Journal of Artificial Intelligence and Technology, **(Ahead of Print)**
https://doi.org/10.37965/jait.2026.1322

RESEARCH ARTICLE

Assessing AI Dependability in Acoustic NDT: Deep Learning Robustness for Gold Purity Verification under Industrial Noise

Mehmet Osman Devrim[1] and Serdar Kirisoglu[2]
[1]Department of Information Technologies, Duzce University, Duzce, Turkiye
[2]Department of Computer Engineering, Faculty of Engineering, Duzce University, Duzce, Turkiye

(Received 09 April 2026; Revised 24 June 2026; Accepted 19 July 2026; Published online 14 August 2026)

Abstract: Acoustic resonance testing (ART) is a widely utilized nondestructive evaluation method for material characterization. However, its measurement reliability significantly degrades under harsh industrial noise, creating undefined operational risks. This study presents a systematic framework to quantify the robustness of deep learning-based acoustic sensing systems against environmental interference. A controlled synthetic noise injection protocol is employed, using white and pink (1/*f*)) noise as standardized proxies for broadband and low-frequency-dominant industrial interference. This protocol enables systematic analysis of measurement performance degradation across varying signal-to-noise ratios (SNRs) under simulated industrial acoustic conditions. The results define the operational boundaries for reliable measurement, identifying a critical SNR threshold of 20 dB where measurement uncertainty significantly increases. Furthermore, we demonstrate that spectral masking and the "Artificial Damping Effect" are the primary physical drivers of measurement error in noisy environments. Although the methodology is demonstrated on gold alloy benchmarks, the proposed reliability envelope and noise compensation approach are applicable to the acoustic evaluation of various homogeneous materials. These findings establish a quantitative robustness envelope and provide a reproducible benchmark for industrial deployment. Further validation under directly recorded field conditions is required before practical implementation.

Keywords: acoustic resonance testing; decision support systems; deep learning; expert systems; measurement uncertainty; signal robustness

I. INTRODUCTION

A. THE NEED FOR RELIABLE GOLD PURITY VERIFICATION

The precise verification of gold purity is essential in both industrial quality assurance and financial security contexts, where even minor compositional deviations can result in significant economic impact. Conventional destructive testing methods such as chemical assays or mechanical hardness tests often compromise sample integrity and are unsuitable for continuous production environments. NDT, particularly acoustic resonance-based approaches, provides a robust, contactless alternative grounded in the physics of elastic wave propagation [1–3]. According to the impulse excitation standards (ASTM E1876 and ASTM E2001), the mechanical resonance spectrum of a solid object reveals intrinsic material properties such as Young's modulus and internal damping [1,2]. These acoustic resonance signatures, when properly excited and measured, are sensitive to density, elasticity, and internal structure [4–6]. Foundational works in acoustics and vibration analysis demonstrate that changes in purity or microstructure alter the natural frequencies and decay characteristics of the material's resonant modes [5–7]. Thus, acoustic methods have emerged as powerful diagnostic tools for metallic characterization, enabling the differentiation of high-purity and alloyed gold specimens through their vibrational response.

B. FROM MANUAL TO DEEP LEARNING APPROACHES

Traditional spectral and statistical models, though interpretable [15, 16], struggle with nonlinear dependencies inherent in acoustic resonance data. The transition toward data-driven learning has been catalyzed by the broader resurgence of deep learning [18], particularly through convolutional neural networks (CNNs) validated on large-scale image recognition tasks [19], and residual architectures (ResNets) [8–10]. The seminal ResNet framework [8] enabled deeper representations of hierarchical features by mitigating gradient vanishing problems that previously limited acoustic feature extraction depth. Optimizers such as Adam [9] and architectures tuned for large-scale audio classification [10–12] have further advanced automated acoustic interpretation.

In this study, ResNet-18 is selected as a representative and stable deep learning baseline owing to its proven balance between expressive capacity and computational tractability. The architecture has been extensively validated in both audio and image domains, offering a controlled experimental benchmark for assessing robustness. Architectural optimization or lightweight deployment (e.g., pruning or quantization) is deliberately excluded from the present scope to maintain comparability and interpretive clarity.

In audio analysis, spectrogram-based CNNs have demonstrated exceptional performance for tasks ranging from environmental

Corresponding author: Mehmet Osman Devrim (e-mail: mehmetosmandevrim@duzce.edu.tr).

sound recognition to music signal processing [11–14]. The use of time-frequency representations, such as the short-time Fourier transform and Mel spectrograms, offers a compact and perceptually relevant domain for machine learning models [20–24]. These advancements collectively redefine acoustic NDT, transforming manual frequency inspection into automated, data-driven classification capable of resolving subtle spectral cues related to purity and structure.

C. LABORATORY SUCCESS vs. INDUSTRIAL ROBUSTNESS

Despite high accuracies reported in controlled laboratory settings, most existing studies — including our previous feasibility work on gold purity classification [27] — remain untested under real-world industrial noise conditions. Works on acoustic material or coin classification have demonstrated the discriminative capacity of resonance-based signatures [25,26], yet these experiments are typically conducted in quiet environments with high signal-to-noise ratios (SNRs).

From a physical standpoint, this limitation is nontrivial. The acoustic response of metallic specimens is governed by their internal damping characteristics, which determine the decay rate or resonance tail of vibrational modes [1,2]. High-purity gold (24k) exhibits exceptionally low internal damping, leading to prolonged, low-amplitude resonance tails. Conversely, lower-purity alloys such as 14k gold, containing copper and silver, display increased internal friction and thus faster decay. The discrimination between these purity levels occurs primarily within the early post-impact decay window, where the resonance amplitude and phase contain material-specific damping information. Beyond this temporal region, the residual signal energy falls below the effective noise floor, and subsequent segments become dominated by environmental and electronic noise.

