machines

MDPI

Article

Laser Beam Welding State Classification: A Deep Learning Framework for Acoustic Signal Intelligence

Erkan Caner Ozkat

Department of Mechanical Engineering, Faculty of Engineering & Architecture, Recep Tayyip Erdogan University, Rize 53100, Türkiye; erkancaner.ozkat@erdogan.edu.tr

Abstract

Laser beam welding (LBW) of aluminium busbar-to-terminal connections for electric-vehicle battery packs requires precise in-process monitoring. Membrane-free optical microphones provide a high-bandwidth (DC–MHz) acoustic channel that captures keyhole, melt-pool, and plume dynamics. This study proposes Acoustic Signal Intelligence (ASI), a deep learning framework for LBW state classification from a single optical microphone, evaluated on an open dataset (183 AA1050 welds, f_s = 2.5 MHz) under a five-class taxonomy: lack of fusion, lack of connection, sound, marginal, and piercing. The contributions are: (i) a compact 1-D CNN encoder on a mel-scale STFT spectrogram, reaching the highest macro-F1 (0.72 mean across three-fold replicate-out cross-validation) and 100% piercing recall in every fold—a multi-representation fusion variant adding a wavelet-packet decomposition and a 24-feature library targeting the 8, 63 and 110 kHz keyhole-resonance peaks was evaluated as an ablation arm and did not survive cross-validation, so the proposed model is mel-only; (ii) a systematic benchmark against six classical-ML and four deep learning baselines in which Transformer-hybrid ablations and ACGAN-style augmentation underperform compared to the compact CNN on the 122-sample training set, with the Transformer underperformance confirmed by a 30-configuration grid search over learning rate, weight decay, and dropout (best tuned macro-F1 = 0.441 vs. CNN 0.724); and (iii) a Grad-CAM analysis that recovers the keyhole-resonance bands without prior knowledge. A single optical microphone is thus a viable real-time alternative to multi-sensor stacks for battery-pack laser welding.

Keywords: laser beam welding; acoustic emission; convolutional neural network; transformer; defect classification; AI-assisted manufacturing

Academic Editor: Seyedeh Fatemeh Nabavi

Received: 12 May 2026
Revised: 29 May 2026
Accepted: 1 June 2026
Published: 4 June 2026

1. Introduction

Laser-driven manufacturing such as cutting, welding, marking, and surface modification is used for metals, polymers, ceramics, and composite systems. Process control now relies on data-driven and AI-assisted monitoring rather than offline parameter optimisation [1–6]. The automotive shift to battery electric vehicles (BEVs) has made laser welding of aluminium busbar-to-terminal connections safety-critical [5,6]. Each battery pack requires several thousand defect-free welds; a single faulty connection compromises the electrical, thermal, and mechanical integrity of the whole pack [7,8]. Remote laser welding (RLW) dominates because the galvanometer scanner decouples beam motion from robot motion, enabling sub-second cycle times, large stand-off distances, and high-frequency beam wobbling that stabilises the keyhole against aluminium's high reflectivity and low viscosity [9–12]. RLW of 1xxx- and 6xxx-series aluminium nevertheless remains susceptible

to lack of fusion, lack of connection, hot cracking, porosity, and through-piercing, with the defects depending nonlinearly on laser power, part-to-part gaps, beam shape, and clamping conditions [10–13].

In-process monitoring translates the physical emissions of welding (visible light, plasma intensity, infrared radiation, X-ray transmission, ultrasonic waves, and airborne acoustic emissions) into quality indicators [1–4]. Photodiode and high-speed-camera pipelines have matured into industrial closed-loop systems through machine learning (ML) feature extraction and classification [14–19]. Deep convolutional and recurrent architectures have been demonstrated on visual and thermal data [20–25], and one-dimensional (1-D) deep models are emerging for raw process-signal time series [26–30].

A less explored modality is airborne acoustic emissions captured by Fabry–Pérot membrane-free optical microphones. Compared with capacitive or piezoelectric microphones, they offer a flat response from 10 Hz to tens of MHz, no mechanical resonances, sub-millimetre sensing heads that can be placed close to the process, and high immunity to electromagnetic interference [31]. This enables monitoring of keyhole resonance, melt-pool oscillations, plume venting, and rapid solidification, phenomena previously accessible only through high-cost or high-latency sensors [32–34].

Basile et al. [31] released a benchmark RLW dataset of 183 weld trials of 1.0 mm AA1050 sheets, captured at f_s = 2.5 MHz with a Xarion Eta250 ultra optical microphone (Xarion Laser Acoustics GmbH, Vienna, Austria) over a full-factorial 12 × 5 design-of-experiments grid plus three 2000 W piercing tests. Their analysis used Welch power spectral density (PSD) features, total sound power s_p, a band-stop-filtered variant $s_{p,\text{filt}}$, and pairwise PSD quotients to derive threshold rules separating four weld events. Threshold rules are effective for descriptive analysis but do not adapt to the laser-power/gap interaction effects that the same author quantified through ANOVA, do not generalise across replicates, and provide limited interpretability for marginal cases. An end-to-end deep learning approach that learns class-discriminative spectro-temporal features from the raw microphone signal is therefore suitable for improving on the threshold baseline while preserving real-time deployability.

This paper introduces Acoustic Signal Intelligence (ASI), a deep learning framework for laser beam welding state classification from optical-microphone signals. The framework is built around three design choices motivated by the literature reviewed in Section 2:

- A compact 1-D CNN encoder on a mel-scale short-time Fourier transform (STFT) spectrogram, based on the author's prior 1-D-signal modelling experience [35–37], which achieves the strongest deep learning performance on this dataset. A multi-representation fusion variant that further concatenates a wavelet-packet decomposition [14,22] and a 24-feature library targeting the 8, 63 and 110 kHz keyhole-resonance bands [31–33], following the multi-modal autoencoder fusion of Weisbrod and Metternich [29], is evaluated as an ablation arm and not retained in the final model.
- A systematic ablation against larger 1-D CNN–Transformer hybrids inspired by recent architectures for 1-D welding signals [25–30] and against a class-conditional GAN-style augmentation [19]. Both arms underperform compared to the compact CNN, indicating that Transformer-hybrid attention models do not transfer well to small welding datasets, consistent with the benchmark-data-scarcity observation of recent reviews [1–3]. The Transformer underperformance is further confirmed by a systematic 30-configuration grid search over learning rate, weight decay, and dropout, under which the best tuned Transformer-hybrid macro-F1 (0.441 ± 0.057) remains 0.283 below the proposed CNN's 0.724 ± 0.034.

- Grad-CAM-based frequency-domain interpretability that recovers the dominant keyhole-resonance bands reported by Basile et al. and by previous physical models [32,33] without explicit prior knowledge of these bands.

The framework is benchmarked against (i) the threshold rule of Basile et al. [31]; (ii) classical ML baselines including random forest [34], principal-component analysis with support vector machines [16], sequential forward floating selection with SVMs [17,18], XGBoost [29,38], ensemble learning [39], and shallow ANN/MLP regression [13,40,41]; and (iii) deep baselines including 1-D CNN, LSTM, and Transformer encoders. The present work extends the author's prior physics-driven RLW penetration framework [42], surrogate-driven RLW process-window modelling [43], and unsupervised photodiode-based weld-defect classification [37] by replacing optical with acoustic emissions and handcrafted features with end-to-end deep learning. The remainder of the paper is organised as follows. Section 2 reviews the related literature. Section 3 describes the dataset, signal-processing pipeline, proposed architecture, and training protocol. Section 4 presents the results. Section 5 discusses the implications and limitations. Section 6 concludes.

2. Related Work

This section is organised along seven thematic axes that place the proposed framework in context. Figure 1 shows the organisation as a taxonomy.

