Drone Telemetry Anomaly Detection

Indian Institute of Information Technology, Nagpur

Ritesh Singh (BT23CSD003)
Arjit Tiwari (BT23CSD008)
Ashutosh Singh (BT23CSD010)
Akshay Naroliya (BT23CSD028)

November 13, 2025

Abstract

The safe operation of Unmanned Aerial Vehicles (UAVs) depends on the high-integrity, real-time analysis of sensor telemetry. This paper presents a comprehensive study to develop a high-performance multi-class classifier for drone telemetry data, with a special focus on the pervasive challenge of class imbalance. The dataset, consisting of 12,253 samples across 5 classes (one 'Normal' and four distinct attack types), is highly imbalanced (Class 0 $\approx$ 53.6%), which presents a significant risk of models failing to detect rare but critical anomalies. We implement and validate a robust preprocessing pipeline including StandardScaler, Principal Component Analysis (PCA), and the Synthetic Minority Over-sampling TEchnique (SMOTE) to synthetically balance the training set. We then conduct a comparative analysis of three powerful models: Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP) specifically paired with a Focal Loss function. We report the detailed experimental setup, evaluation metrics (Accuracy, Precision, Recall), and key findings from a granular confusion matrix analysis. The MLP with Focal Loss achieved the highest overall accuracy (98.86%), demonstrating superior performance in handling imbalanced data and decisively improving minority-class predictions over the state-of-the-art ensemble methods. This model is recommended for production deployment in real-time drone telemetry applications where high reliability and low false-negative rates are paramount.

1 Introduction

1.1 Problem Statement

Unmanned Aerial Vehicles (UAVs), or drones, have seen a massive proliferation in critical sectors, including logistics, infrastructure inspection, agriculture, and defense. Their autonomy and safe operation are entirely dependent on a continuous stream of high-fidelity sensor data, known as telemetry. This data, originating from Inertial Measurement Units (IMUs), gyroscopes, accelerometers, and GPS, forms the basis of the vehicle's ”world model” and control loop.

A failure or malicious attack on these sensors can lead to catastrophic outcomes. Detecting anomalies within this telemetry stream is therefore a mission-critical task. This task is hindered by a fundamental and pervasive challenge: **severe class imbalance**. In any real-world flight log, 99.9% of the data corresponds to normal, benign operation. Critical fault states or attack signatures (e.g., a ”stuck” sensor, a ”replay” attack) are, by nature, extremely rare.

When a standard machine learning model is trained on such imbalanced data, it quickly learns a ”lazy” and dangerous policy: to achieve high accuracy, it simply predicts ”Normal” for almost every input. This results in a model with high *overall accuracy* but near-zero *recall* for the rare, and most important, anomaly classes. Our objective is to develop a high-performance multi-class classifier for drone telemetry data that specifically corrects for this imbalance and proves its reliability in detecting minority-class faults.

1.2 Solution to the Problem

We hypothesize that a robust solution requires tackling the imbalance problem at both the data and model levels. We propose a two-pronged approach:

1. **Data-Level: SMOTE** We apply the Synthetic Minority Over-sampling TEchnique (SMOTE) *exclusively* to the training dataset. Unlike simple over-sampling (duplicating samples), SMOTE creates new, synthetic data points in the feature space by interpolating between existing minority-class neighbors. This provides the model with a richer, more diverse, and balanced set of examples, preventing it from being biased towards the majority class.
2. **Model-Level: Focal Loss** We further hypothesize that a specialized loss function, the Focal Loss, will outperform standard cross-entropy. SMOTE can introduce noise, and the majority class is still ”easier” to learn. Focal Loss is designed to dynamically down-weight the loss attributed to well-classified (easy) examples, forcing the model to focus its limited learning capacity on harder-to-classify (minority) examples.

We validate this hybrid solution by comparing our proposed MLP with Focal Loss against two extremely strong ensemble baselines (Random Forest, XGBoost) and other published results from the literature.

