MethodsX 15 (2025) 103625

Contents lists available at ScienceDirect

MethodsX

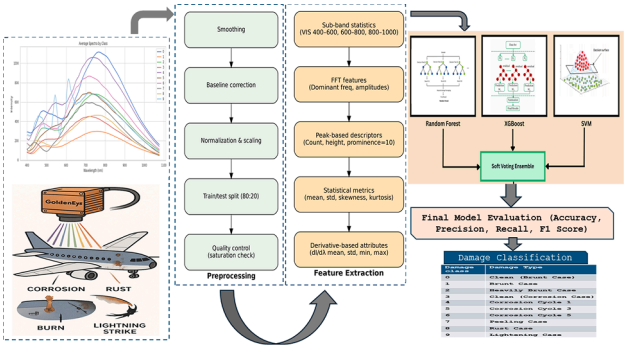
journal homepage: www.elsevier.com/locate/methodsx

Spectral feature extraction and ensemble learning for multiclass aircraft damage identification

Nilima Zade [a,*], Aditya Gupte [b], Pranshu Gupta [a], Nicolas Detalle [b], Anthony Mannion [b], Robert Voyle [b]

[a] Symbiosis Institute of Technology – Pune Campus, Symbiosis International (Deemed University), Pune, India
[b] Aviation Services Research Center, The Hong Kong Polytechnic University, Hong Kong

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Spectral analysis
Aircraft damage detection
Hyperspectral imaging
Feature extraction
Non-destructive testing
Damage Type Identification
Ensemble Learning

ABSTRACT

The research represents a robust methodology for identifying aircraft surface damage using hyperspectral imaging combined with ensemble machine learning. Surface degradation in aircraft, such as corrosion, burn marks, lightning strikes, weld defects, and paint peeling, is often difficult to detect using conventional inspection techniques. By leveraging high-resolution spectral data and domain-specific feature engineering, the proposed method enables accurate classification of ten different damage types using a structured machine-learning framework.

Hyperspectral intensity data were collected from over 500 real and lab-induced samples using the Goldeneye hyperspectral camera, followed by the extraction of handcrafted features across spectral, statistical, and frequency domains. A soft voting ensemble of Random Forest (RF), Extreme Gradient Boost (XGBoost), and Support Vector Machine (SVM) models achieves a peak

* Corresponding author.
E-mail address: nilima.zade@sitpune.edu.in (N. Zade).

https://doi.org/10.1016/j.mex.2025.103625
Received 23 July 2025; Accepted 11 September 2025
Available online 15 September 2025

classification accuracy of 92.6 % with high accuracy across damage classes. This method supports real-time, non-contact, and scalable aircraft inspection workflows and demonstrates strong potential for integration with drone-based or robotic inspection systems in aerospace maintenance.

Key contributions of this methodology:

- A hyperspectral imaging-based pipeline for identifying different aircraft surface damage types
- A robust ensemble model combining RF, XGBoost, and SVM for high-accuracy classification
- Designed for integration into real-world, automated aircraft inspection systems

Specifications table

Subject area	Engineering (Computer Science)
More specific subject area	*Spectral Analytics*
Name of the reviewed methodology	*Spectral feature analysis for aircraft damage detection*
Keywords	*Spectral Analysis; Aircraft Damage Detection; Hyperspectral Imaging; Feature Extraction; Non-destructive Testing; Damage Type Identification; Ensemble Learning*
Name and Reference of the Original Method	*NA*
Resource availability	*NA*

Background

An aircraft experiences a variety of surface damages during its service life, many of which are not easily detected by traditional inspection methods commonly used in the industry [1–2]. These damages include lightning strikes, rust corrosion, salt corrosion, weld imperfections, and paint peeling. Since such deterioration can compromise structural integrity and flight safety, its early and accurate detection during maintenance operations is essential. Traditional approaches such as visual inspection [3] or established Non-Destructive Evaluation (NDE) techniques [4–8] are often time-consuming, subjective, and may miss early-stage defects.

Hyperspectral Imaging (HSI) provides a promising alternative by capturing spectral data beyond the visible range, enabling detection of subtle material variations [9–11]. While this study applies HSI to aircraft surface damage, similar success has been documented across several domains where NDE is critical. In the energy sector, HSI has been used to detect cracks, erosion, and delamination in wind turbine blades; in aerospace, to monitor fatigue, oxidation, and foreign object damage in gas turbine blades; and in agriculture, to detect leaf and seed damage caused by mites [12–15]. Beyond industry, HSI has been adopted in art conservation, distinguishing pigments and restoration layers. These diverse applications highlight the versatility of HSI for inspection, structural integrity assessment, and quality control.

Each damage type exhibits a unique Hyperspectral Signature (HS), which forms the basis for classification. In this research, a GoldenEye hyperspectral camera was used to collect high-dimensional spectral data from aircraft damage samples under controlled laboratory settings. To emulate real-world conditions, the samples were imaged under varied lighting, angles, and distances. The GoldenEye system captures spectral intensity values from 399 to 1100 nm with 141 spectral bands per pixel.

The study aims to establish a systematic process linking specific HS to damage classes. Spectra from known samples were collected and pre-processed using GoldenEye's proprietary software. From these, 23 features were extracted to support pattern recognition. These features include statistical measures (mean, standard deviation), derivative features, peak statistics, and frequency-domain attributes derived using Fast Fourier Transform (FFT). The extracted features were then used to train three machine learning models: SVM, RF, and XGBoost.

To further enhance classification robustness and accuracy, a voting ensemble approach was employed, combining the predictions of all three models [16–17]. Ensemble-based methods have proven effective across domains such as composite structure health monitoring [18] and earthquake-induced bridge damage detection [19], underscoring their adaptability and reliability in addressing diverse challenges. The trained model was subsequently evaluated using spectra from unknown damage samples to assess generalization performance.

The primary motivation of this work is to minimize inspection time while improving accuracy and reliability in damage detection through HSI. By offering objective, data-driven insights, the proposed approach can assist maintenance crews in prioritizing repairs more effectively. Furthermore, the framework demonstrates scalability and can be adapted to other domains where surface monitoring plays a critical role.

Method details

For the present research, six major types with four subcategories of aircraft surface damage were targeted: Burnt case, lightning strike, rust corrosion, salt-induced corrosion, weld damage, and paint peeling. A baseline, undamaged surface was used for reference. The purpose of selecting these damages is their probabilities of occurrence during the aircraft lifecycle and the challenge they pose in early-stage detection through conventional inspection methods. The specimens were selected so that they correspond as closely as

possible to the real-world damages. The samples mainly consisted of lab-induced specimens and naturally damaged components for aged aircraft parts, as shown in Fig. 1.

Imaging was conducted under controlled lighting conditions using a full-spectrum directional light source. Both the camera and the light source were varied in position and angle to replicate realistic inspection variability. The data was acquired and extracted using the camera's proprietary software, with careful exposure calibration to prevent sensor saturation. The Regions of Interest (ROIs) corresponding to visible damage were manually selected, and the extracted spectral signatures were exported as CSV files. These spectra then formed the foundation for a multi-stage ensemble machine learning framework designed to classify the different damage types. The machine learning process included preprocessing (smoothing, baseline correction), feature extraction (statistical, spectral, frequency, and derivative-based), and feature selection using Random Forest. The data imbalance was addressed using SMOTE, and the model training was carried out using an Ensemble model, which consists of using a combination of Random Forest, XGBoost, and SVM methods for increasing the robustness and accuracy of the model.

Data collection

Hyperspectral intensity data were collected from 530 samples using a calibrated GoldenEye hyperspectral camera system. The system captured intensities that cover the visible to Near-Infrared (NIR) spectrum. These measurements provide high-resolution spectral fingerprints necessary for identifying and distinguishing between various types of surface damage. The dataset includes samples from six primary damage categories - burn marks, rust, salt corrosion, weld defects, peeling paint, and lightning strike - as well as a clean reference surface.

Each specimen was imaged under controlled lighting using a full-spectrum directional light source, ensuring consistent illumination across all captures, as shown in Fig. 2. The camera was mounted on a precision tripod, and image acquisition was conducted from multiple distances and angles to simulate field variability. The lighting angle was also varied by $\pm 5^{\circ}$ to capture spectral shifts under different incident conditions. Hyperspectral images were acquired using GoldenEye's proprietary software. The software exported CSV files that contain wavelength-intensity pairs for each ROI. These files formed the raw input for the machine learning pipeline.

