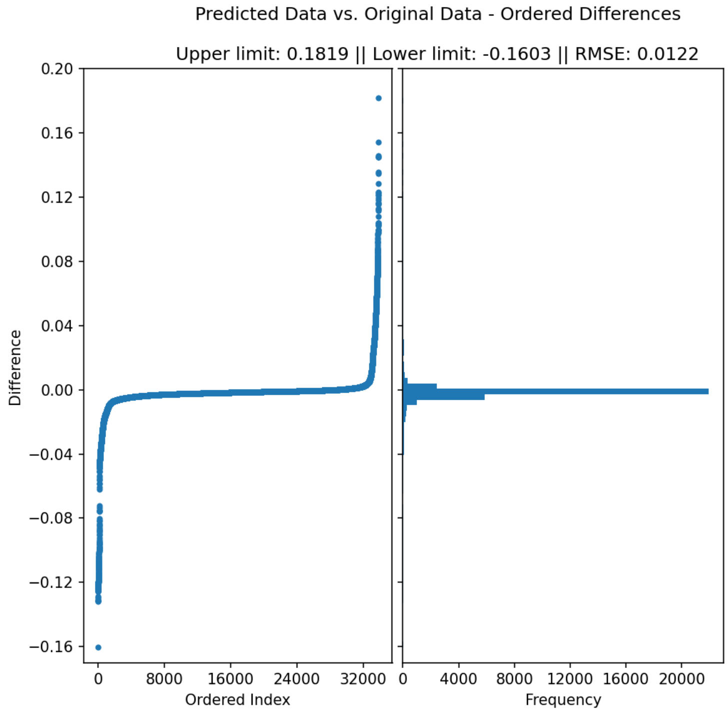

Figure 6: Ordered normalized differences between predictions and original data for Model 1.

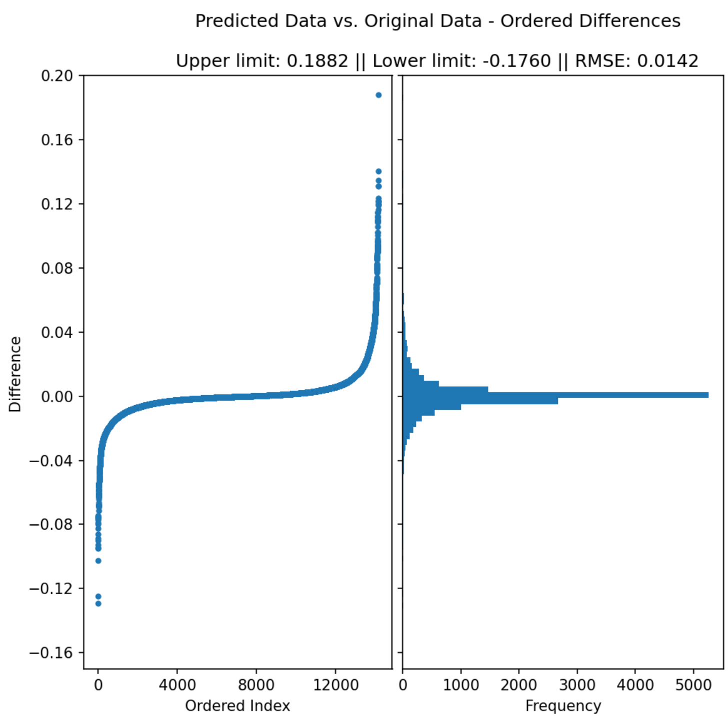

Figure 7: Ordered normalized differences between predictions and original data for Model 2.

tribution of predictions is smooth, continuous, and physically coherent across the entire parameter space, thereby serving as a reliable physical consistency check.

In contrast, Model 1 (Figure 8) exhibits several physically inconsistent behaviors that significantly deviate from the expected aerodynamic trends. For instance, at Mach 6, the model inaccurately predicts a lower thrust at lower altitudes, clearly contradicting the physical principle that a higher air density in these regions should yield a higher thrust (Figure 8, region 1). Another notable inconsistency is observed in the 20 km altitude curve, which initially exhibits a positive slope up to Mach 7.9, followed by a marked flattening at higher Mach values — an unrealistic inversion given the expected monotonic thrust increase with Mach number (Figure 8, region 2). Furthermore, from approximately Mach 8.8 onward (Figure 8, region 3), the derivative/slope of predicted "*F_net*" values increases under high-altitude conditions (40 km), deviating from the more uniform and physically plausible gradient observed in Model 2. Although Model 2 is largely consistent with physical intuition, it presents a minor anomaly at high Mach numbers (>9.5) and low altitudes (20 km), where the thrust curve slightly flattens or inflects (Figure 9, region 1). This subtle inconsistency could be attributed to either limited sample density in that specific parameter region or limitations in the current network architecture (e.g., insufficient depth or nonlinearity in activation functions) to fully capture the exponential nature of thrust variation. Further exploration of these potential architectural enhancements is proposed as future work.

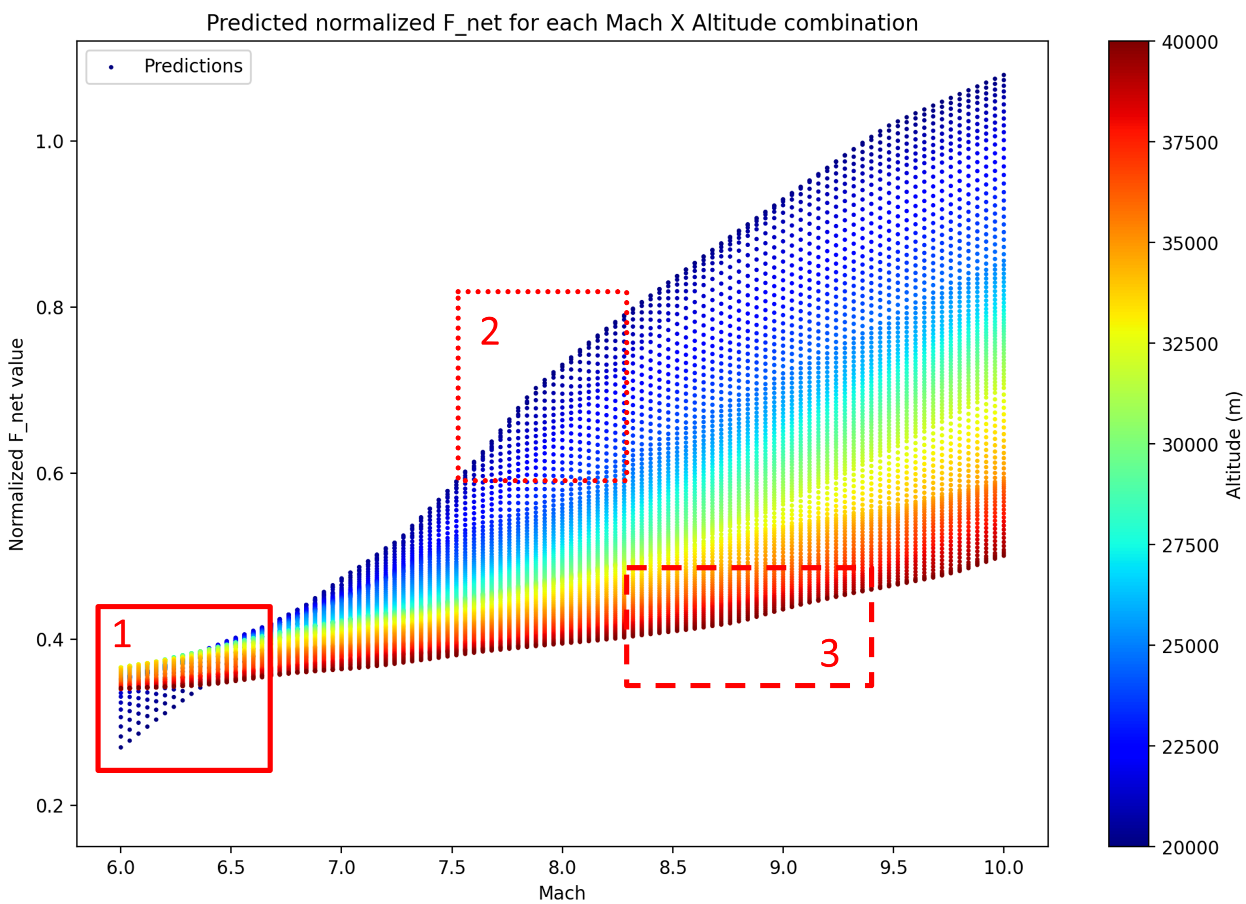

Figure 8: Predicted normalized "*F_net*" for each Mach X Altitude combination for Model 1. Physically inconsistent regions represented: (1 – continuous line) inverted altitude effect at Mach 6; (2 – dotted line) unrealistic flattening of 20 km curve at mid Mach; (3 – dashed line) anomalous slope increase at high altitude for $\geq$ 8.8 Mach.

