Preprints.org

Article

Not peer-reviewed version

Multimodal Sensor-Fusion and Temporal Deep Learning for CNC Toolpath and Condition Classification: A Cross-Validated Ablation Study

Stephen S. Eacuello , Romesh S. Prasad , Manbir S. Sodhi *

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0488.v1

Keywords: CNC machining; multi-modal sensor fusion; sensor modality selection; deep learning; LSTM; temporal modeling; feature ablation; cross-validation; process verification; security monitoring; denoising autoencoder

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Multi-Modal Sensor Fusion and Temporal Deep Learning for CNC Toolpath and Condition Classification: A Cross-Validated Ablation Study

Stephen S. Eacuello *, Romesh S. Prasad and Manbir S. Sodhi

Department of Industrial Systems Engineering, University of Rhode Island, Kingston, RI 02881, USA
* Correspondence: seacuello@uri.edu

Abstract

Determining which sensor modalities carry genuine discriminative signal for CNC monitoring—and how many can be removed before performance degrades—is a practical question that prior work rarely answers quantitatively under leakage-resistant evaluation. We address this through a systematic cross-validated ablation study on a 9-class CNC toolpath and condition classification task, combining three toolpath strategies (adaptive, face, pocket) with three conditions (air-cutting, active-cutting, and damaged-spindle). Using 120 operation files from a desktop CNC mill with six consistently active sensors (17 channels per sensor) plus 8 machine-level electrical features, we evaluate six model families across 690 cross-validated runs spanning five cumulative feature-ablation levels (110–56 features) and ten temporal resolutions. To handle the fusion challenge, we introduce MM-DTAE-LSTM, a multi-modal denoising temporal attention encoder with unidirectional LSTM-based classification that combines learned modality gates, cross-modal attention, and a self-supervised denoising objective. Key findings on this single-machine, single-material dataset: (1) MM-DTAE-LSTM reaches 96.3% $\pm$ 4.7% accuracy at a 98-feature configuration (excluding proximity and pressure), leading all baselines by 3.1–5.2 points, though differences are not statistically significant at n=5 folds; (2) reducing the feature set by 49% (to accelerometer, gyroscope, temperature, RMS audio, and electrical channels) retains 92.5% $\pm$ 8.3% for the encoder while XGBoost drops to 84.4%, a loss of 10.7 points from its full-feature peak; (3) at full features, baselines are competitive (Random Forest: 95.6%, XGBoost: 95.1%); and (4) one-way ANOVA reveals that pressure channels encode session-level barometric confounds ($F > 2{,}000$) rather than machining dynamics, explaining baseline degradation when confound-prone channels are removed. These results suggest that core inertial, acoustic, and machine-level modalities may be sufficient for effective CNC operation classification on similar platforms, providing sensor-selection and temporal-configuration guidance for cost-effective monitoring deployments. Generalization to industrial machines, diverse materials, and production environments requires further validation.

Keywords: CNC machining; multi-modal sensor fusion; sensor modality selection; deep learning; LSTM; temporal modeling; feature ablation; cross-validation; process verification; security monitoring; denoising autoencoder

1. Introduction

Multi-modal sensor fusion—combining heterogeneous measurement modalities with different sampling rates, noise characteristics, and physical bases into a unified learned representation—is a persistent challenge in industrial monitoring. When dozens of channels from inertial, acoustic, environmental, and electrical sensors are available, practitioners face practical questions that sensor-level studies rarely answer quantitatively: which modalities carry genuine discriminative signal, how many can be removed before performance degrades, and what temporal resolution balances classification accuracy against computational cost? Because machining sensor streams are inherently sequential—toolpath geometry unfolds over time as direction changes, step-overs, and plunge transitions produce

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

characteristic temporal patterns—effective fusion must preserve and exploit this temporal structure rather than discard it by temporal flattening. We address these questions through a systematic ablation study in CNC machining, a domain where the answers directly determine retrofit cost, hardware complexity, and monitoring reliability.

Computer Numerical Control (CNC) machines underpin precision manufacturing, executing complex toolpaths with micron-level accuracy across aerospace, automotive, medical device, and defense sectors. A single undetected deviation in spindle speed, feed rate, or depth of cut can scrap parts, damage tooling, or trigger costly unplanned downtime [1]. As factories modernize under the Industry 4.0 paradigm [2], CNC systems are increasingly integrated with networks, remote monitoring, and software-defined workflows—improving productivity and traceability, but also expanding operational complexity and cyber risk.

From a manufacturing perspective, this motivates sensor-based process verification: confirming that the intended operation actually occurred, not just that a controller reported it. From a cybersecurity perspective, it contributes to defense-in-depth strategies: controller logs, alarms, and network telemetry may be incomplete or misleading under misconfiguration, sensor failure, or adversarial manipulation (e.g., altered feed rates, tool offsets, or substituted G-code blocks), particularly when the controller itself is the attack vector [3]. In these settings, external sensors provide an independent physical verification layer—one component of the defense-in-depth approach recommended by OT security guidance [4]—enabling cross-checking between commanded behavior and observed machine response.

Every G-code program induces a characteristic physical signature. Even during air-cutting, toolpath geometry produces distinguishable motion dynamics: adaptive clearing follows trochoidal paths with rapid direction reversals; face milling sweeps in quasi-linear passes with periodic step-overs; and pocket machining alternates between plunges and lateral clearing. When material is engaged, cutting forces, chip formation, and tool–workpiece friction superimpose additional signatures [5]. We target operation verification and condition classification, not high-frequency chatter diagnosis: while tooth-passing harmonics and chatter are central to tool wear monitoring, many verification and security tasks depend primarily on lower-frequency trajectory envelopes, direction-change transients, and structural response modes that persist under modest sampling bandwidth. The DTAE's self-supervised denoising objective draws on representation learning principles from speech and audio processing [6,7], adapted here for multi-modal industrial sensor fusion.

Reliable classification from sensor data is challenging in practice. Real deployments must fuse heterogeneous modalities with different sampling rates, noise floors, and physical measurement bases [8], while handling intermittent telemetry, sensor dropouts, and class imbalance for rare fault modes. In addition, some channels can encode confounds rather than machining physics—for example, environmental measurements may encode session-level variation (*e.g.*, barometric pressure differences across collection days) that correlates with labeled classes and inflates apparent accuracy.

These artifacts are especially problematic in security monitoring settings, where the system must generalize beyond a controlled lab environment and remain robust when specific sensors are unavailable, degraded, or manipulated. Reliable evaluation compounds this challenge: because sliding windows overlap heavily, window-level splits can leak near-duplicate segments across train and test; we therefore avoid window-level data leakage by splitting at the operation-file level.

Prior work in machining monitoring has focused predominantly on tool condition monitoring, chatter detection, and binary fault classification using limited sensor sets [1,5,9]. Deep learning has enabled end-to-end representation learning from raw signals and improved robustness to noise [10,11]. However, there remains limited systematic evidence—under leakage-resistant, file-level validation—quantifying (i) how performance changes as sensor modalities are removed, and (ii) how temporal resolution affects operation classification accuracy and variance. These questions directly determine retrofit cost, compute and latency budgets, and robustness under sensor degradation, especially in changing or adversarial environments emphasized by recent OT security roadmapping [4,12]. We

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

therefore provide deployment guidance on when lightweight baselines suffice and when robust multimodal fusion is required.

In this work, we address a 9-class CNC operation classification task: given multi-modal sensor readings, classify both the toolpath strategy (adaptive clearing, face milling, pocket machining) and operating condition (air-cutting, active cutting, damaged spindle). We evaluate on 120 cleaned operation files collected from a desktop CNC mill instrumented with distributed sensors, and we employ five-fold file-level stratified cross-validation to prevent window-level leakage. Our architecture, MM-DTAE-LSTM, combines cross-modal attention for sensor fusion with a unidirectional LSTM that explicitly models the temporal dynamics of machining operations—capturing direction-change cadences, step-over periodicity, and plunge–clear alternation patterns that distinguish toolpath strategies even during air-cutting. This temporal modeling is a deliberate architectural choice: operation classification is fundamentally a sequence recognition task, and the LSTM's recurrent structure preserves causal ordering required for streaming inference while providing the temporal context that tree-based models discard through feature flattening. Beyond classification, the learned sensor-to-operation mapping established here lays the groundwork for a more comprehensive goal: reconstructing the underlying G-code command sequence directly from sensor signals, enabling full program verification rather than coarse operation-class identification. Our goal is not only high accuracy, but actionable guidance on sensor modality selection and temporal settings for benchtop CNC monitoring, with a methodology designed for transfer to production environments.

We make the following contributions:

1. **Feature ablation for deployment guidance**. We conduct a five-level cumulative modality ablation (110→56 features) that progressively removes low-value or confound-prone modalities and quantifies the minimum feature set that preserves strong performance (Section 3.8).
2. **MM-DTAE-LSTM architecture**. We introduce a Multi-Modal Denoising Temporal Attention Encoder with LSTM-based temporal classification, combining modality-specific projections, learned modality gates, cross-modal attention, and a self-supervised denoising reconstruction objective for noise-robust multimodal representations (Section 3.4).
3. **Temporal resolution analysis**. We evaluate ten window/stride configurations (16–256 timesteps) to characterize the context–latency tradeoff and identify stable performance regimes for streaming inference (Section 3.9).
4. **Leakage-resistant evaluation**. We implement five-fold file-level stratified cross-validation on 120 cleaned operation files from six consistently active sensors, providing unbiased generalization estimates with fold-level variance (Section 3.7).
5. **Baseline benchmarking under identical protocols**. We compare MM-DTAE-LSTM against five baseline model families (XGBoost, Random Forest, Logistic Regression, MLP, SimpleLSTM) under identical splits, preprocessing, and feature configurations, reporting accuracy, macro-F1, and paired fold-level effect sizes (Section 4.4).
6. **Process verification and security implications**. We demonstrate that learned sensor signatures provide independent physical verification of CNC operations and analyze session-level confounds that affect monitoring reliability. Reconstruction-error anomaly scoring serves as a complementary unsupervised indicator, with performance characterized in Section 5.6.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the experimental setup and methods; Section 4 presents results; and Section 5 discusses implications, limitations, and deployment recommendations.

2. Related Work

2.1. CNC Process Monitoring and Multi-Modal Sensor Fusion

Sensor-based process monitoring is well-established for tool wear estimation and chatter detection [1,5,9]. Early work relied on hand-crafted features from vibration signals processed through FFT or wavelet transforms. Teti *et al.* [1] identified force, vibration, acoustic emission, temperature, and motor

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

current as complementary sensing modalities, establishing the foundation for multi-sensor approaches. Serin *et al.* [5] surveyed tool condition monitoring and identified deep learning as a promising direction for end-to-end feature learning [10,11]. CNNs learn spatial features from vibration spectrograms: Janssens *et al.* [13] achieved 93.6% accuracy on bearing fault detection using 1D-CNNs, outperforming random forests (87.3%). Stathatos *et al.* [14] applied 1D-CNNs to raw 8-channel CNC turning signals, achieving $F_1 > 0.97$; Bhandari [15] compared MLPs, CNNs, LSTMs, and Transformers for surface roughness classification, finding Transformers achieved 92.7% on MFCC features. Wang *et al.* [16] combined a denoising Transformer autoencoder with ResNet (DTAE-ResNet) for multi-sensor tool wear classification, demonstrating that DTAE-based fusion outperforms standalone architectures—directly motivating our design.

Multi-modal fusion ranges from early fusion (feature concatenation) to late fusion (classifier combination); intermediate attention-based strategies are most effective for complementary modalities [17,18]. Cao *et al.* [19] demonstrated cross-attention fusion of acoustic and photoelectric signals for laser welding monitoring at 99.7% accuracy, showing cross-modal attention exploits complementary sensor information. Multi-sensor fusion for CNC monitoring has been reviewed by Tsanousa *et al.* [20], though prior work overwhelmingly addresses tool condition rather than operation classification. A principled approach to understanding which modalities matter is systematic ablation. Ablation studies—systematically removing components to quantify contributions—are essential for both interpreting deep learning models [21] and informing cost-effective sensor deployment [22]; our cumulative design removes modality groups progressively to quantify each group's marginal value.

2.2. Temporal Deep Learning for Industrial Sensor Streams

LSTM networks [23] remain widely used for sequential sensor data, capturing long-range dependencies in continuous dynamics. Encoder-decoder architectures with attention [24,25] provide foundations for sequence-to-sequence learning. The Transformer [26] has shown strong time-series performance [27], though LSTMs often retain advantages for lower-dimensional sensor streams with continuous dynamics. Denoising autoencoders [28] develop noise-invariant representations; recurrent architectures have shown temporal modeling advantages for industrial sensor streams, including vibration and process monitoring [29]. Wang *et al.* [16] further showed that coupling a denoising Transformer autoencoder with a downstream classifier improves robustness under sensor noise—a principle directly instantiated in our DTAE component. Our architecture combines both: Transformer-based cross-modal attention for fusion, followed by LSTM for temporal modeling—leveraging each paradigm's complementary strengths.

2.3. Manufacturing Cybersecurity, Anomaly Detection, and Study Positioning

As factories connect to networks, program tampering becomes a credible attack vector [3]. NIST's updated OT security guidelines [4] emphasize defense-in-depth strategies, while federal initiatives like CyManII [30] and NIST roadmapping [12] identify AI for machining cybersecurity as a critical research need. Kimmell *et al.* [31] demonstrated sensor-based detection of control injection attacks in CNC machining using energy data anomalies; Coelho *et al.* [32] applied 1D-CNNs for CNC anomaly detection achieving 91.6% binary accuracy—motivating more fine-grained multi-class approaches. A substantial body of work addresses the security of additive manufacturing processes [33], including G-code inference from side-channel signals on Fused Deposition Modeling (FDM) 3D printers [34–36], with acoustic [37] and video-based [38] attacks achieving high reconstruction fidelity. CNC milling, however, presents distinct challenges: simultaneous multi-axis interpolation and nonlinear force–engagement relationships produce far more complex sensor signatures than single-axis FDM motion.

Table 1 situates our work relative to recent experimental studies. Prior deep learning for CNC monitoring predominantly addresses tool wear or binary fault detection with 2–8 sensor channels and 2–4 classes. Our work differs in three respects: (1) multi-class operation classification (9 classes) rather than binary condition monitoring; (2) substantially higher sensor dimensionality (110 features from 7 modality groups across 6 distributed sensors); and (3) systematic sensor ablation across 690

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

experiments with file-level cross-validation, providing quantitative sensor-selection guidance absent from prior work.

