VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY
**UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI**

# REPORT

**CONTEST**
**“ USTH SCIENTIFIC RESEARCH FOR STUDENTS 2026”**

# Fault diagnosis of Aircraft Engine Fan Blades using ANSYS-based Modal Analysis and Random Forest Algorithm

Subcommittee: A

Student’s name: Bùi Nhật Minh - Nguyễn Hoàng Minh

Student’s ID: 2410581 - 2410601

Department: Department of Technology and Engineering

Major: Aeronautical Engineering

Supervisor: Dr. Bùi Quang Thành

***Hanoi, 2026***

UNIVERSITY OF SCIENCE & TECHNOLOGY OF HANOI
TRƯỜNG ĐẠI HỌC KHOA HỌC & CÔNG NGHỆ HÀ NỘI

**Comments from the supervisor on the scientific contributions of the report:**

The report successfully combines engineering simulation and artificial intelligence to address an important problem in structural health monitoring of aircraft engine. The students developed an automated simulation workflow for generating vibration data, implemented tree-based machine learning models (Random Forest and Gradient Boosting) for crack detection and geometry prediction, and analyzed the physical significance of vibration modes using feature importance. The work demonstrates good technical competence and an ability to integrate knowledge from finite element analysis, vibration engineering, and machine learning. The proposed approach provides a useful basis for future studies involving experimental measurements and larger datasets.

*Hanoi, 15 / 07 / 2026*

**Supervisor**

*(Signature and full name)*

*Bùi Quang Thành*

**ABSTRACT**

Surface cracks on aero-engine fan blades represent a critical safety hazard, and their early detection is the main objective of structural health monitoring. In this article, our work develops a finite-element modal-signature database for training machine-learning models to detect, localize and size such defects. A 3D CATIA model of a fan blade (span 2.88 m, chord 0.9 m, mean thickness 14 mm) was imported into ANSYS Workbench 2024 R1 and discretized into elements. Cracks are simulated by the stiffness-reduction (smeared-crack) approach, degrading the Young's modulus of the element cluster occupying the crack to 0.1% of the original intact value while preserving the mass. Since the blade is strongly curved and twisted, the crack depth was measured along a local surface normal rather than along a fixed global axis. A PyMAPDL script automated Latin-Hypercube sampling of four parameters of the defect (spanwise position, chordwise position, length and depth) and solved a modal analysis for each case, extracting the first ten natural frequencies. About 600 cases were conducted, of which 515 produced a valid crack. The induced frequency shifts proved small (median 0.13%, maximum 5.89% of the intact value), showing that the detectability of a crack on the blade is governed by the ratio of the crack-induced frequency shift to the measurement noise floor. A companion dataset with Gaussian relative noise (0.001, 0.005 and 0.01) and simulated healthy cases was produced so the machine-learning models can be evaluated under realistic measurement uncertainty rather than on idealized data.

**Keywords:** structural health monitoring, modal analysis, crack detection, finite element method, fan blade, machine learning

## I/ INTRODUCTION

### 1.1/ Global context

The fan blade is the first rotating stage of a modern high-bypass turbofan and exposed directly to the outside environment. During the intake of air, every entering objects strike the fan first. Foreign object damage (FOD) is consequently described as both frequent and inevitable for fan blades [1].

The danger lies in the effects it leaves behind. The fatigue stress on an undamaged blade is below the fatigue limit and could be tolerated indefinitely; a nick on the leading edge raises local stress through the notch effect, so load cycles accumulate at the notch until a crack forms and the blade fractures [2]. Approximately half of all component damage in jet engines is attributed to fatigue [3], and the initiating defect is often invisible to visual inspection [1]. The consequences are not hypothetical: 315 uncontained engine failures have been caused by fan blade detachment - about 24.3% of all engine-related accidents [4] - and a recent commercial flight ended fatally when a broken fan blade destroyed an engine and cost the life of a passenger on board [5].

It is therefore vital to detect the crack as early as possible. Current practice relies on scheduled non-destructive testing (NDT), which is sensitive but requires physical access, inspects only small regions, and runs at fixed intervals [5,6]. Structural health monitoring (SHM) based on vibration measurement is the leading alternative: a single dynamic

measurement responds to a defect anywhere on the blade, requires no direct access to the crack, and can be repeated as often as desired.

### 1.2/ Scientific background

For an undamped linear elastic structure, the natural frequencies follow the eigenvalue problem,

$$(K - \omega^2 M)\varphi = 0, \quad f = \omega/2\pi \qquad (1)$$

where K and M are the global stiffness and mass matrices, $\omega$ is the angular eigenfrequency and $\varphi$ is the corresponding mode shape. A crack is a local discontinuity across which load cannot be transmitted, which leads to a local loss of stiffness, so that K becomes K - ΔK, M is unchanged since the surface crack removes a negligible volume of material. Since the eigenvalues are proportional with the stiffness-to-mass ratio, the natural frequencies must therefore decrease [7].

However, detection alone would be insufficient, the key is to determine the position as well as the severity of the crack, which is contained in the pattern of the shifts. Applying a first-order perturbation to (1) gives, for an arbitrary mode i,

$$\Delta f_i/f_i \approx -(\varphi_i^T \Delta K \varphi_i)/(2 \varphi_i^T K \varphi_i) \qquad (2)$$

the relative shift of mode i is the fraction of that mode's strain energy which the crack destroys. Since each mode has a different strain-energy distribution, one crack perturbs the ten modes by ten different amounts, and this is the pattern that carries the information [8]. Equation (2) gives the signed relative frequency shift Δf/f (negative, since stiffness loss lowers frequency); where later sections report a positive frequency drop, this refers to the magnitude |Δf/f| rather than the signed value.

Finding the crack parameters from measured frequencies is an ill-posed inverse problem: different crack configurations can produce similar frequency vectors, and the mapping cannot be inverted analytically. The standard procedure is to solve the forward problem many times, building a database of simulated damage scenarios, then search it for the case best matching the measurement.

### 1.3/ Literature review

Research on vibration-based crack identification has followed three broad strands. The first uses natural frequencies alone - attractive since a single accelerometer suffices and the result is insensitive to sensor placement; two-step methods localizing and sizing damage from frequencies alone have been validated on cracked beams [8], where frequencies fall monotonically with crack depth [7]. The second uses mode shapes or modal curvature, which bring more spatial information but require dense sensor arrays and are far more noise-sensitive [9,10]. The third replaces the explicit database search by machine learning: Random Forest and Artificial Neural Networks have served as search tools over beam damage databases [11], and on wind-turbine blades - the closest published analogue - a Random Forest classifier separated intact from damaged blades with 94.66% accuracy while identifying damage size, type and position [12].