Data preprocessing

The dataset comprises hyperspectral reflectance measurements from 530 aircraft surface samples. Each sample is annotated with one of ten damage types. A Spectra (e.g., Burnt_point_23) and numerical class label (0–9) accompany the 141 spectral features. Fig. 3 summarizes the class distribution, highlighting the imbalance. To ensure robust model performance and reliable classification outcomes, the hyperspectral dataset undergoes meticulous preprocessing to eliminate noise, handle inconsistencies, and align with machine learning requirements.

The process began with data cleaning, where non-numeric and irrelevant columns, such as extraneous metadata, were removed to retain only spectral intensities and class labels, resulting in a compact and interpretable dataset. Missing values in the spectral data were handled in multiple steps. Initially, linear interpolation was applied to fill small gaps by estimating values from neighboring bands; where necessary, average values were used to complete the data, thereby preserving the overall shape of the spectra. To reduce noise, a Savitzky–Golay filter was applied to smooth local fluctuations without losing important spectral details. Smoothing was performed using a window size of 11 points and a third-degree polynomial fit. To confirm the robustness of this choice, a sensitivity analysis was conducted with varying window sizes (7–15) and polynomial orders (2–4). The macro-F1 score varied by <1.5 % across these settings, demonstrating that the filtering approach is stable. Missing bands were handled via linear interpolation between nearest valid neighbors, which preserved continuity of spectral signatures without extrapolating beyond measured ranges. Finally, before applying machine learning models such as SVM and XGBoost, all features were standardized to have zero mean and unit variance, ensuring that no single feature dominated due to its scale. This was done by adjusting each feature so that it has an average of zero and the same amount of variation. This ensures that one feature does not influence the model too much due to bigger numbers.

In the spectral data, some categories, such as Burnt Case, as a higher number of sample numbers compared to others, such as corrosion beginning case. This uneven number of samples can reduce the overall accuracy of the machine learning model. To fix this issue, a method called as Synthetic Minority Over-sampling technique (SMOTE). SMOTE creates new, made-up examples for the underrepresented groups so that all groups have a similar number of records. Instead of simply copying existing samples, SMOTE makes new, slightly different samples by mixing together real samples from the smaller group. It picks one sample and then finds its closest neighbors in the data. It creates new samples that fall in between these points. This helps balance out the groups, making sure the model learns about all the categories fairly rather than being biased towards the ones with more samples. Fig. 4 indicates the data distribution after the SMOTE analysis dataset. We also tested Adaptive Synthetic Sampling (ADASYN) and Borderline-SMOTE. All resampling procedures were applied only within the training folds of cross-validation to prevent any data leakage. Among the evaluated strategies, SMOTE achieved the most stable macro-F1 performance across classes. ADASYN led to unstable outputs in this dataset, while Borderline-SMOTE improved minority recall but reduced overall precision. To avoid potential artifacts, SMOTE was selected for further experiments, with resampling restricted to training folds only.

Finally, the preprocessed dataset is divided into training and testing subsets using an 80:20 stratified split, preserving class proportions to ensure consistent evaluation of classification performance. This comprehensive preprocessing pipeline produces a clean, normalized, and balanced dataset optimized for downstream feature engineering and modeling.

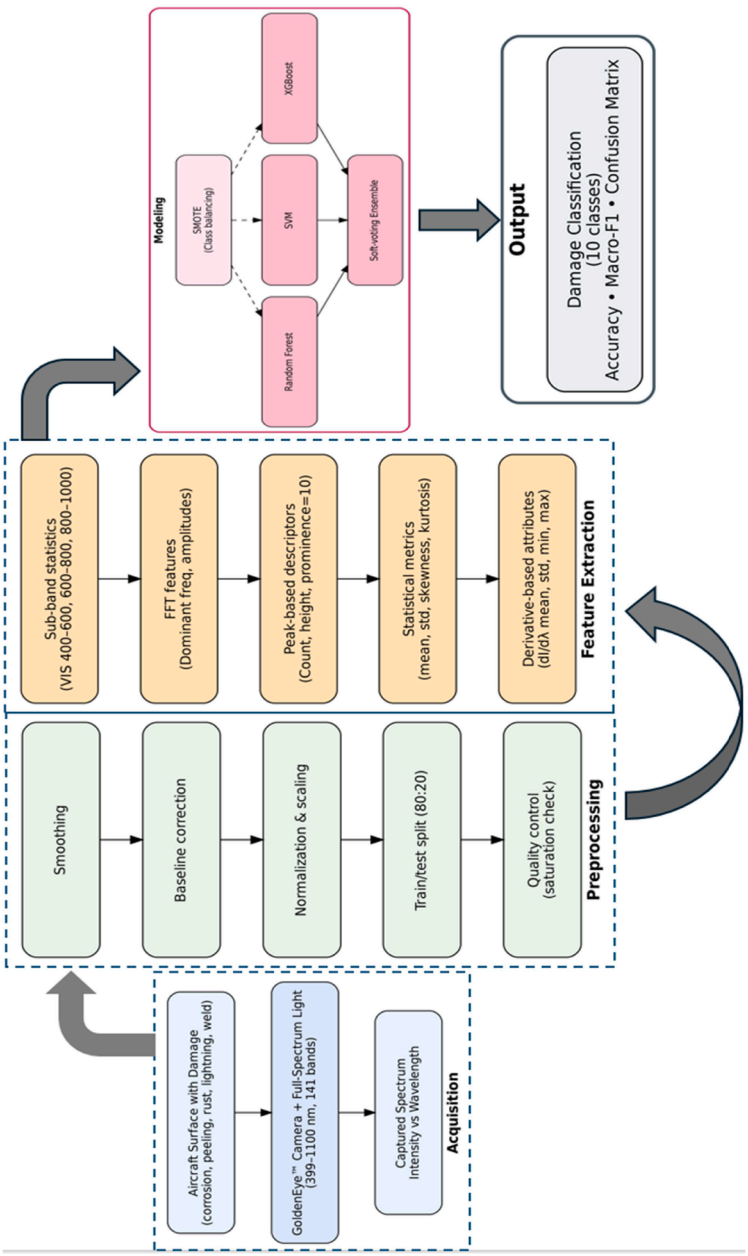


Fig. 1. Overall Process Flowchart.

Fig. 2. Experimental setup for data collection of lab-based samples and aircraft-based samples.

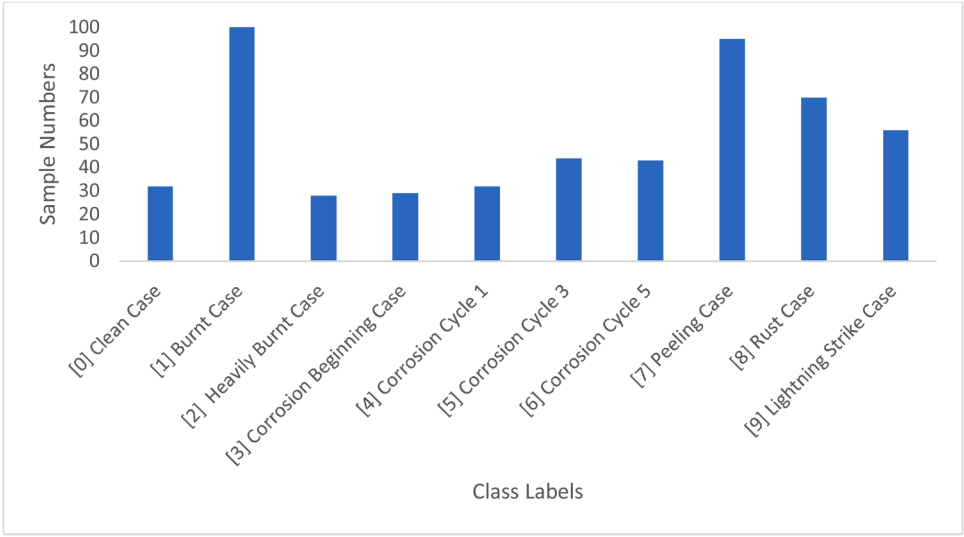


Fig. 3. Summarization of Class Labels and Distribution.

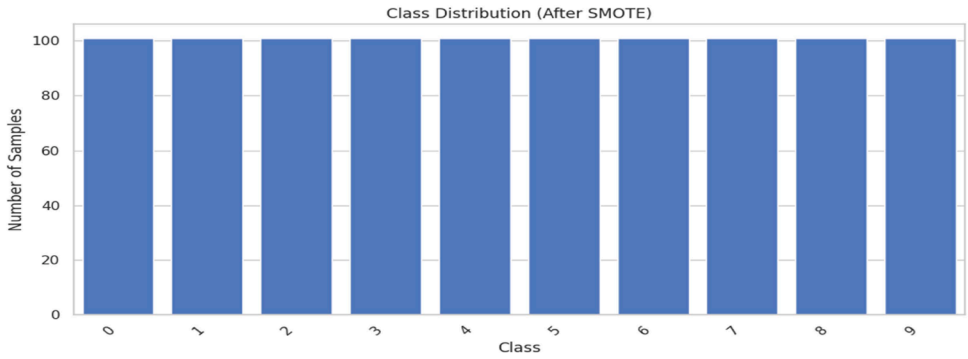


Fig. 4. Balanced class distribution after SMOTE analysis.