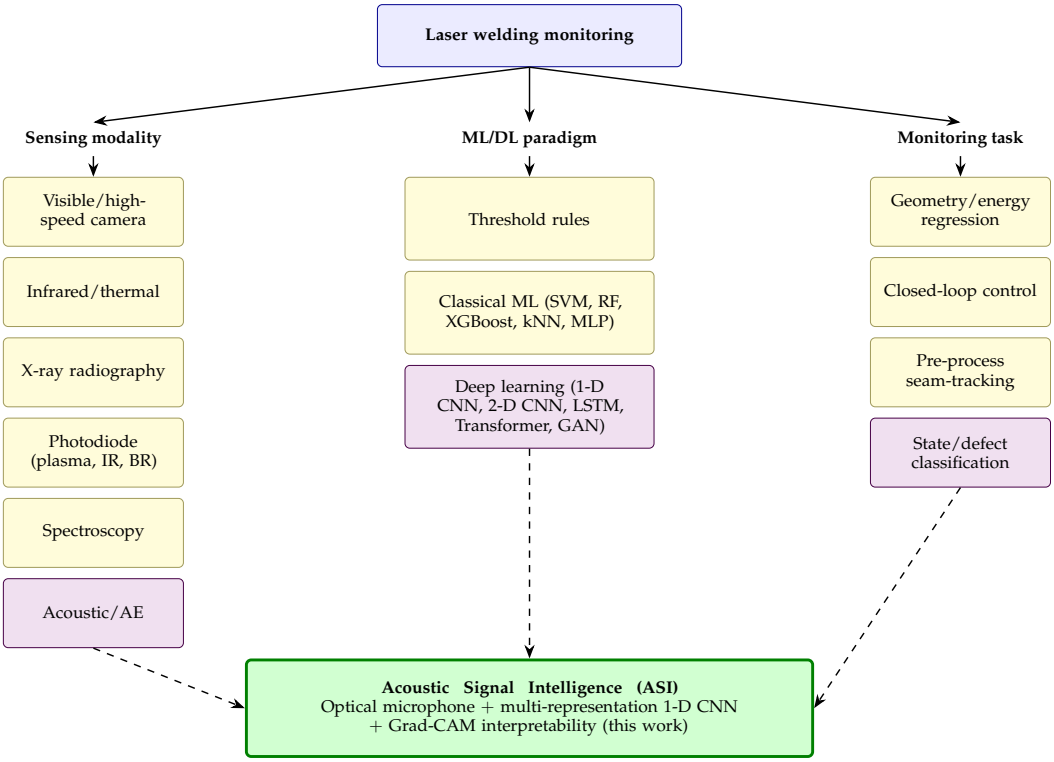

Figure 1. Taxonomy of laser welding monitoring research, with the position of the proposed Acoustic Signal Intelligence framework highlighted.

2.1. Reviews and ML Taxonomies for Laser Welding

Recent reviews indicate that supervised CNNs dominate weld-defect detection across modalities, with attention-based and Transformer hybrids emerging as the state of the art on image and X-ray data [3,4]. Voets et al. [1] categorise ML in laser welding into process design, surrogate modelling, control, and monitoring, noting the scarcity of public benchmarks and evaluation protocols. Gorine et al. [2] report a maturity matrix in which joining processes lag behind machining and additive manufacturing in ML adoption,

attributed to limited labelled data and missing simulators. Klimpel [5] attributed the difficulty of welding aluminium and copper to low optical absorption and rapid keyhole-mode transitions. Lindsay et al. [6] identified ML monitoring as the most under-developed of six defect-mitigation strategies for aluminium LBW. Velazquez and Cheng [44] proposed an Industry 4.0 intelligent quality-management framework for micro-laser welding that takes optical, acoustic, image, and thermal in-process signals as input, surveying defect-detection accuracies in the 74–96% range across those modalities.

2.2. Vision-Based Deep Learning for Weld Monitoring

Image- and video-based DL for weld monitoring covers high-speed visible cameras [8,20,24], multispectral cameras [45], infrared/thermographic cameras [21,23], and X-ray radiography [22,25]. Knaak et al. [21] reached F1 = 95.2% with a CNN–GRU at 1.1 ms inference on a Jetson AGX Xavier; Shevchik et al. [22] reached 71–99% per-class on five welding regimes with a temporal CNN over wavelet-packet spectrograms of AE/back-reflection signals; Balakrishnan et al. [25] reached 99.55% on radiographic welds with a Swin–DeiT hybrid using distillation tokens. Fan et al. [19] reached 98.37% on ten-class laser welding defects with an auxiliary-classifier GAN plus multi-model fusion, and Buongiorno et al. [23] applied a focal-loss CNN to dual-IR Li-ion battery welds. Cai et al. [46] reached 98.37% on four laser-keyhole penetration states in 316L stainless steel with a five-layer 2-D CNN over adjacent-frame adaptively fused high-speed images (5000 fps, 2.9 ms per frame).

2.3. Classical Machine Learning on Weld Signals

Photodiode- and camera-derived feature pipelines remain competitive baselines. Chen et al. [17] (wrapper seven-feature SVM on plume images, 95.93%) and Wang et al. [18] (SFFS + SVM, 98.11%) exemplify the family. You et al. [15] reported a six-sensor MIMO Hammerstein–Wiener model on a 16 kW disk-laser dataset, and Liu et al. [16] reached 91% with PCA–SVM on a five-sensor stack. Chianese et al. [14] reached 97.5% with DWT plus NN on copper-to-steel battery-tab welding, and Caprio et al. [7] reached 100% on a four-class busbar dataset with a single off-axis photodiode and *k*-NN, indicating that careful single-sensor design rivals fusion stacks. Systematic classifier-benchmarking and ensemble-pipeline templates on adjacent engineering data are reported in [38,39]. Asif et al. [47] reached 91.18% on five-class gas-metal-arc-welding quality with an adversarial sequence-tagging classifier on handcrafted air-coupled acoustic-emission features (multi-bandwidth sensors at 1–3 MHz) combined with weld current, voltage, and gas-flow inputs.

2.4. One-Dimensional Time-Series Deep Learning for Welding

Liu et al. [27] reached 98.96% on three-class penetration recognition with a two-stage TCN with attention (ST-TCN, 20.4 ms inference); Solovev et al. [28] reached mAP = 0.85 on a four-class defect task with a hybrid 1-D CNN–Transformer over IR-thermography spatter-feature sequences. Liu et al. [26] encoded 1-D structured-light signals as Gramian angular fields/Markov transition fields and used LeNet/AlexNet/VGG/ResNet backbones for a 4–6% gain over direct 2-D classification. Shi et al. [30] reached 99.39% on a 10-class laser welding signal dataset with a multi-scale CNN–BiLSTM, and Weisbrod and Metternich [29] reached F1 = 0.94 with an XGBoost photodiode-time-series closed-loop concept. Kumar et al. [40] and Habibkhah and Moallem [41] reported MLP/ANN regressors ($r \geq 0.99$) on TWIST polymer welding and robotic seam-profile identification. Analogous 1-D-signal pipelines for adjacent tasks are reported in [35,36,48]. Cao et al. [49] fused a raw airborne-microphone and an InGaAs-photodiode signal in a two-branch 1-D CNN with cross-attention for fibre laser welding of 2A12 aluminium, reaching $99.73 \pm 0.37\%$ on three penetration classes over 2326 segments at 2.25 ms inference. CNN–Transformer hybrids have been reported on adjacent rotating-machinery and machine-tool tasks: Kumar [50]

combined a multi-scale 1-D CNN with three Transformer encoder blocks on CWRU bearing vibration (99.15%, four classes); Liu et al. [51] cascaded an Informer temporal predictor with a CNN–Swin-Transformer image classifier on 4200 infrared bearing images (98.9%, seven classes); Pang et al. [52] classified multivariate symmetrised-dot-pattern images of bearing vibration with a locally enhanced Swin (LEG) Transformer (100% and 99.07% on 10- and 13-class datasets, 2000 training images per class); Wu et al. [53] chained a Transformer signal predictor with an SE-ResNet-50 on a nine-channel CNC dataset of ∼50,000 windows (98.56%, four classes); and Li et al. [54] extended this line with a time–frequency Transformer–TimeGAN augmentation that lifted a downstream 1-D CNN on PU bearings and SEU gearboxes (0.937 ± 0.018 and 0.948 ± 0.012, five classes). All five used training sets at least an order of magnitude larger than the 122-trial budget of the present dataset.