### 1.4/ Problem statement and research questions

Two gaps follow from this body of work. First, almost all of these studies concern beams, blades of the order of one meter, or scaled printed models. The frequency shift caused by

a crack depends on the ratio of the local stiffness loss to the global stiffness of the structure, so a defect easily detected on a one-meter specimen may not be detectable on a three-meter fan blade; moreover, a fan blade is not a beam but a thin, cambered, twisted shell, on which the depth of a surface crack must be measured perpendicular to a surface whose normal changes continuously - a difficulty beam studies never encounter. Second, most published databases are noise-free, while the shift caused by a defect can be of the same order as the uncertainty of a frequency measurement, so a model trained and tested on clean data cannot be trusted in real measurements. Experiment already indicates where the limit lies: on a scaled blade, acceleration transmissibility could not identify defects of 1.0 cm and below that the numerical model could [10]. Without quantifying both the true crack signature on full-scale complex geometries and the limiting impact of measurement noise, it is not known whether a frequency-based system on a blade of this size is possible, nor for which cracks.

Three questions follow:

(i) How can the prescribed position, length and depth of a surface crack be represented on a strongly curved and twisted blade, in a way that is geometrically correct and able to be repeated several times ?

(ii) How large is the resulting frequency signature on a blade of this scale, and how does it vary with the position and size of the crack ?

(iii) Is that signature distinguishable from the noise of a realistic frequency measurement - if yes, for which cracks ?

## II/ OBJECTIVES

The overall objective of this project is to develop a vibration-based system capable of detecting a surface crack on an aero-engine fan blade and of determining its location and its size from the first ten natural frequencies. This requires generating a finite-element modal-signature database in which the crack is represented faithfully on the curved and twisted blade surface, and the magnitude of the resulting frequency signature is characterized. A two-stage pipeline of tree-based classifier and regressor models, selected by grid search under case-grouped cross-validation, is then trained on this database and evaluated against realistic measurement noise, so that detection and sizing performance can be characterised as a function of noise level and crack severity.

## III/ MATERIALS AND METHODS

### 3.1/ Modelling assumptions and scope

The fan assembly comprises twenty-one blades on a spinner; this study models a single blade, clamped at its root and analyzed at rest.

In a complete bladed disc the blades couple through the disc, so the assembly possesses not isolated blade modes but families characterized by nodal diameters. Reproducing these requires a cyclic-symmetry model - prohibitive when several hundred solutions are needed. The single-blade idealization with a clamped root is standard in this field [9,11,12] and is defensible because a crack is a local perturbation of the blade itself: it changes that blade's stiffness, and the resulting change in its dynamics is what the method detects.

### 3.2/ Blade geometry and material

The blade was supplied as a STEP (AP203) file exported from CATIA V5. Geometric interrogation revealed a single solid bounded by seven B-spline faces: two large faces of approximately $2.85 \times 10^6$ mm² forming the suction and pressure surfaces, and five narrow faces forming the leading edge, the trailing edge, the root section and the tip. The span measured along the blade axis is 2.88 m, the chord approximately 0.9 m, the enclosed volume $3.98 \times 10^7$ mm³, and the mean thickness - obtained as the ratio of volume to planform area - 14 mm. The blade is therefore a thin, highly cambered and twisted shell with a thickness-to-chord ratio of the order of 1.5%. The root is a plane section, without a dovetail attachment. The material is Ti-6Al-4V, with the properties: Young's modulus E = 113.8 GPa; Poisson's ratio v = 0.342; density $\rho = 4430$ kg/m³ [13].

**3.3/ Finite-element discretization and mesh convergence**

The blade was meshed in ANSYS Mechanical 2024 R1 (patch-conforming tetrahedra, quadratic order, SOLID187). The mesh is governed by a proximity control (gap factor 2), forcing at least two quadratic elements through the wall, and a curvature control (15°, minimum 1 mm) resolving the edge radius, giving 654,966 elements / 1,115,387 nodes.

Convergence was verified on the intact blade at three densities (maximum element size 0.030, 0.025, 0.015 m). The first ten frequencies proved identical across all three meshes: the solution is fixed by the proximity and curvature controls and is insensitive to further refinement. As an independent check, the model was solved in both ANSYS Mechanical and Mechanical APDL (via PyMAPDL), the two solvers agreeing to the last significant figure. This converged mesh was generated once and re-used unchanged for all six hundred cases, so the discretization error is identical across the database and cannot contribute to the differences between cases.

**3.4/ Crack model**

Cracks were represented by the stiffness-reduction (smeared-crack) approach, damage being represented not by an explicit mesh but by an equivalent degradation of the stiffness of the elements it occupies [14]. Every element whose centroid falls inside the crack footprint receives

$$E\_crack = \alpha \cdot E\_intact, \alpha = 10^{-3}, \rho\_crack = \rho\_intact \quad (3)$$

The density is deliberately unchanged: a surface crack removes a negligible volume, and the effect to capture is the loss of stiffness, not of mass. The factor $10^{-3}$ is small enough that the cracked region no longer transmits load, yet large enough to avoid the spurious local modes that arise when nearly stiffness-free elements retain full inertia [15].

The principal difficulty is the geometry. On a curved, twisted shell the surface normal changes continuously, so crack depth cannot be measured along any fixed global direction. A blade-aligned frame was therefore obtained by principal-component analysis of the nodal coordinates,

$$X - \bar{X} = U \Sigma V^{\mathrm{T}}, [e_1, e_2, e_3] = V \quad (4)$$

and the elements whose centroids **c** lie within a footprint of width *w* spanwise and length *L* chordwise selected,

$$|(c - c_0) \cdot e_1| \leq w/2 \; and \; |(c - c_0) \cdot e_2| \leq L/2 \quad (5)$$

The through-thickness direction was then recomputed locally as the minimum-variance axis n of that specific cluster, and the depth *d* measured along this local normal:

$$(c - \bar{c}) \cdot n \geq max[(c - \bar{c}) \cdot n] - d \quad (6)$$

The depth is thereby measured perpendicular to the actual curved surface at the crack location; a global axis would distort it along the crack wherever the surface twists. This is the difficulty beam studies never encounter, and the answer to research question (i).