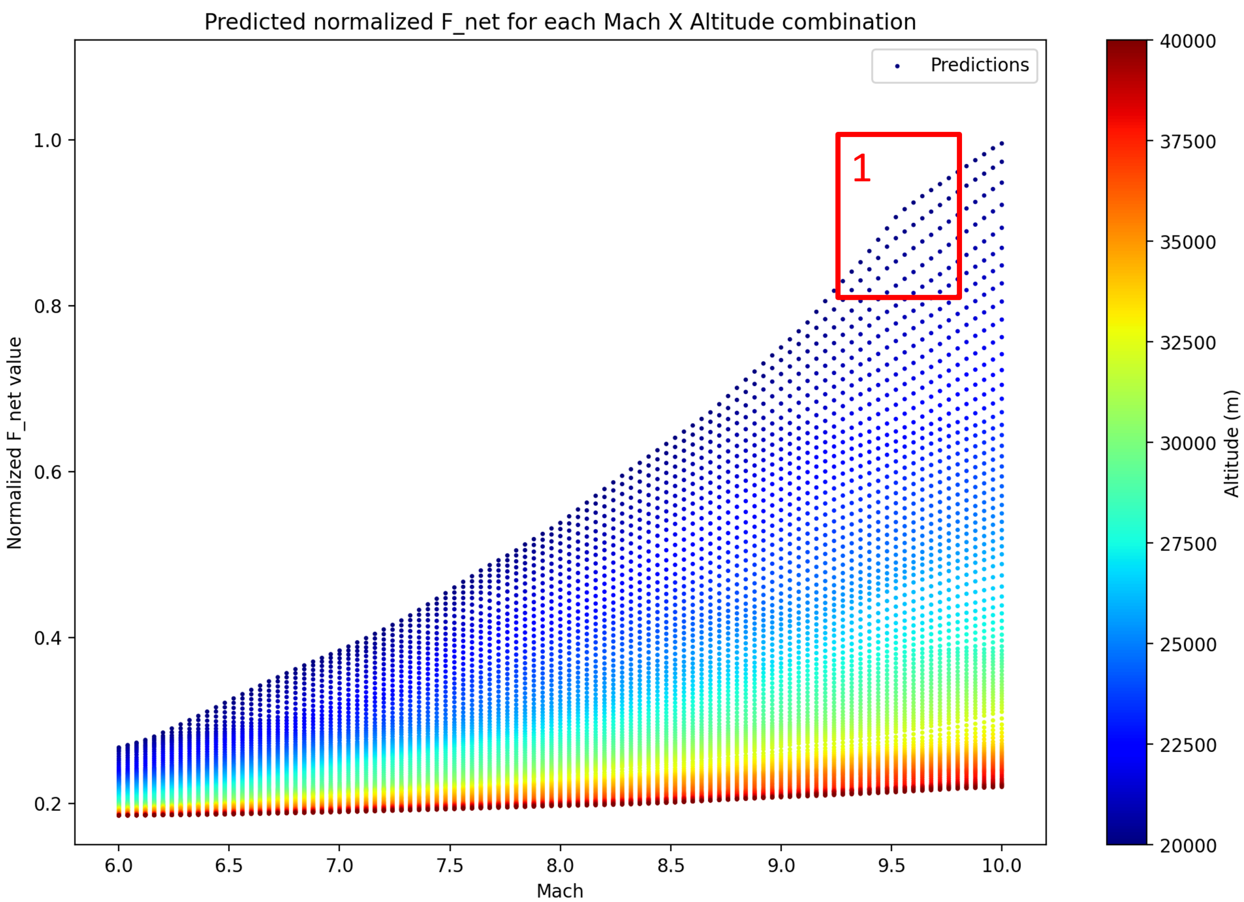

Figure 9: Predicted normalized "F_net" for each Mach X Altitude combination for Model 2. Minor physically inconsistent region represented: (1 – continuous line) slight flattening at low altitude (20 km) and high Mach (>9.5).

4.3 Discussion on Data Imbalance

Model 1 and Model 2 share the exact same neural architecture: identical depth, number of neurons per layer, and activation functions. The sole distinction between them lies in the datasets used during training: unbalanced Dataset 1 (236,640 samples), and balanced Dataset 2 (99,000 samples). The results of Section 4.1 suggest potential model equivalence, as evidenced by statistically similar accuracy scores, RMSE values, and error distributions. However, examination of the predictive outputs reveals significant discrepancies in terms of the observed physical behavior, indicating a divergence from the expected dynamics. As detailed in Section 3, Dataset 1 exhibited uneven coverage of the parameter space, with critical Mach–altitude combinations either underrepresented or entirely absent. The problem is not quantity, that is, the lack of enough samples, but representativeness. This imbalance — particularly the underrepresentation of certain Mach–altitude combinations — directly contributed to the physically inconsistent behaviors observed in the predictions of Model 1 discussed in Subsection 4.2.

The results of this study highlight that the consequences of data imbalance in AI models extend far beyond minor statistical inaccuracies. Using an identical network topology, loss function and optimizer, we find that Model 1 achieves a 14% lower RMSE (0.0122 vs. 0.0142) yet produces thrust profiles that violate first-principles expectations, with local errors exceeding 100% and non-monotonic thrust at low altitude. This outcome mirrors the concerns raised by Ghosh et al. (2024), who reported that headline accuracy can hide bias introduced by the class imbalance, and by Chen et al. (2024) and Hellín et al. (2024), who noted that many studies assume implicitly balanced data and therefore overlook critical failure modes. By isolating dataset balance as the sole experimental variable, the present work provided a quantitative demonstration that rigorous data engineering is a prerequisite for physically consistent deep learning predictions in hypersonic aerodynamics applications.

In practical aerospace design, such mispredictions could propagate into downstream processes, leading to fundamental misjudgments about vehicle performance. In the worst-case scenario, this could compromise the flyability of scramjet-powered systems. More broadly, these findings underscore that in domains governed by physical laws, data is not just input — it encapsulates the structure of the reality being modeled. Therefore, ensuring that the dataset reflects the true distribution of relevant physical regimes is a methodological imperative, not a peripheral concern. Researchers and engineers must treat data balancing as a foundational step in responsible AI development. By doing so, we move toward models that are not only statistically robust but also physically interpretable, less prone to bias, and more trustworthy in critical applications. This mindset is essential for advancing the deployment of AI across disciplines where mathematical models are inherently tied to the laws of nature — not just in aerospace engineering but across a broad range of research domains.

5 Conclusion

Juxtaposing theoretical expectations with empirical evidence, this study has demonstrated that data imbalance can considerably degrade the predictive quality of DNNs in modeling the aerodynamic behavior of scramjet compression sections across broad operational regimes. By comparing two datasets — Dataset 1, which is imbalanced, and Dataset 2, which is balanced — the results showed that data imbalance does not merely affect performance marginally; it can lead to significant and physically implausible outcomes. Although Dataset 1 included 2.4 times more training samples than Dataset 2, Model 1 (trained on Dataset 1) did not exhibit superior global accuracy compared to Model 2 (trained on Dataset 2). Model 1 failed to maintain robustness under flight conditions that deviated from the geometric configuration's optimization baseline.

Even more concerning was Model 1's tendency to produce unrealistic aerodynamic predictions — such as reduced net thrust at lower altitudes or inversions in slope where thrust behavior should have followed established physical trends. These inconsistencies suggest that the model captured artifacts induced by the biased data distribution, leading to unrealistic predictions. In contrast, Model 2 exhibited stable and physically coherent patterns: thrust increased with Mach number and decreased with altitude, as expected from compressible flow dynamics. This difference reinforces the principle that the balance of the dataset is not an auxiliary consideration, but a core requirement to preserve physical integrity in model outputs. This work, situated within an educational context, highlights that even when established data imputation techniques from the literature are applied, the primary issue of class imbalance can persist if absent data instances are not populated rationally.