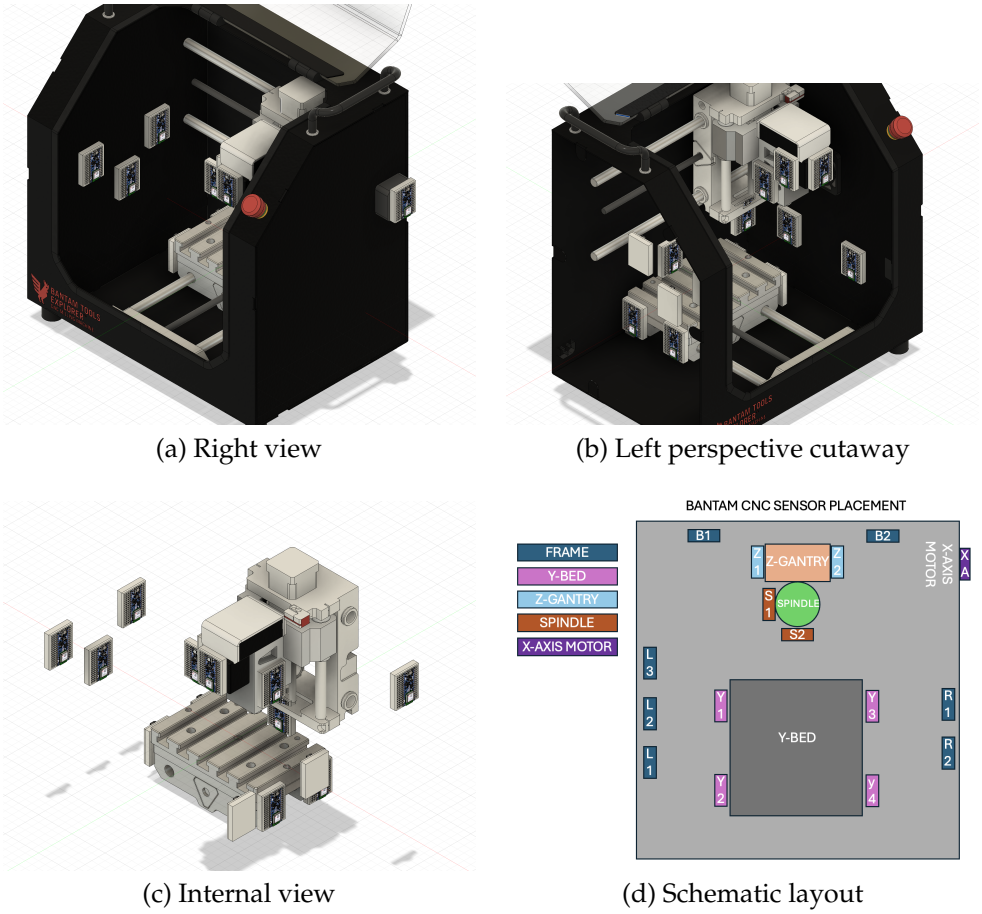
Table 1. Comparison with Recent Experimental Studies in Sensor-Based CNC/Machining Monitoring.

Study	Task	Sensors	Cls.	Best Metric	Method	Validation	Ablation
Janssens [13]	Bearing fault	Accel. (1)	4	93.6% acc	1D-CNN	Train/test	No
Bhandari [15]	Roughness	Sound+force (2)	4	92.7% acc	Transformer	50/25/25	No
Cao [19]	Weld penetr.	Acoustic+photo (2)	3	99.7% acc	Cross-attn	Repeated	No
Wang [16]	Tool wear	Force+vib+AE (7)	3	—	DTAE-ResNet	Train/test	No
Stathatos [14]	Turning qual.	Vib+torque+speed (8)	4	F_1>0.97	1D-CNN	10-fold CV	No
Coelho [32]	CNC anomaly	Current+vib (3)	2	91.6% acc	1D-CNN	80/10/10	No
Kimmell [31]	Attack detect.	Energy (1)	2	—	Threshold	Experimental	No
Ours	**Op. classif.**	**6 sens. (110 feat.)**	**9**	**96.3% acc**	**MM-DTAE-LSTM**	**5-fold CV**	**Yes**

3. Materials and Methods

3.1. Experimental Setup

Data were collected from a Bantam Tools Explorer desktop CNC mill, instrumented with 16 Arduino Nano 33 BLE Sense Lite boards. Each board integrates a 9-axis IMU (LSM9DS1: accelerometer, gyroscope, magnetometer), environmental sensors (LPS22HB barometric pressure/temperature, APDS-9960 color/proximity), and a PDM microphone (MP34DT05) for RMS audio computation. Figure 1 shows sensor mounting locations across the machine structure.



(a) Right view (b) Left perspective cutaway

(c) Internal view (d) Schematic layout

Figure 1. Sensor placement on the Bantam Tools Explorer CNC mill. (a) Right view showing frame-mounted sensors. (b) Left perspective cutaway revealing internal sensor positions. (c) Internal view of spindle and bed sensors. (d) Schematic layout of all 16 sensor positions. Arduino Nano 33 BLE Sense Lite boards are mounted on breadboards affixed with cyanoacrylate adhesive for rigid structural coupling. Of the 16 deployed sensors, 6 maintained ≥93% data consistency and are used for cross-validated experiments (Table 2).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Table 2. Consistently Active Sensors (≥93% Activity Across All Files).

Sensor	Location	Activity (%)	Channels
frame_l2	Frame (left)	>93	17
frame_l3	Frame (left)	>93	17
frame_r2	Frame (right)	>93	17
spindle2	Spindle	>93	17
y_bed_3	Bed	>93	17
y_bed_4	Bed	>93	17
Total sensor features			**102**
+ Electrical (machine-level)			8
Total features (full config)			**110**

3.1.1. Sampling Configuration

The IMU (accelerometer, gyroscope, magnetometer) samples at 119 Hz natively (Nyquist limit 59.5 Hz). Environmental sensors (pressure, temperature, color/proximity) sample at 10 Hz. The on-board PDM microphone computes RMS audio amplitude at the firmware level before transmission. Motor electrical signals (voltage, current) are captured via MCCDAQ at 1 kHz. All streams are aggregated on a Raspberry Pi 4—sensor boards communicate via BLE wireless links or direct USB serial connections—then resampled and synchronized to the G-code execution timeline for operation labeling, yielding an effective aligned rate of ∼4 Hz.

Note on bandwidth: The 400 Hz tooth-passing frequency exceeds the 59.5 Hz IMU Nyquist limit, and the ∼4 Hz aligned rate further reduces temporal resolution. However, discriminative information persists in: (1) low-frequency structural vibration modes (1–50 Hz) excited by gantry motion and, during active cutting, by cutting forces; (2) acceleration transients at toolpath direction changes; and (3) operation-specific trajectory envelopes. Classification succeeds because operation *type* depends on toolpath geometry—distinguishable even during air-cutting—rather than high-frequency cutting dynamics.

3.1.2. Sensor Consistency and Selection

Not all 16 sensors produced consistent data across all files due to intermittent Bluetooth connectivity, USB serial dropouts, and firmware sampling failures. We apply a ≥93% activity threshold (non-null rows across all files) to select reliable sensors:

The six selected sensors span three structural groups—frame (3), spindle (1), and bed (2)—providing spatial diversity across the machine's primary vibration transmission paths.

3.1.3. Data Cleaning

Two cleaning steps reduce the 135 raw files to 120 analysis-ready files:

(1) Sensor consistency filtering. Files missing any of the six required sensors (Table 2) are excluded. Of the 15 dropped files, most lack `frame_l3` or `y_bed_4`—the two sensors nearest the 93% activity threshold—due to intermittent connectivity during those recording sessions. The affected classes are face air-cut (4 files), face active-cut (5), adaptive active-cut (3), pocket active-cut (1), pocket air-cut (1), and face damage (1). This file-level filtering eliminates the need for zero-padding missing sensor columns, which would otherwise create trivial classification shortcuts (models learning sensor *presence* rather than sensor *dynamics*).

(2) Trailing NaN truncation. One file (`face150025_001`) contained 79% NaN-padded rows (3,504 rows total, only 731 with valid sensor data) caused by a sensor communication dropout partway through the machining operation. These trailing NaN rows were truncated to the last valid sensor reading. Without truncation, forward-fill interpolation would propagate the final valid values across 2,772 rows, creating degenerate constant-value windows that corrupt training.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

3.1.4. Data Collection Protocol

Data were collected across two separate days to capture session-to-session variation; the resulting day-to-day environmental differences are analyzed as a potential confound in Section 5.2. Each toolpath strategy was executed for multiple trials under three conditions: air-cutting (spindle running, no material engagement), active cutting (0.025" / 0.635 mm axial depth in UHMW-PE), and damaged-spindle air-cutting (one drive band deliberately removed to induce spindle runout and instability, simulating drive-train degradation).

All operations used a 2-flute HSS endmill (1/4" / 6.35 mm diameter) at 12,000 RPM spindle speed and 1,016 mm/min feed rate (40 IPM); active-cut trials used TIVAR UHMW-PE workpiece material at 0.635 mm (0.025") axial depth, chosen for consistent cutting behavior and minimal tool wear.

3.2. Modality Structure

Each of the six sensors captures 17 channels spanning 8 channel types (accelerometer, gyroscope, magnetometer, pressure, temperature, proximity, color, and RMS audio). For the encoder architecture, the three environmental sub-channels (pressure, temperature, proximity) share a single encoder group, yielding 6 modality encoders for the 102 Arduino sensor features (Table 3). A 7th encoder group processes the 8 machine-level electrical features, yielding 110 input features across 7 encoder groups total.

Table 3. Per-Sensor Channel Structure and Encoder Modality Grouping.

Encoder Group	Channel Type	Components	Per Sensor	6 Sensors
Accelerometer	Accelerometer	Ax, Ay, Az	3	18
Gyroscope	Gyroscope	Gx, Gy, Gz	3	18
Magnetometer	Magnetometer	Mx, My, Mz	3	18
	Pressure	Barometric pressure	1	6
Environmental	Temperature	Ambient temperature	1	6
	Proximity	APDS-9960 proximity	1	6
Color	Color	R, G, B, A	4	24
RMS Audio	RMS Audio	PDM microphone RMS	1	6
Electrical	*Machine-level (voltage, current, spindle)*			8
7 groups			**17**	**110**

3.3. Classification Task

We define a 9-class task combining 3 toolpath strategies with 3 operating conditions (Table 4).

Table 4. 9-Class Classification Task.

Toolpath	Condition	Class Label	Files
Adaptive Clearing	Air-cut	adaptive	20
Adaptive Clearing	Active-cut	adaptive150025	17
Adaptive Clearing	Damage	damageadaptive	5
Face Milling	Air-cut	face	16
Face Milling	Active-cut	face150025	15
Face Milling	Damage	damageface	4
Pocket Machining	Air-cut	pocket	19
Pocket Machining	Active-cut	pocket150025	19
Pocket Machining	Damage	damagepocket	5
Total			**120**

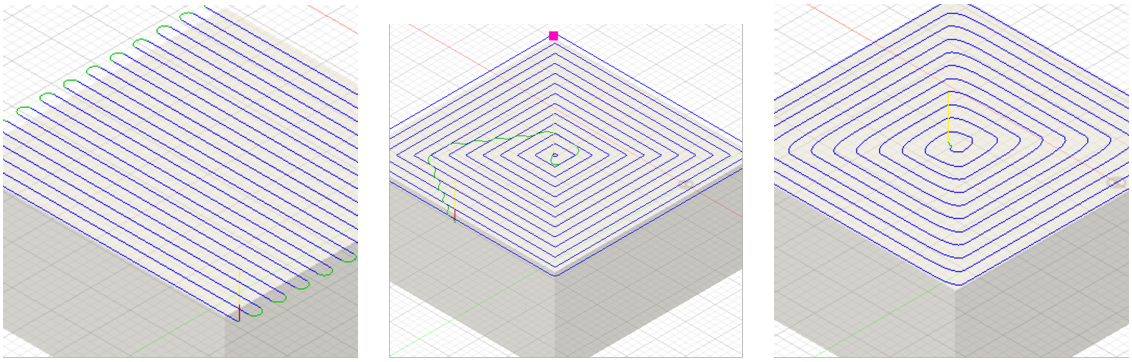
Note: Active-cut class labels (suffix `150025`) encode the 0.025" depth-of-cut parameter from the CAM configuration; damage class labels carry a `damage` prefix.

The three toolpath strategies (Figure 2) represent common CAM operations with distinct geometric trajectories:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

- **Adaptive Clearing**: High-efficiency roughing strategy following trochoidal paths that maintain near-constant radial engagement (typically 10–25% of tool diameter). The resulting trajectory involves frequent direction reversals and high-jerk cornering, producing periodic acceleration transients.
- **Face Milling**: Surface finishing using parallel linear passes with periodic step-over transitions, producing quasi-steady gantry motion punctuated by rapid repositioning at pass boundaries.
- **Pocket Machining**: Enclosed-region clearing combining axial plunge entries and lateral offset or spiral passes, producing bimodal motion signatures with distinct axial and radial frequency content.



(a) Face milling (b) Pocket milling (c) Adaptive milling

Figure 2. CAM toolpath strategies used in this study: (a) face milling sweeps in parallel linear passes with step-over transitions; (b) pocket milling alternates between axial plunges and lateral clearing passes; (c) adaptive (trochoidal) milling follows curved, constant-engagement paths with rapid direction reversals.

Each toolpath was executed under the three conditions described in Section 3.1.4: **air-cut**, **active-cut**, and **damage** (damaged-spindle air-cut). These three conditions, crossed with three toolpath strategies, define the nine classification targets in Table 4.

3.4. The MM-DTAE-LSTM Architecture

The Multi-Modal Denoising Temporal Attention Encoder with LSTM (Figure 3) processes multi-modal sensor data through five stages: (1) modality-specific projection, (2) gated cross-modal fusion with attention, (3) denoising autoencoder for self-supervised representation learning, (4) temporal modeling via LSTM, and (5) classification. We describe each component in detail.

Algorithm 1 summarizes the forward pass. Each stage is detailed in the following subsubsections.