**3.5/ Boundary conditions, modal analysis and parametric study**

A fully-clamped root, a Block Lanczos solve for the first ten modes [16], and element result expansion disabled (only eigenvalues being required). The intact blade gives the baseline $f_0$; for each cracked case the ten frequencies are retained directly as the feature vector, the relative change

$$\Delta f_i / f_{0i} = (f_{0i} - f_i)/f_{0i} \quad (7)$$

being stored for interpretation and for assessing the signature against the noise floor.

The four crack parameters were sampled by Latin-Hypercube design, which partitions each range into N equiprobable intervals and draws one sample from each - far more uniform coverage than random sampling at the same budget, without the exponential cost of a full grid.

The bounds follow from equation (2): the frequency change is proportional to the strain energy stored at the crack location, so two identical cracks give very different signatures depending on where they lie - one at high modal curvature perturbs the eigenvalues strongly, one near a nodal line hardly at all. Since the signature on a blade of this size is already close to the noise floor, sampling predominantly in low-sensitivity regions would have produced a database in which most cases were indistinguishable from an intact blade. The sampled crack parameters are summarized in table below:

| Parameter | Range |
|---|---|
| **Spanwise position** | 2 - 45% of span |
| **Chordwise position** | 5 - 95% of chord |
| **Crack length** | 10 - 200 mm |
| **Crack depth** | 3-12 mm |

Table 1: Crack parameters sampled (Latin-Hypercube).

The whole loop - mesh import, material assignment, element selection, solution, export - was automated in Python with PyMAPDL; six hundred cases were computed, each taking one to two minutes.

**3.6/ Noise-augmented dataset**

The finite-element database is noise-free, whereas a measured frequency carries an uncertainty of the same order as the crack signature itself. Each frequency was therefore perturbed by a multiplicative Gaussian error,

$$\tilde{f}_i = f_i (1 + \varepsilon_i), \varepsilon_i \sim N(0, \sigma^2), \sigma \in \{0.001, 0.005, 0.010\} \quad (8)$$

with thirty independent realizations per case per noise level. One hundred simulated healthy cases were added by perturbing the baseline frequencies themselves, so the intact class is represented by pure measurement noise and carries a binary label, allowing a classifier to separate an intact blade from a cracked one.

### 3.7/ Machine Learning Methodology

**Dataset.** The dataset comprises 55,350 simulated readings from 615 tracked cases: 515 distinct cracked geometries and 100 noise-augmented realisations of a single healthy baseline geometry. Each case carries 90 noisy readings (three synthetic noise levels - 0.001, 0.005, 0.010 - with 30 independent realisations each), approximating the measurement variability of a real accelerometer-based system and allowing repeated-measurement aggregation to be evaluated as an operating mode in its own right (see the aggregation-mode analysis below). Each reading records ten natural frequencies ($f_1$-$f_{10}$, from the modal analysis of Section 3.5), ten shifts relative to a healthy baseline ($\Delta f_1$-$\Delta f_{10}$), the true crack label and, for cracked cases, the crack's spanwise position, chordwise position, length, and depth. No values were missing and all frequencies were verified positive, as physically required. One property is reported as a limitation rather than corrected: the 100 healthy cases are noise-augmented realisations of a single physical healthy geometry (case-level frequency variance 1.0-7.4× lower than the cracked population across the ten modes), not 100 independently simulated healthy blades.

**Feature selection.** Because $\Delta f_i$ is defined relative to a healthy baseline, $f_i$ and $\Delta f_i$ were hypothesised to carry overlapping information; this was tested rather than assumed. Their Pearson correlation was found to be exactly -1.000 for every one of the ten modes, and a numerical fit recovered the underlying exact relation:

$$\Delta f_i = 1 - f_i / \bar{f}_i^{(healthy)}$$

where $\bar{f}_i^{(healthy)}$ is the mean i-th natural frequency over the healthy cases - so each $\Delta f_i$ is a deterministic linear function of its $f_i$ and carries no independent information. Consistent with this, three candidate feature sets (f-only, Δf-only, f+Δf) produced statistically indistinguishable cross-validated performance for the classifier and every regressor. The ten raw frequencies ($f_1$-$f_{10}$) were retained as the sole input feature set for the final models, chosen over Δf because a raw frequency is directly measurable without first knowing a calibrated healthy baseline.

**Leakage prevention.** Because each physical case contributes 90 correlated readings, all data splitting was performed at the case level, never the row level: a case's readings are always kept entirely within the training set or entirely within the test set. The dataset was split once into a training set (492 cases; 44,280 readings) and a held-out test set (123 cases; 11,070 readings), stratified by health status at the case level. The test set was scored for final reporting at each modelling stage; as the methodological caveat below makes explicit, however, the pipeline was developed iteratively and held-out results were consulted more than once across that development process, so the test set is not perfectly “unseen” in the strictest sense. All model selection and tuning proper used five-fold, case-grouped cross-validation on the training set only (stratified by health status for the classifier; restricted to cracked cases for the regressors).

**Model selection and tuning.** A trivial baseline - a classifier predicting the majority class, a regressor predicting the training-set mean - was evaluated first, and every selected model is reported against it. Four classifier families (logistic regression, Random Forest [17], Extra Trees, histogram-based gradient boosting) and five regression candidates (mean

baseline, Ridge, Random Forest [17], Extra Trees, gradient boosting) were compared via grid search under the case-grouped cross-validation above, all implemented with scikit-learn [18]. Classification was scored by average precision, a threshold-independent ranking metric [20]: an initial attempt with threshold-dependent scores (F-beta, balanced accuracy at the default 0.5 cutoff) scored the eventual best model no higher than the trivial baseline, because under the 84%/16% cracked/healthy imbalance most predicted probabilities land above 0.5 regardless of the true label - a single fixed threshold cannot separate a genuinely skilled classifier from one that always predicts the majority class, whereas average precision, evaluating the full precision-recall trade-off across all thresholds, can. Regression was scored by mean absolute error. Class-balanced Random Forest was selected for classification; gradient boosting for spanwise position, chordwise position and length; Random Forest for depth.