Such findings underscore the scientific imperative of addressing data imbalance in deep learning applications, particularly in high-stakes domains governed by physical laws, such as hypersonics. Beyond evident improvements in prediction stability, balanced data ensure that DNNs generalize reliably across the full design space and do not embed or amplify data-driven artifacts. Adequate data balancing strategies can be used to restore representativeness and support interpretability, ultimately allowing the development of models that are not only statistically valid, but physically trustworthy. The broader implication is that balanced datasets directly contribute to the transparency, reproducibility, and credibility of AI-assisted engineering workflows.

In this framework, future studies should extend the investigation beyond the original data imbalance and consider potential imbalances within the folds generated by the S7FCV scheme. Although S7FCV offers a robust structure for model evaluation and selection, imbalanced partitions could still introduce subtle biases that compromise model assessment. Furthermore, exploration of balancing strategies applied to the unbalanced dataset − which encompasses data-driven, algorithmic, or hybrid approaches − is warranted to compare performance against the metaheuristics-balanced dataset. Furthermore, variations in network architecture, such as differences in activation functions, layer depth, or neuron count, merit investigation for potential improvements.

Studies aiming to model the non-linear and potentially exponential behavior of scramjet systems could benefit from more intricate architectures, including self-attention mechanisms or Transformer-based models. Moreover, the integration of AI methodologies should be encouraged not only in compression ramp analysis but also in broader aspects of hypersonic vehicle design. Finally, thoughtful data engineering can enhance model realism, making it an essential component of responsible and rigorous AI deployment in aerospace research.

References

Shaojie Ai, Jia Song, and Guobiao Cai. A real-time fault diagnosis method for hypersonic air vehicle with sensor fault based on the auto temporal convolutional network. *Aerospace Science and Technology*, 119: 107220, dec 2021. ISSN 12709638. doi: 10.1016/j.ast.2021.107220. URL `https://linkinghub.elsevier.com/retrieve/pii/S1270963821007306`.

Anass Akrim, Christian Gogu, Rob Vingerhoeds, and Michel Salaün. Self-Supervised Learning for data scarcity in a fatigue damage prognostic problem. *Engineering Applications of Artificial Intelligence*, 120(January):105837, apr 2023. ISSN 09521976. doi: 10.1016/j.engappai.2023.105837. URL `https://doi.org/10.1016/j.engappai.2023.105837https://linkinghub.elsevier.com/retrieve/pii/S0952197623000210`.

Mohammed Amin Almaiah, Rana Alrawashdeh, Tayseer Alkhdour, Romel Al-Ali, Gaith Rjoub, and Theyazan Aldahyani. Detecting DDoS attacks using machine learning algorithms and feature selection methods. *International Journal of Data and Network Science*, 8(4):2307–2318, 2024. ISSN 25618148. doi: 10.5267/j.ijdns.2024.6.001. URL `http://www.growingscience.com/ijds/Vol8/ijdns{_}2024{_}107.pdf`.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, A. S. Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A. Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H. Al-Timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin Abbosh, and Yuantong Gu. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46, apr 2023. ISSN 2196-1115. doi: 10.1186/s40537-023-00727-2. URL `https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00727-2`.

J. D. Anderson. *Modern Compressible Flow: With Historical Perspective*. Aeronautical and Aerospace Engineering Series. McGraw-Hill Education, 2003. ISBN 9780072424430. URL `https://books.google.com.br/books?id=woeqa4-a5EgC`.

Pedro P. B. Araújo, Roberto Y. Tanaka, Carlos A. Silva, Angelo Passaro, and Paulo G. P. Toro. Multi-objective optimization of a hypersonic airbreathing vehicle. *Physics of Fluids*, 36(2), feb 2024. ISSN 1070-6631. doi: 10.1063/5.0181366. URL /aip/pof/article/36/2/026120/3267168/Multi-objective-optimization-of-a-hypersonichttps://pubs.aip.org/pof/article/36/2/026120/3267168/Multi-objective-optimization-of-a-hypersonic.

Atticus Beachy, Harok Bae, Ian Boyd, and Ramana Grandhi. Emulator embedded neural networks for multi-fidelity conceptual design exploration of hypersonic vehicles. In *AIAA Scitech 2021 Forum*, volume 1 PartF, pp. 1 – 24. American Institute of Aeronautics and Astronautics Inc, AIAA, 2021. ISBN 978-162410609-5. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100303247{&}partnerID=40{&}md5=9adfa4368237d8078c935638cc25eec9.

James Bergstra, James Bergstra Ca, and Yoshua Bengio Ca. Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research*, 13:281–305, 2012. URL http://scikit-learn.sourceforge.net.

Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, 12(6):e0177678, jun 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0177678. URL https://dx.plos.org/10.1371/journal.pone.0177678.

Amrinder Singh Brar and Kawaljeet Singh. A multi-objective stacked regression method for distance based colour measuring device. *Scientific Reports*, 14(1):5530, mar 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-54785-4. URL https://www.nature.com/articles/s41598-024-54785-4.

Steven L. Brunton, J. Nathan Kutz, Krithika Manohar, Aleksandr Y. Aravkin, Kristi Morgansen, Jennifer Klemisch, Nicholas Goebel, James Buttrick, Jeffrey Poskin, Adriana W. Blom-Schieber, Thomas Hogan, and Darren McDonald. Data-Driven Aerospace Engineering: Reframing the Industry with Machine Learning. *AIAA Journal*, 59(8):1–26, jul 2021. ISSN 0001-1452. doi: 10.2514/1.J060131. URL https://arc.aiaa.org/doi/10.2514/1.J060131http://arxiv.org/abs/2008.10740.

Xiangwei Bu, Changchun Hua, Maolong Lv, and Zhonghua Wu. Flight Control of Waverider Vehicles with Fragility-avoidance Prescribed Performance. *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–15, 2023. ISSN 00189251. doi: 10.1109/TAES.2023.3251314. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149384638{&}doi=10.1109{%}2FTAES.2023.3251314{&}partnerID=40{&}md5=267c4cf5c0d01291fd2106f8dc2c03a7.

T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 2014. ISSN 19919603. doi: 10.5194/gmd-7-1247-2014.

Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and C. L. Philip Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6):137, may 2024. ISSN 1573-7462. doi: 10.1007/s10462-024-10759-6. URL https://link.springer.com/article/10.1007/s10462-024-10759-6https://link.springer.com/10.1007/s10462-024-10759-6.

Luyao Cui, Aijun Liu, Changjun Yv, and Taifan Quan. Learning Algorithm for Tracking Hypersonic Targets in Near Space. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, volume 227 LNICST, pp. 206–214. Springer, 2018. ISBN 9783319734460. doi: 10.1007/978-3-319-73447-7_24. URL http://link.springer.com/10.1007/978-3-319-73447-7{_}24.

Maren David Dangut, Zakwan Skaf, and Ian Jennions. Rescaled-LSTM for Predicting Aircraft Component Replacement Under Imbalanced Dataset Constraint. In *2020 Advances in Science and Engineering Technology International Conferences (ASET)*, pp. 1–9. IEEE, feb 2020. ISBN 978-1-7281-4640-9. doi: 10.1109/ASET48392.2020.9118253. URL https://ieeexplore.ieee.org/document/9118253/.

Geraldo Mulato De Lima Filho, Andre Rossi Kuroswiski, Felipe Leonardo Lobo Medeiros, Mark Voskuijl, Herman Monsuur, and Angelo Passaro. Optimization of Unmanned Air Vehicle Tactical Formation in War Games. *IEEE Access*, 10:21727–21741, 2022. ISSN 21693536. doi: 10.1109/ACCESS.2022.3152768.

Jana Doering, Renatas Kizys, Angel A. Juan, Àngels Fitó, and Onur Polat. Metaheuristics for rich portfolio optimisation and risk management: Current state and future trends, jan 2019. ISSN 22147160.

Yutong Dong, Hongkai Jiang, Yunpeng Liu, and Zichun Yi. Global wavelet-integrated residual frequency attention regularized network for hypersonic flight vehicle fault diagnosis with imbalanced data. *Engineering Applications of Artificial Intelligence*, 132(February):107968, 2024. ISSN 09521976. doi: 10.1016/j.engappai.2024.107968. URL `https://doi.org/10.1016/j.engappai.2024.107968`.