Under industrial conditions, broadband noise overlaps with and masks these weak tail regions in the signal spectrum. When such low-amplitude damping cues are buried beneath background noise, it becomes physically impossible to distinguish between the resonance signatures of 24k and 14k gold, even for advanced measurement systems. In this study, the term “industrial noise” refers broadly to adverse acoustic conditions such as broadband mechanical, ambient, or electrical interference rather than to any specific factory soundscape. This acoustic masking effect defines a practical limit to signal interpretability, one that has not yet been systematically quantified in prior deep learning-based NDT research.

In this context, while real-world industrial environments contain complex acoustic mixtures such as machinery hum, conveyor belt friction, and ventilation drone, white and pink noise serve as standardized proxies for “broadband” and “low-frequency-dominant” interference, respectively. Rather than attempting to replicate a specific industrial facility with variable acoustic properties, this study establishes a reproducible stress-testing framework. This controlled approach allows for defining the precise physical robustness limits of deep learning models, isolating the effects of spectral masking from other environmental variables.

D. CONTRIBUTIONS OF THE STUDY

The main contributions of this study are summarized as follows:

- **Controlled quantification of measurement degradation:** A systematic framework is proposed in which white and pink ($1/f$) noise [20] are superimposed on measured resonance signals to evaluate the sensitivity of deep learning models across a continuous SNR range. This approach serves as a reproducible stress test rather than an attempt to mimic specific industrial environments. Moreover, the implications of classification failure extend beyond technical accuracy to financial risk. In the precious metals trade, misclassifying high-purity assets (e.g., 24k bullion) as lower-purity alloys due to environmental noise can lead to significant valuation discrepancies and loss of consumer trust. Therefore, establishing the acoustic operational boundaries of these systems is a prerequisite for their commercial adoption.
- **Data integrity assurance:** A Group-Aware Partitioning strategy is implemented to prevent data leakage, ensuring that recordings from the same physical specimen do not appear in both training and testing sets. This enforces strict sample independence and supports reliable generalization analysis.
- **Quantitative robustness envelope:** Based on controlled degradation experiments, a 20-dB operational threshold is identified as the lower boundary for reliable measurement, marking the onset of resonance-tail masking and subsequent loss of measurement fidelity.
- **Model scope and relevance:** ResNet-18 is adopted as a stable baseline to establish a physics-informed benchmark of robustness. Broader architectural optimization or lightweight deployment considerations are reserved for future investigation.

To the best of our knowledge, this study represents one of the first systematic assessments of deep learning-based acoustic NDT robustness, bridging the gap between idealized laboratory classification and realistic industrial noise conditions.

Unlike traditional classification studies, the novelty of this work lies in experimentally quantifying the physical signal degradation mechanisms (specifically the “Artificial Damping Effect”) caused by industrial noise and establishing a quantitative SNR threshold ($\approx$ 20 dB) to ensure the operational reliability of AI-assisted NDT instrumentation.

The remainder of this paper is organized as follows. Section II reviews the related literature on acoustic NDT, deep learning, and industrial robustness. Section III describes the measurement system and experimental methodology, including the dataset, model architecture, and noise injection protocol. Section IV presents the experimental results and discussion. Section V concludes the paper and outlines directions for future work.

II. RELATED STUDIES

Historically, acoustic resonance testing (ART) was governed by linear vibration theories and standard guidelines such as ASTM E2001 [1], which defined resonance frequencies as key indicators of material integrity. Early computational approaches in this domain predominantly utilized hand-crafted features such as spectral centroids and Mel-frequency cepstral coefficients (MFCCs) coupled with shallow measurement systems like support vector machines or k-nearest neighbors [2,3]. While these methods established the feasibility of acoustic material characterization, they often necessitated noise-free environments and expert-defined feature extraction pipelines that may lack resilience against complex industrial interference.

A. ACOUSTIC NDT AND MATERIAL CLASSIFICATION

Acoustic-based nondestructive evaluation has long been utilized for material characterization, exploiting the sensitivity of vibrational signatures to structural and compositional differences [1–4]. Earlier studies explored acoustic pattern recognition for material and coin classification [25], while foundational works in informatics established the shift toward hierarchical feature learning and deep architectures for complex audio signals [26]. Our prior study on gold purity determination using audio features and machine learning algorithms [27] achieved over 95% measurement accuracy under controlled conditions, validating the feasibility of acoustic-based metal discrimination. However, all these approaches assumed noise-free environments, leaving unresolved questions about robustness under realistic acoustic interference.

B. DEEP LEARNING IN ACOUSTIC ANALYSIS

Recent advances in deep learning have transformed the way acoustic data are modeled and interpreted. CNN architectures [10] and data augmentation strategies [11] have enhanced generalization for environmental and mechanical sounds, while large-scale frameworks such as AudioSet and ESC-50 datasets [12,13] have benchmarked feature representations in noisy conditions. Feature engineering using spectrogram transformations, MFCCs, and psychoacoustic representations [20–22] has proven effective for extracting salient temporal and spectral cues. The combination of residual learning [8], Adam optimization [9], and advanced architectures has enabled robust feature extraction from complex acoustic patterns. Collectively, these developments provide the computational foundation for the deep learning framework adopted in this study.