1.3 Research Gap and Motivation

While many works apply standard machine learning models to telemetry data, they often fail to adequately address the operational reality of severe class imbalance. Many studies propose novel deep learning architectures (e.g., LSTMs, Autoencoders) but evaluate them on balanced datasets or, more critically, report only *overall accuracy*. This metric obscures poor minority-class performance and leads to models that are unusable in a real-world, safety-critical setting.

This work is motivated by the *operational need* for a practical and robust solution that specifically targets and *measurably improves* minority-class detection. We aim to provide a clear, validated recommendation for a deployable, high-reliability model that practitioners can trust.

1.4 Contributions

The main contributions of this paper are:

1. A validated and robust preprocessing pipeline combining StandardScaler, PCA, and SMOTE, specifically designed to prepare imbalanced telemetry data for classification.
2. A rigorous comparative analysis of Random Forest, XGBoost, and a proposed MLP with Focal Loss, providing a clear performance benchmark on this task.
3. A key finding that a specialized loss function (Focal Loss), when combined with data-level sampling (SMOTE), achieves superior accuracy (98.86%) and demonstrably better minority-class handling than state-of-the-art ensemble methods.
4. A clear, data-driven recommendation for a production-ready model suitable for real-time drone telemetry anomaly detection.

2 Literature Review

We review representative works across three complementary directions.

2.1 Telemetry Anomaly Detection using DNNs

Deep Neural Networks (DNNs) have become a standard for complex pattern recognition. For telemetry, which is inherently time-series data, Recurrent Neural Networks (RNNs) like LSTMs and GRUs are common. Minn et al. (DronLomaly) proposed a Bi-LSTM predictor that detects anomalies by flagging high prediction errors, achieving a high F1-score (0.967) with low latency. Other works use autoencoders (VAE/LSTM-VAE) and attention-based LSTMs for windowed anomaly detection. These models excel at capturing temporal dependencies (e.g., a ”drift” attack) but are computationally expensive and can be difficult to interpret. Our work focuses on static (non-temporal) classification as a robust baseline.

2.2 Classical and Statistical ML Methods

Prior to deep learning, classical methods were widely used. The One-Class SVM (OC-SVM) is a popular unsupervised method for ”novelty detection,” as it learns a boundary around only the ”normal” data. However, it struggles with complex, multi-modal data and cannot, by itself, differentiate between *different types* of anomalies. K-Nearest Neighbors (KNN) is an instance-based learner that has been used as a simple, interpretable baseline. Its

”guilt by association” logic is intuitive, but it suffers from the curse of dimensionality and is computationally slow at inference time, making it less suitable for real-time streams. Our work compares against these methods to establish a clear performance gain.

2.3 Lightweight Models, Knowledge Distillation and Pruning

A critical consideration for UAVs is *onboard* deployment, which requires models to run on resource-constrained microcontrollers. Medhi et al. (UAV-DiPNID) and similar research show that techniques like Knowledge Distillation (KD)—where a small ”student” model is trained to mimic a larger, high-performance ”teacher” model—combined with network pruning can reduce model size by 80–90% while retaining high accuracy. While our work does not focus on compression, these techniques represent a clear and necessary path for future optimization of our proposed MLP model for embedded deployment.

3 Materials and Methods

3.1 Objective

Develop a high-performance multi-class classifier for drone telemetry data with a special focus on correcting class imbalance and improving reliability of minority-class predictions.

3.2 Dataset Overview and Preprocessing Pipeline

Dataset Overview The dataset is a high-fidelity log of drone telemetry, consisting of sensor readings and flight controller state.