Berat Doğan and Tamer Ölmez. A new metaheuristic for numerical function optimization: Vortex Search algorithm. *Information Sciences*, 293(February):125–145, feb 2015. ISSN 00200255. doi: 10.1016/j.ins.2014.08.053. URL `http://linkinghub.elsevier.com/retrieve/pii/S0020025514008585`.

John H. Drake, Ahmed Kheiri, Ender Özcan, and Edmund K. Burke. Recent advances in selection hyper-heuristics, sep 2020. ISSN 03772217.

Ken-Li. Du and M. N. S. Swamy. *Search and Optimization by Metaheuristics Techniques and Algorithms Inspired by Nature*. Birkhäuser, 1 edition, 2016.

Chihiro Fujio and Hideaki Ogawa. Deep-learning prediction and uncertainty quantification for scramjet intake flowfields. *Aerospace Science and Technology*, 130:107931, nov 2022. ISSN 12709638. doi: 10.1016/j.ast.2022.107931. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140220589{&}doi=10.1016{%}2Fj.ast.2022.107931{&}partnerID=40{&}md5=1e5e70207bf543a41abc25c69fdddc5ahttps://linkinghub.elsevier.com/retrieve/pii/S1270963822006058`.

Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901, jul 2024. ISSN 0885-6125. doi: 10.1007/s10994-022-06268-8. URL `https://link.springer.com/article/10.1007/s10994-022-06268-8https://link.springer.com/10.1007/s10994-022-06268-8`.

Arjun Gond and Aditi Sengupta. Data-driven optimal design of premixed hydrogen-air combustion in a microcombustor. *AI Thermal Fluids*, 1(March):100007, mar 2025. ISSN 30505852. doi: 10.1016/j.aitf.2025.100007. URL `https://doi.org/10.1016/j.aitf.2025.100007https://linkinghub.elsevier.com/retrieve/pii/S3050585225000060`.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL `https://www.deeplearningbook.org/`.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, sep 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2https://www.nature.com/articles/s41586-020-2649-2`.

Simon S Haykin. *Neural networks and learning machines*. Pearson Education, Upper Saddle River, NJ, third edition, 2009. ISBN 9780131471399.

Haibo He and E.A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, sep 2009. ISSN 1041-4347. doi: 10.1109/TKDE.2008.239. URL `http://ieeexplore.ieee.org/document/5128907/`.

Carlos J. Hellín, Alvaro A. Olmedo, Adrián Valledor, Josefa Gómez, Miguel López-Benítez, and Abdelhamid Tayebi. Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification. *Applied Sciences*, 14(8):3419, apr 2024. ISSN 2076-3417. doi: 10.3390/app14083419. URL `https://www.mdpi.com/2076-3417/14/8/3419/htmhttps://www.mdpi.com/2076-3417/14/8/3419`.

Essam H. Houssein, Mahmoud Khalaf Saeed, Gang Hu, and Mustafa M. Al-Sayed. Metaheuristics for Solving Global and Engineering Optimization Problems: Review, Applications, Open Issues and Challenges. *Archives of Computational Methods in Engineering*, 31(8):4485–4519, dec 2024. ISSN 1134-3060. doi: 10.1007/s11831-024-10168-6. URL https://link.springer.com/10.1007/s11831-024-10168-6.

John D Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55. URL http://ieeexplore.ieee.org/document/4160265/.

Plotly Technologies Inc. Collaborative data science, 2015. URL https://plot.ly.

Ali Can Ispir, Bayindir H. Saracoglu, Thierry Magin, and Axel Coussement. A methodology for estimating hypersonic engine performance by coupling supersonic reactive flow simulations with machine learning techniques. *Aerospace Science and Technology*, 140:108501, sep 2023. ISSN 12709638. doi: 10.1016/j.ast.2023.108501. URL https://doi.org/10.1016/j.ast.2023.108501https://linkinghub.elsevier.com/retrieve/pii/S127096382300398X.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, dec 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0192-5. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5.

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys*, 52(4):1–36, jul 2020. ISSN 0360-0300. doi: 10.1145/3343440. URL https://dl.acm.org/doi/10.1145/3343440.

J Kennedy and R Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pp. 1942–1948. IEEE, 1995. ISBN 0-7803-2768-3. doi: 10.1109/ICNN.1995.488968. URL http://ieeexplore.ieee.org/document/488968/.

Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, aug 2018. ISSN 2162-237X. doi: 10.1109/TNNLS.2017.2732482. URL https://ieeexplore.ieee.org/document/8012579/.

R. H. Korkegi. Comparison of Shock-Induced Two- and Three- Dimensional Incipient Turbulent Seperation. *AIAA Journal*, 13(4):534–535, apr 1975. ISSN 0001-1452. doi: 10.2514/3.49750. URL https://arc.aiaa.org/doi/10.2514/3.49750.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. ISSN 0028-0836. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539.

Julia Ling, Andrew Kurzawski, and Jeremy Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, nov 2016. ISSN 0022-1120. doi: 10.1017/jfm.2016.615. URL https://www.cambridge.org/core/product/identifier/S0022112016006157/type/journal{_}article.

Xiaowen Liu and Huixin Tian. Research on Imbalanced Data Regression Based on Confrontation. *Processes*, 12(2):375, feb 2024. ISSN 2227-9717. doi: 10.3390/pr12020375. URL https://www.mdpi.com/2227-9717/12/2/375.

Jixing Lv, Changhong Wang, and Yonggui Kao. Adaptive fixed-time quantized fault-tolerant attitude control for hypersonic reentry vehicle. *Neurocomputing*, 520:386–399, feb 2023. ISSN 09252312. doi: 10.1016/j.neucom.2022.11.057. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145569342{&}doi=10.1016{%}2Fj.neucom.2022.11.057{&}partnerID=40{&}md5=2dc8d2e7662a59fad719bd04bd294136https://linkinghub.elsevier.com/retrieve/pii/S0925231222014333.

Z. Mao, L. Lu, O. Marxen, T.A. Zaki, and G.E. Karniadakis. DeepM&Mnet for hypersonics: Predicting the coupled flow and finite-rate chemistry behind a normal shock using neural-network approximation of operators. *Journal of Computational Physics*, 447, 2021. doi: 10.1016/j.jcp.2021.110698.

Thomas McCall, Kiarash Seyed Alavi, Loveneesh Rana, and Bernd Chudoba. Artificial Intelligent Research Assistant for Aerospace Design Synthesis—Solution Logic. In *22nd AIAA International Space Planes and Hypersonics Systems and Technologies Conference*, Reston, Virginia, sep 2018. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-577-7. doi: 10.2514/6.2018-5387. URL https://arc.aiaa.org/doi/10.2514/6.2018-5387.

Warren McCulloch and Walter Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity (1943). *Ideas That Created the Future*, 5(4):79–88, feb 2021. doi: 10.7551/mitpress/12274.003.0011. URL https://direct.mit.edu/books/book/5003/chapter/2657032/A-Logical-Calculus-of-the-Ideas-Immanent-in.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a. URL https://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Paul Mooijman, Cagatay Catal, Bedir Tekinerdogan, Arjen Lommen, and Marco Blokland. The effects of data balancing approaches: A case study. *Applied Soft Computing*, 132:109853, jan 2023. ISSN 15684946. doi: 10.1016/j.asoc.2022.109853. URL https://linkinghub.elsevier.com/retrieve/pii/S1568494622009024.

Paulo Moura Oliveira, E. J. Solteiro Pires, José Boaventura-Cunha, and Tatiana Martins Pinho. Review of nature and biologically inspired metaheuristics for greenhouse environment control, aug 2020. ISSN 01423312.

E. Ozbenli, P. Vedula, K. Vogiatzis, and E. Josyula. Numerical solution of hypersonic flows via artificial neural networks. In *AIAA Scitech 2020 Forum*, volume 1 PartF, 2020. ISBN 9781624105951. doi: 10.2514/6.2020-1233.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper{_}files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Ângelo de Carvalho Paulino and Angelo Passaro. Investigação cientométrica sobre aplicações de Redes Neurais em Hipersônica. In *XXV Simpósio de Aplicações Operacionais em Áreas de Defesa*, pp. 1, São José dos Campos, Brasil, 2023.