C. INDUSTRIAL ROBUSTNESS AND FAULT DIAGNOSIS CONTEXT

Within the broader industrial domain, the robustness of deep learning models under noise is an active research frontier. Deep learning has been applied to acoustic fault diagnosis of rotating machinery, bearings, and motors, where fluctuating operational and noise conditions challenge model stability [28–31]. Zhao *et al*. [28] provided a comprehensive review of deep learning applications in machine health monitoring, emphasizing the effectiveness of ResNet architectures in processing vibro-acoustic data for fault detection. Similarly, Chen *et al.* [32] enhanced noise robustness in seal fault detection via a ResNet architecture augmented with data-level sample expansion. Uzel *et al.* [33] demonstrated Mel-spectrogram-based CNNs resilient to low SNR scenarios in electric motor diagnosis, while Kumar *et al.* [34] integrated CNN–LSTM hybrids for denoising partial discharge signals. Collectively, these studies show that deep architectures can sustain performance when explicitly optimized for robustness.

In the field of rail and conveyor diagnostics, CNN and ResNet models have been applied successfully to acoustic emission and vibration data [34,37,38], including transfer learning for structural monitoring [35]. Other works have extended robustness frameworks to powder metallurgy inspection [36], laser-directed additive manufacturing [43], and wind turbine health monitoring [39]. Reviews across NDT and machine learning domains reinforce this shift toward resilience and reliability under environmental variability [40–44]. For example, Wu *et al.* [41] and Scarselli & Nicassio [42] emphasize the necessity of feature redundancy and noise-tolerant training to preserve diagnostic integrity.

This study aligns with these industrial trends by applying established robustness evaluation principles to acoustic purity classification, a niche yet technically relevant domain where minute spectral variations distinguish material quality. By integrating controlled synthetic noise testing with resonance-tail analysis, it provides, to the best of our knowledge, one of the first systematic assessments of deep learning robustness for ART of gold under industrial noise conditions.

III. MEASUREMENT SYSTEM AND EXPERIMENTAL METHODOLOGY

This section outlines the methodological framework established to evaluate the robustness of deep learning models in acoustic purity verification. First, the properties of the gold alloy dataset and the acoustic feature extraction pipeline are described. Second, the ResNet-18 architecture adapted for spectral analysis is detailed. Finally, the rigorous experimental protocols are presented, specifically focusing on the group-aware data partitioning strategy to prevent leakage and the controlled noise injection mechanism designed to simulate adverse industrial environments.

A. DATASET DESCRIPTION

The acoustic dataset consists of impact-response recordings obtained from gold samples belonging to three distinct purity classes (14k, 22k, and 24k). This raw dataset was originally constructed and validated in our previous study [27], which established the baseline feasibility of acoustic purity classification under ideal conditions. In the present work, we extend this dataset by subjecting it to controlled noise stress tests.

Table I summarizes the metallurgical composition and sample distribution of the dataset utilized in this study. The dataset encompasses three distinct purity classes (14k, 22k, and 24k), selected specifically for their divergent acoustic damping characteristics. To prevent bias arising from class imbalance during model training, an equal number of clean impact recordings (400 samples per class) were utilized, resulting in a balanced dataset of 1200 samples.

B. DATA PREPROCESSING AND FEATURE EXTRACTION

Raw audio signals were sampled at 44.1 kHz. To capture the time-frequency characteristics of the resonance decay, log-Mel spectrograms were utilized as the primary input representation. The audio signals were framed using a Hamming window with a hop length of 512 samples and converted into 128 Mel-frequency bands. This

Table I. Description of the gold purity dataset, including metallurgical composition and sample distribution

Purity class	Gold content (%)	Description (physical property)	Sample count (clean)
14k	58.5%	High-impurity alloy, high internal damping	400
22k	91.6%	Standard investment grade, moderate damping	400
24k	99.9%	Pure metal, low damping, long resonance tail	400
Total	-	Balanced dataset	1200

process yields a two-dimensional tensor representation where the vertical axis corresponds to frequency bands and the horizontal axis represents temporal frames.

C. MODEL ARCHITECTURE (ResNet-18)

A ResNet-18 CNN was employed as the backbone classifier. Unlike standard implementations expecting 3-channel RGB images, the first convolutional layer was modified to accept mono-channel spectral inputs. Specifically, the network accepts a tensor of shape $(1 \times 128 \times T)$, where 128 denotes the Mel bands and T represents the time steps. The backbone consists of four residual blocks with skip connections, allowing efficient gradient flow and the learning of deep hierarchical features. Finally, the network terminates in a fully connected layer with a Softmax activation, producing a probability output vector $y \in \mathbb{R}^3$. Each element of this vector corresponds to the predicted probability for the 14k, 22k, and 24k classes, respectively. In this study, the ResNet-18 architecture was specifically selected for the classification of log-Mel spectrograms derived from acoustic impact-response signals. The choice of ResNet-18 is driven by the optimal balance it offers between model complexity and the available dataset size. While deeper architectures such as ResNet-50 or ResNet-101 offer higher theoretical capacity, their significantly larger parameter counts increase the risk of overfitting when training on limited acoustic datasets [17]. Furthermore, the residual connections in ResNet-18 effectively minimize the vanishing gradient problem, ensuring training stability [8]. Crucially, ResNet-18 possesses sufficient depth to capture complex time-frequency patterns in spectrograms while maintaining a lightweight computational footprint, making it suitable for potential integration into handheld industrial NDT devices.

The overall workflow of the proposed robustness analysis framework is illustrated in Fig. 1. As depicted, the system takes raw impact-response signals as input and converts them into log-Mel spectrograms. A critical component of this architecture is the "Noise Injection Module," which acts as a stress-test layer, superimposing controlled white and pink noise profiles onto the feature representations before they are fed into the ResNet-18 classifier. This design allows for the direct evaluation of feature resilience against varying degrees of acoustic interference.

D. GROUP-AWARE DATA PARTITIONING STRATEGY

To ensure methodological rigor and prevent data leakage, a Group-Aware Stratified Split strategy was applied. Recordings originating from the same physical object were confined strictly to either the training or the test set. This ensures that the model learns to generalize based on material properties rather than memorizing the unique acoustic fingerprints of specific samples.