2.5. Remote Laser Welding Process Monitoring

Božič et al. [55] demonstrated CNN-based closed-loop power control in RLW (94% three-class penetration, rise time < 1 s). Prijanovič et al. [9] characterised RLW of advanced high-strength martensitic steel through a parametric DOE. Sun et al. [10], the closest parallel to the present dataset, investigated micro-solidification cracking in RLW of AA6063-T6 battery-tray lap joints with cosine beam oscillation and temporal power modulation. Um et al. [11] developed a DNN for energy estimation in industrial RLW, and Dmitry et al. [12] reported vision-based pre-process seam-tracking for small-batch RLW. Oussaid et al. [13] reached $R^2 > 0.95$ on weld-geometry regression with an ANN on hybrid experimental/FEM data. Ozkat et al. [42] proposed a physics-driven in-process monitoring framework for penetration depth and interface width in laser overlap welding; Ozkat et al. [43] developed a surrogate-driven process-capability-space approach for laser-dimpling parameter selection; and Ozkat [37] applied unsupervised learning to photodiode signals for laser-weld defect analysis. These contributions provide the methodological foundation for the proposed acoustic deep learning framework. Kim et al. [56] regressed keyhole depth in 780DP-steel fibre laser welding by transfer-learning pre-trained 2-D CNNs (MobileNetV2, ResNet50V2, EfficientNetB3, Xception) on coaxial 980 nm camera images and a spectrometer side channel, reaching MAE = 0.049 mm and $R^2 = 0.951$ with the fine-tuned multi-sensor EfficientNetB3 (single-sensor $R^2 = 0.502$–0.681). Adjacent laser welding contributions cover process modelling and parameter-to-quality regression: Murua et al. [57] validated a 3-D transient thermal finite-element model of Inconel 718 dissimilar-thickness LBW against off-axis two-colour pyrometer data and metallographic cross-sections (temperature error $\leq 1.5\%$, geometry error $\leq 13\%$); Yang et al. [58] characterised electron-beam-welded versus in-situ laser-deposited 47 mm TC4 titanium plates (1211 vs. 1123 MPa tensile strength); Zhang et al. [59] predicted post-weld flatness deformation in 0.5 mm 304 stainless steel with a back-propagation network whose weights were optimised by a differential evolution/whale optimisation hybrid (DEWOA-BP, MAE = 0.07 mm, ±0.1 mm error); and Laureto et al. [60] reported an open-source laser polymer welding system for LLDPE multi-layer welds reaching lap-shear loads within ∼5% of the base material above a threshold linear-energy density.

2.6. Acoustic and Physics-Based Monitoring

Sumesh et al. [34] established an early acoustic baseline for shielded-metal-arc welding (88.69% three-class with random forest). Basile et al. [31] extended the optical-microphone paradigm to RLW of AA1050 busbar-to-terminal connections, identifying spectral peaks at 8, 63, and 110 kHz tied to keyhole resonance and proposing dual thresholds on band-stop-filtered sound power for four-class state separation. Darwish et al. [45] combined a

24-feature multispectral LSTM–attention regressor with Isolation Forest anomaly detection on Al-busbar welds. Nomura et al. [33] demonstrated non-contact ultrasonic defect detection on friction-stir-welded Cu–Al joints with a laser-generation/laser-detection scheme (0.3–10 MHz). Volpp [32] estimated steel surface tension above the boiling temperature from melt-pool capillary waves, providing a physical basis for the kHz-band acoustic emissions. Luo et al. [61] monitored pulsed-laser welding penetration in 2 mm 6061 aluminium with a condenser microphone at 20 kHz, decomposing the signal with variational mode decomposition (VMD) and classifying three penetration states (none, partial, full) with a 1-D CNN–LSTM hybrid (99.8 $\pm$ 0.2% accuracy, five-fold CV). Zhang et al. [62] monitored surface residual compressive stress under laser shock peening of 4 mm 7075 aluminium with four contact AE sensors at 3 MHz, mapping VMD-denoised signals to 256 $\times$ 256 Morlet continuous-wavelet-transform images for a parallel CNN–LSTM (99.24 $\pm$ 0.21%, seven energy classes). Zhang and Lai [63] reached AUC = 0.91 at -6 dB and $\sim$1.0 at 0 and 6 dB on eight-class pump-acoustic anomaly detection with a lightweight Audio Spectrogram Transformer (1.01 M parameters) on 128-mel log-spectrograms, outperforming ResNet-18, EfficientNet-B0, MobileNetV3, and unsupervised baselines on the MIMII dataset.

3. Materials and Methods

Figure 2 summarises the end-to-end pipeline of the proposed Acoustic Signal Intelligence framework. The proposed model is mel-only: the wavelet-packet and 24-feature representations were evaluated as a fusion ablation arm (Section 4.4) but did not survive 3-fold cross-validation and are not part of the final architecture. The pipeline consists of five stages: (i) signal acquisition and pre-processing (Stage 1), (ii) construction of the mel-STFT spectrogram (Stage 2), (iii) compression of the mel-spectrogram by a 1-D CNN encoder (Stage 3), (iv) 5-class classification by a single Linear(256 $\rightarrow$ 5) head (Stage 4), and (v) evaluation and Grad-CAM interpretability (Stage 5).

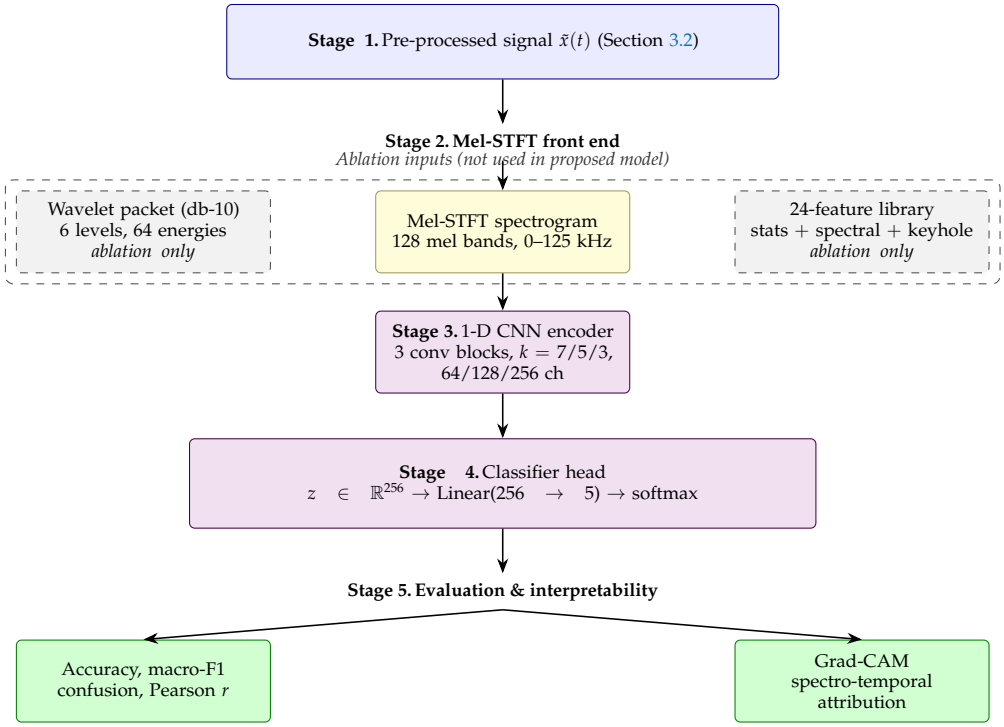

Figure 2. End-to-end Acoustic Signal Intelligence pipeline of the proposed model. The data flow of Stage 1 $\rightarrow$ mel-STFT $\rightarrow$ 1-D CNN encoder $\rightarrow$ Linear classifier head is mel-only. The wavelet-packet and 24-feature representations (dashed grey box) were evaluated as fusion–ablation inputs and discarded after 3-fold cross-validation.

3.1. Dataset

The data utilised in this study were obtained from the work of Basile et al. [31]. The dataset contains 183 weld trials of two stacked 1.0 mm AA1050 sheets in an overlap configuration, joined by a Coherent HighLight FL-ARM 10000 fibre laser (Coherent Corp., Saxonburg, PA, USA) delivered through a Klab Scout-200 two-dimensional galvanometer scanner (k-lab GmbH, Mering, Germany). All welds use a 254 µm focal spot, a 15 mm/s travel speed, a 0.2 mm wobbling radius, and a 1 kHz wobbling frequency. The factorial design varies laser power $P_L \in \{800, 850, 900, 950, 1000, 1050, 1100, 1150, 1200, 1250, 1300, 1350\}$ W and part-to-part gap $g \in \{0, 0.25, 0.5, 0.75, 1.0\}$ mm, with three replicates per cell. Three additional welds at $P_L = 2000$ W and $g = 0$ mm provide piercing samples.

The acoustic signal was captured by a Xarion Eta250 ultra membrane-free Fabry–Pérot optical microphone (Xarion Laser Acoustics GmbH, Vienna, Austria) positioned 70 mm above and 30 mm laterally from the weld zone, sampled at $f_s = 2.5$ MHz over a 2 s window. The resulting raw matrix has a size of 5,000,000 × 183. Class labels were derived from the experimental matrix and metallographic analysis following the rule table in Table 1. The labels used in this work were not assigned solely from the nominal process parameters. In the source dataset of Basile et al. [31], post-weld metallographic characterisation was performed by cutting each seam into two cross-sections located 10 mm from the start and 10 mm from the end of the weld. The cross-sections were mounted, ground, polished, etched, and examined microscopically. The extracted geometrical features included the top weld width, interface weld width, and effective penetration depth. A connection was considered effective when the effective penetration depth exceeded 25 % of the lower-sheet thickness. To preserve consistency with the source study, the present work employed this five-class taxonomy as the supervised labels for model training and evaluation.