3.4.1. Modality Projection

Each modality $M_i \in \mathbb{R}^{T \times C_i}$ is independently projected into a shared d_{model}-dimensional space through a two-layer MLP where each layer applies Linear $\rightarrow$ LayerNorm [39] $\rightarrow$ GELU $\rightarrow$ Dropout [40]:

$$H_i^{(1)} = \text{Drop}\big(\text{GELU}(\text{LN}(M_i W_i^{(1)} + b_i^{(1)}))\big), \quad H_i = \text{Drop}\big(\text{GELU}(\text{LN}(H_i^{(1)} W_i^{(2)} + b_i^{(2)}))\big) \tag{1}$$

Sinusoidal positional encodings and a learned modality identity embedding $\mathbf{e}_i$ are added to preserve temporal ordering and distinguish modality sources for subsequent cross-modal attention:

$$\tilde{H}_i = H_i + \text{PE}(T) + \mathbf{e}_i \tag{2}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

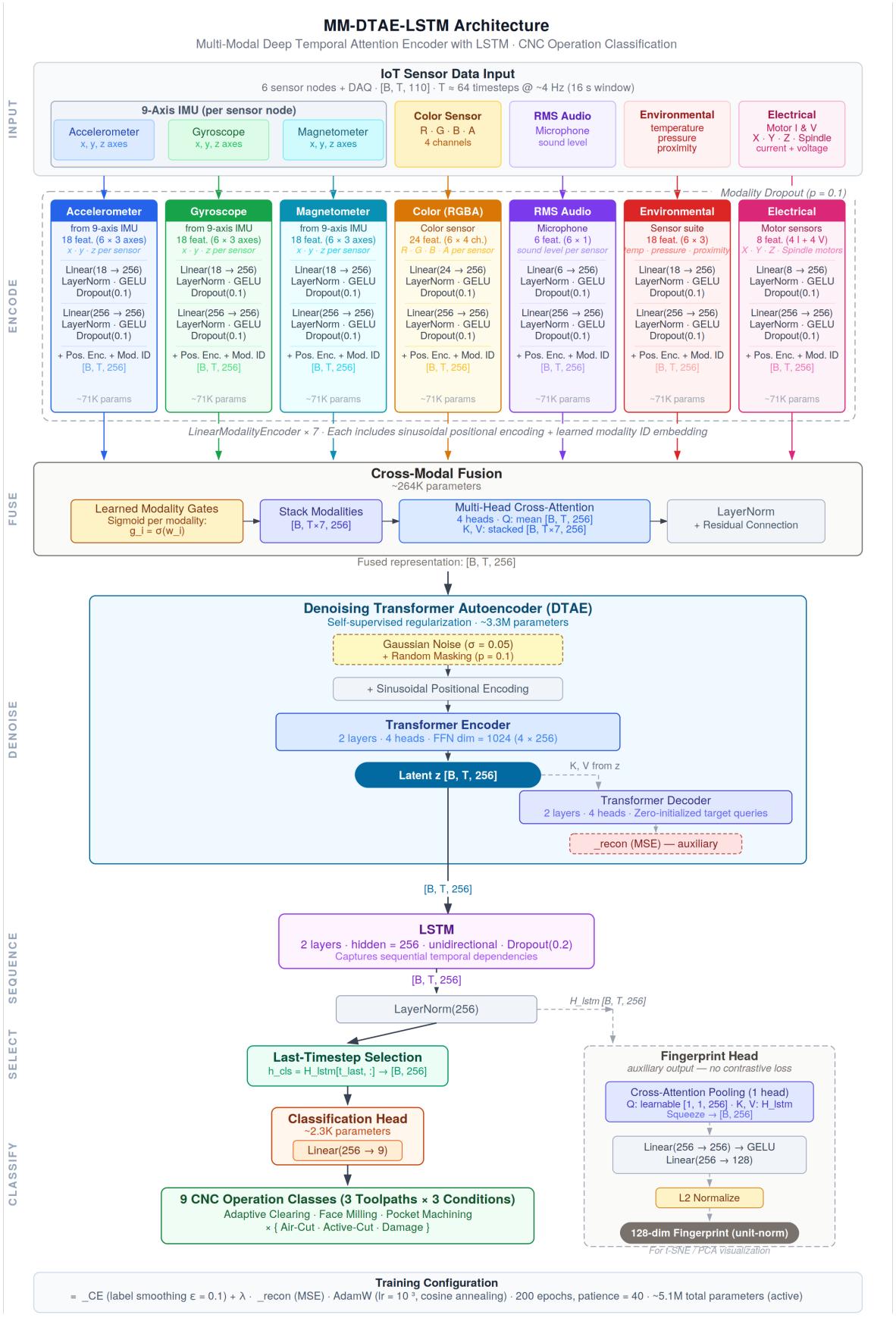


Figure 3. MM-DTAE-LSTM architecture. Modality-specific projections feed gated cross-modal attention, a denoising transformer autoencoder, and a unidirectional LSTM for 9-class classification.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Algorithm 1 MM-DTAE-LSTM Forward Pass

Require: Modalities $\{M_1, \ldots, M_7\}$, sequence lengths ℓ
Ensure: Class prediction $\hat{y} \in \mathbb{R}^9$

Stage 1: Modality Projection
for $i = 1$ to 7 **do**
  $\tilde{H}_i \leftarrow \text{MLP}_i(M_i) + \text{PE}(T) + \mathbf{e}_i$
end for

Stage 2: Cross-Modal Fusion
if training **then**
  Apply modality dropout (p=0.1, keep ≥ 1)
end if
$g_i \leftarrow \sigma(w_i)$ for each modality *(learned gates)*
$H_{\text{fused}} \leftarrow \text{LN}\Big(\text{LN}(\bar{H} + \text{MHA}(\bar{H}, [\tilde{H}_{1:7}], [\tilde{H}_{1:7}])) + \frac{\sum_i g_i \tilde{H}_i}{\sum_i g_i + \epsilon}\Big)$

Stage 3: Denoising Autoencoder
if training **then**
  Corrupt: $\tilde{H} \leftarrow H_{\text{fused}} + \mathcal{N}(0, 0.05^2)$; mask 10%
  $\mathbf{z} \leftarrow \text{TransformerEnc}(\tilde{H})$
  $\hat{H} \leftarrow \text{TransformerDec}(\mathbf{0}, \mathbf{z})$ *(reconstruction)*
else
  $\mathbf{z} \leftarrow \text{TransformerEnc}(H_{\text{fused}})$
end if

Stage 4–5: Temporal Modeling & Classification
$H_{\text{lstm}} \leftarrow \text{LN}(\text{LSTM}(\mathbf{z}))$
$h_{\text{cls}} \leftarrow H_{\text{lstm}}[\ell_b - 1, :]$ *(last timestep)*
$\hat{y} \leftarrow \text{softmax}(W_{\text{cls}} h_{\text{cls}} + b_{\text{cls}})$

return $\hat{y}$; $\mathcal{L}_{\text{recon}} = \text{MSE}(\hat{H}, H_{\text{fused}})$ if training

3.4.2. Cross-Modal Fusion with Learned Gates

The fusion stage combines M=7 encoded modalities through learned importance gates and cross-modal attention.

Learned gates. Each modality receives a learnable scalar importance weight:

$$g_i = \sigma(w_i), \quad \tilde{S}_{\text{gated}} = \frac{\sum_{i=1}^{M} g_i \cdot \tilde{H}_i}{\sum_{i=1}^{M} g_i + \epsilon} \tag{3}$$

Cross-modal attention. Multi-head attention is computed with the mean across modalities as query and all modality-timestep pairs as keys/values:

$$Q = \frac{1}{M} \sum_{i=1}^{M} \tilde{H}_i, \quad KV = [\tilde{H}_1; \ldots; \tilde{H}_M], \quad A = \text{MultiHead}(Q, KV, KV) \tag{4}$$

The fused output combines both mechanisms with layer normalization:

$$H_{\text{fused}} = \text{LN}\big(\text{LN}(Q + A) + \tilde{S}_{\text{gated}}\big) \tag{5}$$

Modality dropout. During training, entire modalities are stochastically zeroed with probability $p_{mod} = 0.1$ (at least one modality always remains active), preventing over-reliance on any single modality and enabling graceful degradation when sensors fail at inference time.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

3.4.3. Denoising Transformer Autoencoder (DTAE)

The DTAE provides a self-supervised learning signal that regularizes the fused representation. During training, the fused features are corrupted with additive Gaussian noise ($\sigma = 0.05$) and random masking (10% of positions zeroed):

$$\tilde{H}_{\text{noisy}} = (H_{\text{fused}} + \mathcal{N}(0, \sigma^2)) \odot (1 - \mathbf{m}), \quad m_{t,j} \sim \text{Bernoulli}(0.1) \tag{6}$$

A 2-layer Transformer encoder encodes the corrupted input; a matching 2-layer Transformer decoder cross-attends to reconstruct the clean fused features:

$$\mathbf{z} = \text{TransformerEncoder}(\tilde{H}_{\text{noisy}}), \quad \hat{H}_{\text{recon}} = \text{TransformerDecoder}(\mathbf{0}, \mathbf{z}) \tag{7}$$

Decoder queries are initialized to zero so that reconstruction at every position is driven purely by cross-attention to the encoder's latent representation $\mathbf{z}$, without any autoregressive or positional prior.

The reconstruction loss $\mathcal{L}_{\text{recon}} = \text{MSE}(\hat{H}_{\text{recon}}, H_{\text{fused}})$ encourages learning robust, noise-invariant representations—particularly valuable when signal quality varies across modalities and operating conditions.

3.4.4. Temporal Modeling and Classification Head

The DTAE latent representation $\mathbf{z}$ is processed by a 2-layer unidirectional LSTM followed by LayerNorm:

$$H_{\text{lstm}} = \text{LN}(\text{LSTM}(\mathbf{z})) \in \mathbb{R}^{T \times d_{model}} \tag{8}$$

We use unidirectional LSTM to preserve causal structure for streaming inference. Classification uses the last valid timestep:

$$\hat{y} = \text{softmax}(W_{\text{cls}} H_{\text{lstm}}[t_{\text{last}}, :] + b_{\text{cls}}) \in \mathbb{R}^9 \tag{9}$$

The total training loss combines classification and reconstruction:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{recon}}, \quad \lambda = 0.1 \tag{10}$$

where $\mathcal{L}_{\text{cls}}$ is cross-entropy with label smoothing [41] ($\epsilon = 0.1$) and inverse-frequency class weights to address class imbalance. The model also produces a 128-dimensional fingerprint embedding via an attention-pooled projection head (Appendix B), used for the embedding visualizations in Section 4.

3.4.5. Parameter Count

Table 5 summarizes the parameter budget. With 7 modality encoder groups at $d_{model} = 256$, the active components total ∼5.1M parameters.

Table 5. Parameter Count Breakdown (Active Components for 9-Class Classification).

Component	Parameters	%
Modality Projections (7×)	∼498K	9.7
Modality Embeddings	1.8K	<0.1
Cross-Modal Fusion (gates + attention)	∼264K	5.2
DTAE (2 encoder + 2 decoder layers)	∼3.3M	64.3
LSTM (2 layers, d_{model}=256)	∼1.05M	20.5
Classification Head (256 → 9)	2.3K	<0.1
Total (active)	**∼5.1M**	**100**

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

3.5. Training Configuration

Table 6. Training Hyperparameters.

Parameter	Value
Model dimension (d_{model})	256
LSTM layers	2
Attention heads	4
DTAE layers (encoder + decoder)	2 + 2
DTAE feedforward dimension	$4 \times d_{model} = 1024$
Dropout (DTAE, LSTM)	0.2
Dropout (modality projections)	0.1
Modality dropout	0.1
Noise level (σ)	0.05
Mask probability	0.1
Reconstruction weight (λ)	0.1
Label smoothing (ϵ)	0.1
Gradient clipping (max norm)	1.0
Optimizer	AdamW [42] (β_1=0.9, β_2=0.999)
Learning rate	10^{-3}
Weight decay	0.01
LR schedule	Linear warmup (10 ep.) + Cosine annealing [43]
Cosine restart period	$T_0 = 30$, $T_{\text{mult}} = 2$
Max epochs	200
Early stopping patience	40
Batch size	32
Random seed	42

All experiments were executed on a workstation with dual NVIDIA RTX A6000 GPUs (48 GB VRAM each), running Python 3.10, PyTorch 2.5 with CUDA 12.1, and Ubuntu Linux. Encoder and neural baseline models trained on a single GPU; traditional ML baselines used CPU (scikit-learn [44], XGBoost [45]).

3.6. Baseline Models

We compare MM-DTAE-LSTM against five baseline classifiers spanning traditional machine learning and neural network approaches:

- **XGBoost** [45]: Gradient-boosted decision trees on flattened window features
- **Random Forest** [46]: Ensemble of 200 decision trees on flattened features
- **Logistic Regression**: L2-regularized multinomial logistic regression on flattened features
- **MLP**: Three-layer feedforward network (512–256–128) with batch normalization, ReLU, and dropout (0.3)
- **SimpleLSTM**: Two-layer unidirectional LSTM (256 hidden) with dropout (0.3) and last-timestep classification, serving as an ablation control that isolates temporal modeling without cross-modal fusion, denoising, or attention

Traditional ML baselines (XGBoost, Random Forest, Logistic Regression) operate on flattened window features ($T \times F$ reshaped to $T \cdot F$), discarding temporal structure. Neural baselines (MLP, SimpleLSTM) receive the same windowed sequences as MM-DTAE-LSTM. All neural baselines share the same AdamW optimizer, label smoothing, gradient clipping, batch size, and early stopping as the encoder (Table 6), ensuring fair comparison.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

3.7. Evaluation Protocol

3.7.1. Five-Fold File-Level Stratified Cross-Validation

To prevent data leakage and provide robust variance estimates, we employ 5-fold cross-validation with file-level stratification. Within each class, files are sorted alphabetically and assigned to folds by deterministic cyclic allocation:

$$\text{fold}(f) = (\text{sorted\_index}(f) \bmod 5) + 1 \tag{11}$$

For fold k: test = files in fold k; validation = files in fold $(k \bmod 5) + 1$; train = all remaining files. This ensures all windows from a single operation file appear in exactly one partition, preventing temporal autocorrelation from inflating accuracy estimates. We report mean $\pm$ standard deviation across the 5 folds.

3.7.2. Data Preprocessing

Raw aligned CSV files undergo a three-stage preprocessing pipeline executed independently for each cross-validation fold, with all statistics computed on training data only to prevent leakage.

1. **Feature extraction, filtering, and imputation.** Continuous sensor and electrical channels are extracted; metadata that would cause leakage (machine position, feed rate, G-code text) are excluded. Three automated filters remove uninformative channels: (a) >50% missing values; (b) zero-variance (constant) signals; (c) highly correlated pairs ($|r| > 0.95$), retaining the higher-variance member. For feature ablation experiments, modality groups are excluded at this stage per Table 7. Missing values are imputed by forward-fill, back-fill, then zero-fill, preserving temporal continuity across brief communication dropouts.
2. **Outlier clipping and normalization.** Extreme values are clipped at $Q_1 - 3{\cdot}\text{IQR}$ / $Q_3 + 3{\cdot}\text{IQR}$ before scaler fitting. A RobustScaler (median centering, IQR scaling) is fitted *only on training-fold data* and applied identically to validation and test folds. RobustScaler is preferred over StandardScaler because sensor distributions are predominantly non-normal with heavy tails characteristic of sensor dropout recovery transients.
3. **Sliding window segmentation.** Normalized time series are segmented into overlapping windows of configurable size (16–256 timesteps) with configurable stride. Only complete windows are created. Each window receives a majority-vote G-code operation label mapped to one of the nine classes.

Table 7. Cumulative Feature Ablation Configurations.

#	Configuration	Removed	Features
1	Full	—	110
2	No Proximity	Proximity	104
3	No Prox + Pressure	Proximity, Pressure	98
4	No Prox + Pres + Color	Proximity, Pressure, Color	74
5	Minimal	Prox, Pres, Color, Magnetometer	56

3.8. Feature Ablation Design

We design a cumulative feature ablation that progressively removes sensor modalities ordered by expected informativeness (Table 7). The removal order reflects domain knowledge: proximity sensors produce near-constant output; barometric pressure varies negligibly in open-air environments; color channels encode gantry position via LED shadows but are environment-dependent; magnetometer captures spindle motor fields but overlaps with accelerometer information.

The minimal configuration (56 features) retains accelerometer, gyroscope, temperature, RMS audio, and electrical channels—the core inertial, acoustic, and machine-level modalities.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

3.9. Temporal Resolution Ablation

We evaluate 10 window/stride combinations spanning 16 to 256 timesteps (Table 8), corresponding to approximately 4 to 64 seconds of machining time at the ∼4 Hz aligned sampling rate.

Table 8. Window/Stride Configurations.

Window	Stride	Overlap	Duration (s)	Ratio
16	8	50%	∼4	2.0×
32	8	75%	∼8	4.0×
32	16	50%	∼8	2.0×
64	16	75%	∼16	4.0×
64	32	50%	∼16	2.0×
64	64	0%	∼16	1.0×
128	32	75%	∼32	4.0×
128	64	50%	∼32	2.0×
256	64	75%	∼64	4.0×
256	128	50%	∼64	2.0×

3.10. Experiment Summary

The full factorial design spans 5 feature configurations × 10 temporal resolutions × 6 model families × 5 CV folds = **1,500 individual training runs**. We report 690 runs from the subset detailed in Table 9.