The classifier search space totalled 302 configurations (5-fold cross-validated): Random Forest and Extra Trees shared `n_estimators` ∈ {200, 500}, `max_depth` ∈ {none, 10, 20}, `min_samples_leaf` ∈ {1, 3, 5} and `max_features` ∈ {√p, 0.5, 1.0}, with `class_weight` ∈ {none, balanced, balanced_subsample} for Random Forest (162 configurations) but {none, balanced} for Extra Trees (108); logistic regression contributed 8 (`C` ∈ {0.01, 0.1, 1, 10} × 2 class weights) and histogram gradient boosting 24 (`max_iter` ∈ {200, 500} × `max_depth` ∈ {none, 6, 10} × `learning_rate` ∈ {0.05, 0.1} × 2 class weights): 162+108+8+24 = 302. Unlimited tree depth was kept for the two ensembles - an early timing check showed it made the Random Forest search alone take ≈ 7.5 hours - while the four regression models' separately tuned grids excluded it, since repeating that cost across four targets was impractical.

**Verification of the metric fix.** The average-precision correction above was first validated cheaply, by re-scoring only the 4 already-tuned configurations plus a class-balanced variant of each; a follow-up experiment therefore re-ran the complete 302-configuration grid from scratch under average precision, to check whether an exhaustive search would have found a materially better model than the inexpensive fix. The result is reported alongside the classification results below.

**Threshold calibration.** The operating threshold was calibrated from out-of-fold training probabilities via the precision-recall curve, maximising precision subject to at least 95% recall on the cracked class - a safety-first priority, a missed crack being costlier than a false alarm. Row-level and case-level thresholds were calibrated separately because they proved non-interchangeable: averaging a case's 90 readings compresses the probability distribution, so a threshold tuned for single readings systematically over-predicts “cracked” on averaged probabilities; the case-level threshold was therefore tuned independently on case-averaged out-of-fold probabilities.

**Evaluation metrics.** Classification is reported using accuracy, balanced accuracy, precision, recall, and F1 for the cracked class, with balanced accuracy emphasised because plain accuracy is uninformative under this imbalance (the majority-class baseline reaches 83.7% accuracy by construction):

$$\textit{Balanced accuracy} = \tfrac{1}{2}\,(TPR + TNR)$$

where TPR and TNR are the true-positive (cracked) and true-negative (healthy) rates. Regression is reported using MAE, RMSE, $R^2$, and a normalised MAE, enabling direct comparison across four targets whose physical ranges differ by more than two orders of magnitude:

$$nMAE = MAE / (max(y\_train) - min(y\_train))$$

**Extrapolation stress test.** A separate experiment tested whether the regressors generalise beyond the crack-severity range seen in training. A combined length-and-depth severity score was computed for all 515 cracked cases; models were retrained on the smallest/medium 80% of cases and evaluated on the largest 20% (and the reverse), with case membership respected so no physical case contributed readings to both sides of this severity split.

## IV/ RESULTS AND DISCUSSION

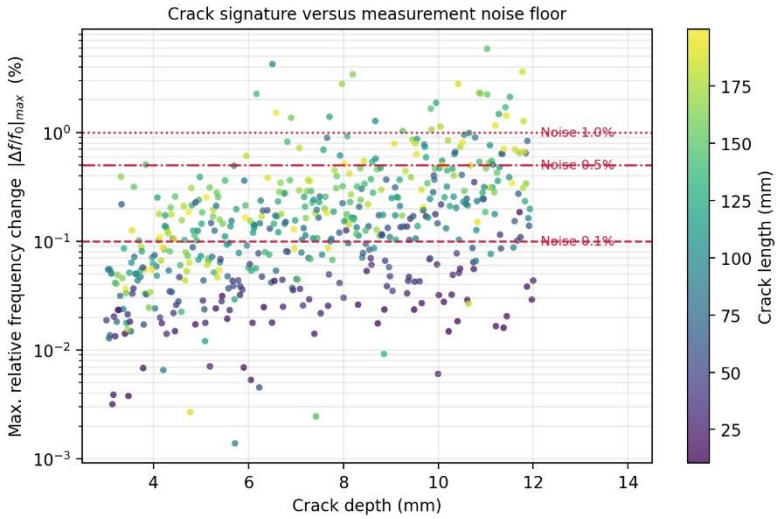
### 1/Results:

### ANSYS Simulation Results

**Modal characteristics of the intact blade**. The intact blade yields the ten natural frequencies, from 2.147 Hz (fundamental flap-wise bending) to 153.25 Hz. The lower modes are bending-dominated, with torsional and higher-order modes appearing from the mid-range; the coexistence of these mode families is what gives each crack location a distinctive signature, as anticipated by equation (2).

**The crack database**. Of six hundred cases, 85 (14.2%) produced a null change - the footprint having fallen on an edge or coarse region and captured no element - and were removed automatically, leaving 515 valid cases. The element count per crack (a few tens to 859) correlates with the frequency change (r = 0.60), confirming that the selection procedure of Section 3.4 responds correctly to crack size.

**Crack signature**. The frequency changes are small: the largest relative change across the ten modes has a median of 0.131%, a 90th percentile of 0.634% and a maximum of 5.893%; only 5% of cases exceed 1%, and 58% exceed 0.1%. The signature grows with crack depth (r = 0.31), length (r = 0.30) and distance from the root (r = 0.38) - the directions expected from equation (2). Figure 1 shows the full population against three assumed noise floors: a large fraction lies at or below 0.1%, and the majority below 0.5%.




b

Figure 1: Crack signature versus measurement noise floor

**Sensitivity varies across modes (Figure 2).** Mode 1 is most responsive (mean 0.20%), the tenth least (0.03%). It is this pattern - varying with crack location - rather than any single change, that makes the inverse problem tractable. For scale: a 200 mm × 12 mm crack (85% of the wall) shifts the frequencies by ~1%, a few-millimeter crack by ~0.01%, while a frequency measurement is itself uncertain to 0.1% (laboratory) or up to 1% (field).