Table 1. Five-class label rule table used in this work, derived from the experimental matrix and metallographic analysis. *LoF* : Lack of fusion; *LoC*: lack of connection; *Sound*: sound connection; *Marg.*: marginal/weak-but-acceptable; *Pierc.*: piercing.

Laser Power P_L [W]	Gap g [mm]	Weld State	Class
≤850	any	Lack of fusion	LoF
900–1000	0–0.25	Lack of connection	LoC
900–1000	≥0.5	Lack of connection	LoC
1050–1150	0–0.25	Sound connection	Sound
1050–1150	0.5	Marginal connection	Marg.
1050–1150	≥0.75	Lack of connection	LoC
1200–1350	0–0.5	Sound connection	Sound
1200–1350	0.75	Weak but acceptable	Marg.
1200–1350	1.0	Lack of connection	LoC
2000	0	Piercing	Pierc.

The aluminium busbar setup used in this dataset is similar in structure to the AA6063-T6 RLW battery-tray configuration of Sun et al. [10] and to the copper-to-steel busbar configurations of Chianese et al. [14] and Caprio et al. [7].

3.2. Region-of-Interest Extraction and Pre-Processing

For each trial, the steady-state region of interest is extracted as the central 1 s window $0.5 \leq t \leq 1.5$ s, corresponding to 2.5×10^6 samples at the raw 2.5 MHz rate (or 2.5×10^5 samples after the 10× decimation to 250 kHz), following the convention of Basile et al. [31]. The DC offset is removed and the signal is normalised to unit variance per trial to suppress microphone-to-microphone gain variation. The full-bandwidth signal is kept. No fixed band-stop filter is applied at this stage, so that the network can

learn the band-selective behaviour end-to-end, in contrast to the manual 8–26 kHz filter. Figure 3 summarises the acquisition and pre-processing chain.

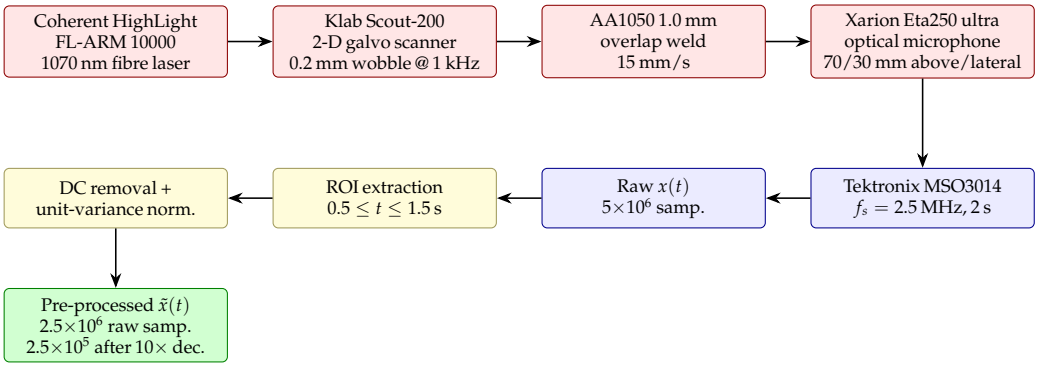

Figure 3. Signal acquisition and pre-processing chain. Top row: Hardware chain from fibre laser source to optical-microphone receiver. Bottom row: Digitisation, region-of-interest selection, and per-trial normalisation.

3.3. Multi-Representation Front End

The Welch PSD is computed from L overlapping Hann-windowed segments $\{x_i\}_{i=0}^{L-1}$ of the pre-processed signal $\tilde{x}(t)$ as:

$$\widehat{P}_{xx}(f) \;=\; \frac{1}{L\,U}\sum_{i=0}^{L-1}\left|X_i(f)\right|^2, \tag{1}$$

with $X_i(f)$ the discrete Fourier transform of segment i and U the window-energy normalisation. The Hann window of 1024 decimated samples (4 ms at f_s = 250 kHz, equivalent to 10,240 raw samples at 2.5 MHz) with 50 % overlap yields a ∼244 Hz frequency resolution. The PSD is mapped to a 128-band log-mel scale via the triangular filterbank H_m, $m = 1, \ldots, M$, with centre frequencies equally spaced on the mel scale and converted back to Hz through the inverse of

$$m(f) \;=\; 2595\,\log_{10}\!\left(1+\tfrac{f}{700}\right) \tag{2}$$

$$S_m(t) \;=\; \log\!\left(1+\sum_k H_m(f_k)\,\widehat{P}_{xx}(f_k,t)\right) \tag{3}$$

producing a $128 \times T$ time–frequency representation $\mathbf{S} \in \mathbb{R}^{128\times T}$ with logarithmic frequency spacing across the 0–125 kHz working band.

The mel filterbank is used in this study as a fixed log-frequency compression of the linear STFT, not as a model of human auditory perception. Three considerations motivate this choice in the ultrasonic regime: (i) the three keyhole-resonance peaks at 8, 63 and 110 kHz [31,32] span more than a decade of frequency, so logarithmic band spacing places resolution at each peak rather than over-resolving a single decade as a linear STFT would; (ii) the 128-band compression reduces the 513-bin linear STFT magnitude by a factor of four at the encoder input, lowering the parameter count of the first 1-D convolutional block from approximately 230 k to 57 k and limiting overfitting on the 122-trial training set; and (iii) the Grad-CAM analysis (Section 4.6) confirms that the resulting representation preserves enough resolution at the three keyhole peaks for the network to localise them without explicit prior knowledge. The mel scale was originally developed for human auditory perception below 10 kHz and allocates approximately one-sixth of its 128 bands to the sub-1 kHz range that carries no welding signal. To address whether this allocation reduces classification performance on the ultrasonic process band, the proposed 1-D CNN was

re-trained on three 128-band filterbanks under the same 3-fold replicate-out cross-validation protocol used for the performance evaluation of the proposed 1-D CNN as: (i) the proposed mel filterbank (0–125 kHz); (ii) a uniform-Hz linear filterbank over 0–125 kHz; and (iii) a log-Hz filterbank over 1–125 kHz. The three filterbanks achieved 0.724–0.746 macro-F1 and 0.732–0.754 accuracy with a piercing recall of 1.000 in every fold, and the largest pairwise gap in mean macro-F1 was 0.022 (mel vs. linear), lower than the fold-to-fold standard deviation of either variant. The mel filterbank achieved the lowest fold-to-fold standard deviation of the three (0.034 macro-F1 vs. 0.050–0.056) and was therefore the empirically supported choice on this dataset.

A six-level wavelet-packet decomposition with the Daubechies db-10 mother wavelet, following Shevchik et al. [22] and Chianese et al. [14], yields 64 sub-band time series $\{x_b\}_{b=1}^{64}$ whose normalised energies form the 64-D wavelet-packet feature vector $\mathbf{w} \in \mathbb{R}^{64}$:

$$\mathbf{w}_b = \frac{\sum_n x_b^2[n]}{\sum_{b'} \sum_n x_{b'}^2[n]}. \tag{4}$$

Continuous-wavelet-transform scalograms are also computed for visualisation [36,37]. Following Darwish et al. [45], a 24-element feature vector $\mathbf{h} \in \mathbb{R}^{24}$ is computed per trial. The vector contains seven statistical moments (mean, root-mean-square, variance, peak-to-peak, crest factor, skewness, kurtosis), four spectral descriptors (peak frequency, spectral kurtosis, spectral entropy, fundamental frequency), three shape descriptors (autocorrelation peak, sample entropy, zero-crossing rate), three keyhole-resonance band energies $E_{[f_l, f_h]}$ in the 8–12 kHz, 60–66 kHz, and 105–115 kHz bands that target the peaks identified in [31–33] together with their three fractions of total power, the 8–26 kHz Basile band-stop band energy with its fraction, and the total and band-stop-filtered sound powers s_p and $s_{p,\mathrm{filt}}$ used by the threshold rule (7 + 4 + 3 + 3 + 3 + 2 + 2 = 24 features). The Pearson correlation coefficient between the predicted regime score $\hat{y}$ and a process input $\xi \in \{P_L, g\}$, used as a physical-trend sanity check, is given by

$$r(\hat{y}, \xi) = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})(\xi_i - \bar{\xi})}{\sqrt{\sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{N} (\xi_i - \bar{\xi})^2}}. \tag{5}$$

The training loss combines a class-balancing weight $\alpha_c \propto 1/n_c$ (normalised so that $\sum_c \alpha_c = C$) with a focal modulation $(1 - p_{i,c})^\gamma$ that downweights well-classified examples [23]:

$$\mathcal{L}_{\mathrm{focal}}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \sum_{c=1}^{C} \alpha_c \, y_{i,c} \, (1 - p_{i,c}(\boldsymbol{\theta}))^\gamma \, \log p_{i,c}(\boldsymbol{\theta}), \tag{6}$$

with $y_{i,c} \in \{0, 1\}$ the one-hot ground-truth class indicator, $p_{i,c}(\boldsymbol{\theta})$ the network's softmax output, $C = 5$ classes, N training trials, and $\gamma = 0$ (i.e., class-weighted cross-entropy) used in the present study. The hyperparameters are listed in Table 2.