Table 9. Experiment Design: Reported Subset of the Full Factorial.

Phase	Feature Configs	Temporal Configs	Models × Folds	Runs
Feature ablation	All 5 levels	1 (default w=64, s=16)	6 × 5	150
Temporal ablation	2 (full + minimal)	9 (remaining)	6 × 5	540
Total (unique runs)				**690**

Phase 1 evaluates all five feature-ablation levels at the default temporal resolution. Phase 2 evaluates the remaining nine window/stride configurations at both the full (110-feature) and minimal (56-feature) configurations. The default temporal resolution at full and minimal features appears in Phase 1, yielding 23 unique (feature, temporal) pairs × 6 models × 5 folds = 690 total runs.

4. Results

4.1. Training Dynamics, Embedding Visualizations, and Raw Signal Analysis

Training converges within ∼30 epochs at the best configuration (no proximity/pressure, w=64, s=16), with a modest generalization gap; training accuracy saturates near 100% while validation and test plateau at ∼95%. Averaged learning curves appear in Figure A1; per-fold learning curves appear in Figure A2 (Appendix A).

Figure 4 presents t-SNE [47] visualization of 128-dimensional fingerprint embeddings from test sets across all 5 folds (3,479 samples). Air-cut classes form tight, well-separated clusters. Active-cut classes cluster near their corresponding air-cut toolpaths, reflecting shared trajectory geometry with additional cutting-force signatures. Damage classes partially overlap their air-cut counterparts, consistent with shared toolpath dynamics differing primarily in spindle vibration characteristics.

PCA of the same embeddings (Figure A4, Appendix A) confirms linear structure: the first two components capture 37.6% of total embedding variance, with toolpath families forming visually distinct regions; five components account for 66.2%; and 21 components reach 95%.

Figure 5 shows accelerometer magnitude signatures for the three air-cut toolpath strategies, illustrating the physically distinct motion dynamics that underpin classification. Additional multi-sensor and multi-modal visualizations appear in Figures A5–A6 (Appendix A).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

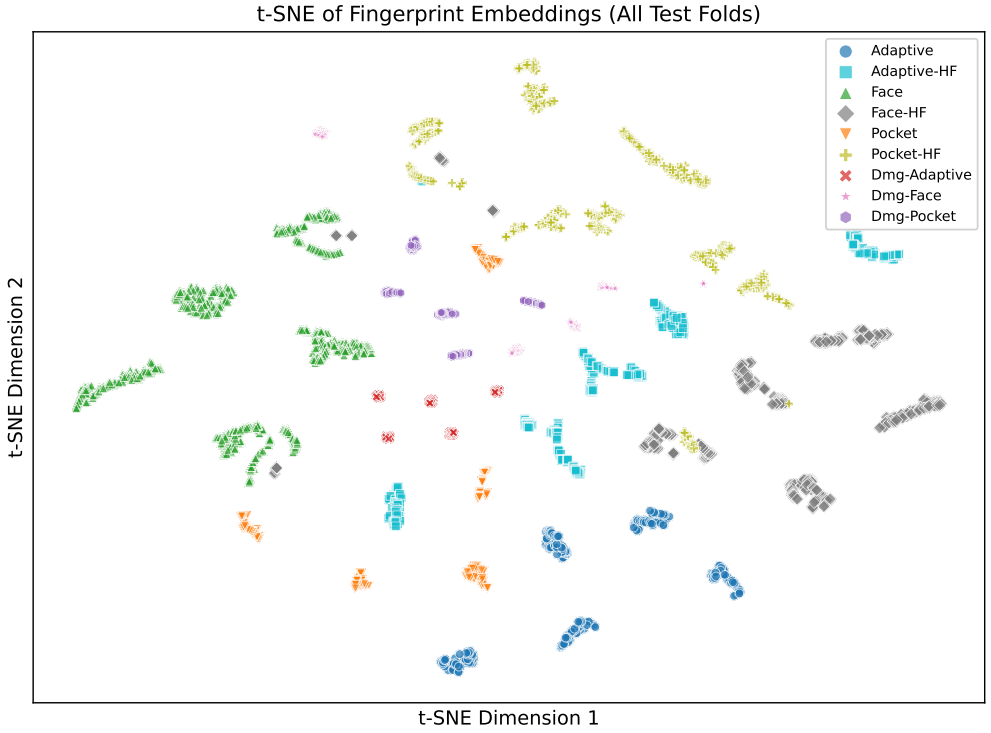


Figure 4. t-SNE visualization of 128-dimensional fingerprint embeddings (projections of LSTM hidden states through the attention-pooled fingerprint head) from MM-DTAE-LSTM, pooled across all 5 test folds (3,479 samples; perplexity = 30). Colors indicate class; marker shapes indicate toolpath strategy. See text for interpretation.

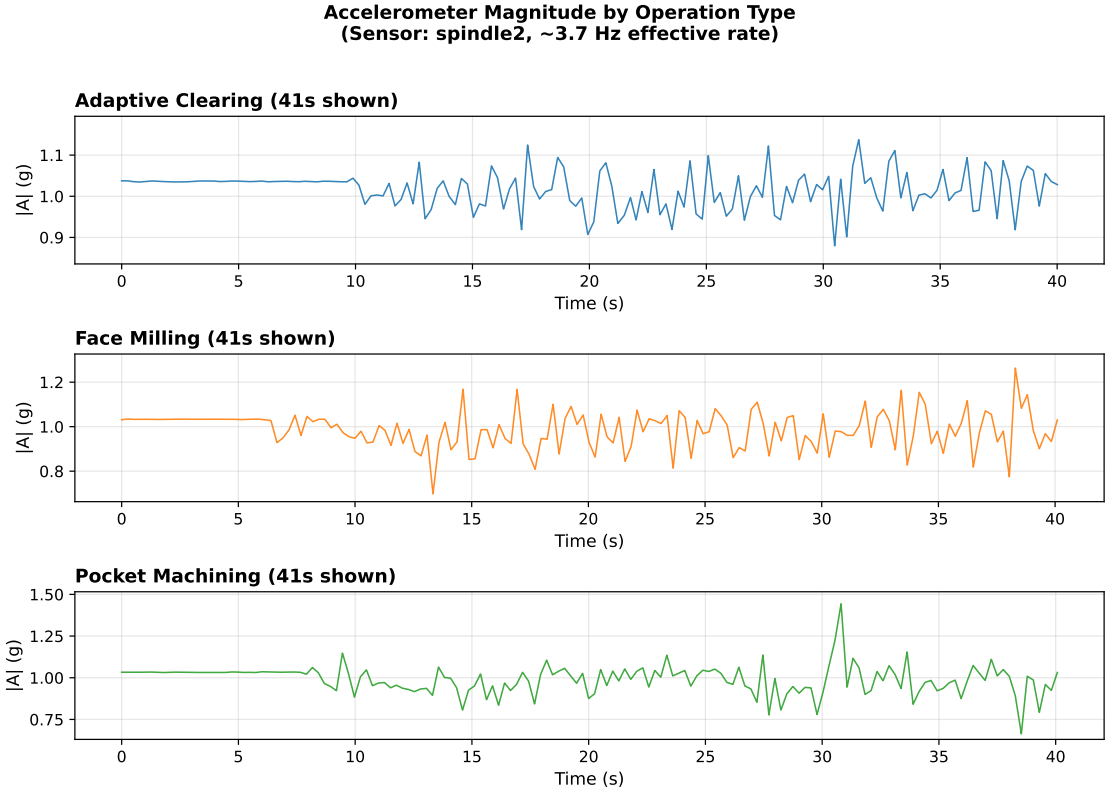


Figure 5. Accelerometer magnitude signatures for three air-cut toolpath strategies (spindle2 sensor). Adaptive clearing shows periodic direction-change patterns; face milling exhibits steady-state gantry motion; pocket machining displays intermittent plunge/lateral transitions.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

4.2. Feature Ablation Results

Table 10 presents classification accuracy across all five feature configurations at the default temporal resolution (window=64, stride=16).

Table 10. Feature Ablation Results (Window=64, Stride=16, 5-Fold CV Mean ± Std).

Config	Feat.	MM-DTAE-LSTM	XGBoost	RF	Logistic	MLP	LSTM
Full	110	93.5±7.2	95.1±4.6	95.6±5.9	92.6±9.4	94.3±8.9	93.3±10.2
No Prox	104	95.2±8.0	95.5±4.8	95.7±5.5	92.7±9.7	92.8±10.4	92.7±9.3
No Prox+Pres	98	**96.3±4.7**	92.4±6.3	91.1±7.9	92.5±10.0	93.2±9.8	92.5±11.5
No Prox+Pres+Col	74	94.7±8.0	92.4±5.3	91.5±7.6	92.0±9.2	93.4±9.6	91.9±10.1
Minimal	56	92.5±8.3	84.4±9.7	88.0±5.9	85.4±11.7	90.2±11.0	88.7±13.0

The encoder achieves its best accuracy (96.3%) after removing proximity and pressure—two low-variance environmental modalities whose removal eliminates channels that the modality gates would otherwise learn to down-weight, freeing model capacity for genuinely discriminative modalities. From this 98-feature configuration to the minimal 56-feature set, the encoder loses only 3.8 points (96.3%→92.5%), while XGBoost loses 8.0 points (92.4%→84.4%) and Logistic Regression loses 7.1 points; measured from full 110 features, XGBoost's total drop is 10.7 points (95.1%→84.4%). The neural baselines (MLP and SimpleLSTM) show intermediate robustness, losing 3.0 and 3.8 points respectively. This asymmetry indicates that tree-based and linear models rely more heavily on easily separable environmental features, while the encoder's cross-modal attention extracts discriminative signal from the remaining inertial and acoustic channels.

4.3. Temporal Resolution Results

Figure 6 shows classification accuracy as a function of window size for all six models at the full 110-feature configuration.

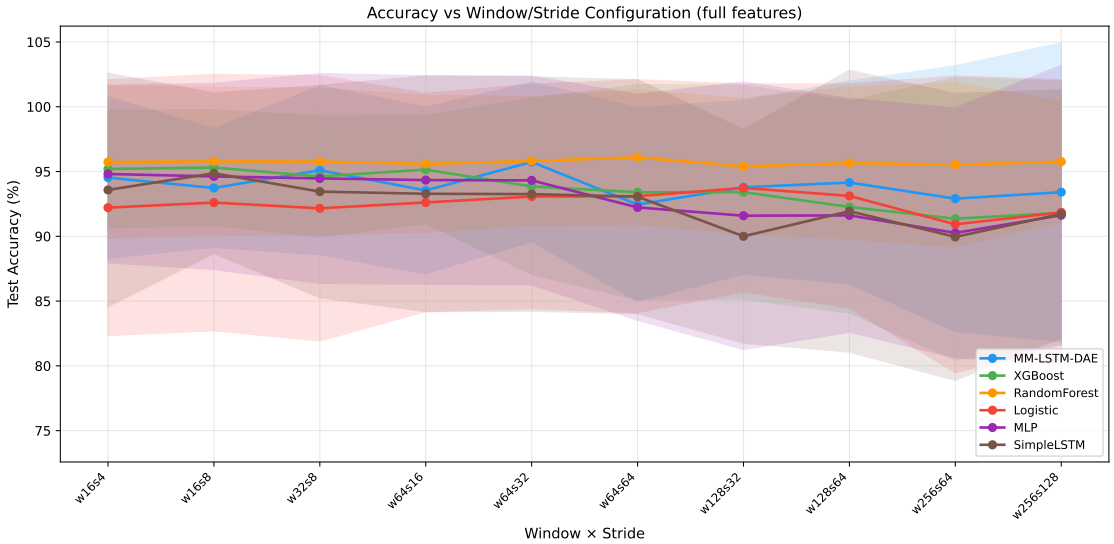


Figure 6. Classification accuracy as a function of window size for each model (full 110-feature configuration). Error bars show ±1 standard deviation across 5 folds. The encoder and neural baselines peak at medium windows (32–64), while Random Forest remains remarkably stable across all window sizes.

Temporal resolution has a modest effect on encoder performance: accuracy ranges from 92.5% (w=64, s=64, no overlap) to 95.7% (w=64, s=32, 50% overlap) with full features. At the ∼4 Hz aligned rate, 64 timesteps span ∼16 seconds—sufficient to capture several direction-change or step-over cycles.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Random Forest maintains 95.5–96.1% regardless of window size, confirming that its feature-level splits are insensitive to temporal extent. The encoder's advantage emerges at medium windows (32–64) where LSTM temporal modeling adds value without excessively reducing sample counts. Full vs. minimal feature comparison and per-model degradation plots appear in Figures A7–A8 (Appendix A).

4.4. Model Comparison

The primary contribution of this comparison is *comprehensive empirical characterization* across 690 experiments, not a narrow claim that the encoder statistically dominates all baselines. At the best configuration, the encoder leads all five baselines by 3.1–5.2 percentage points in accuracy and 0.9–5.7 points in macro F1, with small-to-medium Cohen's d effect sizes ($d = 0.45$–0.74)—though all 95% CIs include zero given n=5. Notably, the encoder outperforms every baseline in *every* comparison; the probability of this consistent direction by chance is $(1/2)^5 = 3.1\%$, providing non-parametric evidence of a genuine advantage.

Table 11. Model Comparison at Best Configuration (No Prox+Pres, w=64, s=16). Cohen's d is computed from 5 paired fold-level accuracy differences (encoder − baseline); 95% CIs use the t-distribution with df=4.

Model	Acc (%)	Macro F1 (%)	Δ Acc	Cohen's d	p (paired t)
MM-DTAE-LSTM	**96.3 ± 4.7**	**91.7 ± 8.0**	—	—	—
XGBoost	92.4 ± 6.3	88.7 ± 11.2	+3.8	0.46 [−0.85, 1.76]	0.314
Random Forest	91.1 ± 7.9	86.0 ± 14.6	+5.2	0.74 [−0.66, 2.14]	0.136
Logistic Reg.	92.5 ± 10.0	90.7 ± 10.4	+3.8	0.56 [−0.77, 1.90]	0.233
MLP	93.2 ± 9.8	90.0 ± 12.9	+3.1	0.47 [−0.84, 1.78]	0.299
SimpleLSTM	92.5 ± 11.5	90.8 ± 11.6	+3.8	0.45 [−0.85, 1.75]	0.323

Note on statistical power: With n=5 folds, paired t-tests have limited power; p-values should be interpreted alongside effect sizes and consistency of direction. *Note on configuration selection*: All models evaluated at the encoder's best feature configuration (No Prox+Pres, 98 features). Baselines may achieve higher accuracy at other configurations—e.g., Random Forest peaks at 95.7% with 104 features (Table 10)—but diverge from the encoder at reduced feature sets.