Mathematical model for feature selection

Given the medium-scale nature of the dataset, a handcrafted feature engineering approach was adopted to extract interpretable and discriminative features, avoiding the data-hungry requirements of deep learning methods. The feature extraction process was designed to capture a comprehensive set of physical, statistical, and frequency-domain attributes from each spectral profile, resulting in a 23-dimensional feature vector per sample. These features have been identified by analyzing and understanding the nature of the average pattern of each class, as shown in Fig. 5. Based on the nature of the patterns, five different categories of the features have been identified; these are structural pattern, overall shape pattern, repeating pattern, brightness pattern, and change pattern. Based on these five types of patterns, 23 different feature categories were computed as represented in Fig. 6.

Peak-Based Features: Peaks were identified with a prominence threshold of 10 to ensure detection of significant spectral features. Extracted features included the number of peaks, maximum and mean peak height, maximum and mean peak width, maximum and mean peak prominence, and the wavelength position of the highest peak. These features captured the presence and intensity of absorption or emission bands associated with specific damage types.

Let $Pi = \{pi, 1, pi, 2, \ldots, pi, ki\}$ be the set of peak indices for sample i, where ki is the number of peaks, and $\lambda p, i, k$ is the wavelength at peak k. The properties (height, width, prominence) are provided by find peaks. The number of peaks ($N_{peaks,i}$) is computed by Eq. (1). Maximum peak Height (H_{max},) is computed by Eq. (2). Mean Peak Height ($H_{mean,}$) is calculated by Eq. (3). Mean Peak Width, $W_{mean,i}$, Mean Peak Prominence ($P_{mean,i}$), Wavelength Position of the Highest Peak, ($\lambda_{\max}$,) are computed by Eqs. (4), 5, and 6, respectively.

$$N_{peaks,} = k_i \tag{1}$$

where Ki is the count of peaks detected in a spectral profile, $Si(\lambda)$

$$H_{max}, = \max_{k=1} Ki\ [Si(\lambda pi, k)] \tag{2}$$

where $Si(\lambda pi, k)$ is the intensity at peak k.

$$H_{mean,} = \frac{1}{Ki} \sum_{k=1}^{Ki} Si(\lambda pi, k) \tag{3}$$

$$W_{mean,i} = \frac{1}{Ki} \sum_{k=1}^{Ki} (\mathrm{wi, k}) \tag{4}$$

(if Ki=0 set to 0).

$$P_{mean,i} = \frac{1}{Ki} \sum_{k=1}^{Ki} promi, k \tag{5}$$

(if Ki=0, set to 0).

$$\lambda_{\max,} = \lambda pi, k* \quad \text{where } k* = \arg[Si(\lambda pi, k)] \tag{6}$$

Statistical Features: To characterize the overall shape and distribution of each spectrum, statistical measures were computed. These included the mean, standard deviation, skewness, and kurtosis of the spectral intensities, providing insights into the central tendency, variability, asymmetry, and peakedness of the spectral profiles. Mean (μi), Standard Deviation (σi), Skewness, a measure of

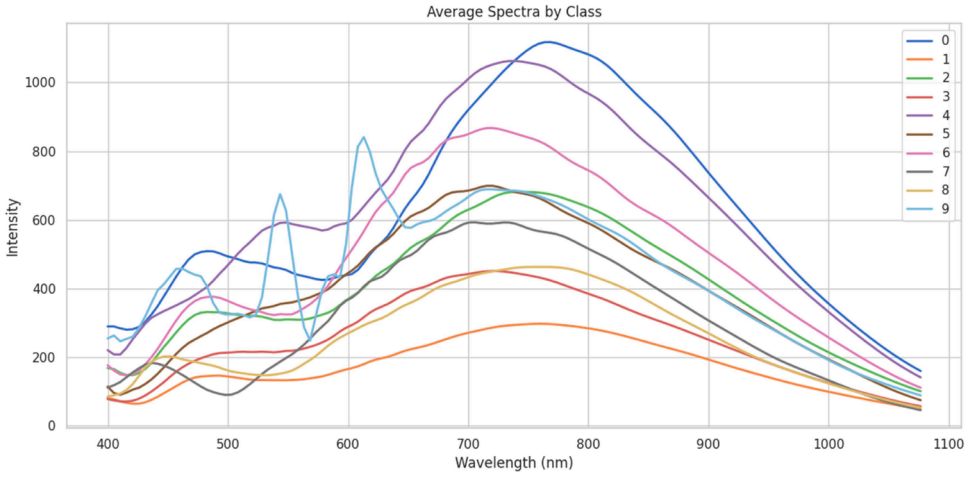


Fig. 5. Average pattern representation of each class.

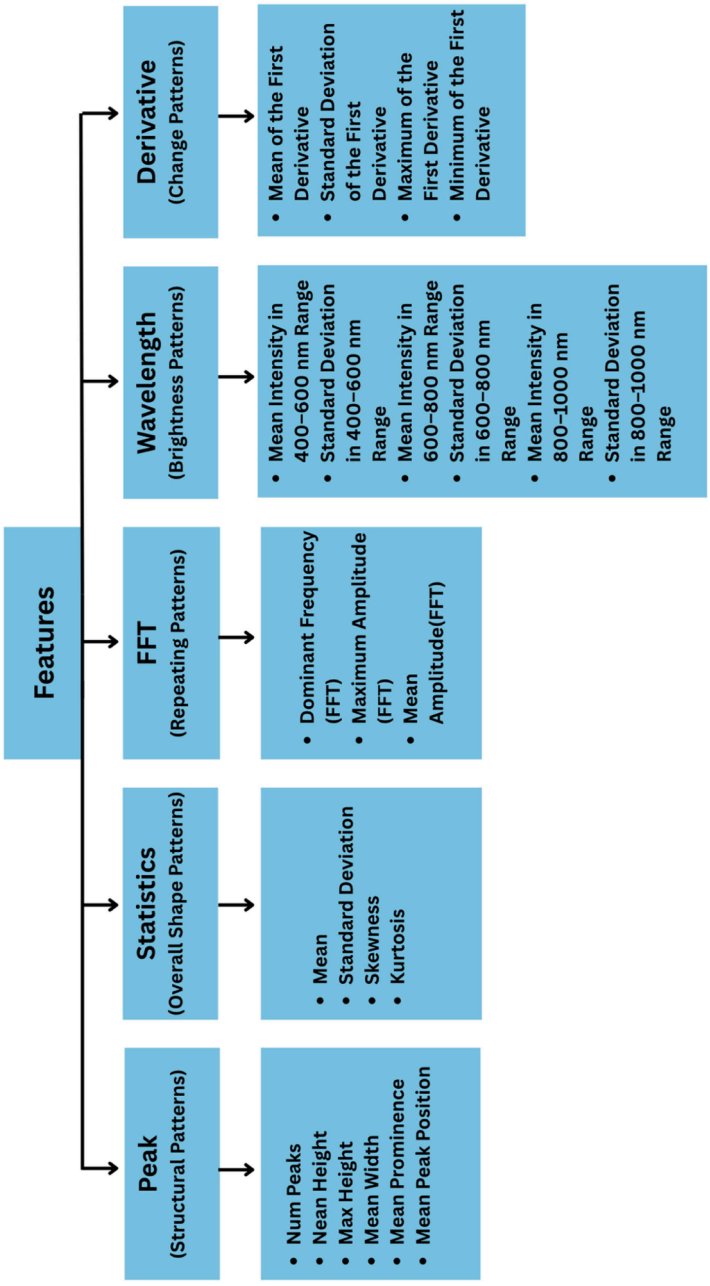


Fig. 6. Summary of features-based pattern types for spectral analysis.

asymmetry (skew*i*), Kurtosis, a measure of peakedness, excess kurtosis (kurt*i*) are computed by Eqs. (7), 8, 9, and 10. In Eq. (10), the subtraction of 3 adjusts for excess kurtosis relative to a normal distribution.