E. CONTROLLED NOISE INJECTION PROTOCOL

To simulate industrial environments, a systematic noise stress test is designed. Two specific noise profiles are selected to simulate distinct real-world interference scenarios encountered in gold workshops. First, white noise is utilized to represent the broadband thermal noise inherent in electronic microphone sensors and high-frequency hiss often present in digital recording chains [6]. Second, pink noise ($1/f$)) is employed to simulate the ambient acoustic environment of a workshop, which is typically dominated by low-frequency mechanical rumble from ventilation systems, heavy machinery, and distant traffic [6,20]. While real-world environments may contain complex reverberations, these standardized noise types provide a reproducible proxy for establishing the baseline robustness of the classifier against wideband and spectral-decay interference.

The noise injection is implemented as an additive process in the time domain. Specifically, for each clean audio sample x(t), α noise segment n(t) of equal length was randomly cropped from the noise source and linearly superimposed to produce the degraded signal y(t):

$$y(t) = x(t) + \alpha \cdot n(t) \tag{1}$$

where α is a real-valued scaling factor that controls the relative energy contribution of the noise component. The additive formulation in Eq. (1) reflects the physical superposition of acoustic waves and ensures that the noise is introduced without altering the temporal structure of the original resonance signal.

The scaling factor α is not fixed but is dynamically computed for each individual audio sample to precisely satisfy the target SNR condition. This per-sample calibration is necessary because different recordings may have different RMS energy levels, and a single fixed α would result in inconsistent actual SNR values across the dataset. Formally, α is derived from the ratio of the root-mean-square (RMS) power of the clean signal ($P_{\{signal\}}$) to that of the noise segment ($P_{\{noise\}}$):

$$\alpha = \sqrt{\frac{P_{\text{signal}}}{\left(P_{\text{noise}} \cdot 10^{\frac{\text{SNR}}{10}}\right)}} \tag{2}$$

Here, the term 10^(−SNR/10) converts the target SNR from the logarithmic decibel scale to a linear power ratio, and the square root accounts for the conversion from power to amplitude. Together, Equations (1) and (2) ensure that the injected noise is physically calibrated to the exact interference level intended, enabling a controlled and reproducible degradation spectrum from 30 dB down to 0 dB SNR.

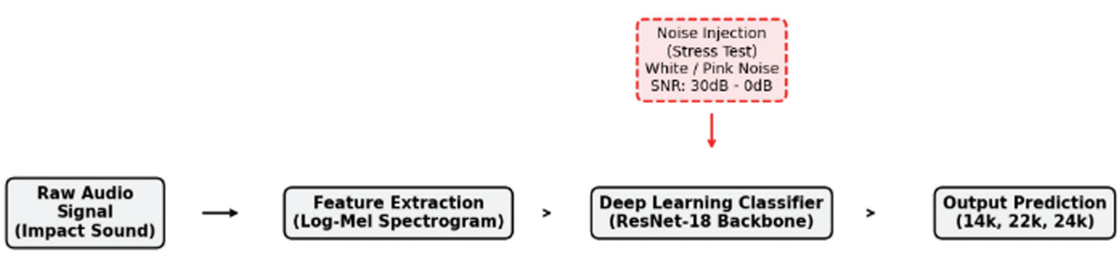

Fig. 1. Block diagram of the proposed robust acoustic classification framework. Raw impact-response signals are first converted into log-Mel spectrograms and then subjected to the Noise Injection Module — a controlled stress-test layer that superimposes white and pink noise at target SNR levels before classification by the ResNet-18 model. This pipeline enables systematic quantification of feature resilience under simulated industrial interference.

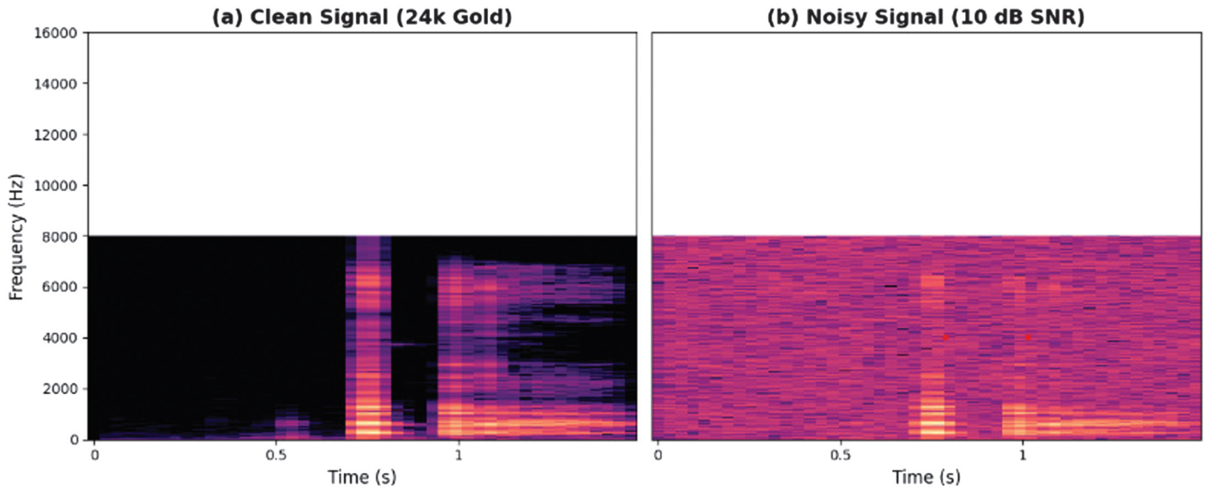

Fig. 2. Log-Mel spectrogram comparison of a 24k gold impact-response signal under (a) clean laboratory conditions and (b) simulated industrial noise at 10 dB SNR. In panel (a), distinct harmonic ridges and a prolonged low-amplitude resonance decay tail — characteristic of high-purity gold's low internal damping — are clearly visible. In panel (b), broadband noise energy raises the spectral floor, effectively obliterating these fine decay features. This "spectral masking" is the primary physical mechanism responsible for classification failures at low SNR levels.