The encoder also achieves the lowest fold-to-fold variance (std = 4.7% accuracy, 8.0% macro F1), compared to 6.3–11.5% and 10.4–14.6% for baselines. The encoder trains in ∼3.3 minutes per fold on a single A6000 GPU—comparable to XGBoost on CPU (∼3.2 min) and substantially faster than MLP (∼11 min) and SimpleLSTM (∼13 min) on the same GPU. The encoder's faster wall time relative to SimpleLSTM is attributed to earlier convergence (early stopping at ∼30 epochs) enabled by the DTAE auxiliary reconstruction loss, which regularizes the representation and stabilizes optimization; SimpleLSTM, without this auxiliary objective, trains longer before the validation criterion plateaus. Training time details by model appear in Table A1 (Appendix A).

4.5. Per-Class Performance

Table 13 reports per-class F1 scores, which account for both precision and recall. The encoder achieves ≥97% F1 on all three air-cut classes and 100% on damageadaptive, confirming reliable classification when ≥16 training files are available. Active-cut classes (adaptive150025, face150025, pocket150025) share toolpath geometry with their air-cut counterparts but add cutting-force signatures, making within-toolpath discrimination harder. The damage-class metrics should be interpreted as preliminary rather than reliable performance estimates. With only 4–5 files per class distributed across 5 folds, some folds contain zero test samples, making per-class accuracy and F1 statistically meaningless for these classes. The reported values (damageface F1 = 65.8±39.4%, damagepocket F1 = 73.6±43.4%) reflect this pathological variance rather than model capability. All models struggle comparably, confirming that data quantity—not model capacity—is the bottleneck. Meaningful damage-class evaluation requires ≥10 files per class.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Table 12. Per-Class Accuracy (%, Mean ± Std Across 5 Folds) at Best Encoder Configuration. Damage-class metrics (bottom three rows) are preliminary: with only 4–5 files, some folds contain zero test samples, producing extreme variance that precludes reliable performance estimation.

Class	Files	MM-DTAE-LSTM	XGBoost	RF	MLP	LSTM
adaptive	20	100.0±0.0	93.9±9.3	91.9±11.5	100.0±0.0	100.0±0.0
adaptive150025	17	99.3±1.1	98.6±3.2	100.0±0.0	99.1±2.0	97.1±6.4
face	16	99.6±1.0	89.4±11.3	93.9±7.9	98.7±1.6	97.0±4.2
face150025	15	95.6±6.9	93.3±9.6	85.8±23.5	91.5±14.0	85.3±20.4
pocket	19	100.0±0.0	95.0±8.5	91.0±12.4	100.0±0.0	98.7±2.1
pocket150025	19	93.6±8.5	94.1±11.0	91.6±10.0	80.6±34.4	82.8±33.8
damageadaptive	5	100.0±0.0	100.0±0.0	81.4±41.5	100.0±0.0	100.0±0.0
damageface	4	76.2±43.4	58.1±53.1	69.4±45.1	77.5±43.7	77.5±43.5
damagepocket	5	70.3±44.6	88.0±26.8	80.0±44.7	81.7±40.9	93.1±15.3

Table 13. Per-Class F1-Score (%, Mean ± Std Across 5 Folds) at Best Encoder Configuration.

Class	Files	MM-DTAE-LSTM	XGBoost	RF	MLP	LSTM
adaptive	20	100.0±0.0	88.3±13.1	93.0±6.3	100.0±0.0	97.3±3.6
adaptive150025	17	99.0±1.5	97.2±4.6	100.0±0.0	99.1±2.2	96.8±7.3
face	16	99.0±1.7	92.4±9.5	94.0±5.8	98.7±1.7	98.2±2.6
face150025	15	94.0±9.3	93.1±6.6	84.8±16.1	91.5±14.8	85.3±20.2
pocket	19	97.8±2.3	91.6±10.7	90.4±9.7	100.0±0.0	93.1±10.4
pocket150025	19	96.0±5.6	95.2±6.1	90.4±8.4	80.6±38.6	83.4±29.1
damageadaptive	5	100.0±0.0	93.9±12.0	82.7±38.8	100.0±0.0	99.3±1.0
damageface	4	65.8±39.4	56.8±52.0	63.0±41.4	77.5±48.9	72.8±42.4
damagepocket	5	73.6±43.4	89.5±18.2	76.1±43.4	81.7±45.7	91.6±16.5

Note: Logistic Regression is omitted from per-class Tables 12–13 for space; its aggregate performance (92.5% ± 10.0% accuracy, 90.7% ± 10.4% macro F1) appears in Table 11.

4.6. Confusion Matrices

Figure 7 presents row-normalized confusion matrices aggregated across 5 folds for all six models.

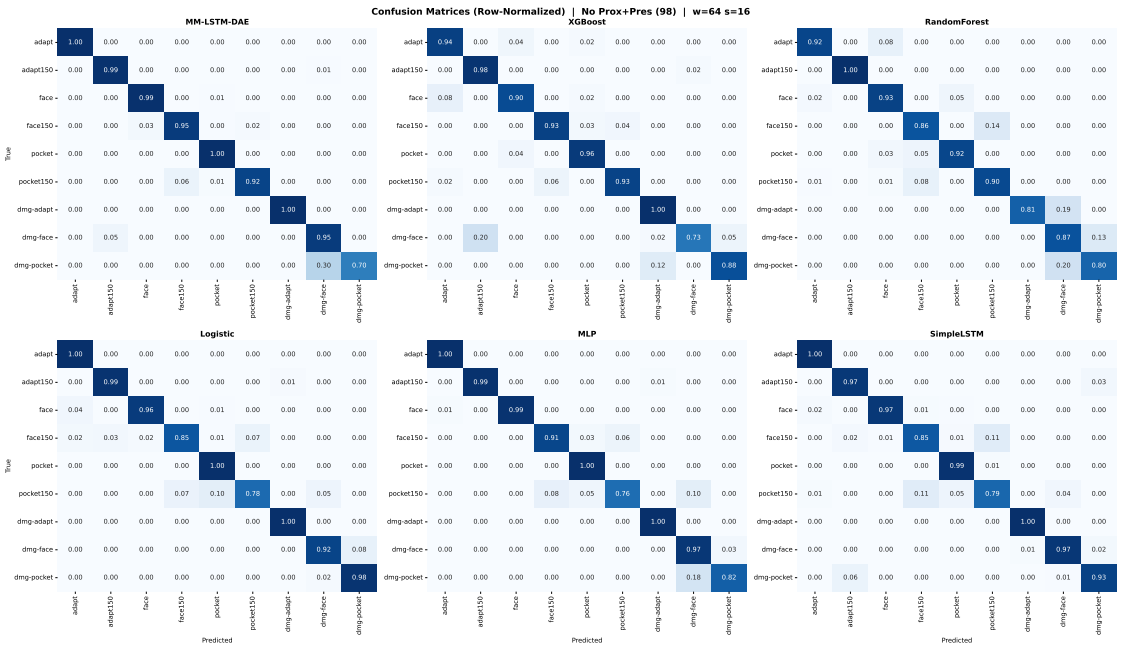


Figure 7. Row-normalized confusion matrices aggregated across 5 folds for all six models at the best configuration (no proximity/pressure, w=64, s=16). The encoder achieves near-diagonal structure for air-cut and active-cut classes while damage classes show diffuse misclassification across all models.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Misclassifications occur primarily *within* toolpath families rather than *across* them—`damagepocket` is confused with `pocket` and `pocket150025`, never with face or adaptive classes. This pattern holds across all models, confirming that toolpath geometry produces fundamentally distinct machine dynamics even during air-cutting, while within-family condition discrimination is the harder subtask. Damage embeddings are dispersed among non-damage clusters rather than forming a distinct region (Figure A9, Appendix A), confirming that spindle degradation produces subtle signature shifts within each toolpath family (see also Figure A3, Appendix A).

5. Discussion

5.1. Feature Importance and Modality Contributions

The ablation results (Table 10) reveal a clear modality hierarchy. Accelerometer and gyroscope channels—which encode the structural vibration transfer function of the machine under different toolpath excitations—are essential. Removing low-variance environmental channels (proximity, pressure) *improves* the encoder by 2.8 points, while further removing color and magnetometer degrades it by only 3.8 points total.

Color channels merit comment: the APDS-9960 captures LED shadow patterns that correlate with gantry position, providing a weak position proxy. Their marginal value atop inertial modalities is small (1.6 points), consistent with the position information already implicit in acceleration sequences.

The minimal 56-feature configuration (accelerometer, gyroscope, temperature, RMS audio, and electrical) retains only inertial, acoustic, and machine-level modalities yet achieves 92.5%—a favorable accuracy–complexity tradeoff for cost-constrained retrofits where each additional sensor board adds hardware, wiring, and maintenance burden.

5.2. Statistical Analysis of Feature Discriminative Power

To quantify the discriminative contribution of individual sensor channels, we performed a one-way ANOVA on per-file summary statistics across the nine operation classes ($df_{between} = 8$, $df_{within} = 111$; group sizes range from n=4 to n=20). For each of the 110 features, we computed the per-file mean, then tested for between-class differences using the F-statistic and effect size (η^2). Significance is assessed at $\alpha = 0.05$ with Bonferroni correction for 110 simultaneous tests ($\alpha_{adj} = 4.5 \times 10^{-4}$); all features reported as significant remain so after correction. Table 14 summarizes results by modality group; Figure 8 visualizes distributions (per-sensor heatmap in Figure A10, Appendix A).

Table 14. ANOVA results by modality group. Median F-statistic and effect size (η^2) across features within each modality. Modalities above the horizontal line were removed during feature ablation.

Modality	**Median** F	**Mean** F	η^2	N_{feat}	N_{sig}
Pressure	2423.2	2067.1	0.994	6	6
Proximity	41.9	44.5	0.749	6	6
Color	17.0	19.8	0.551	24	24
Magnetometer	18.2	167.5	0.568	18	18
RMS Audio	91.4	79.1	0.866	6	6
Gyroscope	11.4	17.9	0.450	18	17
Accelerometer	10.8	26.5	0.438	18	14
Temperature	4.5	4.9	0.246	6	6
Electrical	5.1	5.0	0.271	8	6

Pressure channels exhibit the highest F-statistics ($F > 2{,}000$, $\eta^2 > 0.99$), but this reflects a *session confound*: barometric pressure varies between collection days (∼100.0 vs. ∼101.5 kPa); since certain operation classes were disproportionately collected on specific days, this day-level pressure variation spuriously correlates with class labels. Within-class standard deviation is only ∼0.05 kPa, confirming that pressure encodes *when* data were collected, not *what* was being machined. Proximity channels

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

similarly produce high F from near-constant readings that vary by sensor placement rather than machining dynamics.

This confound is a direct consequence of the two-day collection protocol, in which operation classes were not randomized across collection days. Future studies should either randomize operation–day assignments or collect all classes on every collection day to deconfound environmental variation from machining dynamics—a study-design lesson applicable to any multi-session sensor dataset.

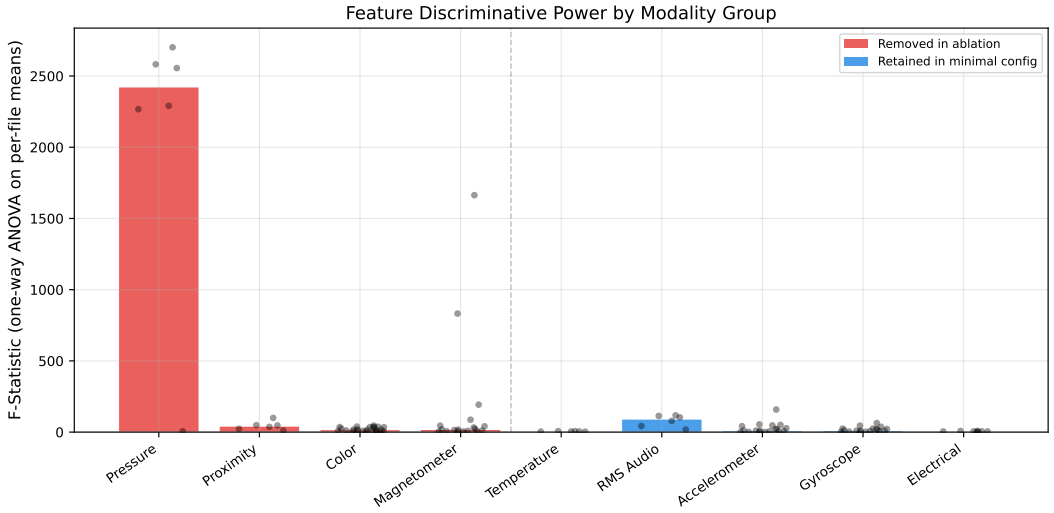


Figure 8. Distribution of ANOVA F-statistics by modality group. Red bars indicate modalities removed during feature ablation; blue bars indicate retained modalities. Individual feature F-values overlaid as scatter points. Pressure channels (off-scale, $F > 2{,}000$) dominate due to session confounds rather than machining physics.

This analysis explains the asymmetric impact of feature ablation. Baseline models (Random Forest, XGBoost) exploit high-F features such as pressure and proximity that provide easy class separation through environmental confounds. Removing only confound features (proximity and pressure) drops XGBoost from 95.1% to 92.4% (-2.7 pp); further removing color and magnetometer costs an additional 8.0 pp (to 84.4%). The encoder, by contrast, learns cross-modal representations from lower-F features (accelerometer, gyroscope: median $F \approx 11$, $\eta^2 \approx 0.44$), maintaining 92.5% at the minimal 56-feature configuration—validating the encoder's ability to extract genuine machining signatures rather than relying on environmental artifacts.

5.3. Temporal Resolution Effects

A broad performance plateau exists between w=32 and w=128 (Figure 6), with no single window size dominating across all models. Tree-based models are insensitive to window size because flattening discards temporal ordering; the encoder's LSTM benefits from additional temporal context at medium windows. Large windows (w=256) increase fold-to-fold variance (std 11–13%) because reduced sample counts exacerbate class imbalance in damage categories. The practical recommendation of w=64 ($\sim$16 s at 4 Hz) balances temporal context against dataset size, capturing several complete direction-change or step-over cycles per window.

5.4. Sensor Modality Contributions

The ANOVA analysis and ablation results together yield a concrete modality ranking, ordered by ablation essentiality rather than univariate F-statistic alone. Among retained modalities: (1) **accelerometer** and **gyroscope** (median $F \approx 11$) are essential—encoding complementary structural vibration and rotational dynamics; removing either degrades classification. (2) **RMS audio** (median $F = 91$) carries the highest discriminative power among retained modalities, critical for air-cut vs. active-cut separation via acoustic emission from material engagement. (3) **Electrical** channels (median

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

$F \approx 5$) provide complementary machine-level signals. (4) **Temperature** ($F \approx 4.5$) is marginal—retained by the ablation design rather than demonstrated per-operation necessity.

Modalities *removed* during ablation each fail for a distinct reason: **pressure** ($F > 2{,}000$) encodes a session confound rather than machining physics; **proximity** produces near-constant readings that vary by sensor placement; **color** provides a weak gantry-position proxy already implicit in acceleration sequences; and **magnetometer** is partially redundant with accelerometer channels. Detailed physical interpretation of each modality's sensing mechanism appears in Appendix C.