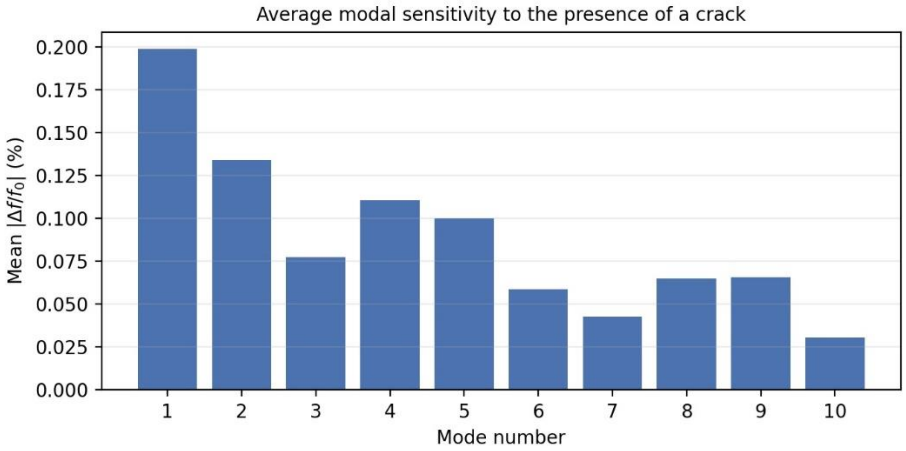



Figure 2: Average modal sensitivity to the presence of a crack

**The noise-augmented dataset**. The simulated healthy cases show apparent deviations of the order of the imposed noise (0.4% mean at the 1% level). Any crack whose signature is smaller is, on frequencies alone, indistinguishable from an intact blade - irrespective of the learning algorithm applied. The dataset deliberately spans both regimes, so the detection threshold can be read from performance rather than assumed. The two datasets are summarized in table below.

| Dataset | Rows | Cracked | Healthy |
|---|---|---|---|
| **Clean** | 516 | 515 | 1 |
| **Noise-augmented** | 55,350 | 46,350 | 9,000 |

*Table 2: Summary of the clean and noise-augmented datasets.*

**Machine Learning Results**

**Classification.** Table 3 reports final test-set performance. Both operating modes reach the targeted recall on cracked blades (≥93%). They diverge sharply in specificity: at row level, only 17.4% of healthy readings are correctly classified, versus 80% of healthy *cases* once 90 readings of the same blade are averaged and evaluated at the independently calibrated case-level threshold.

| Metric | Row level (single reading) | Case level (90-reading average) |
|---|---|---|
| **Threshold** | 0.350 | 0.525 |
| **Accuracy ↑** | 82.3% | 91.1% |
| **Balanced accuracy ↑** | 56.2% | 86.6% |
| **Precision (cracked) ↑** | 85.6% | 96.0% |
| **Recall (cracked) ↑** | 95.0% | 93.2% |
| **F1 (cracked) ↑** | 90.0% | 94.6% |

*Table 3. Classifier performance on the held-out test set (123 cases / 11,070 readings). Majority-class baseline: 83.7% accuracy, 50.0% balanced accuracy.*

**Verification of the metric fix.** A full 302-configuration re-search under average precision was carried out to check whether the inexpensive class-balancing fix could be improved upon; it could not (case-level balanced accuracy 84.6% vs. the deployed 86.6% - see the design-iteration summary, Table 7).

**Regression.** Table 4 reports performance on truly cracked test readings. All four models beat a mean-value baseline, but absolute $R^2$ remains low (0.04-0.14); normalised MAE is more informative, placing all four targets in a narrow 0.22-0.23 band despite differing by more than two orders of magnitude in physical scale. Depth has the lowest $R^2$ and the highest normalised error of the four.

| Target | Model | MAE ↓ | RMSE ↓ | $R^2$ ↑ | Normalised MAE ↓ |
|---|---|---|---|---|---|
| **Spanwise position** | Gradient boosting | 0.250 | 0.300 | 0.144 | 0.215 |
| **Chordwise position** | Gradient boosting | 0.240 | 0.283 | 0.075 | 0.222 |
| **Length** | Gradient boosting | 0.042 | 0.050 | 0.088 | 0.220 |
| **Depth** | Random Forest | 0.0021 | 0.0025 | 0.041 | 0.233 |

*Table 4. Regressor performance on cracked test cases (103 cases / 9,270 readings).*

Aggregating the regressors' raw predictions per case (mean over each case's 90 readings) improves $R^2$ over row-level evaluation for every target (e.g. spanwise position 0.144 → 0.214) - the same benefit repeated-measurement averaging gave the classifier in Table 3, though the per-noise-level analysis below (Table 6) shows this 90-reading figure blends three distinct noise conditions rather than reflecting any single sensor. (Provenance note: this figure and Table 6 were computed before the feature-set simplification of Section 3.7 and use the 20-feature models; the corresponding 10-feature values agree to the third decimal place, so conclusions are unaffected.)

**End-to-end pipeline and error analysis**. Chaining both stages (zero predicted whenever the classifier calls a case healthy) shows the practical cost of classification error relative to the regressor-alone figures of Table 4: row-level MAE worsens for three of four targets - spanwise position 0.250 → 0.321, length 0.042 → 0.050, depth 0.0021 → 0.0029 - while chordwise position improves slightly (0.241 → 0.208), false positives happening to contribute a zero prediction closer to that target's mean than some genuine regressor errors. This gap is invisible if regression is evaluated in isolation.

By noise level, regression error rises monotonically for every target as expected (Table 5), while classification error moves in two directions at once: the false-negative rate falls sharply as noise increases (12.9% → 1.6% → 0.3%) but the false-positive rate rises just as sharply (51.5% → 97.0% → 99.3%) - consistent with noise systematically inflating predicted crack probability regardless of the true label.

| Noise level | Spanwise MAE ↓ | Chordwise MAE ↓ | Length MAE ↓ | Depth MAE ↓ |
|---|---|---|---|---|
| **0.001 (lowest)** | 0.222 | 0.220 | 0.039 | 0.0019 |
| **0.005** | 0.260 | 0.249 | 0.043 | 0.0022 |
| **0.010 (highest)** | 0.269 | 0.252 | 0.043 | 0.0022 |

*Table 5. Regression MAE by simulated noise level (row level, cracked cases only).*

By crack-size tercile, the false-negative rate falls from 8.7% (smallest third) to 3.6% (middle) to 2.5% (largest) - small cracks are missed roughly 3.5× more often than large ones, as expected for a vibration-based method, since a smaller structural change produces a smaller, noisier frequency signature. Length MAE is non-monotonic across terciles (0.054, 0.017, 0.054), dominated by different sources of difficulty at each extreme: near-threshold detectability for the smallest tercile versus within-tercile heterogeneity for the largest.