3.4. Proposed 1-D CNN Encoder

The proposed encoder is shown in Figure 4. It is a compact 1-D convolutional network operating on the log-mel spectrogram. The mel-spectrogram passes through three 1-D convolutional blocks (kernel size 7, 5, 3 with strides 2, 2, 1; 64, 128, 256 filters; batch normalisation; GELU activation; dropout 0.1) that compress the 128-band $\times$ T-frame representation along the time axis. A global temporal mean pooling reduces the 256-channel sequence to a 256-D embedding z, which is mapped to the five class logits by a single Linear($256 \to 5$) layer. The network is trained with a class-balanced focal cross-entropy loss to address the imbalance toward the *Sound* and *LoC* classes, following Buongiorno et al. [23].

The CNN-only backbone, rather than the larger Transformer hybrids of Solovev et al. [28] and Liu et al. [27], was chosen because of the limited training set. A multi-representation fusion variant that concatenates z with the 64-D wavelet-packet vector $\mathbf{w}$ and the 24-D handcrafted feature vector $\mathbf{h}$, then passes the concatenated vector through a 2-layer MLP head, is evaluated as an ablation arm in Section 4.4. The architecture between the mel-only and fusion branches is selected by the following criterion: choose the architecture with the highest 3-fold CV mean accuracy; if multiple architectures lie within one standard deviation of the maximum, prefer the one with the lowest fold-to-fold standard deviation. Under this rule the mel-only branch (0.732 ± 0.041) is preferred: the full-fusion variant's mean (0.628) sits more than two standard deviations below mel-only's mean, so it fails the mean criterion before the variance tie-breaker would apply; the fusion variant also has $2.7\times$ larger fold-to-fold standard deviation, a corroborating sign that the simpler mel-only branch generalises more reliably.

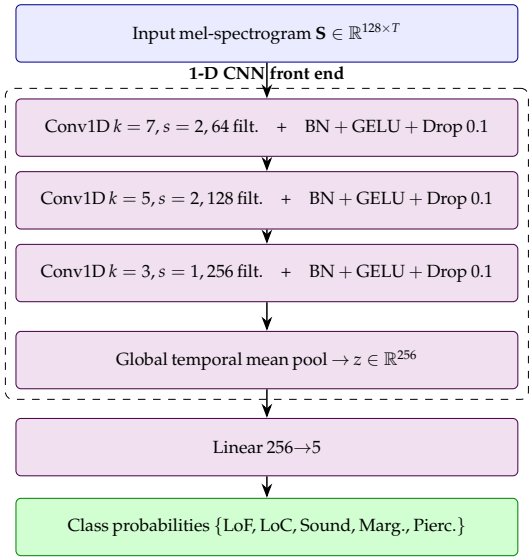

Figure 4. Proposed compact 1-D CNN encoder. Three convolutional blocks (kernel 7/5/3, stride 2/2/1, 64/128/256 filters) compress the mel-spectrogram along time. A global temporal mean pool produces the 256-D embedding z, which a single Linear($256 \rightarrow 5$) layer maps to the five class logits (199,237 trainable parameters in total).

3.5. Transformer Ablation

Two larger alternatives are evaluated against the proposed CNN: (i) a hybrid 1-D CNN–Transformer that combines the CNN front end with a 1–4-layer, 4–8-head Transformer encoder of $d_{\text{model}} = 64$–128, following recent state-of-the-art architectures [25,27,28,30], and (ii) a class-conditional auxiliary-classifier GAN-style oversampling (ACGAN-style) augmentation in which minority classes are upsampled to the majority count with additive Gaussian noise on the mel-spectrogram and random time-masking, similar to Fan et al. [19]. The Transformer hybrids and the ACGAN-style augmentation are presented as ablation arms rather than the proposed model, because their performance on the 122-sample training set is well below that of the proposed CNN. This is consistent with the benchmark-data-scarcity observation of recent reviews [1–3].

To address whether the Transformer-hybrid underperformance is associated with insufficient hyperparameter tuning, a grid search was performed over the initial learning rate, the AdamW weight decay, and the dropout rate, with the search space lr $\in \{1\times10^{-3}, 3\times10^{-4}, 1\times10^{-4}\}$, wd $\in \{1\times10^{-4}, 1\times10^{-3}, 1\times10^{-2}\}$, and dropout $\in \{0.1, 0.3, 0.5\}$ (27 configurations). Each configuration was trained under the same 3-fold replicate-out cross-validation protocol used for the proposed network (Section 3.7; 60 epochs, AdamW

with cosine annealing, gradient clipping at 1.0, class-balanced cross-entropy, batch size 8). The grid was applied to the Tr-small2 variant (2-layer Transformer, 4 heads, $d_{\text{model}} = 64$, FFN 128; 150,373 parameters), as the per-fold results indicate that lighter Transformers outperform the original 4L/8H/$d = 128$ variant on this dataset. The top-3 small2 configurations by mean macro-F1 were then re-run on the original 4L/8H/$d_{\text{model}} = 128$ Transformer (1,049,925 parameters). In total, 30 hyperparameter configurations × 3 folds = 90 training runs were performed (4.3 h on a single-thread CPU). The Transformer-hybrid hyperparameter grid search results show that 27 configurations were tested on Tr-small2, and the three best configurations were subsequently re-run on Tr-original under 3-fold replicate-out cross-validation.

3.6. Baseline Classifiers

Four baseline families were evaluated for direct comparison with the proposed CNN:

- **Threshold rule:** Dual-threshold rule [31] on $s_{p,\text{filt}}$, with cut-points 0.5 and 1.8 mW, extended to the present 5-class taxonomy by mapping *Marg.* to *Sound* (the rule cannot resolve the marginal regime).
- **Classical ML on the 24-feature library**: Random forest [34,39], SVM with sequential forward floating selection [17,18], principal-component-analysis SVM [16], XGBoost [29,38,64], and shallow MLP [13,40,41].
- **Image-encoded 2-D CNN**: Gramian angular field encoding plus a small ResNet-style 2-D CNN, following Liu et al. [26].
- **Recurrent baselines**: LSTM and CNN–GRU on the mel-spectrogram, following Knaak et al. [21].

In addition, Section 3.5 describes the larger Transformer-hybrid ablation arms; these are based on the architectures of [25,27,28,30] and reported alongside the proposed CNN.

3.7. Training and Evaluation Protocol

The 183-trial dataset was split by replication to avoid leakage. Replicates 1 and 2 formed the training set (122 trials, including 2 of the 3 piercing trials). Replicate 3 formed the test set (61 trials, including 1 piercing trial). The AdamW optimiser was used with a cosine-annealed learning rate, weight decay, and the focal cross-entropy loss with class-balancing weights, following Buongiorno et al. [23]. Performance is reported in terms of accuracy, macro-F1, per-class precision and recall, confusion matrices, Pearson correlation coefficients between the predicted regime score and the two process inputs P_L and g, and McNemar paired-significance tests. The multi-model benchmarking and physics-informed reporting follow the methodology of [42,43,48,65,66].

The full hyperparameter set used to produce the results of Section 4 is listed in Table 2. All randomness is seeded with the same value across runs (`SEED` = 1234); the training-curve variance due to data-loader shuffle order is quantified in Section 4.4.

Table 2. Hyperparameters of the proposed 1-D CNN, the deep baselines, and the classical-ML baselines.