Features are computed over the entire spectral profile Si(λ), where L is the number of wavelength points.

$$\mu i = \frac{1}{L}\sum\nolimits_{\lambda}(Si(\lambda)) \tag{7}$$

$$\sigma i = \sqrt{\frac{1}{L}\sum_{\lambda}\left((\mathrm{Si}(\lambda)-\mu \mathrm{i})^{2}\right)} \tag{8}$$

$$skewi = \frac{\frac{1}{L}\sum_{\lambda}(\mathrm{Si}(\lambda)-\mu \mathrm{i})^{3}}{\left(\frac{1}{L}\sum_{\lambda}(\mathrm{Si}(\lambda)-\mu \mathrm{i})^{2}\right)^{\frac{3}{2}}} \tag{9}$$

$$kurti = \frac{\frac{1}{L}\sum_{\lambda}(\mathrm{Si}(\lambda)-\mu \mathrm{i})^{4}}{\left(\frac{1}{L}\sum_{\lambda}(\mathrm{Si}(\lambda)-\mu \mathrm{i})^{2}\right)^{2}} - 3 \tag{10}$$

Fast Fourier Transform (FFT) Features: The FFT function was applied to transform each spectrum into the frequency domain, capturing periodic patterns in the spectral signal. Extracted features included the dominant frequency, maximum amplitude, and mean amplitude of the FFT components, which were particularly useful for identifying repetitive spectral structures associated with certain damage types. Dominant Frequency (*f*dom,), Maximum Amplitude (*A*max,), and Mean Amplitude (*A*mean,) are computed using Eqs. (11), 12, and 13, respectively.

The FFT transforms Si(λ) into the frequency domain. Let Fi(f) be the FFT magnitude, where f represents frequency components, and only the first half of the FFT (positive frequencies) is considered due to symmetry.

$$f_{\mathrm{dom}\ ,} = \arg\max_{f}|Fi_{\mathrm{i}}(\mathrm{f})| \tag{11}$$

Fi is the magnitude of the FFT at frequency f.

$$A_{\mathrm{max}\ ,} = \max_{f}|Fi(\mathrm{f})| \tag{12}$$

$$A_{\mathrm{mean}\ ,} = \frac{1}{M}\sum\nolimits_{M}^{f=1}|Fi(\mathrm{f})| \tag{13}$$

where $M = L/2$ is the number of frequency components in the positive half.

Wavelength Sub-Range Features: To capture localized spectral variations, the wavelength range was divided into three sub-bands: 400–600 nm (visible), 600–800 nm (visible to near-infrared transition), and 800–1000 nm (near-infrared). This was done to link localized spectral variations to known material and chemical characteristics relevant to surface degradation on aircraft components. These sub-bands were chosen based on established hyperspectral literature in corrosion detection, remote sensing, and pigment analysis, where materials like rust, paint, and thermally altered surfaces exhibit distinguishable spectral features in specific regions [14]. The mean and standard deviation of intensities were calculated for each sub-band, providing region-specific descriptors that highlighted differences in reflectance properties across damage types. By computing mean and standard deviation within these bands, we capture coarse-level but meaningful indicators of degradation. For example, rusted regions show increased reflectance variability in the 400–600 nm band due to surface roughness and pigmentation loss. This domain-informed sub-band partitioning, as shown in Table 1, ensures that the extracted statistical features are not arbitrarily defined, but rather linked to real physical and chemical transitions occurring on the aircraft surface. It also improves the interpretability of machine learning predictions, as feature importance scores can be traced back to sub-bands that reflect specific material behaviors.

Mathematically these feature are represented as follows. Let Λ*j*, be the subset of wavelengths in sub-band j (where $j = 1,2,3$) for sample i, and Si(λ) the corresponding intensities. Mean Intensity in 400–600 nm $\mu 400-600$, *i*: Standard Deviation in 400–600 nm ($\sigma 400-600,i$), Mean Intensity in 600–800 nm ($\mu 600-800$, *i*), Standard Deviation in 600–800 nm ($\sigma 600-800$, *i*), Mean Intensity in 800–1000 nm ($\mu 800-1000$, *i*), Standard Deviation in 800–1000 nm ($\sigma 800-1000,i$) are computed by Eqs. (14),15,16,17,18 and 19

Table 1
Domain informed sub-band partitioning.

Sub-Band (nm)	Spectral Region	Relevant Aircraft Material/Damage Signatures
400–600 nm	Visible (VIS)	Paint color, charring effects, Fe_2O_3 (iron oxide) pigments (rust), peeling paint
600–800 nm	VIS–NIR transition	Oxide layer changes, mild corrosion, organic binders, oxidized aluminum
800–1000 nm	NIR	Surface roughness, moisture, and carbonized material from burnt or lightning strike
1000–1100 nm	NIR tail	Low-reflectance metallic absorption, surface chemistry shifts

respectively. If a sub-band has no data, set the mean and standard deviation to 0.

$$\mu 400-600,=\frac{1}{|\Lambda 1,\mathrm{i}|}\sum_{\lambda\in\Lambda_1,Si}Si_{\mathrm{i}}(\lambda) \tag{14}$$

$$\sigma 400-600,=\sqrt{\frac{1}{|\Lambda 1,\mathrm{i}|}\sum_{\lambda\in\Lambda_1,S}(Si(\lambda)-\mu 400-600,\mathrm{i})^2} \tag{15}$$

$$\mu 600-800,=\frac{1}{|\Lambda 2,\mathrm{i}|}\sum_{\lambda\in\Lambda_1,S}Si_{\mathrm{i}}(\lambda) \tag{16}$$

$$\sigma 600-800,=\sqrt{\frac{1}{|\Lambda 2,\mathrm{i}|}\sum_{\lambda\in\Lambda_1,S}(Si(\lambda)-\mu 600-800,\mathrm{i})^2} \tag{17}$$

$$\mu 800-1000,=\frac{1}{|\Lambda 3,\mathrm{i}|}\sum_{\lambda\in\Lambda_1,S}Si_{\mathrm{i}}(\lambda) \tag{18}$$

$$\sigma 800-1000,=\sqrt{\frac{1}{|\Lambda 3,\mathrm{i}|}\sum_{\lambda\in\Lambda_{3,\mathrm{j}}}(\lambda\mathrm{i}(\lambda)-\mu 800-1000,\mathrm{i})^2} \tag{19}$$

Derivative-Based Features: The first derivative of each spectrum was computed to emphasize abrupt changes in intensity, which are indicative of damage boundaries or material transitions. Features extracted from the derivative included the mean, standard deviation, maximum, and minimum values, enhancing the model's sensitivity to subtle spectral shifts. Mean of Derivative ($D_{\mathrm{mean},i}$), Standard Deviation of Derivative ($D_{std,}$), Maximum of Derivative (Dmax,i), and Minimum of Derivative ($D_{\mathrm{min},i}$) are computed using Eqs. (20), 21, 22, and 23. This feature engineering strategy resulted in a robust and interpretable representation of the spectral data, capturing both global and local characteristics relevant to damage classification.

The first derivative $\mathrm{Di}(\lambda)$ is computed as the numerical gradient of $Si(\lambda)$ with respect to λ.

$$D_{\mathrm{mean},i}=\frac{1}{L}\sum_{\lambda}Di(\lambda) \tag{20}$$

where $\mathrm{Di}(\lambda)=\frac{\partial Si(\lambda)}{\partial\lambda}\approx\frac{Si(\lambda+\Delta\lambda)-Si(\lambda)}{\Delta\lambda}$

$$D_{std,}=\sqrt{\frac{1}{L}\sum_{\lambda}(Di(\lambda)-\mathrm{Dmean},^{\mathrm{i}})^2} \tag{21}$$

$$D_{\mathrm{max},i}=Di(\lambda) \tag{22}$$

$$D_{\mathrm{min},}=Di(\lambda) \tag{23}$$

Conceptual design

The current method is based on the idea that combining detailed spectral analysis with an ensemble machine learning model can improve the accuracy of classifying surface damage on aircraft using hyperspectral images. Hyperspectral data is useful because it captures fine differences in material properties, but it can also be very noisy, very high in dimensionality, and hard to interpret. These challenges are especially important when trying to classify different types of damage, such as burning, corrosion, or lightning strike, which often look similar in regular images. To handle these challenges, a structured framework was developed that uses both custom feature engineering and multiple machine learning models that work together. The main goal was to build a system that is not only accurate but also reliable and easy to interpret. This is especially important in aviation, as even small errors in damage detection can have serious consequences.

The method begins by processing the raw hyperspectral data (between wavelengths from 399.85 to 1100 nm) to remove noise and normalize the values across the 141 bands. Feature extraction is then carried out to capture the key patterns in the data – such as peaks in intensity, statistical trends, and frequency-domain characteristics. These features are tailored to reflect material properties, like oxidation or carbonization, which are known signs of corrosion or burn damage. For example, variations observed in the 600–800 nm wavelength range are especially useful for detecting early signs of electrochemical degradation.