Ângelo de Carvalho Paulino and Angelo Passaro. Estudo do Desempenho Aerodinâmico de Scramjets usando Redes Neurais Profundas. In *XXVI Simpósio de Aplicações Operacionais em Áreas de Defesa*, São José dos Campos, Brasil, oct 2024. URL https://www.sige.ita.br/edicoes-anteriores/2024/st/243329{_}1.pdf.

Ângelo de Carvalho Paulino, Lamartine Nogueira Frutuoso Guimarães, and Elcio Hideiti Shiguemori. Hybrid Adaptive Computational Intelligence-based Multisensor Data Fusion applied to real-time UAV autonomous navigation. *Inteligencia Artificial*, 22(63):162–195, may 2019. ISSN 1988-3064. doi: 10.4114/intartif.vol22iss63pp162-195. URL https://journal.iberamia.org/index.php/intartif/article/view/237.

Ângelo de Carvalho Paulino, Lamartine Nogueira Frutuoso Guimaraes, and Elcio Hideiti Shiguemori. Assessment of Noise Impact on Hybrid Adaptive Computational Intelligence Multisensor Data Fusion Applied to Real-Time UAV Autonomous Navigation. *IEEE Latin America Transactions*, 18(02):295–302, feb 2020. ISSN 1548-0992. doi: 10.1109/TLA.2020.9085283. URL `https://ieeexplore.ieee.org/document/9085283/`.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Resa Septiani Pontoh, Muhammad Rivaldi Saiful Rafi, Chrysentia Clarissa Clorinda, Absalom Zakharia Ady Ena, Mo-hamad Naufal Farras, Restu Arisanti, Toni Toharudin, and Farhat Gumelar. Deep learning approaches to predict sea surface height above geoid in Pekalongan. *International Journal of Data and Network Science*, 8(2):743–752, 2024. ISSN 25618148. doi: 10.5267/j.ijdns.2024.1.004. URL `http://www.growingscience.com/ijds/Vol8/ijdns{_}2024{_}4.pdf`.

Jihuan Ren, Xiang Wu, Yi Liu, Feng Ni, Yuming Bo, and Changhui Jiang. Long-Term Trajectory Prediction of Hypersonic Glide Vehicle Based on Physics-Informed Transformer. *IEEE Transactions on Aerospace and Electronic Systems*, 59(6):9551–9561, dec 2023. ISSN 0018-9251. doi: 10.1109/TAES.2023.3322977. URL `https://ieeexplore.ieee.org/document/10286284/`.

Gennaro Russo, Claudio Voto, and Raffaele Savino. S4 - A demonstrator of HYPLANE, a single stage suborbital spaceplane and a hypersonic business jet. *Acta Astronautica*, 183:244–254, jun 2021. ISSN 00945765. doi: 10.1016/j.actaastro.2021.03.025. URL `https://linkinghub.elsevier.com/retrieve/pii/S0094576521001387`.

Wakim Boulos Saba. *Framework de otimização baseada em metaheurísticas*. PhD thesis, ITA, 2017.

Nicolas Santana, Nicolas Santana, Alexandre Caldeira, Angelo Passaro, and Lamartine Guimaraes. Optimization of a microreactor core's dimensions using metaheuristic methods. *Brazilian Journal of Radiation Sciences*, 10(3A (Suppl.)), oct 2022. ISSN 2319-0612. doi: 10.15392/2319-0612.2022.1903. URL `https://bjrs.org.br/revista/index.php/REVISTA/article/view/1903`.

Giuseppe Scarlatella, Jacopo Guadagnini, Leonardo Paniccià, Gabriele De Zaiacomo, Martin Tajmar, Christian Bach, and Dario Pastrone. Assessment of Mission Capabilities of a Reusable Heavy-lift Launch Vehicle Concept with Aerospike Engine. In *AIAA SCITECH 2024 Forum*, pp. Orlando, Reston, Virginia, jan 2024. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-711-5. doi: 10.2514/6.2024-2122. URL `https://arc.aiaa.org/doi/10.2514/6.2024-2122`.

Jürgen Schmidhuber. Deep Learning in neural networks: An overview. *Neural Networks*, 61:85–117, jan 2015. ISSN 18792782. doi: 10.1016/J.NEUNET.2014.09.003.

Neelima Sharma, V.K. Chawla, and N. Ram. Comparison of machine learning algorithms for the automatic programming of computer numerical control machine. *International Journal of Data and Network Science*, pp. 1–14, 2020. ISSN 25618148. doi: 10.5267/j.ijdns.2019.9.003. URL `http://www.growingscience.com/ijds/Vol4/ijdns{_}2019{_}32.pdf`.

Cunxiao Shen, Hanyuan Zhang, Songping Meng, and Chengdong Li. Augmented data driven self-attention deep learning method for imbalanced fault diagnosis of the HVAC chiller. *Engineering Applications of Artificial Intelligence*, 117:105540, jan 2023. ISSN 09521976. doi: 10.1016/j.engappai.2022.105540. URL `https://linkinghub.elsevier.com/retrieve/pii/S0952197622005309`.

Kenan Shen and Dongbiao Zhao. A Fault Diagnosis Method under Data Imbalance Based on Generative Adversarial Network and Long Short-Term Memory Algorithms for Aircraft Hydraulic System. *Aerospace*, 10(2):164, feb 2023. ISSN 2226-4310. doi: 10.3390/aerospace10020164. URL `https://www.mdpi.com/2226-4310/10/2/164`.

Jianlin Shi, Jinbo Wang, Linfeng Su, Zhenwei Ma, and Hongbo Chen. A Neural Network Warm-Started Indirect Trajectory Optimization Method. *Aerospace*, 9(8):435, aug 2022. ISSN 2226-4310. doi: 10.3390/aerospace9080435. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137191067{&}doi=10.3390{%}2Faerospace9080435{&}partnerID=40{&}md5=b6c56c59201c2068fd54c05d136b4097https://www.mdpi.com/2226-4310/9/8/435.

Yingxin Shou, Tian Yan, Bin Xu, and Fuchun Sun. Integrated guidance and control of hypersonic flight vehicle with coordinated mission requirement and input constraint. *International Journal of Robust and Nonlinear Control*, 2023. ISSN 10498923. doi: 10.1002/rnc.6607. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146981121{&}doi=10.1002{%}2Frnc.6607{&}partnerID=40{&}md5=b823e215221fefa6114fa502332c94cb.

Breno Silva, Ângelo de Carvalho Paulino, Israel da Silva Rêgo, Élcio J de Oliveira, and Angelo Passaro. Optimal preliminary design of hypersonic waverider using multiple metaheuristics. In *72th International Astronautical Congress*, Dubai, 2021. 72th International Astronautical Congress.

L. H. S. M. Soares, P. P. B. Araújo, R. B. C. Facchin, R. Carneiro, P. G. P. Toro, and A. Passaro. An Optimization Approach for Conceptual and Preliminary Design of Scramjet Engines. In *HiSST: 2nd International Conference on High-Speed Vehicle Science Technology*, volume 1, pp. 1–15, Brussels, 2022. Council of European Aerospace Societies (CEAS).

Kenneth Sörensen. Metaheuristics—the metaphor exposed. *International Transactions in Operational Research*, 22(1):3–18, jan 2015. ISSN 0969-6016. doi: 10.1111/itor.12001. URL http://doi.wiley.com/10.1111/itor.12001https://onlinelibrary.wiley.com/doi/10.1111/itor.12001.

Richard Speier, George Nacouzi, Carrie Lee, and Richard Moore. *Hypersonic Missile Nonproliferation: Hindering the Spread of a New Class of Weapons.* RAND Corporation, Santa Monica, CA, 2017. ISBN 9780833099167. doi: 10.7249/RR2137. URL https://www.rand.org/pubs/research{_}reports/RR2137.html.

Christian Staerk, Hannah Klinkhammer, Tobias Wistuba, Carlo Maj, and Andreas Mayr. Generalizability of polygenic prediction models: how is the R2 defined on test data? *BMC Medical Genomics*, 17(1):132, may 2024. ISSN 1755-8794. doi: 10.1186/s12920-024-01905-8. URL http://creativecommons.org/licenses/by/4.0/.TheCreativeCommonsPublicDomainDedicationwaiverhttp://creativecom-mons.org/publicdomain/zero/1.0/https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-024-01905-8.