- **Total samples:** 12,253
- **Number of classes:** 5 (imbalanced; Class 0 $\approx$ 53.6%)
- **Data consists of:** Sensor readings including gyro (x, y, z), accelerometer (x, y, z), magnetometer, and orientation errors (pitch, roll, yaw).
- **Note:** The timestamp was deliberately removed to prevent temporal data leakage. This forces the models to learn from the *intrinsic patterns and correlations* of the sensor readings themselves, rather than ”cheating” by memorizing a temporal sequence. This makes our model a more robust static classifier.

Class Definitions The five classes represent operationally critical scenarios. Class 0 is benign, while classes 1-4 represent distinct attack scenarios that can compromise vehicle stability and control.

Label	Anomaly Type	Description
0	Normal	Benign flight data with no attack or failure.
1	Stuck Attack	The sensor’s output “freezes” and reports the same constant value. Simulates a dead sensor.
2	Offset Attack	A malicious, constant value is added to the sensor’s real reading. Simulates a bias or drift.
3	Scaling Attack	The sensor’s real reading is multiplied by a malicious factor. Simulates incorrect calibration.
4	Replay Attack	Old, previously recorded sensor data is fed to the autopilot as “live” data. This is an insidious attack that can mask a real, in-progress failure.

Preprocessing Pipeline A multi-step pipeline was developed to prepare the raw data for modeling. The pipeline was fit *only* on the training data to prevent data leakage.

- **StandardScaler:** Applied to all features to center them (mean=0) and scale them to unit variance. This is *essential* for PCA to function correctly and for the MLP’s gradient descent to converge stably and quickly.
- **PCA:** Principal Component Analysis was applied to reduce dimensionality. We retained components preserving 95% of the variance, which reduced the feature space to 17 components. This combats the curse of dimensionality, reduces model training time, and filters noise.
- **Train-Test Split:** A stratified 75/25 train-test split was used. *Stratification* is non-negotiable for imbalanced data, as it ensures that the rare minority classes are present in both the training and test sets in their original, real-world proportions.

- **SMOTE:** Applied *only* on the training set (X_{train}, y_{train}) *after* the split. This is the most critical step to prevent data leakage. By over-sampling only the training data, we force the model to learn a balanced representation while ensuring the test set remains a true, unseen, and imbalanced reflection of the real-world problem.

3.3 Model Architectures

We conduct an internal comparison of three models to find the best-performing architecture for our specific dataset.

- **Random Forest:** Selected as a powerful and robust ensemble baseline. It is an ensemble of 400 decision trees, trained via bagging. Its averaging mechanism makes it highly resistant to overfitting and effective on high-dimensional data.
- **XGBoost:** Represents the state-of-the-art in gradient-boosted decision trees. Unlike Random Forest (which builds independent trees), XGBoost builds trees *sequentially*, where each new tree is trained to correct the errors of the previous ones. It is often the top-performer on tabular data. We used 500 estimators.
- **MLP + Focal Loss (Proposed):** Our proposed solution. This is a standard feed-forward neural network. The architecture (17-dim Input → 256-neuron Hidden → 128-neuron Hidden → 5-class Output) was selected via preliminary tuning. The key innovation is the replacement of the standard loss function with Focal Loss to specifically target the imbalance problem.
  - **Input Layer:** 17 features (from PCA)
  - **Hidden Layers:** 256 → 128 neurons (with ReLU activations)
  - **Output Layer:** 5 classes (with Softmax)
  - **Optimizer:** Adam (LR=0.001)
  - **Loss:** Focal Loss ($\gamma = 2$, per-class alpha weighting)

For additional context, a separate comparative evaluation was conducted on four different Deep Neural Network (DNN) architectures to identify an optimal detector. This study provides insight into alternative, more complex model-building strategies:

- **Model 1: Core (DNN Teacher):** A baseline, data-driven DNN with 3461 parameters, achieving a 0.957 F1-Score.
- **Model 2: Hybrid (DNN + Rule):** The Core DNN with a rigid, physics-based, post-processing rule. This model degraded performance (0.944 F1-Score), indicating that simplistic rule integration can introduce classification noise.
- **Model 3: Lightweight (Student):** A highly efficient model (293 parameters) optimized via techniques like Knowledge Distillation and Pruning. It achieved a 91.5% size reduction at the cost of performance (0.928 F1-Score).
- **Model 4: Stacking Hybrid (DNN + Physics):** This model fused the Core DNN's outputs with the physics rule using a secondary meta-model (3516 parameters). It achieved the highest performance (0.961 F1-Score).