To reduce the number of features and avoid overfitting, a method consisting of feature selection using RF importance scores is applied. Thresholds in the range of 0.01–0.05 were tested to determine a cutoff for feature selection. An elbow analysis indicated that a threshold of 0.02 provided the best compromise between retaining informative features and reducing redundancy. This ensures only the features that show the most difference between damage types are considered, while making the model more efficient to run. The selected features include the number of peaks, mean peak height, maximum peak height, mean peak width, mean peak prominence, mean peak position, mean intensity, skewness, kurtosis, dominant frequency, maximum amplitude, and mean amplitude. In addition, mean and standard deviation values were extracted from the spectral sub-bands 400–600 nm, 600–800 nm, and 800–1000 nm, along

with derivative-based statistics such as mean, standard deviation, and minimum.

The classification step uses an ensemble of three machine learning models: RF, XGBoost, and SVM. Each model is trained independently on the same set of extracted features. These models were chosen for their different strengths: RF is good for reducing noise and for understanding which features matter most; XGBoost works well with complex, non-linear patterns; and SVM helps separate damage types in high-dimensional datasets. The outputs from these models are then combined using a soft-voting ensemble method, where each model gives a probability for each damage class. The final prediction is a weighted combination of these probabilities, with weights based on each model's validation accuracy. This strategy improves consistency, reduces overfitting, and boosts performance on difficult damage types such as lightning strikes.

The whole system was designed to be modular and easy to understand, so that it can be used in practical aircraft inspection tasks and other forms of non-destructive testing. It also generalizes well, performing reliably under different inspection conditions. Fig. 7 shows the model architecture that was used for spectral analysis and machine learning. The method deals effectively with the challenges of hyperspectral data and meets the strict reliability standards needed in aviation safety procedures. It is built to detect many types of surface damage in aircraft materials with high confidence, and it can be extended in the future as new types of damage or sensor technologies emerge.

Random forest model

Random Forest is a group-based learning method where several decision trees are built during training. Each tree is trained on a different random sample of the data, and at each step, it considers only a random set of features, which helps reduce similarity between trees and improves the model's ability to generalize. The final prediction is determined by majority voting across all trees, providing a robust and interpretable output. The architecture is shown in Fig. 8.

RF is particularly suitable for hyperspectral damage classification due to its resistance to overfitting, especially in noisy datasets, and its minimal assumptions about data distributions. Its ability to handle high-dimensional inputs efficiently—by considering only a subset of features at each split—makes it ideal for the 23-dimensional feature set, where features like peak prominence and sub-band variance vary significantly across damage types.

RF was used not only as a classifier, but also to identify which features contributed the most to the model's decision. These feature importance scores were based on how much each feature helped reduce the classification error during tree splits. A cutoff value of 0.02 was chosen to keep the features that were most helpful for distinguishing between different spectral patterns. This value was selected based on initial tests that showed a good trade-off between including useful features and avoiding unnecessary ones. A lower threshold (0.01) led to including too many irrelevant features, and a higher threshold (0.05) led to the removal of important signals, reducing overall model performance.

To improve the RF models' accuracy, key hyperparameters such as the number of trees, depth of each tree, and the smallest number

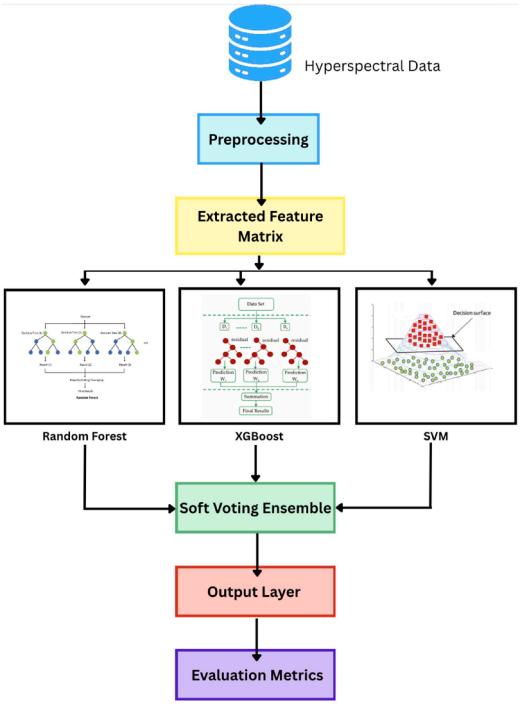


Fig. 7. Model architecture for spectral analysis.

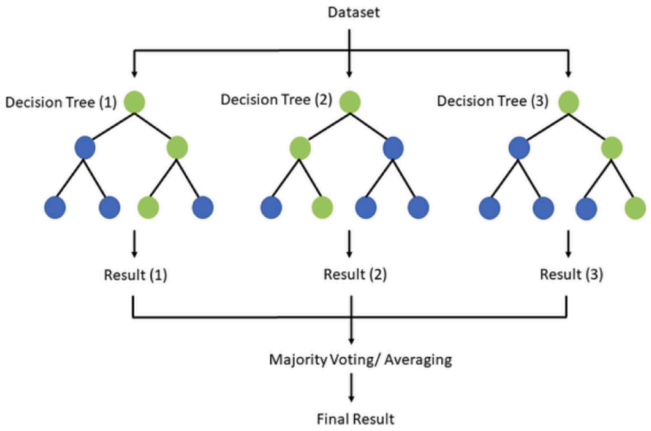


Fig. 8. Architecture of RF model.

of samples allowed in a node split were fine-tuned using GridSearchCV with cross-validation on the training data. RF was especially valuable in highlighting key spectral features linked to specific damage types, making it a key part of the ensemble modelling approach.

Extreme gradient boosting model

In the current research, we used XGBoost, a machine learning method that builds a series of decision trees, where each new tree improves on the mistakes of the previous ones. This step-by-step process helps the model learn better patterns in the data. One of the strengths of XGBoost is that it includes a mechanism to reduce overfitting, which makes it well-suited for working with structured data such as hyperspectral features. It is also efficient in terms of computation, which is important when dealing with large datasets. XGBoost's learning strategy makes it particularly useful for picking up subtle hyperspectral differences such as those seen between burnt and heavily burnt metal surfaces, where slight changes in intensity play a key role.

Hyperparameters, including learning rate (0.01–0.3), number of estimators (100–300), and maximum depth (3–10), were tuned using GridSearchCV to optimize test accuracy. XGBoost's ability to prioritize predictive power and handle multi-class tasks makes it a vital component of the ensemble, complementing RF's interpretability with enhanced modeling of intricate spectral relationships. Fig. 9 shows the architecture of the XGBoost model.

Support vector machine model

SVM is a discriminative classifier that seeks the optimal hyperplane to maximize the margin between classes in feature space. For the multi-class hyperspectral dataset, a one-vs-one strategy is employed, constructing binary classifiers for each pair of the ten damage

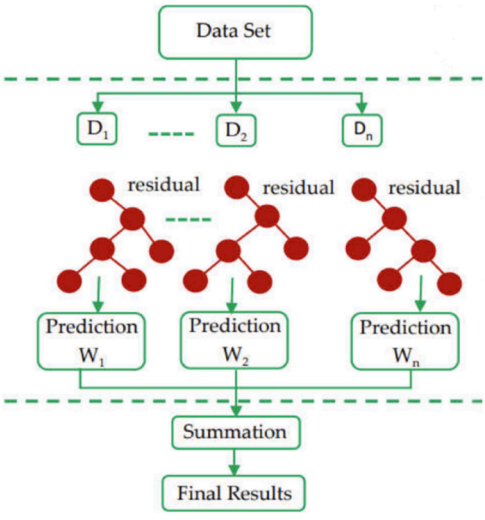


Fig. 9. Architecture of XGBoost.

types. This study uses a radial basis function (RBF) kernel to map features into a higher-dimensional space. SVM is particularly effective for high-dimensional datasets with a moderate number of samples, as it captures complex decision boundaries, such as those separating spectrally similar classes like Rust Point and Corrosion. The penalty term (C) (0.1–10) balances margin maximization with misclassification tolerance, while γ(0.001–1) was optimized via cross-validation to ensure precise boundary delineation. Fig. 10 indicates the architecture of SVM.

SVM's robustness to imbalanced data, when combined with SMOTE, enhances its performance on minority classes like Lightning Strike. Although SVM lacks native probabilistic outputs, its integration into the soft-voting ensemble allows it to contribute confidence-weighted predictions, leveraging its strength in delineating subtle spectral variations. SVM's inclusion in the ensemble ensures sensitivity to boundary cases, complementing the tree-based approaches of RF and XGBoost. Table 2 indicates the summary of the model parameters implemented.