5.5. Model Complexity vs. Performance

At full sensor coverage, baselines achieve competitive accuracy (up to 95.6% for Random Forest), indicating that the data appear well-structured for classification at full features—though, as the ANOVA reveals (Section 5.2), part of this structure reflects session-level confounds rather than machining physics alone. The encoder's advantage emerges under sensor reduction: at 110 features, Random Forest leads the encoder by 2.1 points (95.6% vs. 93.5%), but at 56 features the encoder (92.5%) leads every baseline—including MLP (90.2%)—while tree-based models degrade sharply (RF to 88.0%, XGBoost to 84.4%). This crossover suggests two deployment regimes: (1) when all sensors are available, lightweight models (Random Forest, XGBoost) provide excellent accuracy at minimal cost; (2) when retrofitting with limited sensor access, the encoder's cross-modal attention maintains performance where classical models degrade.

A natural question is whether 5.1M parameters are justified for 120 files and 9 classes. The bulk of the parameter budget (64.3%, ∼3.3M) resides in the DTAE Transformer layers, which serve a regularization role via self-supervised denoising rather than adding classification capacity. The classification-critical pathway—modality projections (498K), cross-modal fusion (264K), LSTM (1.05M), and classification head (2.3K)—totals ∼1.8M parameters. Whether a smaller d_{model} (e.g., 128 vs. the current 256) would achieve comparable accuracy is an open question; we did not perform a model-size ablation and acknowledge this as a limitation (Section 5.8).

5.6. Implications for Process Verification and Security Monitoring

Operation classification at 96.3% accuracy demonstrates the feasibility of sensor-based process verification and security monitoring on a benchtop CNC platform; extending to production manufacturing systems requires revalidation across industrial machines, materials, and environments. As CNC systems become network-connected, the attack surface expands: controller firmware may be compromised, G-code tampered during transmission, or tool offsets injected [3]. Controller logs and network monitoring fail when the controller itself is the attack vector.

Sensor signatures provide independent physical verification. Learned representations encode expected physical patterns; deviations indicate anomalous execution regardless of what controller logs report, contributing an independent physical verification layer consistent with NIST SP 800-82 [4].

Fingerprint embeddings quantify operation identity. Mean intra-class cosine similarity (sample to own class centroid, computed on the 20-component PCA projection capturing 95% of embedding variance) is 0.81 ± 0.18, while mean inter-class centroid similarity is -0.08 ± 0.18—class centroids are nearly orthogonal in embedding space. Air-cut classes achieve the tightest clusters (adaptive: 0.92, face: 0.88, pocket: 0.94), while damage classes show lower cohesion (0.65–0.78), consistent with limited training data.

Minimal-sensor deployment enables low-cost monitoring. The 56-feature minimal configuration (92.5% accuracy) requires only accelerometer, gyroscope, temperature, audio, and electrical channels—all available on commodity MEMS boards (approximately $33 per board; Table 15) and a standard data acquisition module, enabling independent verification on legacy machines without controller modification.

Reconstruction error as an anomaly signal. The DTAE's reconstruction loss provides an unsupervised anomaly score without requiring damage-specific training labels. Per-window reconstruction MSE discriminates damage windows (n=443) from non-damage windows (n=3,036) with AUROC

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

$= 0.69 \pm 0.07$ (Figure 9)—insufficient for standalone anomaly detection. This is a negative but informative result: the DTAE's denoising objective is optimized to reconstruct the fused modality embedding, which encodes toolpath geometry rather than spindle health. Consequently, reconstruction error reflects how well the model recognizes the *operation type*, not whether the spindle is damaged. This architectural insight suggests that anomaly-sensitive representations require either a dedicated health-aware training objective or explicit fault-condition supervision, rather than repurposing a task-optimized reconstruction loss.

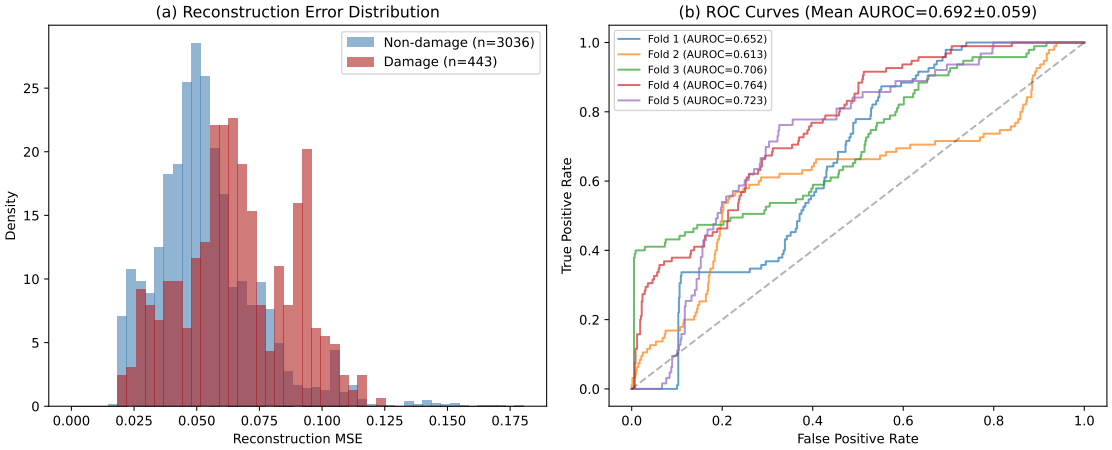


Figure 9. Reconstruction-error anomaly detection. (a) Distribution of per-window reconstruction MSE for non-damage (blue) and damage (red) test samples pooled across 5 folds. Distributions overlap substantially, indicating that damage signatures are subtle in the fused representation space. (b) Per-fold ROC curves; mean AUROC = 0.69 ± 0.07.

5.7. Deployment Recommendations

- **Feature configuration**: The no-proximity/no-pressure configuration (98 features) achieves the best accuracy and removes channels with minimal discriminative value. For cost-constrained deployment, the minimal 56-feature configuration retains 92.5% accuracy.
- **Temporal resolution**: Window size 64 with stride 16–32 provides the best accuracy–latency tradeoff (∼16 s at 4 Hz). Stride 32 (50% overlap) halves computational load with minimal accuracy reduction.
- **Model selection**: When all sensors are available and environmental conditions match the training distribution, Random Forest provides ∼96% accuracy with sub-second training. However, the ANOVA analysis reveals that much of this accuracy derives from session-level barometric confounds; robustness under changing conditions or limited sensor coverage favors the encoder.
- **Accuracy thresholds**: Toolpath classification (>95% for air-cut and active-cut classes) enables process verification. Fault detection via damage classes requires >10 files per class for reliable performance; the current 4–5 files per damage class are insufficient for production-grade fault detection.

Table 15. Approximate Sensing Hardware Costs (USD, 2025 Pricing).

Component	Cost
Arduino Nano 33 BLE Sense Lite (×6)	$198
MCC USB-202 DAQ (electrical)	$169
Raspberry Pi 4 (aggregator)	$75
Cabling, breadboards, mounting	∼$50
Total	**∼$492**

All six sensor boards are required for both the full (110-feature) and minimal (56-feature) configurations; the feature ablation removes channel types in software, not physical boards. No controller modification required.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

5.8. Limitations

1. **Single machine**: All data were collected from one Bantam Tools Explorer desktop CNC mill. Industrial machines (Haas, DMG Mori, Mazak) differ in structural stiffness, spindle drive type, and vibration propagation paths. The *methodology* transfers directly; accuracy values require per-machine re-validation.
2. **Single material and depth**: Active-cutting used UHMW-PE at 0.025" depth only. Metals produce fundamentally different cutting forces, thermal loads, and acoustic emissions.
3. **Limited operation types**: Three CAM strategies tested. Drilling, threading, and multi-axis contouring remain untested.
4. **Controlled environment**: Laboratory setting with minimal external vibration and consistent ambient temperature. Production floors introduce floor-borne vibration, coolant dynamics, and tool wear progression as confounding factors.
5. **Class imbalance**: Damage classes have only 4–5 files versus 15–20 for air-cut/active-cut classes. With 4 files across 5 folds, some folds receive zero test samples, inflating variance.
6. **Sensor subset**: Only 6 of 16 deployed sensors met the 93% consistency threshold. The full array may yield higher accuracy but cannot be evaluated under cross-validation.
7. **Sampling bandwidth**: The ∼4 Hz aligned rate captures only low-frequency dynamics. Higher-bandwidth acquisition (1–10 kHz) would access tooth-passing harmonics and chatter frequencies.
8. **Statistical power**: With n=5 fold-level observations, no standard paired significance test can achieve $p < 0.05$ (the minimum achievable p for a two-sided sign test is $2\times(1/2)^5 = 0.0625$); results are therefore characterized by effect size and consistency of direction rather than statistical significance.
9. **No model-size ablation**: We did not evaluate smaller architectures (e.g., d_{model}=128). The 5.1M-parameter design may be overparameterized for the 120-file dataset.
10. **No inference latency benchmark**: Training time was reported but per-window inference latency was not measured. Edge deployment requires characterizing inference cost across models.

6. Conclusions

We presented MM-DTAE-LSTM, a multi-modal denoising temporal attention encoder for 9-class CNC operation classification on a benchtop platform (Bantam Tools Explorer, UHMW-PE workpiece), and evaluated it across 690 cross-validated experiments spanning feature ablation, temporal resolution, and baseline comparison.

Key findings:

1. **Process verification and security monitoring**: Learned sensor embeddings provide operation-specific fingerprints (intra-class cosine similarity 0.81 ± 0.18, inter-class -0.08 ± 0.18) suitable for independent physical verification. Reconstruction-based anomaly detection proved ineffective (AUROC = 0.69 ± 0.07), confirming that the denoising objective encodes toolpath geometry rather than fault signatures—an architectural insight relevant to future anomaly detection designs.
2. **Feature robustness and model comparison**: The encoder achieves 96.3±4.7% at its best configuration, leading all baselines by 3.1–5.2 points at the encoder's best feature configuration (98 features); at full features (110), Random Forest (95.6%) and XGBoost (95.1%) both exceed the encoder (93.5%). At full features, baselines achieve competitive accuracy (up to 95.6%), but the encoder's advantage is most pronounced under feature reduction: from its 98-feature optimum to the minimal 56-feature set, the encoder loses only 3.8 points (96.3%→92.5%), whereas XGBoost drops 10.7 points across the full 110→56 range (95.1%→84.4%). Accelerometer, gyroscope, and RMS audio channels carry the dominant discriminative signal.
3. **Temporal resolution**: Window size 64 (∼16 s at 4 Hz) provides the best accuracy; stride 32 (50% overlap) performs comparably and halves computational cost.
4. **Fault detection (preliminary)**: Damage-class evaluation is limited by insufficient data—with only 4–5 files per class, some cross-validation folds contain zero test samples, and the resulting metrics

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

(F1 ranging from 65.8% to 100%) reflect pathological variance rather than model performance. Data quantity is the primary limitation; reliable fault detection requires $\geq$10 files per damage class.

Practical recommendations: For maximum accuracy, deploy the encoder with accelerometer, gyroscope, temperature, color, magnetometer, RMS, and electrical channels (98 features, no proximity/pressure) at w=64, s=16. For edge deployment under stable environmental conditions, Random Forest with full features achieves >95% accuracy, though its performance may degrade under distribution shift given its reliance on session-level confounds (Section 5.2); the encoder with the minimal 56-feature set achieves 92.5% and is more robust to such variation. Fault detection requires expanded datasets (>10 files per damage class) before production deployment.

Future work will address:

- **Generalization across machines and materials**: Transfer learning and few-shot fine-tuning to industrial machines (Haas, DMG Mori) across diverse materials (aluminum, steel, titanium), depths of cut, and extended operation taxonomies (drilling, threading, multi-axis contouring).
- **Expanded fault datasets**: Collecting >10 files per damage class with varied degradation modes (bearing wear, misalignment, tool breakage) to enable reliable production-grade fault detection.
- **High-bandwidth acquisition and real-time edge deployment**: Direct IMU sampling at 1–10 kHz to capture tooth-passing harmonics and chatter onset, paired with edge-optimized inference for continuous streaming process verification.
- **Sensor-to-G-code reconstruction**: Extending the architecture from operation classification to full G-code sequence prediction via a sequence-to-sequence decoder—recovering commanded toolpath, feed rate, and spindle parameters directly from sensor signals for complete program verification.

Author Contributions: Conceptualization, S.S.E. and R.S.P.; methodology, S.S.E.; software, S.S.E.; validation, S.S.E.; formal analysis, S.S.E.; investigation, S.S.E.; resources, M.S.S.; data curation, S.S.E. and R.S.P.; writing—original draft, S.S.E.; writing—review & editing, R.S.P. and M.S.S.; visualization, S.S.E.; supervision, R.S.P. and M.S.S.; project administration, R.S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors thank the Department of Industrial Systems Engineering at the University of Rhode Island for providing laboratory facilities and computational resources.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability Statement: The sensor datasets and analysis code generated during this study are publicly available at https://github.com/stephenseacuello/gcode_fingerprinting.

Software Availability Statement: Source code for the MM-DTAE-LSTM architecture, experiment pipeline, and analysis scripts is available at https://github.com/stephenseacuello/gcode_fingerprinting under the MIT License.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Appendix A. Supplementary Figures and Tables

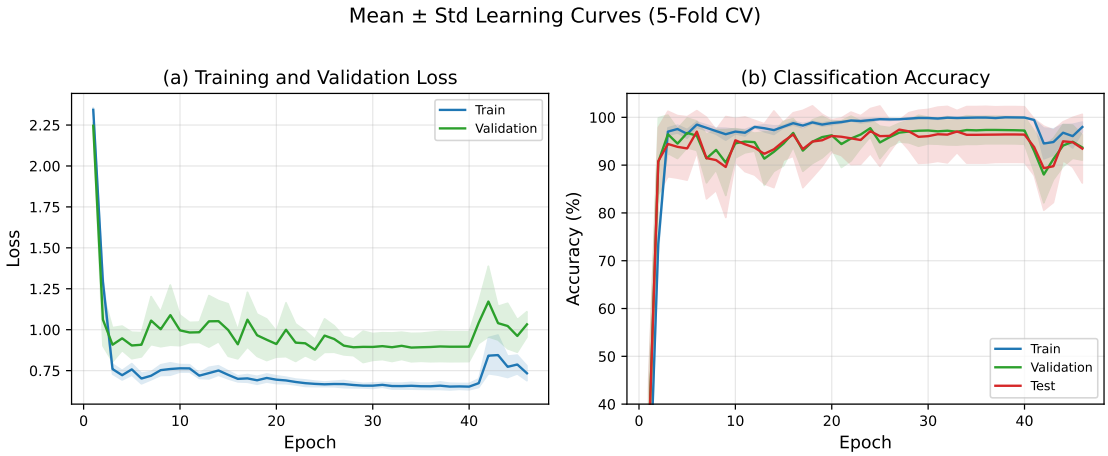


Figure A1. Averaged training and validation curves for MM-DTAE-LSTM at the best configuration (no proximity/pressure, w=64, s=16), with ±1 standard deviation shading. (a) Loss converges within ∼30 epochs with a modest generalization gap. (b) Training accuracy saturates near 100% while validation and test accuracy plateau at ∼95%, indicating mild overfitting.