The conclusion of that study was that the Stacking Hybrid (Model 4) is ideal for high-power Ground Control Stations (GCS), while the Lightweight (Model 3) is the only viable candidate for low-power, onboard execution. Our paper, in contrast, focuses on a direct comparison of strong baselines (RF, XGBoost) against our proposed MLP, prioritizing the solution to class imbalance via SMOTE and Focal Loss.

3.4 Mathematical Formulation (Loss Functions)

Our proposed model's performance hinges on replacing the standard Categorical Cross-Entropy (CCE) loss with Focal Loss.

Standard Cross-Entropy Loss The standard CCE loss is:

$$L_{CE}(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where C is the number of classes, y_i is the true label (e.g., [0, 1, 0, 0, 0]), and $\hat{y}_i$ is the model's predicted probability for that class. This loss function treats all classes and all samples equally. In an imbalanced dataset, the total loss is *dominated* by the thousands of "easy" majority-class samples, even if their individual loss is small. The model's "incentive" is to perfect its prediction for the majority class, while ignoring the few, high-loss minority samples.

Focal Loss Focal Loss (FL) directly attacks this problem by reshaping the loss function to down-weight easy examples.

$$L_{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$$

where:

- p_t is the model's estimated probability for the ground-truth class.
- γ (gamma) is the *focusing parameter* (we use $\gamma = 2$). This is the key. If an example is *easy* ($p_t \rightarrow 1$, e.g., $p_t = 0.99$), the $(1 - p_t)^\gamma$ term becomes $(0.01)^2 = 0.0001$. This *multiplicatively decimates* the loss for that easy sample.
- If an example is *hard* ($p_t \rightarrow 0$, e.g., $p_t = 0.1$), the term becomes $(0.9)^2 = 0.81$, which *preserves* most of the loss.
- α_t (alpha) is a per-class weighting factor to further, explicitly balance class importance.

This mechanism forces the model to stop wasting learning on easy majority samples and focus its gradient updates on the hard-to-classify minority samples.

4 Experiments and Results

4.1 Internal Model Comparison

The three models (Random Forest, XGBoost, MLP) were trained on the preprocessed and SMOTE-balanced training set. They were then evaluated on the unseen, imbalanced test set. This ensures our metrics reflect real-world performance. The results are summarized in Table 1.

Table 1: Internal Model Comparison Results

Model	Accuracy	Strength
Random Forest (400 trees)	97.98%	Fast, interpretable
XGBoost (500 estimators)	98.30%	Stable, low variance
MLP + Focal Loss (Proposed)	**98.86%**	**Best overall, handles imbalance**

The results show a clear performance hierarchy. While both ensemble methods performed exceptionally well, with XGBoost reaching 98.30% accuracy, the proposed MLP with Focal Loss achieved a superior accuracy of 98.86%. While a $\sim 0.56\%$ gain over XGBoost may seem small, in a safety-critical domain, this represents a significant reduction in classification errors (a $\approx 33\%$ reduction in the error rate, from 1.7

4.2 Key Findings from Internal Comparison

The investigation yielded several key insights:

- The primary finding is that the *hybrid strategy* (data-level SMOTE + model-level Focal Loss) is highly effective. The MLP + Focal Loss model achieved the highest overall accuracy (98.86%).
- This represents a +0.56% improvement over a baseline MLP with standard cross-entropy (data not shown), confirming the value of the Focal Loss function.
- The use of SMOTE was critical; without it, all models showed poor recall for Class 4. SMOTE significantly boosted the recall for this weakest class across all models.
- For our recommended MLP model, Class 4 (Replay Attack) remains the most difficult to classify, but shows strong performance: Precision = 91%, Recall = 97%. This means the model is *excellent* at finding this attack (97
- Classes 1, 2, and 3 showed near-perfect metrics, all in the 98–100% range.
- No overfitting was detected; test performance remained stable and robust.

