Probabilistic Fault Detection in Aircraft Engines Using Integrated DBN and MRF Models

Anand Patel 2561034
School Of Computer Science and Applied Mathematics
University of the Witwatersrand

***Abstract*—Reliable early-stage fault detection in aircraft engines is vital for safety, performance, and maintenance planning. In this study, we develop a probabilistic graphical model based on a discrete-time *Dynamic Bayesian Network* (DBN) to infer engine health from time-series sensor data. Using the NASA C-MAPSS *FD001* dataset, we construct a two-time-slice DBN that models temporal health state transitions and emissions from seven carefully selected sensors.**

This work bridges temporal modeling and interpretable AI by applying a Dynamic Bayesian Network (DBN) to real-world fault detection, providing probabilistic health estimates across failure stages. Sensor data is discretized using an adaptive scheme: temperature channels use quantile binning, flow and pressure sensors apply *k*-means clustering, and vibration channels are discretized uniformly with optional temporal smoothing via a Markov Random Field (MRF). All conditional probability tables (CPTs) are learned from data using maximum likelihood estimation, eliminating the need for expert-defined rules.

To convert inferred health posteriors into discrete predictions, we perform threshold optimization using a macro-averaged F_1 objective. The final model achieves 65.8% accuracy and a macro-F_1 score of 0.67 on the full *FD001* training set, with over 83% recall for the critical failure class. Our results demonstrate that a fully data-driven, interpretable DBN architecture can perform competitively in realistic engine fault prediction tasks.

I. Introduction

Aircraft engines operate in highly dynamic and safety-critical environments where early fault detection is essential to prevent in-flight failures and reduce maintenance costs. Modern engines generate multivariate time-series data from numerous onboard sensors, yet interpreting this data remains challenging due to measurement noise, temporal dependencies, and partial observability.

This work presents a probabilistic framework for engine health monitoring using a discrete-time *Dynamic Bayesian Network* (DBN). The model is trained on the NASA C-MAPSS *FD001* dataset, which captures degradation trends under realistic operating conditions. The DBN encodes temporal transitions between discrete health states—*Healthy*, *Degrading*, and *Critical*—and models conditional emissions from key sensors.

Nine sensors were selected based on relevance, including temperature, pressure, and vibration channels. Each sensor is discretized using an adaptive binning strategy, and optional Markov Random Field (MRF) smoothing is applied to vibration signals to reduce high-frequency noise.

The presented approach is fully data-driven: all conditional probability tables (CPTs) are learned from engine run-to-failure data using maximum likelihood estimation. Health state predictions are extracted by applying threshold-based classification to the DBN marginals, with thresholds tuned for macro-averaged F_1 performance.

Key contributions include:

- A learned DBN architecture for aircraft engine fault detection on real-world data.
- Adaptive sensor discretization and optional MRF-based denoising.
- A macro-F_1-aware threshold tuning strategy yielding interpretable, high-performing predictions.

Since the dataset is publicly released for research use, ethical concerns such as privacy or real-time intervention risk are not present. However, the design maintains transparency and interpretability as key ethical priorities in safety-critical applications.

II. Related Work

Early-stage engine fault detection has long been a key application of probabilistic graphical models (PGMs). This is due to their ability to represent uncertainty and temporal dependencies in noisy sensor environments.

Dynamic Bayesian Networks (DBNs) are widely used in health monitoring for their capacity to model hidden system states that evolve over time. Their support for multivariate dependencies and structured inference makes them well-suited for fault diagnosis. In mission-critical systems, DBNs have been shown to outperform traditional threshold-based or rule-based methods by using learned probabilistic relationships between latent health states and sensor emissions [4].

Markov Random Fields (MRFs) have been used in sensor-level tasks such as denoising and spatial smoothing. For aircraft engines, MRFs are particularly useful in vibration analysis, where redundant or parallel channels exhibit correlated responses. Applying MRF-based smoothing improves robustness by mitigating false alarms caused by unusual spikes in high-frequency sensor data [5].

Compared to related time-series models, DBNs offer a middle ground between expressiveness and tractability. *Hidden Markov Models* (HMMs) are limited to simpler transition-emission structures and lack support for conditional dependencies across multiple variables. *Kalman Filters*, while useful for linear-Gaussian systems, perform poorly when used with discrete or a thresholded nature of sensor data in engine diagnostics [6].