Position error also varies with location: spanwise-position MAE is 0.112 for cracks in the middle third of the observed span range, versus 0.242 and 0.397 in the low and high thirds respectively - a more than threefold difference - and chordwise-position MAE shows the same pattern (0.084 mid-range vs. 0.287 and 0.320 at the extremes). Predictions are least reliable near the boundaries of the spatial range represented in the training data.

**Feature importance.** Permutation importance [17] - the drop in performance when a single feature's values are randomly shuffled - was computed on the held-out test set for the classifier and each regressor, prior to the feature-set simplification described in Section 3.7, over the full f/Δf feature set. The fundamental bending mode's raw frequency, $f_1$ (combined with $\Delta f_1$ where the collinear pair both ranked highly), is the single most important individual input for detection and for both position regressors - roughly 10× the next-ranked mode for spanwise position - consistent with the original report's own feature-importance analysis, which identified the same mode as most diagnostic for crack length and width using an independently generated, five-times-smaller dataset [19]. Length depends most on $f_2$ alone. Depth shows no comparably dominant input among any of the ten frequencies (every feature's importance $\leq 0.00003$), numerically consistent with its weak $R^2$ in Table 4: no single scalar frequency carries strong information about depth specifically.

**Aggregation-mode analysis**. The case-level figures in Table 3 average all 90 readings of a case, blending the dataset's three deliberately distinct noise conditions in a fixed 1:1:1 proportion no single real sensor would produce. A follow-up experiment therefore compared three operating modes at the (case, noise level) granularity of 30 same-condition readings, using the deployed models unmodified: predicting each reading independently (row-level), averaging 30 predicted outputs (prediction-averaging), and averaging the 30 frequency vectors before predicting once (feature-averaging). Each (mode, noise level) pair received its own threshold, calibrated on training out-of-fold probabilities within that noise level under the same recall ≥ 95% policy - defensible because a deployed system knows its sensor's noise class from calibration. A first variant pooling one threshold across noise

levels was rejected: it satisfied the recall floor only on average while silently violating it within the low-noise level (Table 7, row 6); all figures below enforce the floor per level.

| Noise level | Row-level ↑ | Prediction-averaging ↑ | Feature-averaging ↑ |
|---|---|---|---|
| **0.001 (lowest)** | 0.604 (95.3%) | **0.876** (95.2%) | 0.615 (98.1%) |
| **0.005** | 0.520 (94.2%) | **0.617** (88.4%) | 0.595 (99.0%) |
| **0.010 (highest)** | 0.503 (93.7%) | **0.538** (82.5%) | 0.521 (94.2%) |

*Table 6. Classification balanced accuracy (and realised test recall on cracked cases, in parentheses) by aggregation mode, per noise level; 123 cases per level, thresholds calibrated per (mode, noise level) on training data only. Realised recall can drift below the 95% calibration target on the higher noise levels because only 103 cracked cases per level are available for testing.*

Two conclusions follow. First, detection is dominated by sensor noise, not aggregation strategy: prediction-averaging is best at every noise level, reaching 0.876 balanced accuracy on low-noise readings but degrading to 0.62 and 0.54 as noise rises - a spread no aggregation strategy closed. The low-noise figure (0.876) essentially reproduces Table 3's 90-reading blended figure (0.866): averaged over all 90 readings, the 30 low-noise readings dominate the mean probability, so the blended figure describes what a low-noise sensor would deliver. Feature-averaging did not help classification: its nearly noise-free averaged inputs are unlike anything the noise-trained classifier saw - a distribution shift that hurts its probability estimates.

Second, for regression - which involves no threshold - the picture is unambiguous: feature-averaging gives the lowest MAE in all 12 target-by-noise combinations and the highest $R^2$ in 11 of 12 (spanwise position at noise 0.005: $R^2$ 0.09 row-level → 0.14 prediction-averaging → 0.36 feature-averaging; depth at the same level: ≈0 → 0.03 → 0.16), matching the measurement-theory expectation that averaging N same-condition readings suppresses noise by ~√N at the signal level. Feature-averaging is accordingly recommended for crack characterisation, prediction-averaging for detection, whenever repeated readings are available.

**Extrapolation.** Under the severity-based stress test, length and depth predictions collapse ($R^2$ between -4.1 and -4.8 in both directions) while spanwise and chordwise position remain comparatively robust. This matches a long-recognised property of tree-based regressors: each leaf predicts a constant, so an ensemble of trees cannot output values outside the target range it was trained on - a limitation noted as early as the original regression-tree literature [21]. A follow-up check with Ridge regression, which is not bounded to the training target range, performed no better ($R^2$ -4.5 to -5.3): the limit is not fixable by model choice.

**Design-iteration summary.** The pipeline was not produced in one pass: several design alternatives were built, evaluated, and explicitly rejected or superseded. Table 7 summarises each decision and its evidence, so the final configuration can be audited rather than taken on trust; every rejection was decided by the same primary metric (case-level balanced accuracy, or the relevant stage's validation score, with the recall ≥ 95% safety floor as a hard constraint). Two further checks, described below the table, were verification

runs confirming that an already-adopted cheap fix could not be improved upon - both confirmations held.

| # | Design decision tested | Evidence | Outcome |
|---|---|---|---|
| **1** | F-beta/accuracy scoring at default 0.5 threshold | Tied with always-predict-cracked baseline (≈0.963) | **Rejected** - replaced with average precision |
| **2** | Class-balanced RF, inexpensive re-ranking (8 configs, ~25 min) | AP 0.914; case-level balanced accuracy 0.876 | **Adopted** |
| **3** | Row-level threshold reused for case-level probabilities | Case-level accuracy 0.500 (all 20 healthy cases missed) | **Rejected** - dedicated case-level threshold (0.876) |
| **4** | Full f+Δf (20 features) vs. f-only (10) | Perfect collinearity (r = -1.000); scores within CV noise (0.876 vs. 0.866) | **Adopted f-only** - half the features, no baseline needed |
| **5** | Feature- vs. prediction-averaging at fixed noise level | Regression: best in 12/12 MAE, 11/12 $R^2$; classification: prediction-averaging best but noise-dependent (0.876 → 0.54) | **Feature-avg for regression; prediction-avg for detection** |
| **6** | Pooled vs. per-noise-level thresholds | Pooled gave better low-noise figure (0.937) at only 87% recall - violated safety floor | **Rejected** - per-noise calibration adopted (Table 6) |

*Table 7. Major design alternatives evaluated during development, with the deciding evidence; row 6 corrected the initial variant of row 5's experiment.*

Two further checks confirmed existing decisions rather than changing them: the exhaustive 302-configuration re-search under average precision (row 2's alternative) reached higher average precision (0.918) but lower case-level accuracy (0.846) than the adopted fix, and the Ridge-regression check is reported under Extrapolation above.