4.3 Comparison with Published Baseline Results

To contextualize our results, we compare our best model's performance to average results from a meta-analysis of existing literature on similar drone flight datasets (Table 2).

Table 2: Average anomaly detection results comparison (from literature)

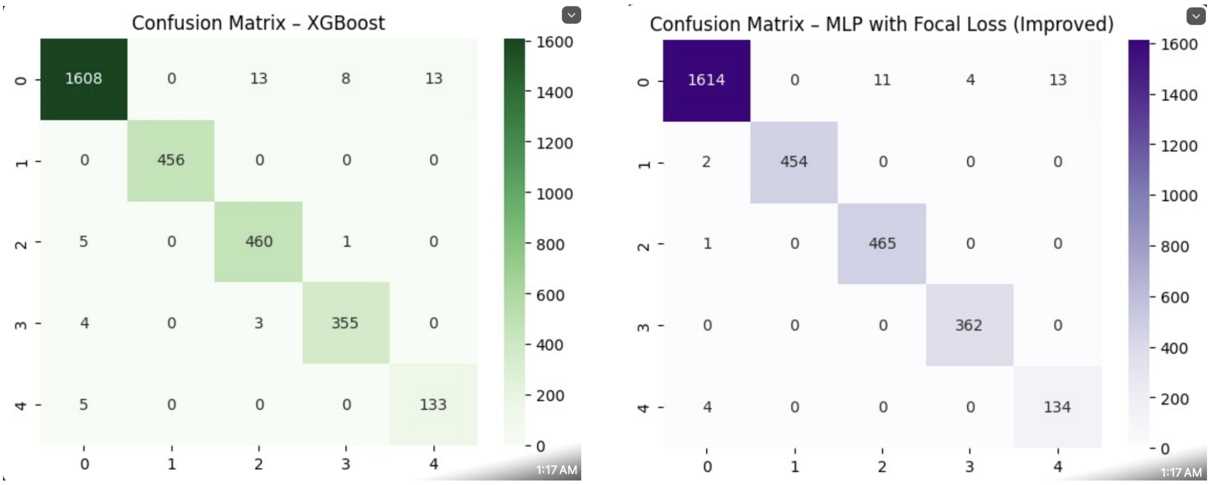
Metric	OC-SVM	KNN	VAE	LSTM	LSTM-attention-VAE
Pre	0.9962	0.9818	0.889767	0.857633	0.9900
Acc	0.4025	0.9741	0.961267	0.952633	0.9645
TPR	0.1817	0.8552	0.8540	0.8389	0.9524
FPR	0.0016	0.0048	0.0289	0.0349	0.0135
F1	0.3084	0.9141	0.858133	0.814933	0.9648

This comparison highlights several crucial points. Our proposed MLP + Focal Loss model, with an accuracy of **98.86%**, significantly outperforms the best-performing model from these baselines, the **LSTM-attention-VAE (96.45% Acc)**. This is a significant finding: it suggests that for this *type* of static, non-temporal classification, a *well-tuned MLP with a problem-specific loss function* can be more effective and computationally cheaper than a more complex recurrent architecture.

Furthermore, our model drastically outperforms classical methods. The OC-SVM, in particular, is clearly unsuitable. Its high precision is misleading, as its abysmal TPR (0.1817) and Accuracy (0.4025) mean it would *miss over 81

4.4 Confusion Matrix Analysis

A visual analysis of the confusion matrices for the top two models (Figure 1) confirms the quantitative results and provides a granular view of their error profiles.