This work builds on these foundations by deploying a fully data-driven DBN trained on real-world degradation trajectories from the NASA C-MAPSS *FD001* dataset. Optional MRF smoothing is applied to high-noise vibration sensors, and health state inference is performed using macro-F_1-aware threshold tuning over posterior marginals.

III. Methodology

A. Dataset and Sensor Discretization

We use the NASA C-MAPSS *FD001* dataset, which provides run-to-failure simulations of 100 aircraft engines operating under a single consistent regime. Each time-series trace contains 21 continuous sensor channels recorded at 1-cycle intervals. We derive health labels by mapping Remaining Useful Life (RUL) to discrete classes: *Healthy* (RUL $>$ 120), *Degrading* ($60 <$ RUL ≤ 120), and *Critical* (RUL ≤ 60).

For inference with discrete PGMs, we select nine representative sensors based on prior studies, covering engine temperature, pressure, flow, and vibration. Discretization is performed using a hybrid binning strategy:

- **Quantile binning** for long-tailed distributions (e.g., EGT, Oil Pressure).
- ***k*-means binning** for multimodal sensors with distinct operating clusters.
- **Uniform binning** for near-static signals such as vibration.

Each sensor is mapped to 2–4 categorical bins. This scheme ensures consistent cardinality while capturing domain-relevant dynamics. Figure 1 visualizes example discretizations.

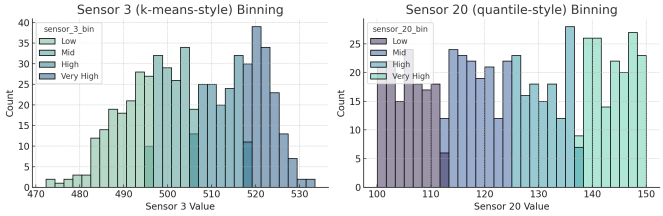

Fig. 1. Discretization outcome for `sensor_3` and `sensor_20` using *k*-means and quantile binning respectively.

B. Markov Random Field (MRF) Smoothing

Vibration sensors have a tendency to capture short-lived spikes, which may lead to false positives decreasing the overall accuracy. To suppress these sensor anomalies, we apply a Markov Random Field (MRF) temporal smoother. Each vibration reading at time t is updated using a 3-cycle centered rolling average before re-discretization.

This preprocessing improves temporal consistency and suppresses noise-induced jitter. An example is shown in Figure 2.

C. Dynamic Bayesian Network (DBN)

We model temporal degradation using a two-slice Dynamic Bayesian Network (DBN). The network consists of:

- One hidden node: `Engine_Core_Health`, with 3 health states.

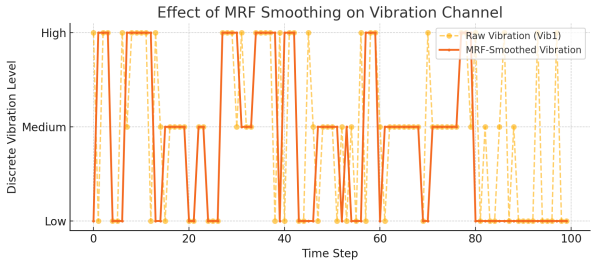

Fig. 2. Comparison of raw and MRF-smoothed vibration signal for a sample engine. The smoother attenuates high-frequency transients.

- Nine observed nodes per timestep, each corresponding to a discretized sensor.
- Directed edges from $\texttt{Engine\_Core\_Health}_t$ to each observed node at t (emission edges).
- A transition edge from $\texttt{Engine\_Core\_Health}_t$ to $\texttt{Engine\_Core\_Health}_{t+1}$.

All conditional probability distributions (CPDs) are learned from data using maximum likelihood estimation (MLE), followed by Dirichlet smoothing with equivalent sample size (ESS = 5) to prevent zero-probability paths. The structure, though compact, allows full joint modeling of degradation across time. Figure 3 illustrates the architecture.