Proposed system

The final system employs a soft-voting ensemble model to classify ten aircraft surface damage types from hyperspectral reflectance data, integrating predictions from three supervised classifiers: RF, XGBoost, and SVM. The soft-voting ensemble aggregates probability outputs from each classifier, weighted by their respective test accuracies, to produce a robust and interpretable damage class prediction with associated confidence scores. The class with the highest score is selected as the final prediction. This mechanism leverages RF's robustness to noise, XGBoost's ability to capture complex non-linear interactions, and SVM's precise boundary delineation to achieve final accuracy.

The ensemble approach was chosen to address the challenges of hyperspectral data, such as high dimensionality and subtle spectral variations between classes like Burnt Case and Heavily Burnt Case. RF provides interpretable feature importance, aiding in understanding spectral signatures like peak prominence in thermal degradation. XGBoost excels at modeling intricate relationships, such as those between derivative features and corrosion stages, while SVM's radial basis function kernel ensures precise separation of spectrally similar classes. To construct the ensemble, classifier weights were optimized using an inner 5-fold cross-validation. The resulting weights were 0.4 for RF, 0.4 for XGB, and 0.2 for SVM, reflecting their relative predictive strengths. Two strategies were evaluated: hard voting (majority rule) and soft voting (probability-based aggregation). Soft voting was ultimately selected, as it incorporates probability confidence and therefore better captures minority-class information. The soft-voting ensemble mitigates individual model limitations—RF's potential underfitting, XGBoost's overfitting risk, and SVM's parameter sensitivity—enhancing robustness and equitable performance across imbalanced classes.

Fig. 11 depicts a soft voting ensemble model that shows i^{th} multiple classifiers (C1 to Cn) trained on the same training data. For a validation input, each classifier provides the probability of the input being in class L0 and class L1. All probabilities of participants of each of these are averaged separately for L0 and L1. These averages are compared and, if the average probability of L0 is larger than that of L1, the input is classified as L0; otherwise, it is classified as L1. This technique uses the power of all the models together to make better predictions.

Method validation

The validation of the proposed methodology for aircraft surface damage classification using HCI involved rigorous quantitative and comparative analysis. The goal was to establish the model's robustness, generalization ability, and accuracy through ensemble learning. Three widely used evaluation metrics, Accuracy, F1-Score, and the Confusion Matrix, to assess the predictive quality across both individual and ensemble classifiers, are employed. To begin with, performance evaluation was conducted individually for three baseline models: RF, XGBoost, and SVM. Each of these models exhibits distinct learning characteristics: RF leverages bagging over decision trees, XGBoost uses gradient boosting to minimize bias, while SVM separates feature space with optimal hyperplanes. Subsequently, these models were fused using a soft-voting ensemble classifier, enabling the system to harness diverse perspectives from each algorithm for improved reliability.

For reproducibility and consistency in validation, a Stratified K-Fold Cross-Validation strategy with five folds was employed. This ensured that the class proportions were maintained across folds, especially important due to the imbalance present in the original dataset. After applying SMOTE to address class imbalance, the model demonstrated substantial gains in generalization. When comparing resampling methods, SMOTE consistently achieved the most stable macro-F1 across folds. Borderline-SMOTE improved

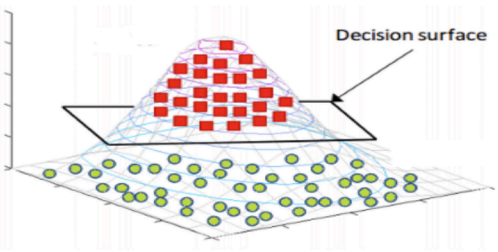


Fig. 10. Architecture of SVM Model.

Table 2
Implementation details of the models.

Component	Parameter	Value
Model Training	Data Split (Train/Test)	80 %/20 %
	Stratification	stratify=*y*
	Cross-validation folds (GridSearch)	5
	Cross-validation folds (Final Evaluation)	10
Feature Selection	Selection Model	RandomForestClassifier
	Importance Threshold	0.02
RF Model	max depth	15
	min samples leaf	2
	min samples split	2
	n estimators	300
XGBoost Model	Learning rate	0.1
	max depth	3
	n estimators	200
SVM Model	C	5
	Gamma	Scale
	Kernel	Rbf

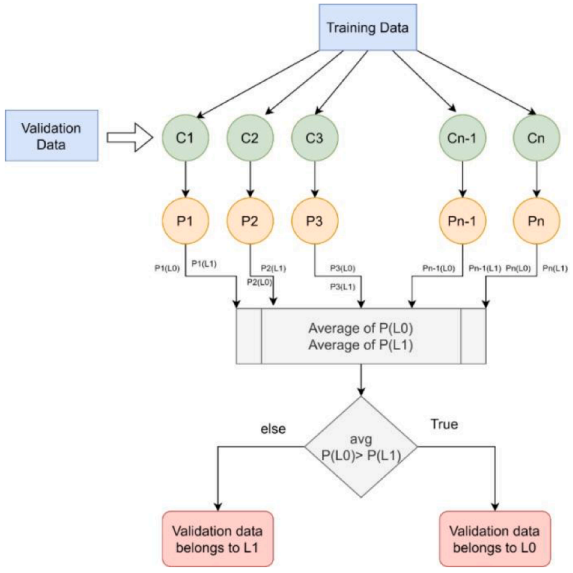


Fig. 11. Architecture of the Ensemble model.

minority recall slightly but reduced precision, while ADASYN results were unstable and are therefore not reported in Table 3. The comparative outcomes are summarized in Table 3.

The ensemble configuration of [0.4, 0.4, 0.2] for RF, XGB, and SVM achieved the highest balance of performance. Soft voting consistently outperformed hard voting by improving recall for minority classes while maintaining competitive overall accuracy. Cross-validation results of the individual base models are provided in Table 4, demonstrating that the ensemble improved stability and generalization.

The ensemble model showed the highest accuracy, as indicated in Table 4, benefiting from the complementarity of the component models. To benchmark against deep learning methods, a 1D Convolutional Neural Networks (CNN) was implemented. The CNN achieved an overall accuracy of 85.8 %; its macro-F1 score was lower than that of the ensemble, reflecting challenges posed by limited training data and class imbalance. By comparison, the proposed ensemble achieved a stronger balance across all classes. A comparative summary of proposed ensemble, recent benchmark work, and deep learning models is provided in Table 5. When compared with

Table 3
Comparison of resampling methods on classification performance.

Method	Macro-F1 (%)	Precision (%)	Recall (%)	Std Dev	Notes
SMOTE	0.868	0.876	0.868	0.049	Stable across folds
Borderline-SMOTE	0.881	0.889	0.882	0.057	Higher recall, lower precision

Table 4
Cross-validation performance of individual base classifiers with the proposed model.

Model	Test Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
RF	88.2	89.3	88.3	88.1
XGBoost	86.8	87.7	86.7	86.5
SVM	86.4	87.0	86.4	86.1
Voting Ensemble	92.6	91.7	91.6	91.4

existing benchmark works, the ensemble shows competitive or superior performance. For instance, [14] reported RF and SVM classifiers achieving up to 92 % accuracy and 91–92 % macro-F1 on a binary cucumber leaf dataset with 440 samples. Similarly, [15] achieved 88 % accuracy and macro-F1 using RF on 3200 sunflower seed samples, also limited to binary classification. Unlike these studies, the present work addresses a multiclass aircraft damage classification problem, where balancing class-wise performance is more challenging. The ensemble's consistent macro-F1 underscores its robustness and adaptability, outperforming CNN baselines and aligning favorably against prior RF- and SVM-based approaches in related hyperspectral domains. The CNN baseline achieved 85.8 % accuracy but exhibited a lower macro-F1 (80.9 %), reflecting its sensitivity to class imbalance and data size. In contrast, the ensemble classifier provided consistent macro-F1, highlighting its robustness in imbalanced settings. These findings align with recent reports in hyperspectral classification literature, where ensemble methods have been shown to outperform deep networks when training data is constrained. Recent studies have explored advanced deep learning strategies to mitigate these challenges. For instance, shrinkage-based relation networks with out-of-distribution (OOD) augmentation have been proposed for fault detection under zero-faulty or highly imbalanced conditions [20–21]. Such approaches demonstrate that incorporating shrinkage mechanisms and OOD augmentation can substantially enhance robustness in domains with scarce or imbalanced data. While this study focuses on ensemble-based learning, future work could integrate similar augmentation and shrinkage-inspired strategies into HSI pipelines to further strengthen generalization in aerospace applications with limited training samples.