To visualize the physical mechanism behind classification failures, Fig. 2 provides a comparative spectral analysis. As observed in Panel (a), the high-purity (24k) gold sample exhibits distinct harmonic ridges and a prolonged resonance decay tail under clean conditions. However, as shown in Panel (b), the introduction of noise at 10 dB SNR creates a broadband energy floor that effectively masks these low-amplitude decay features. This "spectral masking" phenomenon confirms that the drop in model accuracy is not due to a lack of learning capacity, but rather the physical obliteration of the discriminative acoustic signatures required for purity verification.

IV. EXPERIMENTAL RESULTS

This section presents a quantitative evaluation of the proposed ResNet-18 framework under both ideal and adverse acoustic conditions. First, the baseline performance is established using the clean test set to verify the feasibility of the feature extraction pipeline. Subsequently, the system's robustness is challenged through the controlled injection of white and pink noise across a degradation spectrum from 30 dB down to 0 dB. The results are analyzed in terms of overall measurement accuracy, Macro-F1 scores, and class-specific failure modes, with a particular focus on identifying the operational boundaries for reliable industrial deployment.

Table II quantitatively demonstrates the performance degradation of the ResNet-18 model as a function of the SNR. The model maintains stability at 30 dB SNR (mild noise), achieving over 95% accuracy for both noise types. However, a sharp performance collapse is observed when the SNR drops below 20 dB. Specifically, at 10 dB and below, accuracy rates approach chance level, proving that 20 dB SNR is a safe operating threshold for the system to function reliably with raw spectrogram data.

Table III summarizes the performance gap between the "quiet environment" and the "industrial environment." The drop from a flawless score of 1.0000 on clean data to 0.6023 when noisy scenarios are included demonstrates how sensitive the features learned by the deep learning model are to noise. This decline suggests that the model focuses not only on the "fundamental frequency" but also on "fine spectral tails," which are easily affected by noise.

Table II. Accuracy (%) vs. SNR

SNR	White (%)	Pink (%)
30	95.42	96.25
20	73.75	80.42
10	40.83	57.50
5	38.33	42.92
0	37.50	36.67

Table III. Macro-F1

Condition	Macro-F1
Clean	1.0000
Noisy	0.6023
Overall	0.6402

Table IV reveals "Class-wise Asymmetry," one of the most critical findings of the study. Under noisy conditions, the model correctly identifies 14k gold at a rate of 82% (recall), while this rate drops to around 49% for 24k and 22k gold. This indicates that the error is not randomly distributed; instead, the model exhibits a systematic bias towards predicting "lower purity" (14k) in moments of uncertainty.

The curves in Fig. 3 illustrate that the degradation due to noise is nonlinear. The most prominent feature in the graph is the

Table IV. Purity-level recall rates under noisy conditions (averaged across all SNR levels from 0 dB to 30 dB and both noise types)

Class	Recall
14k	0.82
22k	0.49
24k	0.49

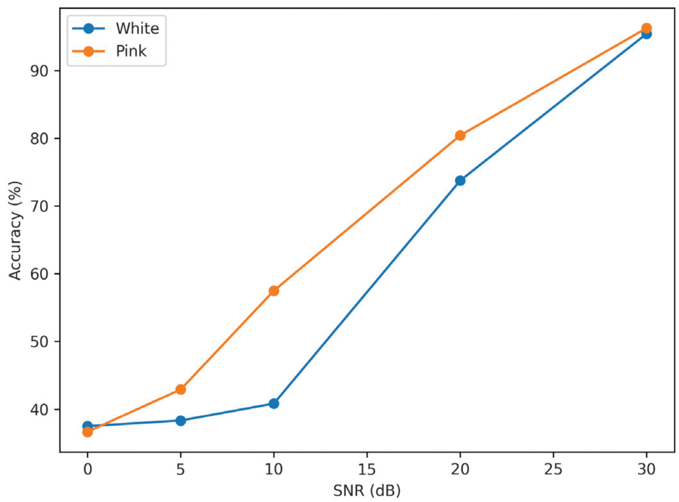

Fig. 3. Measurement accuracy (%) as a function of signal-to-noise ratio (SNR) for white and pink noise conditions. Both curves remain stable above 20 dB SNR but exhibit a sharp nonlinear collapse between 20 dB and 10 dB — the critical "knee-point" region where resonance decay tails fall below the noise floor and the model loses access to discriminative acoustic features. Below 10 dB, accuracy approaches chance level for both noise types.

"knee-point" occurring between 20 dB and 10 dB. The steepening of the curve in this range indicates the critical region where acoustic features fall below the noise floor, and the model loses access to discriminative resonance frequencies.