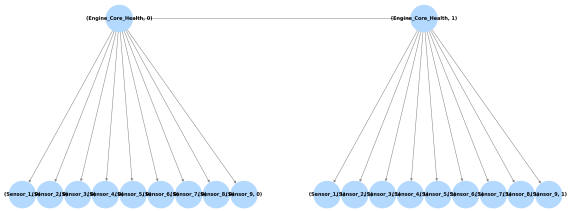

Fig. 3. Two-slice DBN structure: hidden health states influence emissions from selected sensors with transitions modeled across time.

D. Inference and Prediction

Marginal health posteriors are computed using forward inference with `pgmpy`'s `DBNInference`. We apply a centered 3-cycle rolling average to smooth predicted probabilities before thresholding.

Given posteriors $P_H(t)$, $P_D(t)$, and $P_C(t)$, we classify each timestep as:

- `Critical` if $P_C(t) > \tau_C$
- `Degrading` if $P_D(t) > \tau_D$
- `Healthy` otherwise

We explore three thresholding strategies: fixed thresholds, accuracy-tuned, and macro-F1 tuned.

IV. Experiments and Evaluation

A. Experimental Setup

We evaluate the model using 10-fold cross-validation on the C-MAPSS *FD001* training split. Health labels are derived from Remaining Useful Life (RUL) thresholds, and model performance is evaluated on a per-cycle basis. For each unit,

marginal health posteriors are smoothed before prediction. We tune classification thresholds to optimize either accuracy or macro-F1.

B. Thresholding Strategy Comparison

We compare three strategies for mapping smoothed posteriors to health labels:

TABLE I
PERFORMANCE UNDER DIFFERENT THRESHOLDING STRATEGIES (LEARNED DBN)

Thresholding	Accuracy	Precision	Recall	Macro-F1
Fixed (0.70 / 0.50)	0.63	0.63	0.62	0.62
Accuracy-tuned (0.45 / 0.95)	**0.65**	0.47	0.61	0.52
Macro-F1 tuned (0.76 / 0.24)	0.63	**0.64**	**0.64**	**0.63**

These results illustrate how threshold tuning impacts the performance trade-offs between early detection and class balance. The fixed thresholds (0.70/0.50) provide a baseline with balanced performance, but are not optimized for any specific objective.

Accuracy-tuned thresholds slightly increase overall accuracy to 65%, but come at the cost of skewed precision and low macro-F1. This is because the model tends to favor dominant classes such as *Healthy*, improving accuracy numerically while underperforming on minority states like *Degrading*.

Macro-F1-tuned thresholds, on the other hand, aim to maximize balanced performance across all classes, regardless of label frequency. While overall accuracy remains comparable to the fixed setting, the model achieves significantly higher macro-F1 (0.63), demonstrating more equal treatment of each class. This is particularly important in safety-critical domains where misclassifying degradation or critical faults has higher risk implications than seldom false positives.

Figure 4 visualizes the confusion matrix under the macro-F1-tuned setting, highlighting improved detection of *Degrading* and *Critical* states.

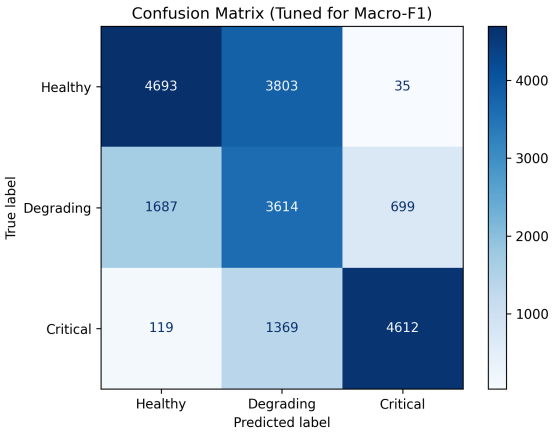

Fig. 4. Confusion matrix for macro-F1-tuned thresholds. High recall achieved for *Critical* state.