El-Ghazali Talbi. *Metaheuristics: From Design to Implementation.* John Wiley & Sons, Inc., Hoboken, NJ, 1 edition, 2009.

Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, mar 2020. ISSN 00200255. doi: 10.1016/j.ins.2019.11.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0020025519310497.

Jianying Wang, Yuanpei Wu, Ming Liu, Ming Yang, and Haizhao Liang. A Real-Time Trajectory Optimization Method for Hypersonic Vehicles Based on a Deep Neural Network. *Aerospace*, 9(4), 2022. ISSN 22264310. doi: 10.3390/aerospace9040188. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85128224111{&}doi=10.3390{%}2Faerospace9040188{&}partnerID=40{&}md5=bc0d79831e4ab8e9d842bdcaa5bbb720.

Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, volume 2016, pp. 4368–4374. IEEE, jul 2016. ISBN 978-1-5090-0620-5. doi: 10.1109/IJCNN.2016.7727770. URL http://ieeexplore.ieee.org/document/7727770/.

Weicheng Wang and Jinye Ma. A review: Applications of machine learning and deep learning in aerospace engineering and aero-engine engineering. *Advances in Engineering Innovation*, 6(1):54–72, feb 2024. ISSN 2977-3903. doi: 10.54254/2977-3903/6/2024060. URL https://aei.ewapublishing.org/article/721acfa0a6cd40f5be50b413afdb9e80.

Vitor Werner de Vargas, Jorge Arthur Schneider Aranda, Ricardo dos Santos Costa, Paulo Ricardo da Silva Pereira, and Jorge Luis Victória Barbosa. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, 65(1):31–57, jan 2023. ISSN 0219-1377. doi: 10.1007/s10115-022-01772-8. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC9645765/https://link.springer.com/10.1007/s10115-022-01772-8`.

Da Zhang, Hao Yang, Junyu Gao, and Xuelong Li. Imbalanced Flight Test Sensor Temporal Data Anomaly Detection. *IEEE Transactions on Aerospace and Electronic Systems*, PP(1):1–11, 2024. ISSN 0018-9251. doi: 10.1109/TAES.2024.3471488. URL `https://ieeexplore.ieee.org/document/10707114/`.

Zehui Zhao, Laith Alzubaidi, Jinglan Zhang, Ye Duan, and Yuantong Gu. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, 242(December 2023):122807, may 2024a. ISSN 09574174. doi: 10.1016/j.eswa.2023.122807. URL `https://doi.org/10.1016/j.eswa.2023.122807https://linkinghub.elsevier.com/retrieve/pii/S0957417423033092`.

Zida Zhao, Shilong Sun, Dong Wang, and Wenfu Xu. MCVAE-GAN: A Novel Diagnosis Method for Solving Harmonic Reducer Fault Signal Scarcity and Sample Diversity Problem. *IEEE Transactions on Instrumentation and Measurement*, 73:1–10, 2024b. ISSN 0018-9456. doi: 10.1109/TIM.2024.3472781. URL `https://ieeexplore.ieee.org/document/10706936/`.

Ai, Shaojie, Song, Jia, & Cai, Guobiao. (2021). A real-time fault diagnosis method for hypersonic air vehicle with sensor fault based on the auto temporal convolutional network. Aerospace Science and Technology, 119, 107220. https://doi.org/10.1016/j.ast.2021.107220

Akrim, Anass, Gogu, Christian, Vingerhoeds, Rob, & Salaün, Michel. (2023). Self-Supervised Learning for data scarcity in a fatigue damage prognostic problem. Engineering Applications of Artificial Intelligence, 120(January), 105837. https://doi.org/10.1016/j.engappai.2023.105837

Almaiah, Mohammed Amin, Alrawashdeh, Rana, Alkhdour, Tayseer, Al-Ali, Romel, Rjoub, Gaith, & Aldahyani, Theyazan. (2024). Detecting DDoS attacks using machine learning algorithms and feature selection methods. International Journal of Data and Network Science, 8(4), 2307–2318. https://doi.org/10.5267/j.ijdns.2024.6.001

Alzubaidi, Laith, Bai, Jinshuai, Al-Sabaawi, Aiman, Santamaría, Jose, Albahri, A. S., Al-dabbagh, Bashar Sami Nayyef, Fadhel, Mohammed A., Manoufali, Mohamed, Zhang, Jinglan, Al-Timemy, Ali H., Duan, Ye, Abdullah, Amjed, Farhan, Laith, Lu, Yi, Gupta, Ashish, Albu, Felix, Abbosh, Amin, & Gu, Yuantong. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. Journal of Big Data, 10(1), 46. https://doi.org/10.1186/s40537-023-00727-2

Anderson, J. D. (2003). Modern Compressible Flow: With Historical Perspective. McGraw-Hill Education. https://books.google.com.br/books?id=woeqa4-a5EgC

Araújo, Pedro P. B., Tanaka, Roberto Y., Silva, Carlos A., Passaro, Angelo, & Toro, Paulo G. P. (2024). Multi-objective optimization of a hypersonic airbreathing vehicle. Physics of Fluids, 36(2). https://doi.org/10.1063/5.0181366

Beachy, Atticus, Bae, Harok, Boyd, Ian, & Grandhi, Ramana. (2021). Emulator embedded neural networks for multi-fidelity conceptual design exploration of hypersonic vehicles. AIAA Scitech 2021 Forum, 1 PartF, 1 – 24. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100303247&partnerID=40&md5=9adfa4368237d8078c935638cc25eec9

Bergstra, James, Ca, James Bergstra@umontreal, & Ca, Yoshua Bengio@umontreal. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. Journal of Machine Learning Research, 13, 281–305. http://scikit-learn.sourceforge.net.

Boughorbel, Sabri, Jarray, Fethi, & El-Anbari, Mohammed. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE, 12(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Brar, Amrinder Singh, & Singh, Kawaljeet. (2024). A multi-objective stacked regression method for distance based colour measuring device. Scientific Reports, 14(1), 5530. https://doi.org/10.1038/s41598-024-54785-4

Brunton, Steven L., Nathan Kutz, J., Manohar, Krithika, Aravkin, Aleksandr Y., Morgansen, Kristi, Klemisch, Jennifer, Goebel, Nicholas, Buttrick, James, Poskin, Jeffrey, Blom-Schieber, Adriana W., Hogan, Thomas, & McDonald, Darren. (2021). Data-Driven Aerospace Engineering: Reframing the Industry with Machine Learning. AIAA Journal, 59(8), 1–26. https://doi.org/10.2514/1.J060131

Bu, Xiangwei, Hua, Changchun, Lv, Maolong, & Wu, Zhonghua. (2023). Flight Control of Waverider Vehicles with Fragility-avoidance Prescribed Performance. IEEE Transactions on Aerospace and Electronic Systems, 1–15. https://doi.org/10.1109/TAES.2023.3251314

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3). https://doi.org/10.5194/gmd-7-1247-2014

Chen, Wuxing, Yang, Kaixiang, Yu, Zhiwen, Shi, Yifan, & Chen, C. L. Philip. (2024). A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review, 57(6), 137. https://doi.org/10.1007/s10462-024-10759-6

Cui, Luyao, Liu, Aijun, Yv, Changjun, & Quan, Taifan. (2018). Learning Algorithm for Tracking Hypersonic Targets in Near Space. In Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST: Vol. 227 LNICST (pp. 206–214). https://doi.org/10.1007/978-3-319-73447-7_24

David Dangut, Maren, Skaf, Zakwan, & Jennions, Ian. (2020). Rescaled-LSTM for Predicting Aircraft Component Replacement Under Imbalanced Dataset Constraint. 2020 Advances in Science and Engineering Technology International Conferences (ASET), 1–9. https://doi.org/10.1109/ASET48392.2020.9118253

De Lima Filho, Geraldo Mulato, Kuroswiski, Andre Rossi, Medeiros, Felipe Leonardo Lobo, Voskuijl, Mark, Monsuur, Herman, & Passaro, Angelo. (2022). Optimization of Unmanned Air Vehicle Tactical Formation in War Games. IEEE Access, 10, 21727–21741. https://doi.org/10.1109/ACCESS.2022.3152768