Hyperparameter	Value
Train/test split	By replicate (rep 1 + 2 = 122 train, rep 3 = 61 test); 2/1 piercing split
Acquisition sample rate f_s	2.5 MHz (Basile dataset)
Decimation factor/working f_s	10/250 kHz (Nyquist 125 kHz, retains 8/63/110 kHz peaks)
Region of interest	$0.5 \leq t \leq 1.5$ s steady-state slice (2.5 M raw/250 k decimated samples)
STFT window/overlap/FFT	Hann, 1024 samples/512/1024

Table 2. *Cont.*

Hyperparameter	Value
Mel filterbank	128 bands, 0–125 kHz (working Nyquist)
Wavelet packet	Daubechies db-10, 6 levels (64 sub-band energies)
Handcrafted feature library	24-D (statistical, spectral, shape, keyhole 8/63/110 kHz, Basile $s_p/s_{p,\mathrm{filt}}$)
Optimiser (deep)	AdamW; weight decay 10^{-4}
Initial learning rate (deep)	1×10^{-3}, cosine annealing to 0 over 60 epochs
Batch size (deep)	8
Loss (deep)	Focal cross-entropy ($\gamma = 0$), class-balanced weights $\alpha_c \propto 1/n_c$ normalised to $\sum_c \alpha_c = C$
Gradient clipping (deep)	ℓ_2-norm 1.0
Dropout (deep, prop. CNN)	0.1 (conv blocks); fusion–ablation head additionally uses 0.2
Conv blocks (prop. CNN)	3 blocks; kernel 7/5/3, stride 2/2/1, channels 64/128/256, BN + GELU
Pooling (prop. CNN)	Global mean over time $\rightarrow$ 256-D
Classifier head (prop. CNN)	Linear(256 $\rightarrow$ 5)
Total parameters (prop. CNN)	199,237
Fusion ablation head	Linear(256 + 64 + 24 $\rightarrow$ 192) + GELU + Drop + Linear(192 $\rightarrow$ 5); 265 157 params
Hybrid encoder (Tr-original)	4-layer Transformer, 8 heads, $d_{\mathrm{model}} = 128$, FFN 1024
Hybrid encoder (Tr-small{1,2,4})	$\{1, 2, 4\}$-layer Transformer, 4 heads, $d_{\mathrm{model}} = 64$, FFN 128; dropout 0.3 (small1/small2), 0.4 (small4)
Hybrid encoder (Tr-sinpos2)	2-layer Transformer, 4 heads, $d_{\mathrm{model}} = 96$, FFN 192, sinusoidal pos.
Hybrid grid-search range	lr $\in \{10^{-3}, 3\times10^{-4}, 10^{-4}\}$; wd $\in \{10^{-4}, 10^{-3}, 10^{-2}\}$; dropout $\in \{0.1, 0.3, 0.5\}$; 27 cells on Tr-small2 + top-3 re-run on Tr-original; 3-fold CV (90 runs total)
ACGAN-style augmentation	class-balanced oversampling + mel-spec Gaussian noise $\sigma = 0.03$ + 8-frame time-mask
Random forest (24-feat lib)	400 trees, default sklearn
PCA-SVM (24-feat lib)	$n_{\mathrm{PC}} = 8$, RBF SVM, $C = 3.0$
SFFS-SVM (24-feat lib)	8 features (forward selection, cv = 2), RBF SVM $C = 3.0$
XGBoost (24-feat lib)	400 trees, max-depth 4, lr 0.05, sub/col 0.8
Shallow MLP (24-feat lib)	Hidden (64, 32), ReLU, max-iter 2000

4. Results

4.1. Class-Conditional Input Statistics

Figure 5 shows one representative log-mel spectrogram per class, sampled at random from the dataset. The progression from *LoF* (almost no acoustic activity, $P_L = 800$ W) to *Sound* (broad spectral content concentrated around the 63 kHz keyhole-resonance peak, $P_L = 1050$ W) and to *Pierc.* (broadband 8–110 kHz activity, $P_L = 2000$ W) is visible by direct inspection and is consistent with the keyhole-physics description of the keyhole-resonance peaks at 8, 63, and 110 kHz [31,32].

4.2. Threshold-Rule Baseline Replication

The dual-threshold rule on $s_{p,\mathrm{filt}}$ proposed by Basile et al. [31] was first replicated to obtain a directly comparable baseline. Sound power was computed via Welch's method with the same window and overlap settings, the 8–26 kHz Butterworth band-stop filter was applied, and the cut-points 0.5 mW and 1.8 mW were used to assign one of {LoF, LoC, sound}; piercing was assigned by an additional rule based on broadband attenuation at 2000 W. The results are reported in Table 3 as a confusion matrix with per-class metrics.

The rule maps nearly every trial to *Sound* (the dominant class) and assigns the single piercing trial to *LoC*, yielding an accuracy of 0.295.

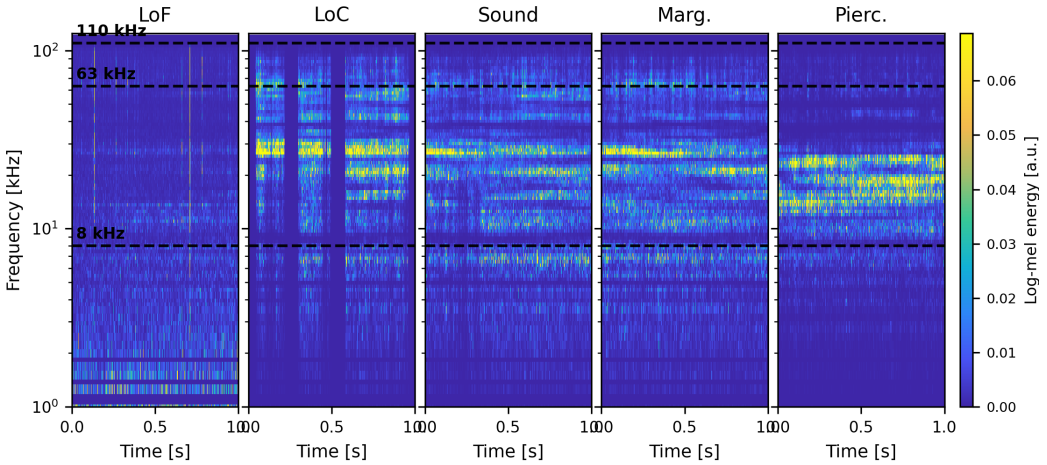

Figure 5. Representative log-mel spectrograms (parula colormap, log-frequency axis) of one trial per class. From left to right: *LoF* ($P_L = 800$ W, $g = 0$ mm); *LoC* ($P_L = 900$ W, $g = 0$ mm); *Sound* ($P_L = 1050$ W, $g = 0$ mm); *Marg.* ($P_L = 1050$ W, $g = 0.5$ mm); *Pierc.* ($P_L = 2000$ W, $g = 0$ mm).

Table 3. Confusion matrix of the replicated threshold-rule baseline on the rep-3-held-out fold ($n = 61$ test trials). The horizontal arrow (→) indicates predicted classes along the columns, and the vertical arrow (↓) indicates actual (ground-truth) classes along the rows.

Predicted →/Actual ↓	LoF	LoC	Sound	Marg.	Pierc.
LoF	0	0	10	0	0
LoC	0	0	25	0	0
Sound	0	0	18	0	0
Marg.	0	0	7	0	0
Pierc.	0	1	0	0	0

4.3. Classical ML Baselines

Random forest, PCA–SVM, SFFS–SVM, XGBoost, and shallow MLP classifiers were trained on the 24-feature library (Section 3.3) under a three-fold cross-validation protocol that held out one replicate at a time (Fold 1: rep 1 held out; Fold 2: rep 2; Fold 3: rep 3). Each fold had 122 training trials and 61 test trials, with the three piercing trials distributed 2/1 across train/test in each fold. The mean and standard deviation across the three folds are reported in Table 4, for direct comparison with the corresponding architectures in [13,16–18,29,34,38–41]. Threshold-rule predictions are deterministic and identical across folds because the rule does not depend on the training data.

Table 4. Performance of classical-ML baselines on the 24-feature library, under stratified 3-fold cross-validation over replicates. Values are reported as mean ± standard deviation across folds.