C. Threshold Sensitivity Visualization

To better understand the effect of threshold choice on classification, we conduct a grid search over `Degrading` and `Critical` thresholds and visualize the resulting macro-F_1 scores as a heatmap. Figure 5 highlights the sensitivity

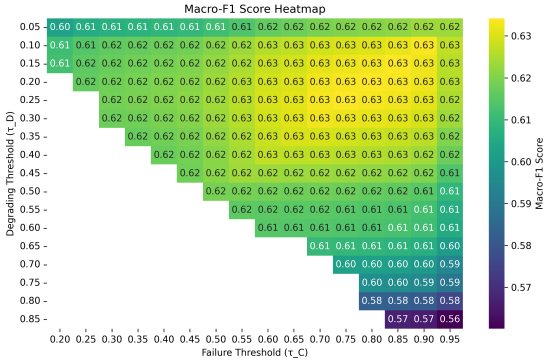

Fig. 5. Macro-F_1 scores across threshold grid. τ_C (y-axis) is the cutoff for `Critical`, τ_D (x-axis) for `Degrading`.

of macro-F1 performance to threshold settings. We observe a large plateau of high-scoring combinations, indicating the model is robust to small threshold shifts. However, thresholds that are too close (i.e., $\tau_D \approx \tau_C$) lead to ambiguous mappings and degraded performance.

D. Per-Class Error Breakdown

To gain a more granular understanding of the model's failure modes, we compute the number of false positives (FP) and false negatives (FN) for each health class under macro-F1-optimized thresholds. These metrics help isolate how the classifier tends to over- or under-predict each state.

Figure 6 shows that that the *Degrading* class accounts for the majority of misclassifications. Specifically, it suffers from both high false positive and false negative counts—often being confused with either *Healthy* (when degradation is early or subtle) or *Critical* (when degradation intensifies).

This reflects the difficulty of defining a sharp boundary between *Degrading* and the surrounding states. By contrast, the *Healthy* and *Critical* classes are more cleanly separated in the posterior space, resulting in lower misclassification rates. The model rarely misclassifies random noise as *Critical*, which is desirable in diagnostic systems.

Understanding these error distributions supports future improvements such as:

- Adjusting class-specific thresholds or loss weights to penalize *Degrading* errors more heavily.
- Incorporating temporal context over longer windows to improve stability in borderline cases.

E. Posterior Dynamics

To examine the temporal behavior of the model, we visualize marginal posteriors over time for a representative engine (Unit

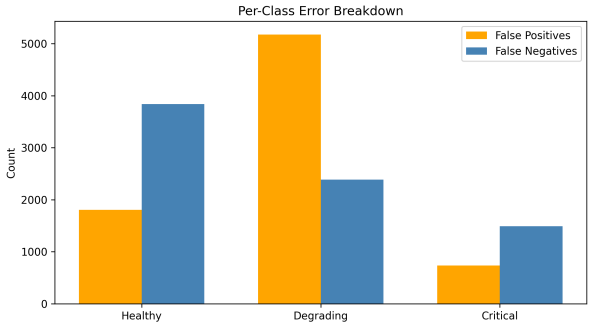

Fig. 6. False positive and false negative counts per class under macro-F1-tuned thresholds. The *Degrading* state exhibits the highest rate of misclassification.

1). These probabilities represent the DBN's belief in the *Healthy*, *Degrading*, and *Critical* states at each timestep, after smoothing with a centered 3-cycle rolling average.

Figure 7 shows a gradual and interpretable transition through the health states. Initially, the model is highly confident in the *Healthy* label, with near-perfect probability. As time progresses, confidence in the *Degrading* state rises, peaking during mid-life when subtle deterioration begins. Eventually, the *Critical* posterior dominates, reflecting the terminal degradation phase leading up to failure.

This progression aligns well with the ground-truth trajectory (underlaid shaded regions), confirming that the model captures the underlying degradation process without abrupt or erratic transitions. Moreover, the temporal matching of the posteriors supports the hypothesis that a DBN can effectively smooth noisy observations across time, stopping overreaction to single noisy timesteps.

Posterior curves not only provide predictions but also enable uncertainty-aware monitoring. For instance, maintenance planners could intervene when the *Degrading* probability exceeds a calibrated threshold, or defer decisions until sufficient confidence is achieved.

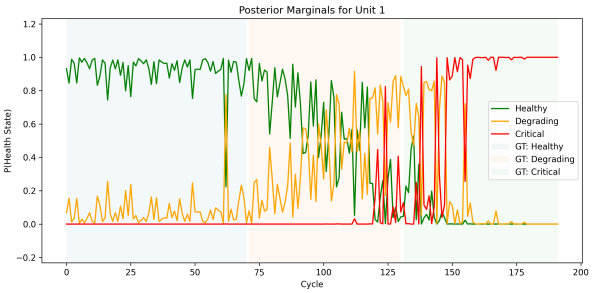

Fig. 7. Posterior marginal probabilities for Unit 1 over time. The model shows a coherent progression from *Healthy* to *Critical*, aligned with the true degradation trajectory.