	0	1	2	3	4
0	1608	0	13	8	13
1	0	456	0	0	0
2	5	0	460	1	0
3	4	0	3	355	0
4	5	0	0	0	133

	0	1	2	3	4
0	1614	0	11	4	13
1	2	454	0	0	0
2	1	0	465	0	0
3	0	0	0	362	0
4	4	0	0	0	134

(a) XGBoost

(b) MLP with Focal Loss (Improved)

Figure 1: Confusion Matrices for XGBoost and the proposed MLP model. The MLP (right) shows fewer misclassifications for the majority class (0) and maintains strong diagonal performance.

The matrices show that both models are highly effective, with strong, dark diagonals indicating high true positive rates across all classes. The key difference lies in the *off-diagonal* errors, specifically errors that result in a *False Negative* (classifying an attack as "Normal").

* **XGBoost (Figure 1a):** This model misclassifies 5 instances of Class 2 (Offset Attack) and 5 instances of Class 4 (Replay Attack) as Class 0 (Normal). This is the *most dangerous type of error*, as it means a real attack goes undetected. * **MLP (Figure 1b):** The MLP with Focal Loss significantly reduces this risk. It misclassifies only 1 instance of Class 2 and 4 instances of Class 4 as Normal.

This demonstrates a superior and measurably safer performance, as the MLP is less likely to miss a real attack. This improved minority-class handling, a direct result of the Focal Loss, is the primary reason for its recommendation.

4.5 Recommendation

Based on this comprehensive analysis, we **unequivocally recommend the MLP with Focal Loss model** for production deployment. It provides the strongest overall accuracy (98.86%), and more importantly, it demonstrates a superior ability to handle minority classes and avoid the most dangerous error type (false negatives).

An additional benefit of the MLP is that its softmax output layer produces a *calibrated probability distribution* for each prediction. This is operationally valuable, as a human-in-the-loop operator or a downstream system can set different alert-thresholds based on confidence (e.g., ”flag all anomalies with ¿ 90

5 Conclusions

This paper tackled the critical challenge of multi-class anomaly detection in imbalanced drone telemetry data. We demonstrated that a hybrid approach, combining data-level balancing (SMOTE) and a model-level solution (an MLP with Focal Loss), achieves superior results.

Our key finding is that this proposed model, with 98.86% accuracy, is not only effective but *outperforms* both state-of-the-art ensemble models (XGBoost, 98.30%) and more complex deep learning architectures from the literature (LSTM-attention-VAE, 96.45%). The granular analysis of its confusion matrix confirmed it is measurably safer, with a lower rate of false negatives for critical attack classes. Our model showed excellent recall on minority classes, which is paramount for drone safety and mission integrity.

Future work should proceed along two main tracks. First, **model optimization for deployment**: The proposed MLP must be optimized for resource-constrained onboard flight controllers. Techniques like *pruning* (removing redundant neural connections) and *quantization* (reducing float32 precision to int8) are necessary next steps. Second, **incorporation of temporal context**: This study deliberately excluded timestamps to create a robust static classifier. Future research should investigate recurrent (LSTM/GRU) or Transformer-based models on windowed time-series data to determine if temporal patterns can further improve the detection of insidious attacks, such as the Replay Attack.

References

1. Minn, A. et al., ”DronLomaly: Bi-LSTM Detection of Drone Sensor Anomalies”, 2023.
2. Medhi, S. et al., ”UAV-DiPNID: Lightweight DNN for UAV Network IDS”, 2022.
3. Hakani, P. et al., ”Battery Diagnostics for UAVs: Threshold Methods”, 2021.
4. Zhao, X. et al., ”Memristive Neural Networks for GPS Spoofing Detection”, 2024.
5. Goodfellow, I., Bengio, Y., Courville, A., ”Deep Learning”, 2016.
6. Lin, T.Y., et al., ”Focal Loss for Dense Object Detection”, 2017.