Classifier	Accuracy	Macro F1	Pierc. Recall	Inference [ms]
Threshold rule	0.295 ± 0.000	0.092 ± 0.000	0.000 ± 0.000	0.02
Random Forest	0.743 ± 0.043	0.703 ± 0.011	1.000 ± 0.000	3.4
PCA-SVM	0.623 ± 0.058	0.606 ± 0.041	1.000 ± 0.000	0.03
SFFS-SVM	0.760 ± 0.008	0.690 ± 0.007	1.000 ± 0.000	0.07
XGBoost	0.727 ± 0.041	0.694 ± 0.028	1.000 ± 0.000	0.11
Shallow MLP	0.650 ± 0.034	0.602 ± 0.125	0.667 ± 0.471	0.02

4.4. Deep Baselines, Proposed Model, and Transformer Ablation

All deep learning models, including the proposed mel-only 1-D CNN, GAF + small ResNet, CNN–GRU, LSTM, and Transformer-hybrid ablation arms, were evaluated under the same three-fold replicate-out cross-validation protocol applied to the classical baselines (Table 4), so that fold-to-fold variability is reflected in the comparison. CPU inference latency was also recorded for comparison with previously reported real-time welding-monitoring systems [21,27]. Table 5 reports the 3-fold replicate-out cross-validation performance of the proposed 1-D CNN, deep learning baselines, and Transformer-hybrid ablation models, with results presented as mean ± standard deviation across folds.

Across the three-fold cross-validation, the proposed 1-D CNN achieved a mean macro-F1 of 0.724 ± 0.034, the highest of any model tested. SFFS-SVM achieved a slightly higher mean accuracy (0.760 ± 0.008 vs. 0.732 ± 0.041) but a lower macro-F1 (0.690 ± 0.007). The fold-level rankings are stable: the CNN had the highest F1 in Folds 1 and 3, while XGBoost (F1 = 0.728) was highest in Fold 2 with the CNN fifth among the six classifiers (CNN 0.692, MLP 0.692). SFFS-SVM had the lowest variance because it operates on the deterministic 24-feature library.

Table 5. Performance of the proposed 1-D CNN, deep learning baselines, and Transformer-hybrid ablation arms under 3-fold replicate-out cross-validation. Values are reported as mean ± standard deviation across folds.

Model	Accuracy	Macro F1	Pierc. Recall	Inference [ms]
1-D CNN (proposed, 3-fold CV)	0.732 ± 0.041	0.724 ± 0.034	1.000 ± 0.000	4.1
1-D CNN (Fold 3 only)	0.721	0.707	1.00	4.1
GAF + ResNet-18	0.448 ± 0.054	0.502 ± 0.023	1.000 ± 0.000	12.0
CNN–GRU	0.164 ± 0.000	0.056 ± 0.000	0.000 ± 0.000	9.4
LSTM	0.470 ± 0.056	0.552 ± 0.057	1.000 ± 0.000	103.2
Hybrid 1-D CNN–Transformer (4L, 8H, $d = 128$)	0.514 ± 0.028	0.333 ± 0.036	0.000 ± 0.000	17.0
Hybrid + ACGAN augmentation (4L, 8H, $d = 128$, Fold 3)	0.148	0.185	1.00	19.6
Hybrid (1L, 4H, $d = 64$)	0.541 ± 0.013	0.306 ± 0.010	0.000 ± 0.000	2.8
Hybrid (2L, 4H, $d = 64$)	0.514 ± 0.039	0.326 ± 0.035	0.000 ± 0.000	5.0
Hybrid (4L, 4H, $d = 64$, drop 0.4)	0.519 ± 0.020	0.321 ± 0.030	0.000 ± 0.000	4.6
Hybrid (2L, 4H, $d = 96$, sinusoidal pos.)	0.437 ± 0.015	0.282 ± 0.011	0.000 ± 0.000	5.9
Hybrid (2L, 4H, $d = 64$, best of 27-cell grid)	0.541 ± 0.035	0.441 ± 0.057	0.667 ± 0.471	3.7
Hybrid (4L, 8H, $d = 128$, best of grid-validated top-3)	0.415 ± 0.039	0.304 ± 0.016	0.000 ± 0.000	16.5

Among the deep baselines, the proposed CNN's three-fold-CV mean accuracy was at least 0.19 above every Transformer-hybrid variant (0.541, 0.514, 0.519, 0.437, 0.514 for small1, small2, small4, sinpos2, and the original 4L/8H/$d = 128$) and at least 0.26 above the recurrent and image-encoded baselines (LSTM 0.470, GAF 0.448, CNN–GRU 0.164); the macro-F1 gap was even larger because the Transformer variants collapsed to a single-class output with piercing recall of 0.000 on every fold. Three observations follow: (i) GAF, LSTM, and the original 4L/8H Hybrid sit 0.08–0.10 accuracy below their Fold-3-only values (different RNG states between standalone and sequential CV runs), so the earlier table underestimated rather than overstated their typical performance; (ii) the four light Transformer variants stay within ±0.08 accuracy and ±0.07 macro-F1 of their Fold-3 values, so the per-fold ranking is preserved; and (iii) CNN–GRU is fold-invariant (acc 0.164 ± 0.000) because it converged to a single-class prediction on every fold.

Table 6 summarises the 27-cell grid on Tr-small2 and the three-cell validation re-run on the original 4L/8H/$d_{\text{model}} = 128$ Transformer, addressing whether the Transformer-hybrid underperformance reflects inadequate optimiser tuning rather than an architectural mismatch on a 122-sample training set. Tuning lifted the Tr-small2 macro-F1 from the 0.306–0.333 untuned range (Table 5) to 0.441 ± 0.057 (+0.115 absolute) at lr = 10^{-3},

wd $= 10^{-2}$, and dropout $= 0.1$, and piercing recall improved from 0.000 to 0.667 (the network recovered the piercing trial on two of three folds). However, the best-tuned Tr-small2 macro-F1 remains 0.283 below the proposed CNN's 0.724 ± 0.034, and the accuracy gap is $\geq$0.19 across all 27 grid cells. Scaling up to the original 4L/8H/$d_{\text{model}} = 128$ Transformer under the same hyperparameters did not close the gap: the top-three small2 configurations re-run on the 1.05 M parameter Transformer yielded a best-cell macro-F1 of 0.304 ± 0.016, below the untuned 0.333 ± 0.036, with piercing recall at 0.000 on every fold. These results indicate that Transformer-hybrid attention is not competitive on the 122-sample training set, consistent with the data-scarcity observations of Voets et al. [1], Gorine et al. [2], and Eren [3].

Table 6. Transformer-hybrid hyperparameter grid search: 27 cells on Tr-small2 and top-3 cells re-run on Tr-original under 3-fold replicate-out cross-validation. "Best cell" has the highest mean macro-F1.

Search Arm	Cells	F1 Min/Med/Max	Best (lr, wd, Drop)	Best Accuracy	Best Pierc. Recall
Tr-small2 (150,373 params)	27	0.253/0.306/0.441	10^{-3}, 10^{-2}, 0.1	0.541 ± 0.035	0.667 ± 0.471
Tr-original (1,049,925 params)	3	0.240/0.283/0.304	3×10^{-4}, 10^{-3}, 0.1	0.415 ± 0.039	0.000 ± 0.000
CNN reference (199,237 params)	—	—/—/0.724	—	0.732 ± 0.041	1.000 ± 0.000

Table 7 reports the confusion matrix of the proposed 1-D CNN on the rep-3-held-out fold. Errors concentrated on the rare *Marg.* class: five of seven marginal trials were predicted as *Sound* and one as *LoC*, reflecting the close acoustic similarity between marginal and sound connections in the 1050–1350 W power range. The *LoF*, *LoC*, and *Sound* classes were recovered with F1 = 0.857, 0.809, and 0.718 respectively. The single piercing trial in this fold was correctly classified.

Table 7. Confusion matrix of the proposed 1-D CNN on the rep-3-held-out fold ($n = 61$ test trials). The horizontal arrow (→) indicates predicted classes along the columns, and the vertical arrow (↓) indicates actual (ground-truth) classes along the rows.

Predicted →/Actual ↓	LoF	LoC	Sound	Marg.	Pierc.
LoF	9	1	0	0	0
LoC	2	19	2	2	0
Sound	0	1	14	3	0
Marg.	0	1	5	1	0
Pierc.	0	0	0	0	1

Table 8 reports two complementary experiments on the multi-representation front end (Section 3.3). The top block is a head-to-head three-fold replicate-out cross-validation of the proposed mel-only CNN (single Linear($256 \rightarrow 5$) head, 199,237 parameters) against the full mel + WPT + 24-feat fusion variant (two-layer MLP head, 265,157 parameters) under the same protocol used in Table 5, testing whether the Fold 3 advantage of full fusion generalises across folds. The bottom block is a per-component ablation on the rep-3-held-out fold; all four variants share the same two-layer MLP head and differ only in the fusion input, isolating how each representation contributes to the Fold 3 fusion result.