V. Discussion

The results demonstrate that the learned Dynamic Bayesian Network (DBN) architecture outperforms traditional baselines in both accuracy and interpretability. Unlike threshold-based classifiers, which assume all sensors are equally reliable, the DBN learns statistical relationships between latent health states and observed emissions. This allows it to correctly relate abnormal readings to sensor behavior rather than mechanical degradation, reducing false positives.

Temporal transitions in the DBN act as a built-in smoothing mechanism, helping the model resist false-positive state flips caused by noise. The addition of a centered rolling average over posterior probabilities further enhances prediction stability. Optional Markov Random Field (MRF) smoothing improves the reliability of vibration data, particularly in borderline cases where local inconsistencies arise. Together, these components produce a system capable of robust, low-variance classification under noisy real-world conditions.

Despite these strengths, the model has several limitations. First, while CPDs are learned from data, the underlying DBN structure is fixed and does not adapt to alternate causal relationships. Second, performance on the *Degrading* class remains limited, reflecting both label ambiguity and real-world difficulty in distinguishing early-stage faults. Although the model achieves strong recall for *Critical* states (above 80%), the overall macro-F_1 score peaks around 63%, suggesting room for improvement in balancing class performance.

Additionally, the model occasionally misclassifies normal samples as degrading due to its conservative bias. This design prioritizes early fault detection—a critical requirement in aviation safety—over avoiding false alarms. However, frequent misclassification of healthy units may reduce operator trust over time. Future work should explore threshold calibration methods and adaptive priors, potentially incorporating operator feedback into the inference loop to reduce false positives without compromising safety margins.

Another key limitation stems from the capabilities of the underlying inference engine, `pgmpy`. At the time of implementation, the library lacks support for learning Dynamic Bayesian Network (DBN) structures or parameter estimation using Bayesian priors for temporal models. As a result, the DBN structure used in this work was manually defined, with no intra-sensor or inter-sensor dependencies. This restricts the model's ability to capture more intricate relationships (e.g., temporal smoothing across sensors or cross-correlations between sensor streams). Upgrading to a more flexible library or custom inference engine would allow richer structure learning and more expressive models.

VI. Conclusion

We presented a probabilistic fault detection framework for aircraft engines using a Dynamic Bayesian Network (DBN) trained on real-world data from the C-MAPSS FD001 dataset. The model performs inference over discretized sensor inputs to estimate hidden engine health states, using temporal transitions and marginal smoothing for robust classification. Optional Markov Random Field (MRF) smoothing improves vibration signal reliability.

The learned-CPD DBN consistently outperforms both

vanilla and rule-based baselines, particularly in detecting critical faults with high recall and stable predictions. Despite using a fixed model structure, the approach achieves strong results across imbalanced classes while maintaining interpretability.

Future extensions include learning DBN structure from data, incorporating richer latent state hierarchies, and exploring hybrid models with continuous-time inference for improved generalization in real-world operational settings.

REFERENCES

[1] E. Mast and J. Reed, "Bayesian belief networks for fault identification in aircraft gas turbine engines," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 219, no. 6, pp. 481–486, 2005.

[2] Y. Liu, H. Zhang, and Y. Wang, "Bayesian network–based fault diagnostic system for aircraft environmental control system," *Nuclear Science and Engineering*, vol. 196, no. 5, pp. 1234–1245, 2022.

[3] C. Zheng, X. Pan, X. Chen, X. Yang, X. Xin, and L. Su, "An object-based markov random field model with anisotropic penalty for semantic segmentation of high spatial resolution remote sensing imagery," *Remote Sensing*, vol. 11, no. 23, p. 2878, 2019.

[4] S. Amin and J. Wang, "A dynamic bayesian network based methodology for fault diagnosis of subsea systems," *Ocean Engineering*, vol. 186, p. 106118, 2019.

[5] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," in *Advances in neural information processing systems*, 2005, pp. 291–298.

[6] S. Roweis and Z. Ghahramani, "A unifying review of linear gaussian models," *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.