Figure 4 depicts the impact of noise color on measurement performance. At intermediate noise levels (20 dB), performance under pink noise (80.42%) is noticeably higher than under white noise (73.75%). The physical reason for this is that white noise distributes energy equally across all frequencies, more aggressively masking the high-frequency harmonics that allow us to distinguish

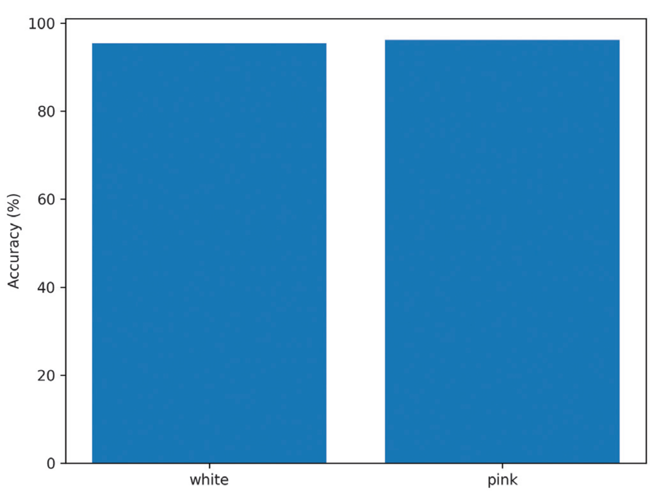

Fig. 4. Comparison of classification accuracy under white noise and pink (1/*f*) noise across SNR levels. Pink noise consistently yields higher accuracy at intermediate SNR values (e.g., 80.42% vs. 73.75% at 20 dB), as its energy is concentrated in lower frequencies and therefore preserves high-frequency harmonic resonance details. White noise, distributing energy uniformly across all frequencies, more aggressively masks the metallic ringing signatures critical for purity discrimination.

metallic ringing. Pink noise, on the other hand, concentrates energy at lower frequencies, allowing some of the resonance details in the upper frequencies to be preserved.

Figure 5 visualizes how noise disrupts decision boundaries. The perfect diagonal structure seen in the clean matrix (1) is distorted in the noisy matrix (2), with a particular concentration forming in the bottom-left corner (True: 24k, Predicted: 14k). This visual evidence confirms that noise causes high-purity classes to be confused with lower-purity classes.

Figure 6 analyzes how 24k (True 24k) samples are misclassified under noise. The shift of a large portion of samples to the 14k class can be explained by the Spectral Masking theory. Physically, pure gold (24k) has lower internal damping and a longer resonance tail. However, intense noise masks this long and thin decay tail. The model associates this "truncated" or "buried" signal with 14k gold, which exhibits inherently shorter and more attenuated resonance characteristics due to its higher internal damping. We can term this the "Artificial Damping Effect." This finding proves that noise in industrial environments not only reduces accuracy but also creates the risk of making pure gold appear less valuable.

V. CONCLUSION AND FUTURE WORK

In this study, the robustness of deep learning-based acoustic gold purity measurement systems under industrial noise conditions was systematically analyzed. Experiments conducted using the ResNet-18 architecture and log-Mel spectrograms quantitatively demonstrated how the flawless performance achieved in laboratory conditions (100% Macro-F1) degraded under uncontrolled industrial environments. The results indicated that the system maintained its reliability, with accuracy exceeding 95%, at noise levels of 20 dB SNR and above. However, below this threshold, particularly at 10 dB, performance exhibited a sharp collapse, rendering the system unreliable for industrial standards without further mitigation. Furthermore, the analysis revealed that noise did not distribute measurement errors randomly. Instead, it obscured the resonance decay tails of high-purity (24k) gold through spectral masking. This phenomenon caused the model to misclassify pure gold as lower-purity (14k) gold, which inherently possessed higher intrinsic damping. This "Artificial Damping Effect" highlighted a significant commercial risk where high-value assets could be systematically undervalued. Consequently, while deep learning models held high potential in ART, direct industrial deployment with raw spectrogram data presented significant risks, necessitating a pivot toward signal enhancement and noise-robust feature extraction techniques in future developments.

A. LIMITATIONS OF THE STUDY

While this study established a clear physical robustness baseline, certain limitations must be acknowledged to contextualize the findings. First, the investigation relied on ResNet-18 as a representative deep convolutional benchmark. However, results may vary with emerging architectures such as transformer-based models or attention-enhanced networks, which warranted further comparative investigation. Second, and most critically, the robustness evaluation was based entirely on controlled synthetic noise injection using white and pink noise as standardized proxies for industrial interference. While this approach ensured reproducibility and allowed the isolation of specific degradation mechanisms, it did not fully capture the acoustic complexity of real industrial environments. Real-world settings introduced nonstationary and time-varying noise sources — including impulsive mechanical