Doering, Jana, Kizys, Renatas, Juan, Angel A., Fitó, Àngels, & Polat, Onur. (2019). Metaheuristics for rich portfolio optimisation and risk management: Current state and future trends. In Operations Research Perspectives (Vol. 6). Elsevier Ltd. https://doi.org/10.1016/j.orp.2019.100121

Doğan, Berat, & Ölmez, Tamer. (2015). A new metaheuristic for numerical function optimization: Vortex Search algorithm. Information Sciences, 293(February), 125–145. https://doi.org/10.1016/j.ins.2014.08.053

Dong, Yutong, Jiang, Hongkai, Liu, Yunpeng, & Yi, Zichun. (2024). Global wavelet-integrated residual frequency attention regularized network for hypersonic flight vehicle fault diagnosis with imbalanced data. Engineering Applications of Artificial Intelligence, 132(February), 107968. https://doi.org/10.1016/j.engappai.2024.107968

Drake, John H., Kheiri, Ahmed, Özcan, Ender, & Burke, Edmund K. (2020). Recent advances in selection hyper-heuristics. In European Journal of Operational Research (Vol. 285, Issue 2, pp. 405–428). Elsevier B.V. https://doi.org/10.1016/j.ejor.2019.07.073

Du, Ken-Li., & Swamy, M. N. S. (2016). Search and Optimization by Metaheuristics Techniques and Algorithms Inspired by Nature (1st ed.). Birkhäuser.

Fujio, Chihiro, & Ogawa, Hideaki. (2022). Deep-learning prediction and uncertainty quantification for scramjet intake flowfields. Aerospace Science and Technology, 130, 107931. https://doi.org/10.1016/j.ast.2022.107931

Ghosh, Kushankur, Bellinger, Colin, Corizzo, Roberto, Branco, Paula, Krawczyk, Bartosz, & Japkowicz, Nathalie. (2024). The class imbalance problem in deep learning. Machine Learning, 113(7), 4845–4901. https://doi.org/10.1007/s10994-022-06268-8

Gond, Arjun, & Sengupta, Aditi. (2025). Data-driven optimal design of premixed hydrogen-air combustion in a microcombustor. AI Thermal Fluids, 1(March), 100007. https://doi.org/10.1016/j.aitf.2025.100007

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/

Harris, Charles R., Millman, K. Jarrod, van der Walt, Stéfan J., Gommers, Ralf, Virtanen, Pauli, Cournapeau, David, Wieser, Eric, Taylor, Julian, Berg, Sebastian, Smith, Nathaniel J., Kern, Robert, Picus, Matti, Hoyer, Stephan, van Kerkwijk, Marten H., Brett, Matthew, Haldane, Allan, del Río, Jaime Fernández, Wiebe, Mark, Peterson, Pearu, … Oliphant, Travis E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Haykin, Simon S. (2009). Neural networks and learning machines (Third). Pearson Education.

He, Haibo, & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hellín, Carlos J., Olmedo, Alvaro A., Valledor, Adrián, Gómez, Josefa, López-Benítez, Miguel, & Tayebi, Abdelhamid. (2024). Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification. Applied Sciences, 14(8), 3419. https://doi.org/10.3390/app14083419

Houssein, Essam H., Saeed, Mahmoud Khalaf, Hu, Gang, & Al-Sayed, Mustafa M. (2024). Metaheuristics for Solving Global and Engineering Optimization Problems: Review, Applications, Open Issues and Challenges. Archives of Computational Methods in Engineering, 31(8), 4485–4519. https://doi.org/10.1007/s11831-024-10168-6

Hunter, John D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Inc., Plotly Technologies. (2015). Collaborative data science. Plotly Technologies Inc. https://plot.ly

Ispir, Ali Can, Saracoglu, Bayindir H., Magin, Thierry, & Coussement, Axel. (2023). A methodology for estimating hypersonic engine performance by coupling supersonic reactive flow simulations with machine learning techniques. Aerospace Science and Technology, 140, 108501. https://doi.org/10.1016/j.ast.2023.108501

Johnson, Justin M., & Khoshgoftaar, Taghi M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kaur, Harsurinder, Pannu, Husanbir Singh, & Malhi, Avleen Kaur. (2020). A Systematic Review on Imbalanced Data Challenges in Machine Learning. ACM Computing Surveys, 52(4), 1–36. https://doi.org/10.1145/3343440

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Khan, Salman H., Hayat, Munawar, Bennamoun, Mohammed, Sohel, Ferdous A., & Togneri, Roberto. (2018). Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482

Korkegi, R. H. (1975). Comparison of Shock-Induced Two- and Three- Dimensional Incipient Turbulent Seperation. AIAA Journal, 13(4), 534–535. https://doi.org/10.2514/3.49750

LeCun, Yann, Bengio, Yoshua, & Hinton, Geoffrey. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Ling, Julia, Kurzawski, Andrew, & Templeton, Jeremy. (2016). Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. Journal of Fluid Mechanics, 807, 155–166. https://doi.org/10.1017/jfm.2016.615

Liu, Xiaowen, & Tian, Huixin. (2024). Research on Imbalanced Data Regression Based on Confrontation. Processes, 12(2), 375. https://doi.org/10.3390/pr12020375

Lv, Jixing, Wang, Changhong, & Kao, Yonggui. (2023). Adaptive fixed-time quantized fault-tolerant attitude control for hypersonic reentry vehicle. Neurocomputing, 520, 386–399. https://doi.org/10.1016/j.neucom.2022.11.057

Mao, Z., Lu, L., Marxen, O., Zaki, T. A., & Karniadakis, G. E. (2021). DeepM&Mnet for hypersonics: Predicting the coupled flow and finite-rate chemistry behind a normal shock using neural-network approximation of operators. Journal of Computational Physics, 447. https://doi.org/10.1016/j.jcp.2021.110698

McCall, Thomas, Seyed Alavi, Kiarash, Rana, Loveneesh, & Chudoba, Bernd. (2018, September 17). Artificial Intelligent Research Assistant for Aerospace Design Synthesis—Solution Logic. 22nd AIAA International Space Planes and Hypersonics Systems and Technologies Conference. https://doi.org/10.2514/6.2018-5387

McCulloch, Warren, & Pitts, Walter. (2021). A Logical Calculus of the Ideas Immanent in Nervous Activity (1943). In Ideas That Created the Future (Vol. 5, Issue 4, pp. 79–88). The MIT Press. https://doi.org/10.7551/mitpress/12274.003.0011

McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Mooijman, Paul, Catal, Cagatay, Tekinerdogan, Bedir, Lommen, Arjen, & Blokland, Marco. (2023). The effects of data balancing approaches: A case study. Applied Soft Computing, 132, 109853. https://doi.org/10.1016/j.asoc.2022.109853

Oliveira, Paulo Moura, Solteiro Pires, E. J., Boaventura-Cunha, José, & Pinho, Tatiana Martins. (2020). Review of nature and biologically inspired metaheuristics for greenhouse environment control. In Transactions of the Institute of Measurement and Control (Vol. 42, Issue 12, pp. 2338–2358). SAGE Publications Ltd. https://doi.org/10.1177/0142331220909010

Ozbenli, E., Vedula, P., Vogiatzis, K., & Josyula, E. (2020). Numerical solution of hypersonic flows via artificial neural networks. AIAA Scitech 2020 Forum, 1 PartF. https://doi.org/10.2514/6.2020-1233

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, Desmaison, Alban, Kopf, Andreas, Yang, Edward, DeVito, Zachary, Raison, Martin, Tejani, Alykhan, Chilamkurthy, Sasank, Steiner, Benoit, Fang, Lu, … Chintala, Soumith. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Paulino, Ângelo de Carvalho, Guimaraes, Lamartine Nogueira Frutuoso, & Shiguemori, Elcio Hideiti. (2020). Assessment of Noise Impact on Hybrid Adaptive Computational Intelligence Multisensor Data Fusion Applied to Real-Time UAV Autonomous Navigation. IEEE Latin America Transactions, 18(02), 295–302. https://doi.org/10.1109/TLA.2020.9085283

Paulino, Ângelo de Carvalho, Guimarães, Lamartine Nogueira Frutuoso, & Shiguemori, Elcio Hideiti. (2019). Hybrid Adaptive Computational Intelligence-based Multisensor Data Fusion applied to real-time UAV autonomous navigation. Inteligencia Artificial, 22(63), 162–195. https://doi.org/10.4114/intartif.vol22iss63pp162-195

Paulino, Ângelo de Carvalho, & Passaro, Angelo. (2023). Investigação cientométrica sobre aplicações de Redes Neurais em Hipersônica. XXV Simpósio de Aplicações Operacionais Em Áreas de Defesa, 1.