**2/ Discussion:**

**ANSYS Simulation Discussion**

**Signature small by physics, not artefact**. The signature is faint (median 0.13%) even for cracks penetrating 85% of the wall, because a frequency is a global property while a crack is a local loss of stiffness: on a 2.88 m blade, a centimeter-scale defect removes a tiny fraction of the total stiffness, and equation (2) fixes the change accordingly. Its ordered

growth with depth, length and root distance confirms the model responds correctly; only the absolute level, set by the blade's size, is small. This answers research question (ii).

**Detectability set by signal-to-noise ratio**. The limit is therefore the signature relative to measurement noise (Figure 1): at 0.1% uncertainty about two cases in five already fall below the floor, at 1% the majority do. Hence the outcome is framed as a detection threshold, not a single accuracy: below a certain crack size the signature is erased by noise, and no algorithm can recover it. The healthy cases confirm this - their noise-only deviations match the signatures of the smaller cracks. This answers question (iii).

**Consistent with the literature**. Frequency methods are established on beams and meter-scale specimens [8,9,12], but studies on realistic blades report the same difficulty at small sizes: on a scaled blade, defects ≤1.0 cm could not be identified experimentally though the model identified them [10]. This work reproduces that on a full-scale blade and quantifies it across 515 configurations. The enabling contribution, and the answer to research question (i), is the curved-geometry treatment: measuring depth along a locally recomputed normal (equations 4-6) rather than a global axis, which on a twisted shell would misrepresent it.

**Limitations**. The blade is modelled in isolation with a clamped root, so blade-to-blade coupling and mistuning are absent; it is analyzed at rest, omitting centrifugal stiffening (a conservative omission, since load opens the crack); the crack is an open-crack idealization, not a non-linear breathing crack; depth resolution - and the 3 mm lower bound - is limited by two elements through the wall; the noise is idealized as independent Gaussian error, without systematic components such as thermal drift; and the study is numerical, validated against a second solver but not against experiment.

**Machine Learning Discussion**

The addition of an explicit crack-detection stage is the main structural improvement over the earlier single-stage regression approach [19]: it prevents the system from reporting a spurious crack geometry for a blade that is, in fact, healthy. The trade-off this introduces is new information the earlier pipeline could not have revealed: detection reliability depends heavily on whether the system can aggregate multiple readings of the same blade before deciding (balanced accuracy 56% vs. 87%), a distinction that only exists because this dataset, unlike the original 165-case dataset used in [19], contains repeated noisy readings of each physical case.

Regression accuracy did not improve over the earlier report [19] in absolute terms - if anything, $R^2$ is lower here (0.04-0.14 vs. the earlier 46-63% for length, depth, and width). This is not read as a regression in model quality: the earlier evaluation used a single, non-group-aware train/test split on a much smaller dataset, and the present pipeline's rigorous case-level validation was specifically designed to expose exactly the kind of optimistic bias a naive split can produce. The lower $R^2$ here is judged the more trustworthy estimate of performance on genuinely unseen blade configurations.

The extrapolation failure is the most consequential new finding on the ML side, since it was not investigated in [19] at all. That length and depth predictions collapse outside the training severity range, and that switching to a model family without the same architectural bound does not fix it, indicates a genuine limit of the frequency-only approach for these

two targets - not a limitation of Random Forest specifically, since other tree-based algorithms share the same leaf-averaging constraint [21]. Combined with the modest $R^2$ even within the training distribution, this suggests ten scalar natural frequencies may be an information bottleneck for precisely localising crack depth and severity, with further gains more likely from additional sensing modalities than from further tuning of the same ten inputs. This gap is the clearest bridge back to the ANSYS side: it argues for widening the simulation's crack-severity range in a future campaign, since the model cannot be trusted to extrapolate and the FE database is the only lever that can remove that limit.

The noise-level breakdown (Table 5) has a direct practical implication beyond confirming the expected degradation of regression accuracy: the false-negative and false-positive rates move in opposite directions as noise increases, rather than both simply worsening. If the sensing hardware or environment used in an eventual physical deployment is noisier than the 0.001-0.010 range simulated here, this pipeline should be expected to miss fewer cracks but raise substantially more false alarms - a shift a maintenance team would need to be told to expect, rather than treating an increase in inspection call-outs as a sign the model has failed. The location-dependent error pattern (spanwise/chordwise MAE lowest near the centre of the observed range, more than threefold higher at its edges) points to the same underlying issue as the severity extrapolation limit discussed above: predictions are only as reliable as the region of input space they were trained on, and this holds for position, not only for the severity dimensions the dedicated stress test targeted directly.

One methodological caveat should be acknowledged. Although every threshold and hyperparameter was tuned strictly on training-set out-of-fold predictions, the pipeline was developed iteratively: some corrected problems (the scoring-metric collapse, the case-level threshold mismatch) were first noticed in held-out results, and the choice between the exhaustively searched and inexpensively fixed classifiers likewise consulted held-out performance. The reported test figures may therefore carry a small optimistic bias; a confirmation run on a fully untouched dataset - for example, a new batch of simulations - would be required for a strictly unbiased estimate.

## V/ CONCLUSION & PERSPECTIVE

### 5.1/ ANSYS Simulation Conclusion and Perspective

**Conclusion**

This work built and validated a finite-element workflow generating a modal-signature database for a full-scale fan blade, as the foundation of a vibration-based crack-detection system.

1. An automated PyMAPDL workflow represents a crack by local stiffness reduction on a single fixed mesh, producing 515 valid cases; the mesh is fully converged (identical frequencies across three densities) and agrees with an independent ANSYS Mechanical solution to the last significant figure.
2. The curved, twisted geometry was handled by measuring crack depth along a locally recomputed surface normal.
3. The signature was quantified - median 0.13%, maximum 5.9%, growing with depth, length and root distance - and, being comparable to real measurement uncertainty,

shows detectability to be governed by the signal-to-noise ratio: a property of the structure, not a flaw of the method.