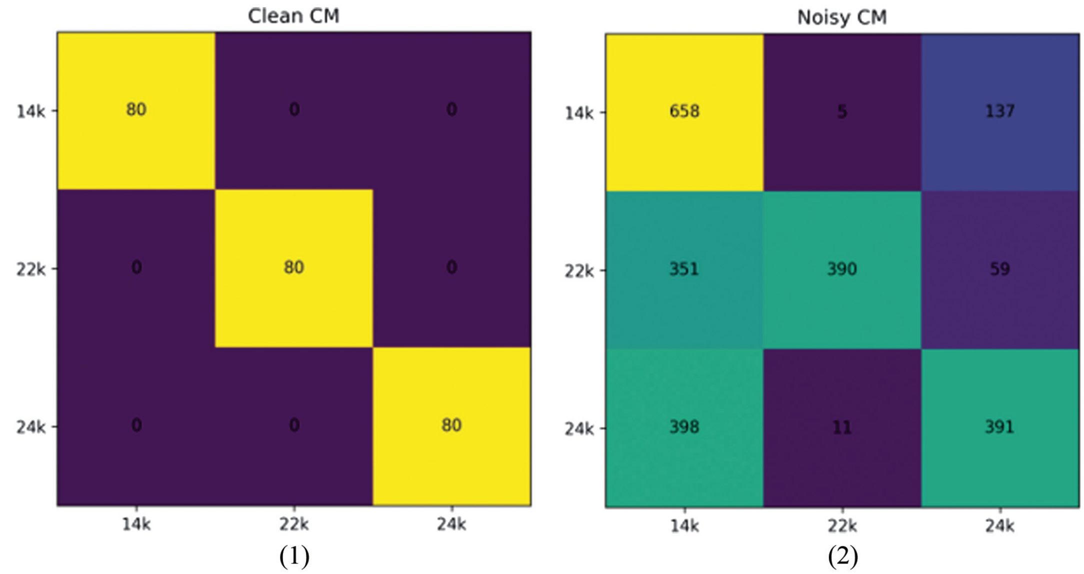

Fig. 5. Confusion matrices under (1) clean laboratory conditions and (2) averaged noisy conditions across all SNR levels and noise types. The perfect diagonal structure in panel (1) confirms flawless purity discrimination under ideal acoustic conditions. In panel (2), the diagonal degrades significantly, with a pronounced off-diagonal concentration in the 24k→14k cell, visually confirming that noise systematically biases the model toward predicting lower-purity classes.

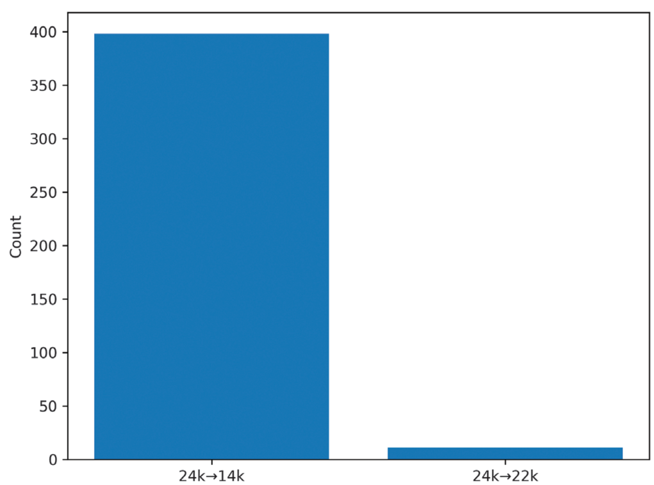

Fig. 6. Systematic misclassification analysis of 24k gold samples under increasing noise levels, illustrating the "Artificial Damping Effect." As SNR decreases, a growing proportion of true 24k samples are predicted as 14k gold. Physically, broadband noise masks the prolonged, low-amplitude resonance tail that distinguishes high-purity gold's low internal damping — causing the model to interpret the noise-truncated signal as the naturally shorter, more damped acoustic signature of 14k alloy. This bias introduces a commercially significant risk of undervaluing high-purity gold assets in industrial inspection scenarios.

shocks, reverberant room acoustics, and intermittent electrical interference — whose spectral and temporal characteristics differed substantially from the stationary noise profiles employed here. As a result, the 20 dB SNR operational threshold identified in this study should be interpreted as a controlled baseline boundary rather than a directly transferable deployment specification.

The generalizability of this threshold to specific industrial environments therefore remained an open research question. Future work should prioritize two complementary directions: (i) extending the noise stress-test framework to include nonstationary and impulsive interference profiles, and (ii) conducting on-site validation experiments using recordings collected directly from gold refinery or workshop environments. Such field validation would confirm whether the identified robustness envelope holds under real acoustic conditions and, if necessary, recalibrate the operational threshold accordingly. Until such validation is completed, deployment in uncontrolled industrial settings should be approached with appropriate caution and supplemented by hardware-level noise isolation or signal preprocessing measures.

AUTHOR CONTRIBUTIONS

M.O.D.: Conceptualization, Methodology, Software, Validation, Data Curation, Writing — Original Draft, Writing — Review & Editing.

S.K.: Methodology, Writing — Original Draft, Writing — Review & Editing, Supervision, Project Administration.

All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest to report regarding the present study.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICAL APPROVAL

This study does not involve human or animal participants. All procedures followed scientific and ethical principles.

USE OF AI TOOLS

During the preparation of this work, the authors used Artificial Intelligence (AI) tools only for the purposes of improving readability, grammatical editing, and formatting citations.

REFERENCES

[1] ASTM E1876-22, "Standard Test Method for Dynamic Young's Modulus, Shear Modulus, and Poisson's Ratio by Impulse Excitation of Vibration," ASTM International, West Conshohocken, PA, 2022.

[2] ASTM E2001-20, "Standard Guide for Resonant Ultrasound Spectroscopy for Defect Detection in Both Metallic and Non-metallic Parts," ASTM International, West Conshohocken, PA, 2020.

[3] C. J. Hellier, *Handbook of Nondestructive Evaluation*, 3rd ed. New York: McGraw-Hill Education, 2020.

[4] J. L. Rose, *Ultrasonic Guided Waves in Solid Media*. Cambridge, UK: Cambridge University Press, 2014.

[5] K. F. Graff, *Wave Motion in Elastic Solids*. New York, NY: Dover Publications, 1991.

[6] L. E. Kinsler *et al.*, *Fundamentals of Acoustics*, 4th ed. New York, NY: Wiley, 2000.

[7] M. Müller, *Fundamentals of Music Processing*. Cham, Switzerland: Springer, 2015.

[8] K. He *et al.*, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[9] D. P. Kingma and J. Ba "Adam: A method for stochastic optimization," *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[10] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 131–135.

[11] J. Salamon and J. P. Bello "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process Lett.*,vol. 24, no. 3, pp. 279–283, 2017, DOI: 10.1109/LSP.2017.2657381.

[12] K. J. Piczak, "ESC: Dataset for environmental sound classification," *Proc. ACM Int. Conf. Multimed.*, 2015, pp. 1015–1018, DOI: 10.1145/2733373.2806390.