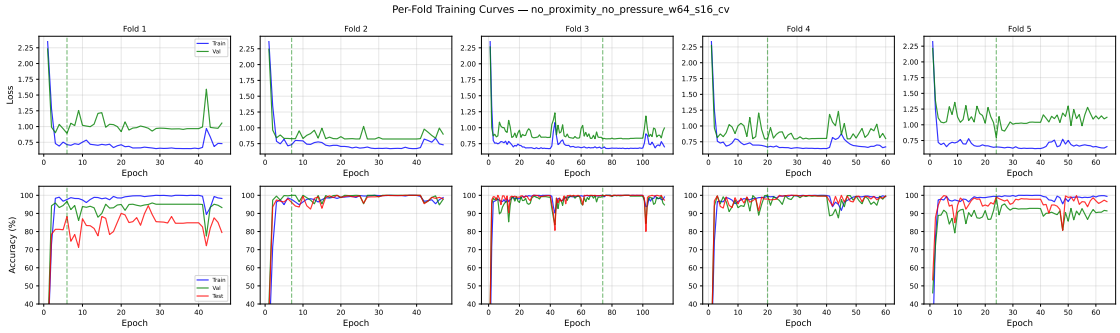


Figure A2. Per-fold training curves for MM-DTAE-LSTM at the best configuration (no proximity/pressure, w=64, s=16). Top row: loss; bottom row: accuracy (train, validation, test). Folds 1, 2, 4, and 5 converge within 40–60 epochs; Fold 3 trains to ∼100 epochs before early stopping, reflecting greater difficulty in its particular train/test partition. Fold-to-fold variance in final test accuracy (85–100%) is driven primarily by the uneven distribution of low-count damage classes across folds.

Table A1. Training Time per Fold at Best Configuration (No Prox+Pres, w=64, s=16).

Model	Time (s)	Relative to Encoder
MM-DTAE-LSTM	198 ± 85	1.0× (GPU)
XGBoost	191 ± 15	1.0× (CPU)
Random Forest	1.7 ± 0.2	116× faster
Logistic Reg.	10.9 ± 3.0	18× faster
MLP	672 ± 240	3.4× slower (GPU)
SimpleLSTM	755 ± 234	3.8× slower (GPU)

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

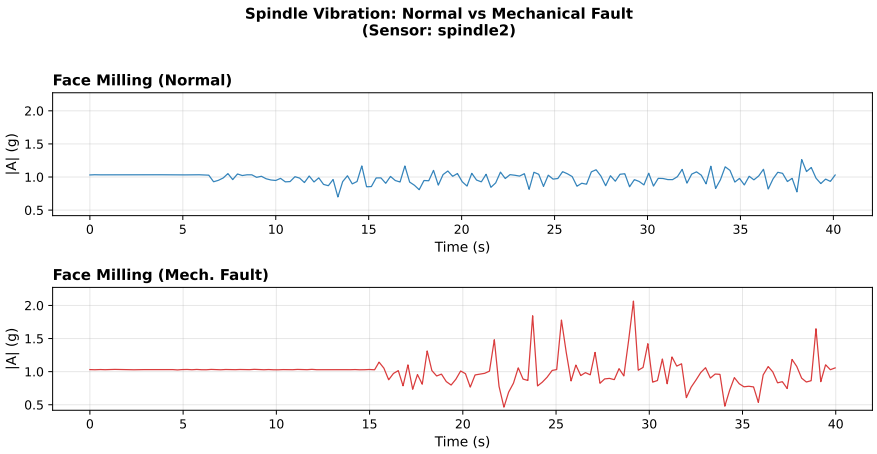


Figure A3. Spindle sensor vibration magnitude comparing air-cut face milling (top) versus damaged-spindle air-cut face milling (bottom). The damage condition shows elevated vibration amplitude due to spindle runout from the removed drive band.

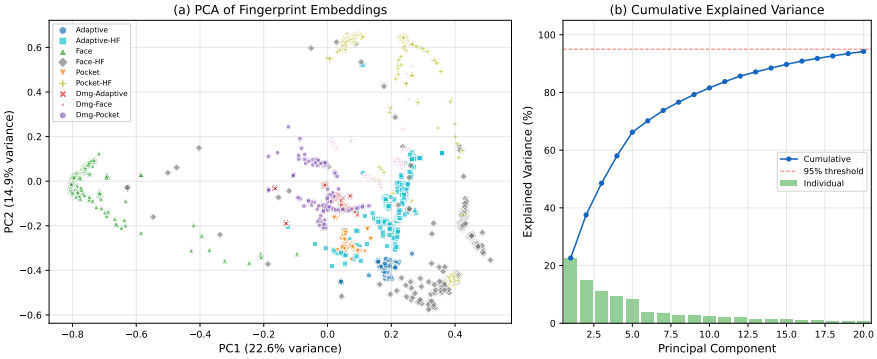


Figure A4. PCA [48] of 128-dimensional fingerprint embeddings (3,479 test samples across 5 folds). (a) PC1 vs. PC2 scatter shows linear separability among the three toolpath families. (b) Cumulative explained variance: 21 components capture 95% of embedding variance; PC1–PC5 account for 66.2%.

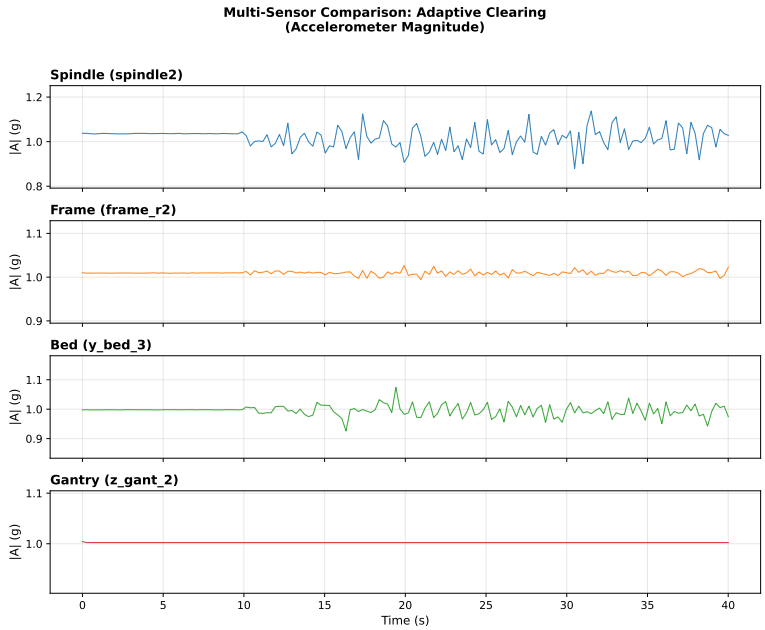


Figure A5. Multi-sensor comparison during adaptive clearing (air-cut). Spindle sensors show highest amplitude due to direct spindle vibration; bed sensors travel with the workpiece; frame sensors act as structural filters. Signal strength varies by sensor location, motivating multi-sensor deployment.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

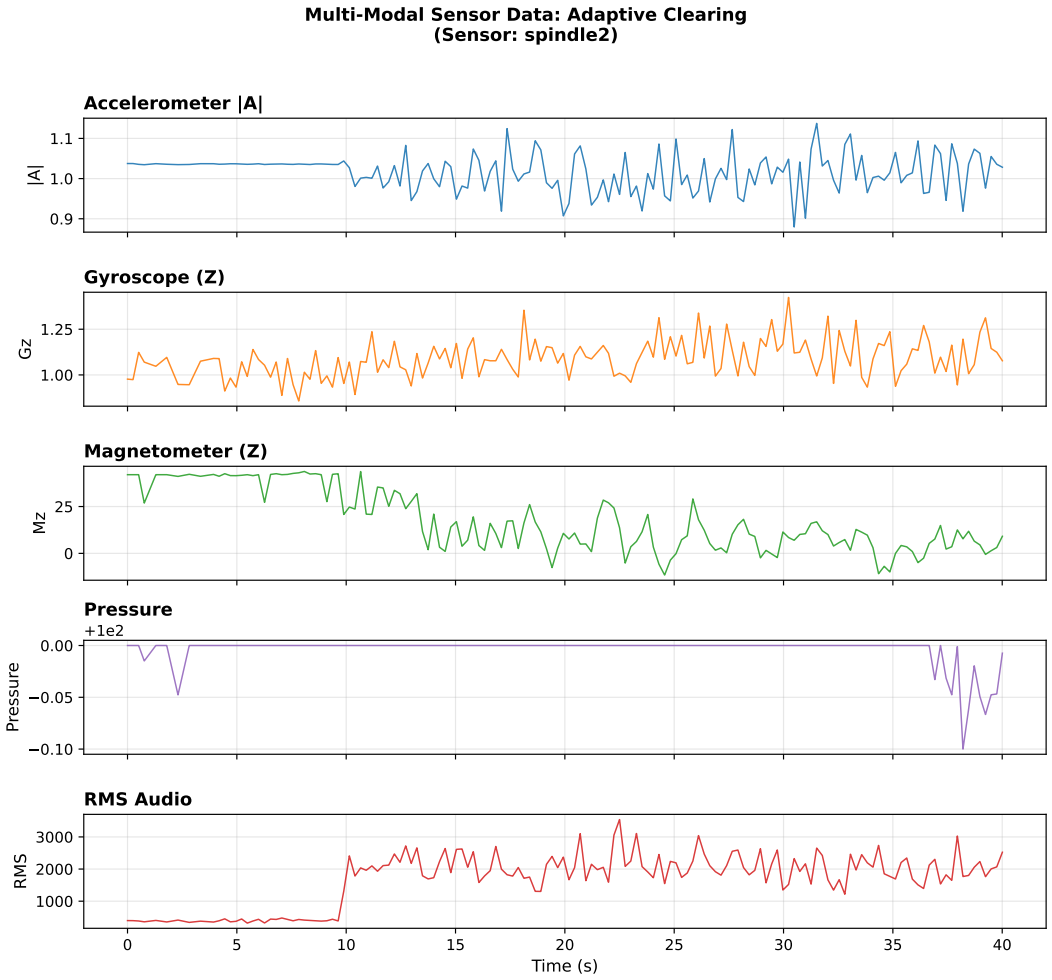


Figure A6. Multi-modal sensor data from spindle2 during adaptive clearing (air-cut). Each modality captures different physical phenomena: accelerometer magnitude reflects gantry acceleration and structural vibration; gyroscope captures rotational dynamics; magnetometer senses spindle motor fields; pressure and RMS audio provide environmental and acoustic context.

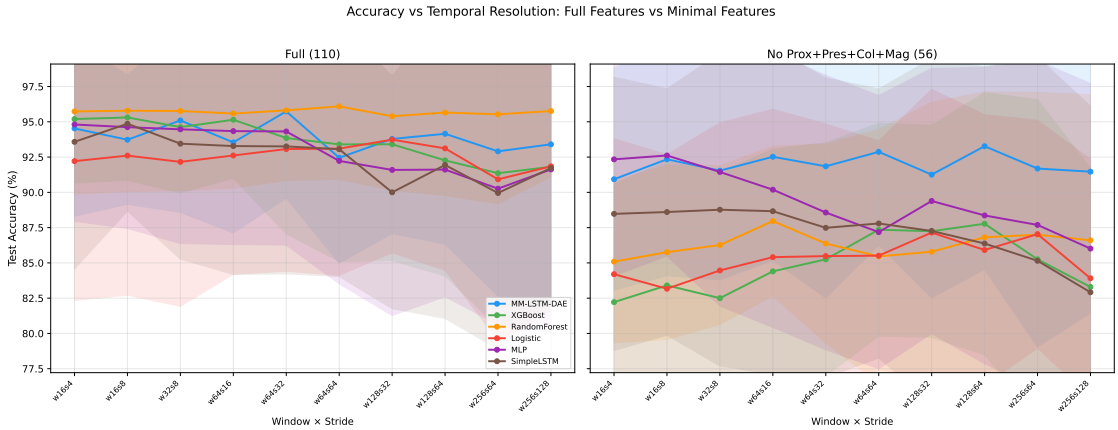


Figure A7. Comparison of full (110 features) versus minimal (56 features) configurations across temporal resolutions. The encoder maintains consistent performance while classical baselines exhibit substantial degradation at the minimal feature set.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

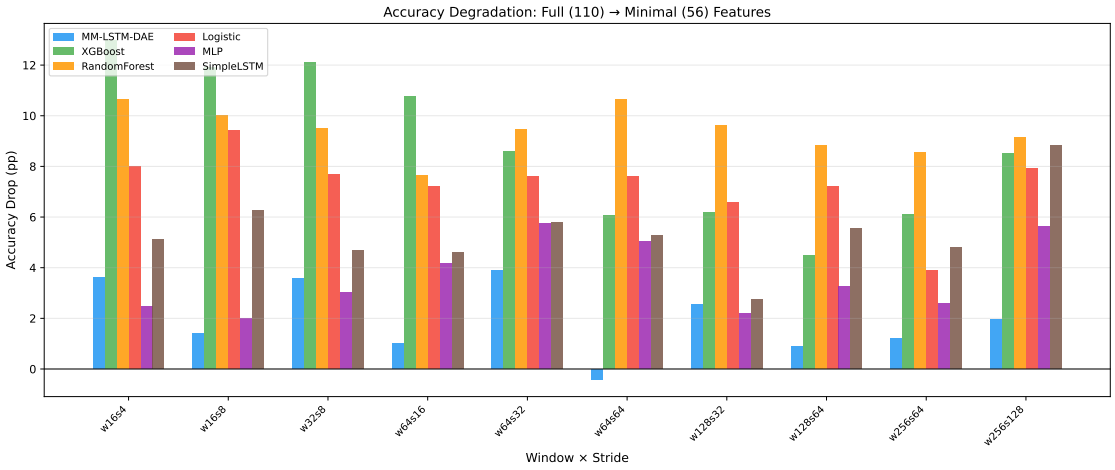


Figure A8. Accuracy degradation (full − minimal) per model across temporal resolutions. XGBoost suffers the largest degradation (8–11 points), while the encoder consistently degrades least (<3 points at most window sizes).

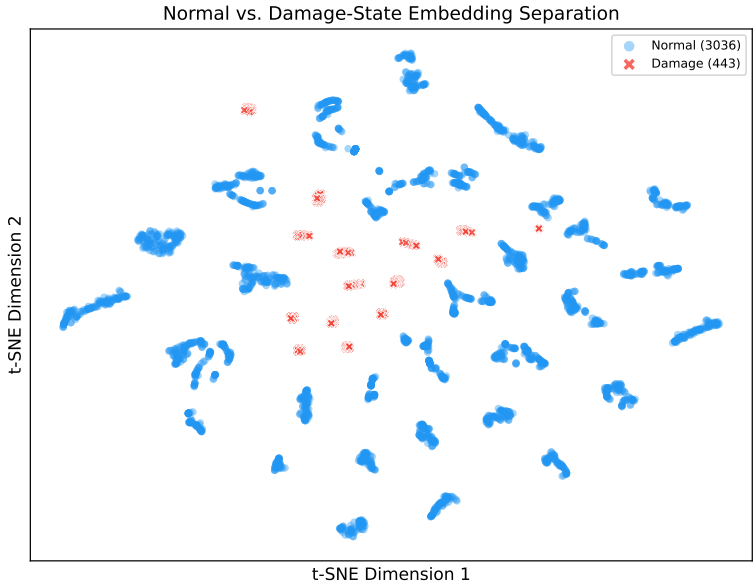


Figure A9. t-SNE visualization of damage vs. non-damage embeddings (3,479 test samples: 3,036 air-cut/active-cut, 443 damage). Damage samples (red markers) are dispersed among non-damage clusters rather than forming a distinct region, confirming that spindle degradation produces subtle signature shifts within each toolpath family.