4. A noise-augmented dataset with three noise levels and simulated intact blades lets any model trained on it be assessed under realistic uncertainty, its detection threshold determined rather than assumed.

Together these establish both the feasibility and the intrinsic limits of frequency-based crack detection on a large blade, and deliver a database ready for the machine-learning stage.

**Perspective**

In the near term, the smeared crack could be cross-checked against a discrete geometric crack, and mode-shape features added to strengthen localization. In the medium term, a pre-stressed modal analysis across rotational speeds would extend the method to the operating blade - natural, since centrifugal opening makes the open-crack idealization more valid. In the longer term, a cyclic-symmetry model of the full disc would capture mistuning, a breathing-crack model the crack's opening and closing, and - decisively - experimental modal testing would validate the database against physical measurement.

**5.2/ Machine Learning Conclusion and Perspective**

On the machine learning side, the classifier reliably detects cracked blades (95% recall at row level; 93.2% at case level) and, when multiple readings of the same blade can be aggregated, correctly identifies 80% of healthy blades as well - a substantial improvement in specificity over relying on any single noisy reading. Crack localisation and sizing show weak predictive performance ($R^2$ 0.04-0.14) and are shown, for the first time in this project, to fail outright outside the range of crack severities represented in training, a limitation that persists across model families and is therefore attributed to the information content of frequency-only features rather than to a specific algorithmic choice.

Two perspectives follow directly. First, the healthy-class and severity-range limitations identified here are primarily data limitations, not modelling ones: augmenting the simulation dataset (which is the ANSYS side's responsibility to extend) with independently generated healthy geometries and a wider range of crack severities is expected to directly address both. Second, the aggregation analysis establishes concrete design guidance for any physical deployment: collect multiple readings per blade rather than relying on a single measurement, average the predictions for detection but the feature vectors for crack-geometry estimation (best regression accuracy in 12/12 tested combinations), calibrate the decision threshold for the specific number of readings and sensor noise class in use, and treat sensor noise quality as a first-order system requirement - detection reliability ranges from 0.88 balanced accuracy at the lowest simulated noise level to near-chance at the highest, a spread no aggregation strategy was able to close. This weak explanatory power should be read as evidence for where future effort is best spent: it is not a failure of the tree-based models tried here, but a signal that the ten scalar frequencies alone may not carry enough information to size a crack precisely, regardless of which algorithm is applied to them.

**REFERENCES**

[1] Kırca, A.İ., Diltemiz, S.F., Yumrukaya, S. and Batar, A. (2023) 'Evaluation of foreign object damage on the fan blades with microscopic techniques', Duzce University Journal of Science and Technology, 11(5), pp. 2341-2351. DOI: 10.29130/dubited.1374721.
[2] Aero Engine Safety, Graz University of Technology (2020) 'Fundamentals and damage mechanisms - blade damage', section 5.2.1.1.
[3] Cowles, B.A. (1996) 'High cycle fatigue in aircraft gas turbines - an industry perspective', International Journal of Fracture, 80, pp. 147-163.
[4] Yang, W. et al. (2025) 'Failure analysis of the premature fan blade-off in aeroengine containment test', Engineering Failure Analysis. doi:10.1016/j.engfailanal.2025.109888.
[5] Aust, J. and Pons, D. (2019) 'Taxonomy of gas turbine blade defects', Aerospace, 6(5), 58. doi:10.3390/aerospace6050058.
[6] Palei, T. (2021) 'Aircraft engine blade damages and contribution to engine health', Aeronautics Today (Medium/LinkedIn publication, not a peer-reviewed source), 21 March 2021.
[7] Barad, K.H., Sharma, D.S. and Vyas, V. (2013) 'Crack detection in cantilever beam by frequency based method', Procedia Engineering, 51, pp. 770-775.
[8] Chen, X.F., He, Z.J. and Xiang, J.W. (2005) 'Experiments on crack identification in cantilever beams', Experimental Mechanics, 45(3), pp. 295-300.
[9] Haseeb, S.A. and Krawczuk, M. (2026) 'A state-of-the-art review of structural health monitoring techniques for wind turbine blades', Journal of Nondestructive Evaluation, 45(1), article 4. DOI: 10.1007/s10921-025-01296-5.
[10] Henderson, R., Azhari, F. and Sinclair, A. (2024) 'Natural frequency transmissibility for detection of cracks in horizontal axis wind turbine blades', Sensors, 24(14), 4456. DOI: 10.3390/s24144456.
[11] Gillich, N., Tufisi, C., Sacarea, C., Rusu, C.V., Gillich, G.-R., Praisach, Z.-I. and Ardeljan, M. (2022) 'Beam damage assessment using natural frequency shift and machine learning', Sensors, 22(3), 1118. DOI: 10.3390/s22031118.
[12] Regan, T., Beale, C. and Inalpolat, M.(2017) 'Wind turbine blade damage detection using supervised machine learning algorithms', Journal of Vibration and Acoustics.
[13] Boyer, R., Welsch, G. and Collings, E.W. (1994) Materials Properties Handbook: Titanium Alloys. Materials Park, OH: ASM International.
[14] Alava, M.J., Nukala, P.K.V.V. and Zapperi, S. (2006) 'Statistical models of fracture', Advances in Physics, 55(3-4), pp. 349-476. arXiv:cond-mat/0609650.
[15] Hou, C. and Lu, Y. (2016) 'Identification of cracks in thick beams with a cracked beam element model', Journal of Sound and Vibration, 385, pp. 104-124. DOI: 10.1016/j.jsv.2016.09.009.
[16] ANSYS Inc. (2024) ANSYS Mechanical APDL Theory Reference, Release 2024 R1. Canonsburg, PA.
[17] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.
[18] F. Pedregosa, G. Varoquaux, A. Gramfort et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[19] [Group 2, USTH Department of Aeronautics], “Fault detection of fan blades on aircraft engine using ANSYS simulation and Random Forest,” Measurement and Condition Monitoring Project Report, USTH, 2026.
[20] J. Davis and M. Goadrich, “The Relationship Between Precision-Recall and ROC Curves,” in *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, Pittsburgh, PA, USA, 2006, pp. 233-240. DOI: 10.1145/1143844.1143874.
[21] J. R. Quinlan, “Learning with Continuous Classes,” in *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence (AI '92)*, Hobart, Australia, 1992, pp. 343-348.