[13] B. McFee *et al.*, "Librosa: Audio and music signal analysis in Python," *Proc. 14th Python Sci. Conf. (SciPy)*, 2015, pp. 18–25.

[14] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *Proc. 1st Int. Workshop Deep Learn. Music*. 2017, pp. 37–41.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer, 2009.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, DOI: 10.1038/nature14539.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[20] R. F. Voss and J. Clarke, "1/f noise in music: Music from 1/f noise," *J. Acoust. Soc. Am.*, vol. 63, no. 1, pp. 258–263, 1978.

[21] A. V. Oppenheim and R. W Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[22] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 3rd ed. Berlin, Germany: Springer, 2007.

[23] B.Boashash ed. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*, 2nd ed. Oxford, UK: Academic Press, 2016.

[24] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process*, vol. 28, no. 4, pp. 357–366, 1980.

[25] X. Jin *et al.*, "Construction and recognition of acoustic ID of ancient coins based on deep learning for audio signals," *Herit. Sci.*, vol. 11, p. 46, 2023, DOI: 10.1186/s40494-023-00891-x.

[26] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Feature learning and deep architectures: New directions for music informatics," *J. Intell. Inf. Syst.*, vol. 41, pp. 461–481, 2013, DOI: 10.1007/s10844-013-0248-5.

[27] M. O. Devrim and S. Kirisoglu, "Determination of gold purity degrees using audio features with machine learning algorithms," *Appl. Acoust.*, vol. 240, p. 110887, 2025, DOI: 10.1016/j.apacoust.2025.110887

[28] R. Zhao *et al.*, "Deep learning and its applications to machine health monitoring," *Mech Syst. Signal Process*, vol. 115, pp. 213–237, 2019, DOI: 10.1016/j.ymssp.2018.05.050.

[29] Z. Zhao *et al.*, "Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study," *ISA Trans.*, vol. 107, pp. 224–255, 2020, DOI: 10.1016/j.isatra.2020.08.010.

[30] Y. Zhang *et al.*, "Fault diagnosis of rotating machinery based on recurrent neural networks," *Measurement*, vol. 171, p. 108774, 2021, DOI: 10.1016/j.measurement.2020.108774.

[31] J. Tong *et al.*, "A fault diagnosis method of rolling bearing based on improved deep residual shrinkage networks," *Measurement*, vol. 206, p. 112282, 2023, DOI: 10.1016/j.measurement.2022.112282.

[32] J. Chen *et al.*, "Fault state identification method with noise robustness of dry gas seals based on sample-augmented MA1D-ResNet," *Sensors*, vol. 25, no. 22, p. 7005, 2025, DOI: 10.3390/s25227005.

[33] H. Uzel *et al.*, "Acoustic-based fault diagnosis of electric motors using Mel spectrograms and convolutional neural networks," *Sci. Rep.*, vol. 15, p. 33269, 2025, DOI: 10.1038/s41598-025-33269-z.

[34] C. Kumar *et al.*, "Multi-scale CNN–LSTM network for denoising acoustic partial discharge signal in an electrical apparatus," *Arab. J. Sci. Eng.*, vol. 51, pp. 8461–8479, 2026, DOI: 10.1007/s13369-025-10678-w.

[35] S. X. Chen *et al.*, "An acoustic-homologous transfer learning approach for acoustic emission–based rail condition evaluation," *Struct. Health Monit.*, vol. 20, no. 4, p. 1475921720976941, 2021, DOI: 10.1177/1475921720976941.

[36] H. H. Hsiao *et al.*, "Advanced eco-friendly non-destructive testing in powder metallurgy through AI-based acoustic methodology," *J. Jpn. Soc. Powder Powder Metall.*, vol. 72, no. Suppl, pp. S899–906, 2025, DOI: 10.2497/jjspm.15P-T11-01

[37] X. Li *et al.*, "Belt conveyor idler fault diagnosis method based on multi-scale feature fusion and residual mask convolution attention," *Insight - Non-Destr. Test Cond. Monit.*, vol. 66, no. 2, pp. 82–91, 2024, DOI: 10.1784/insi.2024.66.2.82.

[38] A. Rzeszowska, R. Błażej, and L. Jurdziak, "Non-destructive testing for conveyor belt monitoring and diagnostics: A review," *Appl. Sci.*, vol. 15, no. 24, p. 13272, 2025, DOI: 10.3390/app152413272.

[39] S. Zhang *et al.*, "UAV based defect detection and fault diagnosis for static and rotating wind turbine blade: A review," *Nondestruct. Test*

Eval., vol. 40, no. 4, pp. 1691–1729, 2025, DOI: 10.1080/10589759.2024.2395363.

[40] A. Prakash and T. Tadepalli, "A review of acoustic signal-based detection of damage in railway tracks," *Proc IEEE Conf Eng Informatics (ICEI)*, 2024, pp. 1–13, DOI: 10.1109/ICEI64305.2024.10912263.

[41] G. Wu *et al.*, "A review and progress of insulation fault diagnosis for cable using partial discharge approach," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 32, no. 3, pp. 1639–1652, 2025, DOI: 10.1109/TDEI.2024.3524332.

[42] G. Scarselli and F. Nicassio, "Machine learning for structural health monitoring of aerospace structures: A review," Sensors, vol. 25, no. 19, p. 6136, 2025, DOI: 10.3390/s25196136.

[43] L. Chen, "Multi-sensor monitoring for in-situ defect detection and quality assurance in laser-directed energy deposition [thesis]," Nanyang Technol. Univ., 2024. Available: https://dr.ntu.edu.sg/handle/10356/180195.

[44] Y. Zhu *et al.*, "Current status of machine learning-based acoustic emission technology [preprint]," *SSRN*, 2024. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5084492.