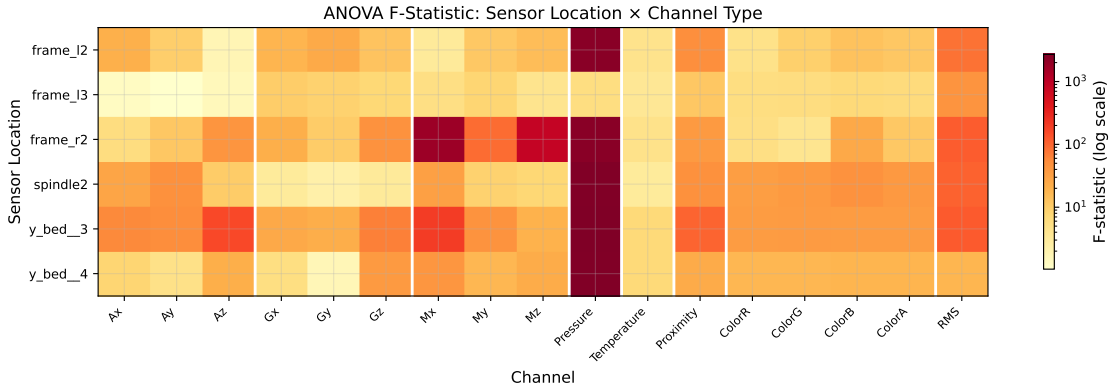


Figure A10. ANOVA F-statistic heatmap by sensor location and channel type. Pressure channels show uniformly extreme F across all sensors (session confound). Magnetometer channels on `frame_r2` are outliers ($F > 800$), likely reflecting a consistent magnetic field gradient at that location. Accelerometer and gyroscope channels show moderate, spatially varying F-values reflecting genuine location-dependent sensitivity to machine dynamics.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Appendix B. Fingerprint Embedding Details

In addition to the classification head, the model produces a 128-dimensional fingerprint embedding via a dedicated projection head. Attention pooling over the LSTM hidden states yields a fixed-length summary, which is mapped through a two-layer MLP ($256 \to 256 \to 128$) with GELU activation, followed by L2 normalization:

$$\mathbf{f} = \frac{\mathrm{MLP}_{\mathrm{fp}}(\mathrm{AttnPool}(H_{\mathrm{lstm}}))}{\|\mathrm{MLP}_{\mathrm{fp}}(\mathrm{AttnPool}(H_{\mathrm{lstm}}))\|_2} \in \mathbb{R}^{128} \tag{A1}$$

In the current training configuration, the fingerprint head is not optimized with a dedicated contrastive loss; instead, it projects the LSTM representations—trained via classification and reconstruction objectives—through an attention-pooled MLP. The resulting embeddings inherit their class-discriminative structure from the LSTM hidden states and are used for the t-SNE (Figure 4) and PCA (Figure A4) visualizations. The fingerprint head is designed for extension toward contrastive G-code identity learning in future work.

Appendix C. Physical Interpretation of Modality Contributions

Appendix C.1. Why Accelerometers Dominate

Accelerometers capture broadband structural vibration signatures along three orthogonal axes, encoding both the forced response from servo-driven axis motion and, during active cutting, the dynamic cutting force components. Even during air-cutting, toolpath strategies produce distinct acceleration profiles: adaptive clearing generates high-jerk transients at trochoidal corners; face milling produces quasi-periodic motion with acceleration pulses at step-over transitions; pocket machining alternates between axial plunge and lateral clearing with bimodal frequency content. During active cutting, additional signatures emerge from cutting forces at the tooth-passing frequency ($f_t =$ 400 Hz for our 2-flute endmill at 12,000 RPM)—above the 59.5 Hz Nyquist limit, but discriminative information persists in the low-frequency force envelope, structural resonance excitation (10–50 Hz), and modulation of servo tracking error by cutting load.

Appendix C.2. Why Gyroscopes Complement Accelerometers

Gyroscopes measure angular rate about three axes, sensing phenomena invisible to accelerometers: spindle angular velocity fluctuations from belt slip or runout, gantry pitch and yaw during rapid direction reversals, and bed-mounted rotational modes excited by off-center cutting loads. This orthogonal physical basis—translational versus rotational—suggests that each modality captures information invisible to the other; both are retained in the minimal configuration, and their similar median F-statistics (accelerometer: 10.8, gyroscope: 11.4) indicate comparable but non-redundant discriminative contributions. Isolating the individual contribution of each modality would require a finer-grained ablation not performed in this study.

Appendix C.3. The Role of Acoustic (RMS) Features

RMS audio captures airborne acoustic emission—physically distinct from structure-borne vibration measured by accelerometers. Air-cutting produces tonal spindle noise dominated by the belt-driven fundamental and harmonics; active cutting adds broadband stochastic emission from chip formation, shearing, and tool–workpiece friction. This makes RMS audio particularly discriminative for the air-cut versus active-cut distinction (median $F = 91$, the highest among retained modalities).

Appendix C.4. Temperature as a Slow Diagnostic

Temperature channels respond on timescales of minutes rather than milliseconds, encoding thermal equilibrium state rather than dynamic machining events. Their modest discriminative power ($F \approx 4.5$, comparable to electrical channels) likely reflects session-level thermal variation—spindle warm-up duration, ambient workshop temperature—rather than per-operation physics. Their retention

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

in the minimal configuration reflects the ablation design's modality-group granularity rather than a claim of per-operation diagnostic value.

References

1. Teti, R.; Jemielniak, K.; O'Donnell, G.; Dornfeld, D. Advanced Monitoring of Machining Operations. *CIRP Ann.* **2010**, *59*, 717–739. https://doi.org/10.1016/j.cirp.2010.05.010.
2. Zhong, R.Y.; Xu, X.; Klotz, E.; Newman, S.T. Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering* **2017**, *3*, 616–630. https://doi.org/10.1016/J.ENG.2017.05.015.
3. Zonouz, S.; Rrushi, J.; McLaughlin, S. Detecting Industrial Control Malware Using Automated PLC Code Analytics. *IEEE Secur. Priv.* **2014**, *12*, 40–47. https://doi.org/10.1109/MSP.2014.113.
4. Stouffer, K.; Pease, M.; Tang, C.; Zimmerman, T.; Pillitteri, V.; Lightman, S.; Hahn, A.; Saravia, S.; Sherule, A.; Thompson, M. NIST Special Publication 800-82 Rev. 3: Guide to Operational Technology (OT) Security; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. https://doi.org/10.6028/NIST.SP.800-82r3.
5. Serin, G.; Sener, B.; Ozbayoglu, A.M.; Unver, H.O. Review of Tool Condition Monitoring in Machining and Opportunities for Deep Learning. *Int. J. Adv. Manuf. Technol.* **2020**, *109*, 953–974. https://doi.org/10.1007/s00170-020-05449-w.
6. Graves, A.; Mohamed, A.-r.; Hinton, G. Speech Recognition with Deep Recurrent Neural Networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947.
7. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 12449–12460.
8. Gao, R.; Wang, L.; Teti, R.; Dornfeld, D.; Kumara, S.; Mori, M.; Helu, M. Cloud-Enabled Prognosis for Manufacturing. *CIRP Ann.* **2015**, *64*, 749–772. https://doi.org/10.1016/j.cirp.2015.05.011.
9. Zhu, K.; Wong, Y.S.; Hong, G.S. Wavelet Analysis of Sensor Signals for Tool Condition Monitoring: A Review and Some New Results. *Int. J. Mach. Tools Manuf.* **2009**, *49*, 537–553. https://doi.org/10.1016/j.ijmachtools.2009.02.003.
10. Wang, J.; Ma, Y.; Zhang, L.; Gao, R.X.; Wu, D. Deep Learning for Smart Manufacturing: Methods and Applications. *J. Manuf. Syst.* **2018**, *48*, 144–156. https://doi.org/10.1016/j.jmsy.2018.01.003.
11. Zhao, R.; Yan, R.; Chen, Z.; Mao, K.; Wang, P.; Gao, R.X. Deep Learning and Its Applications to Machine Health Monitoring. *Mech. Syst. Signal Process.* **2019**, *115*, 213–237. https://doi.org/10.1016/j.ymssp.2018.05.050.
12. Vogl, G.W.; et al. Research Needs for Cyberphysical Systems in Machining and Machine Tools: FMNet 2024 Report; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2024. Available online: https://www.nist.gov/system/files/documents/2024/12/12/FMNET_2024Report.pdf.
13. Janssens, O.; Slavkovikj, V.; Vervisch, B.; Stockman, K.; Loccufier, M.; Verstockt, S.; Van de Walle, R.; Van Hoecke, S. Convolutional Neural Network Based Fault Detection for Rotating Machinery. *J. Sound Vib.* **2016**, *377*, 331–345. https://doi.org/10.1016/j.jsv.2016.05.027.
14. Stathatos, E.; Tzimas, E.; Benardos, P.; Vosniakos, G.-C. Convolutional Neural Networks for Raw Signal Classification in CNC Turning Process Monitoring. *Sensors* **2024**, *24*, 1390. https://doi.org/10.3390/s24051390.
15. Bhandari, B. Comparative Study of Popular Deep Learning Models for Machining Roughness Classification Using Sound and Force Signals. *Micromachines* **2021**, *12*, 1484. https://doi.org/10.3390/mi12121484.
16. Wang, H.; Wang, S.; Sun, W.; Xiang, J. Multi-Sensor Signal Fusion for Tool Wear Condition Monitoring Using Denoising Transformer Auto-Encoder ResNet. *J. Manuf. Process.* **2024**, *124*, 1227–1238. https://doi.org/10.1016/j.jmapro.2024.07.002.
17. Lahat, D.; Adali, T.; Jutten, C. Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects. *Proc. IEEE* **2015**, *103*, 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697.
18. Ramachandram, D.; Taylor, G.W. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. https://doi.org/10.1109/MSP.2017.2738401.
19. Cao, L.; Li, J.; Zhang, L.; Luo, S.; Li, M.; Huang, X. Cross-Attention-Based Multi-Sensing Signals Fusion for Penetration State Monitoring During Laser Welding of Aluminum Alloy. *Knowl.-Based Syst.* **2023**, *261*, 110212. https://doi.org/10.1016/j.knosys.2022.110212.
20. Tsanousa, A.; Bektsis, E.; Kyriakopoulos, C.; Gómez González, A.; Leturiondo, U.; Gialampoukidis, I.; Karakostas, A.; Vrochidis, S.; Kompatsiaris, I. A Review of Multisensor Data Fusion Solutions in Smart Manufacturing: Systems and Trends. *Sensors* **2022**, *22*, 1734. https://doi.org/10.3390/s22051734.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

21. Meyes, R.; Lu, M.; de Puiseau, C.W.; Meisen, T. Ablation Studies in Artificial Neural Networks. *arXiv* **2019**, arXiv:1901.08644.
22. Jia, F.; Lei, Y.; Lin, J.; Zhou, X.; Lu, N. Deep Neural Networks: A Promising Tool for Fault Characteristic Mining and Intelligent Diagnosis of Rotating Machinery with Massive Data. *Mech. Syst. Signal Process.* **2016**, *72–73*, 303–315. https://doi.org/10.1016/j.ymssp.2015.10.025.
23. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
24. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.
25. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
27. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in Time Series: A Survey. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 6778–6786. https://doi.org/10.24963/ijcai.2023/759.
28. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
29. Malhotra, P.; Vig, L.; Shroff, G.; Agarwal, P. Long Short Term Memory Networks for Anomaly Detection in Time Series. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Bruges, Belgium, 22–24 April 2015; pp. 89–94.
30. Cybersecurity Manufacturing Innovation Institute (CyManII). National Cybersecurity Manufacturing Roadmap; University of Texas at San Antonio: San Antonio, TX, USA, 2022. Available online: https://cymanii.org/wp-content/uploads/2022/05/CYMANII-CYBERSECURITY-MANUFACTURING-ROADMAP-2022-PUBLIC-VERSION-FINAL.pdf.
31. Kimmell, J.C.; Orlyanchik, V.; Sturm, L.; Dawson, J.; Taylor, C. Detecting Control Injection Attacks Using Energy Data Anomalies in Computer Numerical Control Machining. In *Critical Infrastructure Protection XVIII*; IFIP Advances in Information and Communication Technology; Springer: Cham, Switzerland, 2025; Volume 725, pp. 43–63. https://doi.org/10.1007/978-3-031-81888-2_3.
32. Coelho, P.J.; Athar, A.; Mozumder, M.A.I.; Ali, S.; Kim, H.-C. Deep Learning-Based Anomaly Detection Using One-Dimensional Convolutional Neural Networks (1D CNN) in Machine Centers (MCT) and Computer Numerical Control (CNC) Machines. *PeerJ Comput. Sci.* **2024**, *10*, e2389. https://doi.org/10.7717/peerj-cs.2389.
33. Yampolskiy, M.; King, W.E.; Gatlin, J.; Belikovetsky, S.; Brown, A.; Skjellum, A.; Elovici, Y. Security of Additive Manufacturing: Attack Taxonomy and Survey. *Addit. Manuf.* **2018**, *21*, 431–457. https://doi.org/10.1016/j.addma.2018.03.015.
34. Al Faruque, M.A.; Chhetri, S.R.; Wan, J.; Canedo, A. Acoustic Side-Channel Attacks on Additive Manufacturing Systems. In Proceedings of the 7th ACM/IEEE International Conference on Cyber-Physical Systems, Vienna, Austria, 11–14 April 2016; pp. 1–10. https://doi.org/10.1109/ICCPS.2016.7479068.
35. Song, C.; Lin, F.; Ba, Z.; Ren, K.; Zhou, C.; Xu, W. My Smartphone Knows What You Print: Exploring Smartphone-Based Side-Channel Attacks Against 3D Printers. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 895–907. https://doi.org/10.1145/2976749.2978300.
36. Chhetri, S.R.; Canedo, A.; Al Faruque, M.A. KCAD: Kinetic Cyber-Attack Detection Method for Cyber-Physical Additive Manufacturing Systems. In Proceedings of the 35th IEEE/ACM International Conference on Computer-Aided Design, Austin, TX, USA, 7–10 November 2016; Article 74, pp. 1–8. https://doi.org/10.1145/2966986.2967050.
37. Jamarani, A.; Tu, Y.; Hei, X. Practitioner Paper: Decoding Intellectual Property: Acoustic and Magnetic Side-Channel Attack on a 3D Printer. In *Security and Privacy in Cyber-Physical Systems and Smart Vehicles. SmartSP 2024*; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0488.v1

Engineering; Springer: Cham, Switzerland, 2025; Volume 622, pp. 33–52. https://doi.org/10.1007/978-3-031-93354-7_3.
38. Chattopadhyay, T.; Ceschin, F.; Garza, M.E.; Zyunkin, D.; Chhotaray, A.; Stebner, A.P.; Zonouz, S.; Beyah, R. One Video to Steal Them All: 3D-Printing IP Theft through Optical Side-Channels. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, Taipei, Taiwan, 13–17 October 2025. https://doi.org/10.1145/3719027.3744837.
39. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
40. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
41. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
42. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
43. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
44. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
45. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785.
46. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
47. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
48. Jolliffe, I.T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202. https://doi.org/10.1098/rsta.2015.0202.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.