Beyond accuracy, the macro-averaged F1-score was used to measure balanced performance across all classes, including minority damage types like "Peeling" and "Rust Points". The ensemble achieved the highest F1-score across these metrics, demonstrating its robustness. Furthermore, a confusion matrix shown in Fig. 10 heat map was plotted to visualize misclassifications, revealing that classes with overlapping spectral signatures, such as "Burnt Case" and "Heavily Burnt," had occasional confusion but overall high separability. The final model's prediction confidence was bolstered by class-specific metrics and minimal overfitting as inferred from the consistency in train-test and cross-validation accuracy. To support generalizability, feature selection was enforced based on feature importance scores from the RF classifier, using a threshold of 0.02. This ensured only the most discriminative spectral features (such as peak intensity and regional spectral variance across 600–800 nm) were retained, reducing dimensionality and boosting interpretability.

The confusion matrix in the given Fig. 12 indicates that the model achieves high accuracy across most classes, with several—such as Clean (burnt case), Heavily Burnt Case, Clean Corrosion Point, and Corrosion cycle-5 being predicted with perfect or near-perfect accuracy. However, there are a few notable misclassifications: Burnt Case (Class 1) is misclassified as Rust Point and Lightning Strike in several instances, and Corrosion Cycle 3 shows minor confusion with Clean Corrosion Point. These errors suggest that certain damage types, especially those involving burnt or corroded surfaces, may share overlapping spectral patterns, making them harder to distinguish. Despite these overlaps, the overall distribution is tightly concentrated along the diagonal, reflecting the strong generalization of the model. This indicates that the ensemble classifier is effective in capturing key spectral features for most material conditions, with room for improvement in differentiating subtly varying surface states. Details evaluation of the model can be understood from Table 6. 'F1'- score of the classes 0, 2, 3, 4, 5, 6, 7, and 9 is above 0.9, suggesting that the model balances precision and recall well. These classes are being classified consistently and accurately. It suggests that the model is not overfitting. For classes 1 and 8, an 'F1'-score in the range of 0.8 means less reliability for detecting these classes. For class 1, recall is also low, which means that true positives are missed. For class 8, it predicts more false positives but recovers most true positives. The reason behind this could be the similarity in these patterns, which hinders differentiation from other patterns. Further analysis of misclassifications revealed that classes 1 and 8 consistently underperformed. As shown in Fig. 13, t-Distributed Stochastic Neighbor Embedding (t-SNE), the feature distributions of these two classes strongly overlap in the visible spectral range, limiting the ability of classifiers to separate them reliably.

The analysis, as shown in Table 7, sheds light on how individual models perform when used alone, offering a fresh perspective on their contributions. With all 19 features and optimal settings, the model achieves a baseline accuracy of 92.6 % and a macro F1-score of 91.9 %, serving as the reference point. When relying solely on Random Forest, accuracy drops to 88.7 % and the macro F1-score to 88.5 %, indicating a solid but less robust performance. Using XGBoost alone results in an accuracy of 89.3 % and a macro F1-score of 89.2 %, showing a slight edge over Random Forest. The Support Vector Machine, when used independently, yields an accuracy of 89.9 % and a macro F1-score of 89.8 %, suggesting it holds its own among the single models. Removing SMOTE has a profound effect, reducing accuracy to 75.1 % and macro F1-score to 69.8 %, underscoring its necessity for balancing the dataset. Feature-wise, excluding wavelength features leads to a significant drop to 83.0 % accuracy and 82.5 % macro F1-score, while omitting peak features lowers performance to 90.1 % and 90.0 %, respectively. It was found that removing derivative-based features improved the model's accuracy to 91.2 % with a macro F1-score of 91.1 %. This change to evaluating one model at a time suggests that, although each classifier brings something different, the wavelength and the peak features are the most important. The use of SMOTE also proved essential for maintaining consistent results. These findings show that combining carefully selected spectral features with ensemble classifiers can lead to a reliable model for detecting aircraft damage in real-time for hyperspectral images. The final validation confirms that blending

Table 5
Comparison of the ensemble classifier with the CNN baseline and recent benchmark work.

Model		Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)	Remark
Ensemble (RF+XGBoost+SVM)		92.6	91.7	91.6	91.4	Balanced across classes
1D CNN		85.8	85.0	80.7	80.9	Sensitive to class imbalance and size of data
[14] 2024	RF	92	93	90	92	440 cucumber leaves samples of binary class
	SVM	91	88	94	91	
[15] 2025	RF	88	88	88	88	3200 sunflower seed samples of binary class
	SVM	89	89	89	89	
	XGBoost	85	85	85	85	

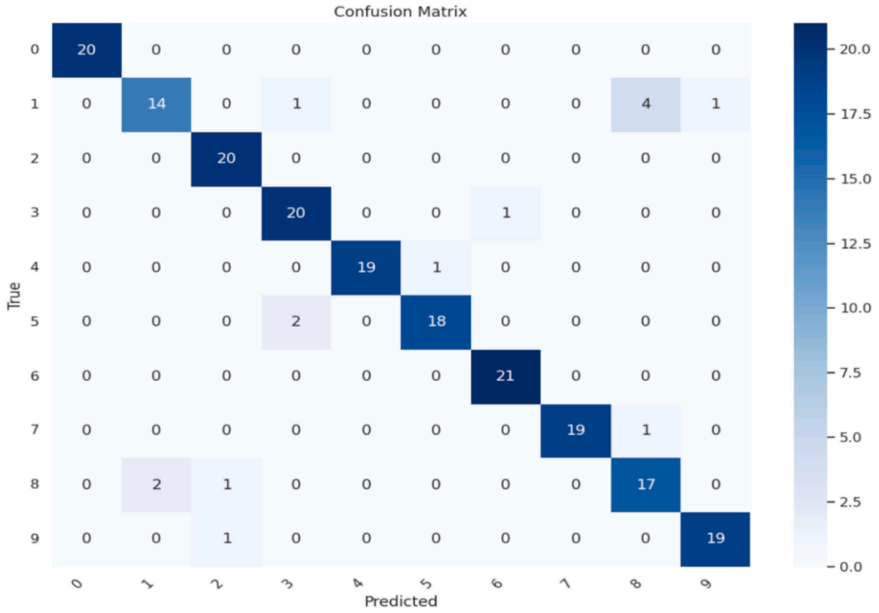


Fig. 12. Confusion Matrix of the damage classification.

Table 6
Comparative results of the damage classification.

Class	Precision	Recall	F1-Score	Accuracy
0	1.0	1.0	1.000	1.00
1	0.882	0.75	0.811	0.70
2	0.904	0.95	0.927	1.00
3	0.875	1.0	0.933	1.00
4	1.0	0.95	0.974	0.95
5	0.947	0.9	0.923	0.85
6	0.954	1.0	0.977	1.00
7	0.947	0.9	0.923	0.95
8	0.809	0.85	0.829	0.90
9	0.944	0.85	0.895	0.85
Accuracy				0.926

domain-specific spectral feature engineering with ensemble learning produces a classifier that is not only accurate and consistent but also suitable for practical deployment in real-world aircraft damage detection.

Limitations

This study presents a practical method for identifying aircraft damage using HCI together with a combined machine learning model. Spectral data were collected from real and simulated damage samples. The data was then analyzed using an ensemble model made up of RF, XGBoost, and SVM. By using a soft voting ensemble model approach, an overall accuracy of 92.6 % was achieved in classifying the different types of damage. The combination of these models helps to improve both performance and understanding of

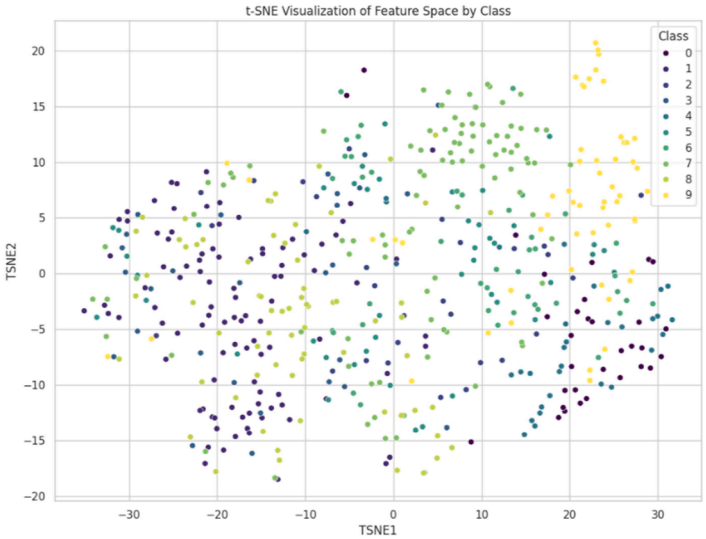


Fig. 13. t-SNE embedding of spectral features.