Paulino, Ângelo de Carvalho, & Passaro, Angelo. (2024, October 3). Estudo do Desempenho Aerodinâmico de Scramjets usando Redes Neurais Profundas. XXVI Simpósio de Aplicações Operacionais Em Áreas de Defesa. https://www.sige.ita.br/edicoes-anteriores/2024/st/243329_1.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Pontoh, Resa Septiani, Rafi, Muhammad Rivaldi Saiful, Clorinda, Chrysentia Clarissa, Ena, Absalom Zakharia Ady, Farras, Mo-hamad Naufal, Arisanti, Restu, Toharudin, Toni, & Gumelar, Farhat. (2024). Deep learning approaches to predict sea surface height above geoid in Pekalongan. International Journal of Data and Network Science, 8(2), 743–752. https://doi.org/10.5267/j.ijdns.2024.1.004

Ren, Jihuan, Wu, Xiang, Liu, Yi, Ni, Feng, Bo, Yuming, & Jiang, Changhui. (2023). Long-Term Trajectory Prediction of Hypersonic Glide Vehicle Based on Physics-Informed Transformer. IEEE Transactions on Aerospace and Electronic Systems, 59(6), 9551–9561. https://doi.org/10.1109/TAES.2023.3322977

Russo, Gennaro, Voto, Claudio, & Savino, Raffaele. (2021). S4 - A demonstrator of HYPLANE, a single stage suborbital spaceplane and a hypersonic business jet. Acta Astronautica, 183, 244–254. https://doi.org/10.1016/j.actaastro.2021.03.025

Saba, Wakim Boulos. (2017). Framework de otimização baseada em metaheurísticas. ITA.

Santana, Nicolas, Santana, Nicolas, Caldeira, Alexandre, Passaro, Angelo, & Guimaraes, Lamartine. (2022). Optimization of a microreactor core's dimensions using metaheuristic methods. Brazilian Journal of Radiation Sciences, 10(3A (Suppl.)). https://doi.org/10.15392/2319-0612.2022.1903

Scarlatella, Giuseppe, Guadagnini, Jacopo, Paniccià, Leonardo, De Zaiacomo, Gabriele, Tajmar, Martin, Bach, Christian, & Pastrone, Dario. (2024). Assessment of Mission Capabilities of a Reusable Heavy-lift Launch Vehicle Concept with Aerospike Engine. AIAA SCITECH 2024 Forum, Orlando. https://doi.org/10.2514/6.2024-2122

Schmidhuber, Jürgen. (2015). Deep Learning in neural networks: An overview. Neural Networks, 61, 85–117. https://doi.org/10.1016/J.NEUNET.2014.09.003

Sharma, Neelima, Chawla, V. K., & Ram, N. (2020). Comparison of machine learning algorithms for the automatic programming of computer numerical control machine. International Journal of Data and Network Science, 1–14. https://doi.org/10.5267/j.ijdns.2019.9.003

Shen, Cunxiao, Zhang, Hanyuan, Meng, Songping, & Li, Chengdong. (2023). Augmented data driven self-attention deep learning method for imbalanced fault diagnosis of the HVAC chiller. Engineering Applications of Artificial Intelligence, 117, 105540. https://doi.org/10.1016/j.engappai.2022.105540

Shen, Kenan, & Zhao, Dongbiao. (2023). A Fault Diagnosis Method under Data Imbalance Based on Generative Adversarial Network and Long Short-Term Memory Algorithms for Aircraft Hydraulic System. Aerospace, 10(2), 164. https://doi.org/10.3390/aerospace10020164

Shi, Jianlin, Wang, Jinbo, Su, Linfeng, Ma, Zhenwei, & Chen, Hongbo. (2022). A Neural Network Warm-Started Indirect Trajectory Optimization Method. Aerospace, 9(8), 435. https://doi.org/10.3390/aerospace9080435

Shou, Yingxin, Yan, Tian, Xu, Bin, & Sun, Fuchun. (2023). Integrated guidance and control of hypersonic flight vehicle with coordinated mission requirement and input constraint. International Journal of Robust and Nonlinear Control. https://doi.org/10.1002/rnc.6607

Silva, Breno, Paulino, Ângelo de Carvalho, Rêgo, Israel da Silva, Oliveira, Élcio J. de, & Passaro, Angelo. (2021). Optimal preliminary design of hypersonic waverider using multiple metaheuristics. 72th International Astronautical Congress.

Soares, L. H. S. M., Araújo, P. P. B., Facchin, R. B. C., Carneiro, R., Toro, P. G. P., & Passaro, A. (2022). An Optimization Approach for Conceptual and Preliminary Design of Scramjet Engines. HiSST: 2nd International Conference on High-Speed Vehicle Science Technology, 1, 1–15.

Sörensen, Kenneth. (2015). Metaheuristics—the metaphor exposed. International Transactions in Operational Research, 22(1), 3–18. https://doi.org/10.1111/itor.12001

Speier, Richard, Nacouzi, George, Lee, Carrie, & Moore, Richard. (2017). Hypersonic Missile Nonproliferation: Hindering the Spread of a New Class of Weapons. RAND Corporation. https://doi.org/10.7249/RR2137

Staerk, Christian, Klinkhammer, Hannah, Wistuba, Tobias, Maj, Carlo, & Mayr, Andreas. (2024). Generalizability of polygenic prediction models: how is the R2 defined on test data? BMC Medical Genomics, 17(1), 132. https://doi.org/10.1186/s12920-024-01905-8

Talbi, El-Ghazali. (2009). Metaheuristics: From Design to Implementation. In Metaheuristics: From Design to Implementation (1st ed.). John Wiley & Sons, Inc.

Thabtah, Fadi, Hammoud, Suhel, Kamalov, Firuz, & Gonsalves, Amanda. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Wang, Jianying, Wu, Yuanpei, Liu, Ming, Yang, Ming, & Liang, Haizhao. (2022). A Real-Time Trajectory Optimization Method for Hypersonic Vehicles Based on a Deep Neural Network. Aerospace, 9(4). https://doi.org/10.3390/aerospace9040188

Wang, Shoujin, Liu, Wei, Wu, Jia, Cao, Longbing, Meng, Qinxue, & Kennedy, Paul J. (2016). Training deep neural networks on imbalanced data sets. 2016 International Joint Conference on Neural Networks (IJCNN), 2016, 4368–4374. https://doi.org/10.1109/IJCNN.2016.7727770

Wang, Weicheng, & Ma, Jinye. (2024). A review: Applications of machine learning and deep learning in aerospace engineering and aero-engine engineering. Advances in Engineering Innovation, 6(1), 54–72. https://doi.org/10.54254/2977-3903/6/2024060

Werner de Vargas, Vitor, Schneider Aranda, Jorge Arthur, dos Santos Costa, Ricardo, da Silva Pereira, Paulo Ricardo, & Victória Barbosa, Jorge Luis. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. Knowledge and Information Systems, 65(1), 31–57. https://doi.org/10.1007/s10115-022-01772-8

Zhang, Da, Yang, Hao, Gao, Junyu, & Li, Xuelong. (2024). Imbalanced Flight Test Sensor Temporal Data Anomaly Detection. IEEE Transactions on Aerospace and Electronic Systems, PP(1), 1–11. https://doi.org/10.1109/TAES.2024.3471488

Zhao, Zehui, Alzubaidi, Laith, Zhang, Jinglan, Duan, Ye, & Gu, Yuantong. (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Systems with Applications, 242(December 2023), 122807. https://doi.org/10.1016/j.eswa.2023.122807

Zhao, Zida, Sun, Shilong, Wang, Dong, & Xu, Wenfu. (2024). MCVAE-GAN: A Novel Diagnosis Method for Solving Harmonic Reducer Fault Signal Scarcity and Sample Diversity Problem. IEEE Transactions on Instrumentation and Measurement, 73, 1–10. https://doi.org/10.1109/TIM.2024.3472781