Table 7
Ablation analysis.

Configuration	SMOTE	Model	Accuracy	Macro F1-Score
Baseline	Yes	Voting	0.926	0.919
No SMOTE	No	Voting	0.751	0.698
RF Only	Yes	RF	0.887	0.885
XGBoost Only	Yes	XGBoost	0.893	0.892
SVM Only	Yes	SVM	0.899	0.898
No Scaling	Yes	Voting	0.899	0.898
All except Peak	Yes	Voting	0.901	0.9
All except Stats	Yes	Voting	0.916	0.915
All except FFT	Yes	Voting	0.926	0.925
All except Wavelength	Yes	Voting	0.83	0.825
All except Derivative	Yes	Voting	0.912	0.911

the results. This technique shows promise for use in automated aircraft inspection systems. The pronounced overlap of classes 1 and 8 observed in the t-SNE space reflects their inherent spectral similarity in the visible domain. This explains their reduced classification performance. Future work could explore augmentation strategies such as spectral jitter, mix-up, and hybrid approaches like SMOTE-Edited Nearest Neighbors to increase class separability and improve performance on these weaker categories.

The limitations of the present study are as follows:

a) The dataset, while diverse, is limited to 530 samples, which may not fully represent the variability in the damage patterns across the different aircraft types and environmental conditions.
b) The data was collected in a controlled laboratory condition, which may or may not entirely replicate the complexities of the real-world inspections conducted under variable lightning, weather, or structural surface geometries.
c) Although the ensemble model increases accuracy with the imbalanced dataset, it becomes computationally complex. Complexity can be reduced if a balanced dataset is available.

Ethics statements

Participant data has been fully anonymized to ensure privacy, and the data redistribution policies of Kaggle have been fully complied with, adhering to ethical guidelines for data usage.

CRediT author statement

Nilima Zade: Writing – original draft, Writing – review & editing, Investigation, Validation, project administration
Aditya Gupte: Conceptualization, Investigation, Validation, Writing – original draft, formal analysis, Resources.
Pranshu Gupta: Writing –Methodology, Writing – review & editing, Implementation.
Nicolas Detalle: Writing review & editing, Resources, Conceptualization.

Anthony Mannion: Writing review & editing, Resources, Conceptualization.
Robert Voyle: Writing review & editing, Resources, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The project was largely funded by the Government of the Hong Kong Special Administrative Region of the People's Republic of China through the Innovation and Technology Commission, Innovation and Technology Support Programme, as project ITS/017/22FP 'Aerostructure Digital Twin', in addition to the consortium members of the ASRC. The authors extend their deepest gratitude to the authorities and staff of Aviation Services Research Centre (ASRC), The Hong Kong Polytechnic University (PolyU), and Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India, whose collaborative efforts were instrumental in this work. We sincerely express our gratitude to the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India, for providing the research support fund and the invaluable platform and support that made this research work possible.

Data availability

Data will be made available on request.

References

[1] A. Katunin, M. Nagode, S. Oman, A. Cholewa, K. Dragan, Monitoring of Hidden Corrosion Growth in Aircraft Structures Based on ᴅ-Sight Inspections and Image Processing, Sensors 22 (19) (2022).
[2] A. Wronkowicz-Katunin, A. Katunin, K. Dragan, Reconstruction of Barely Visible Impact Damage in Composite Structures Based on Non-Destructive Evaluation Results, Sensors 19 (21) (2019).
[3] P. Lafiosca, I.-S. Fan, Review of non-contact methods for automated aircraft inspections, Insight - Non-Destruct. Test. Condit. Monitor. 62 (12) (2020) 692–701.
[4] A. Katunin, K. Dragan, M. Dziendzikowski, Damage identification in aircraft composite structures: a case study using various non-destructive testing techniques, Compos. Struct. 127 (September 2015) 1–9.
[5] A. Cugniere, O. Tusch, A. Mosenbacher, How fraud detection technologies can help to detect damages in aircraft structures, Procedia Struct. Integr. 38 (2022) 168–181.
[6] J. Amraei, A. Katunin, D. Wachla, K. Lis, Damage assessment in composite plates using extended non-destructive self-heating based vibrothermography technique, Measurement 241 (February 2025).
[7] T. Berghout, Military Training Aircraft Structural Health Monitoring Leveraging an Innovative Biologically Inspired Feedback Mechanism for Neural Networks, Machines 13 (3) (2025).
[8] A. Tabatabaeian, B. Jerkovic, P. Harrison, Barely visible impact damage detection in composite structures using deep learning networks with varying complexities, Compos. Part B 264 (2023).
[9] Y. an, J. Ren, H. Zhao, J.F.C. Windmill, W. Ijomah, J. de Wit, J. von Freeden, Non-Destructive Testing of Composite Fiber Materials With Hyperspectral Imaging—Evaluative Studies in the EU H2020 FibreEUse Project, IEEe Trans. Instrum. Meas. 71 (2022) 1–13.
[10] L. Qing-song, D. Wen-jing, L. Bo, L. Kai-ge, T. Shen-lan, Near Infrared Hyperspectral Identification of Surface Damage on Aircraft Wings, Spectroscopy Spectral Anal. 44 (11) (2024) 3069.
[11] J. Długosz, P.B. Dao, W.J. Staszewski, T. Uh, Damage Detection in Glass Fibre Composites Using Cointegrated Hyperspectral Images, Sensors 24 (6) (2024).
[12] Imad Gohar, Weng Kean Yew, Abderrahim Halimi, John See, Review of state-of-the-art surface defect detection on wind turbine blades through aerial imagery: challenges and recommendations, Eng. Appl. Artif. Intell. 144 (2025) 109970, https://doi.org/10.1016/j.engappai.2024.109970. ISSN 0952-1976.
[13] Taspia Shawkat Chowdhury, Fatima Mohsin, Mita Mita, Mohammad Ehsan, A Critical Review on Gas Turbine Cooling Performance and Failure Analysis of Turbine Blades, Int. J. Thermofluids" 18 (2023) 100329, https://doi.org/10.1016/j.ijft.2023.100329.
[14] B. Mandrapa, K. Spohrer, D. Wuttke, U. Ruttensperger, C. Dieckhoff, J. Müller, Machine learning-based hyperspectral wavelength selection and classification of spider mite-infested cucumber leaves, Exp. Appl. Acarol. 93 (3) (2024) 627–644, https://doi.org/10.1007/s10493-024-00953-0.
[15] Bright Mensah, Jarrad Prasifka, Brent Hulke, Ewumbua Monono, Xin Sun, Detection of insect-damaged sunflower seeds using near-infrared hyperspectral imaging and machine learning, Smart Agricult. Technol. 12 (2025) 101110, https://doi.org/10.1016/j.atech.2025.101110. ISSN 2772-3755.
[16] Aaditya Ahire, Nilima Zade, Umeed Mujawar, Dimple Mehta, Ketan Kotecha, Meteorological drought severity forecasting utilizing blended modelling, MethodsX. 15 (2025) 103456, https://doi.org/10.1016/j.mex.2025.103456. ISSN 2215-0161.
[17] N. Krishnasamy, N. Zade, D. Khambholia, R. Henry, A. Gupte, Ensemble Deep Learning Framework for Hybrid Facial Datasets Using Landmark Detection: state-of-the-Art Tools, J. Computat. Cognit. Eng. (2025), https://doi.org/10.47852/bonviewJCCE52024451.
[18] O. Schackmann, O.A.M. Reyes, V. Memmolo, D. Lozano, J. Prager, J. Moll, P. Kraemer, Machine learning strategies with ensemble voting for ultrasonic damage detection in composite structures under varying temperature or load conditions, Struct. Health Monit. (2025).
[19] D. Gautam, A. Bhattarai, R. Rupakhety, Machine learning and soft voting ensemble classification for earthquake induced damage to bridges, Eng. Struct. 303 (2024).
[20] Zuoyi Chen, Hong-Zhong Huang, Zhongwei Deng, Jun Wu, Shrinkage mamba relation network with out-of-distribution data augmentation for rotating machinery fault detection and localization under zero-faulty data, Mech. Syst. Signal. Process. 224 (2025) 112145, https://doi.org/10.1016/j.ymssp.2024.112145. ISSN 0888-3270.
[21] Zuoyi Chen, Hong-Zhong Huang, Jun Wu, Yuanhang Wang, Zero-faulty sample machinery fault detection via relation network with out-of-distribution data augmentation, Eng. Appl. Artif. Intell. 141 (2025) 109753, https://doi.org/10.1016/j.engappai.2024.109753. ISSN 0952-1976.