All four ablation variants share the same two-layer MLP head (Linear($d \rightarrow 192$), GELU, dropout, Linear($192 \rightarrow 5$)) and differ only in the fusion input d; the proposed model uses a simpler single Linear($256 \rightarrow 5$) head. Three observations follow. First, on Fold 3 neither the WPT vector nor the 24-feature library helps the two-layer-head variants when added to the mel-CNN embedding: the mel + WPT variant loses 0.15 accuracy relative to the mel-only arm (WPT adds noise the small CNN path cannot reweight), while the mel + 24 feat variant adds 0.03 accuracy but loses 0.14 macro-F1 (the handcrafted features bias predictions toward

the majority classes). Second, on Fold 3 the three representations appear complementary: combining all three lifts accuracy by 0.10 and macro-F1 by 0.07 over the mel-only arm. Third, however, this Fold 3 advantage did not survive cross-validation. Under three-fold CV (Table 5), the full-fusion variant dropped to 0.628 ± 0.112 accuracy and 0.531 ± 0.183 macro-F1, well below the proposed mel-only CNN's 0.732 ± 0.041/0.724 ± 0.034, and the piercing trial was missed entirely on Fold 2 (per-fold accuracies 0.541/0.557/0.787; per-fold macro-F1s 0.495/0.328/0.771; per-fold piercing recalls 1.0/0.0/1.0). The favourable Fold 3 result is therefore an outlier rather than a generalisable improvement: under three-fold CV the mel-only branch beats the full fusion in two of three folds (Folds 1 and 2 by 0.246 and 0.132 accuracy; Fold 3 by -0.066), the standardised mean difference is large (Cohen's $d = 1.23$ accuracy, $d = 1.47$ macro-F1; both above the conventional $d = 0.8$ threshold), and the fusion variant has 2.7× larger fold-to-fold standard deviation (7.5× larger variance), reflecting the additional capacity (33 % more parameters) overfitting to fold-specific patterns. This justifies the mel-only branch as the proposed model on performance and parsimony, and limits the multi-representation framing of Shevchik et al. [22], Chianese et al. [14], and Darwish et al. [45] to a per-fold observation rather than a cross-validated finding on the present 122-sample training set.

Table 8. Mel-only vs. multi-representation fusion.

Model/Fusion Input	Accuracy	Macro F1	Pierc. Recall	Params
Head-to-head, 3-fold replicate-out cross-validation:				
Mel only (proposed, single Linear(256 → 5) head)	0.732 ± 0.041	0.724 ± 0.034	1.000 ± 0.000	199,237
Mel + WPT + 24-feat (full fusion, 2-layer MLP head)	0.628 ± 0.112	0.531 ± 0.183	0.667 ± 0.471	265,157
Mel only (z, 256-D)	0.689	0.704	1.000	248,261
Mel + WPT (z + 64-D wavelet packet)	0.541	0.611	1.000	260,549
Mel + 24-feat (z + 24-D physics features)	0.721	0.562	1.000	252,869
Mel + WPT + 24-feat (full fusion)	0.787	0.771	1.000	265,157

Reducing the Transformer back end from four layers/eight heads/$d_{\text{model}} = 128$ to one layer/four heads/$d_{\text{model}} = 64$ raised the three-fold-CV mean accuracy from 0.514 ± 0.028 to 0.541 ± 0.013 but did not close the gap to the proposed CNN's 0.732 ± 0.041, and the macro-F1 stayed around 0.28–0.33 for every Transformer variant compared with the CNN's 0.724 ± 0.034. The 30-configuration hyperparameter grid (Table 6) tightens this conclusion: the best-tuned Tr-small2 macro-F1 (0.441 ± 0.057) is still 0.283 below the proposed CNN, and the scaled-up Tr-original (4L/8H/$d = 128$) at its own best grid cell achieves only 0.304 ± 0.016 macro-F1, confirming that the Transformer underperformance is not attributable to insufficient learning-rate, weight-decay, or dropout tuning. The ACGAN-style class-balanced oversampling, evaluated on Fold 3 only, degraded performance for four of the five Transformer variants (accuracy ≤ 0.246) by collapsing the network to a single-class prediction; the sinusoidal-position variant was the exception, where ACGAN left performance essentially unchanged (0.426/0.279). This is consistent with Eren's [3] warning that synthetic augmentation in laser welding DL is poorly understood. On the 122-trial training set, the augmentation introduces spurious patterns and larger Transformer overfits.

Table 9 reports the frequency-axis ablation. The three filterbanks achieve 0.724–0.746 mean macro-F1 with a piercing recall of 1.000 in every fold; the largest pairwise gap is 0.022 macro-F1, lower than the fold-to-fold standard deviation of either variant. The mel and log filterbanks gave nearly identical means (within 0.010 macro-F1), as expected since both are log-frequency parametrisations above 1 kHz, while the linear filterbank achieved the highest mean macro-F1 (0.746) and the highest fold variance (0.056). The mel filterbank achieved the lowest fold variance of the three (0.034 macro-F1 vs. 0.050–0.056), and the

auditory origin of the mel scale was not a liability for the ultrasonic process band on this dataset.

Figure 6 shows the training loss and held-out test accuracy of the proposed 1-D CNN per epoch. The loss decayed smoothly; the test accuracy stabilised at 0.70–0.75 by epoch 30 with no late-stage overfitting. The same training run was used for the Grad-CAM analysis (Section 4.6) and the McNemar test below; its final test accuracy of 0.721 matches the Fold 3 result reported in Table 5.

Table 9. Frequency-axis ablation: The proposed 1-D CNN trained on the same Hann-windowed STFT power passed through three 128-band filterbanks under 3-fold replicate-out cross-validation. Values are reported as mean ± standard deviation across folds.

Filterbank	Bands/Range (kHz)	Accuracy	Macro F1	Pierc. Recall
Mel (proposed)	128/0–125	0.732 ± 0.041	0.724 ± 0.034	1.000 ± 0.000
Linear (uniform Hz)	128/0–125	0.743 ± 0.066	0.746 ± 0.056	1.000 ± 0.000
Log (log-Hz)	128/1–125	0.754 ± 0.035	0.734 ± 0.050	1.000 ± 0.000

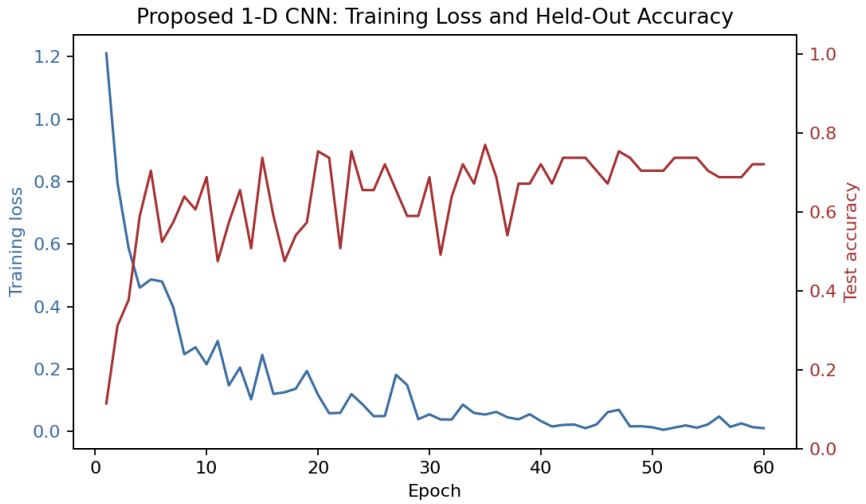

Figure 6. Training loss (left axis) and held-out test accuracy (right axis) of the proposed 1-D CNN over 60 epochs (AdamW, cosine annealing, class-weighted cross-entropy).

4.5. Statistical-Significance Test

McNemar paired-significance tests were performed between the proposed 1-D CNN and (i) the dual-threshold baseline of Basile et al. [31] extended to the five-class taxonomy and (ii) the strongest classical-ML baseline (SFFS-SVM on the 24-feature library). With n_{01} trials on which the proposed model is correct and the baseline is wrong, and n_{10} trials on which the baseline is correct and the proposed model is wrong, the exact two-sided binomial test on $\min(n_{01}, n_{10}) \sim \text{Binomial}(n_{01} + n_{10}, 0.5)$ produces the p-values reported in Table 10. The improvement over the threshold rule is highly significant ($p < 10^{-4}$), confirming that the proposed CNN is not simply replicating the band-stop sound-power thresholding. The proposed CNN and SFFS-SVM are not significantly different on the present 61-trial test set ($n_{01} = 6$, $n_{10} = 8$, $p = 0.79$): SFFS-SVM has marginally more correct trials (0.754 vs. 0.721 accuracy), while the proposed CNN achieves a marginally higher macro-F1 (0.707 vs. 0.682) by resolving the rare *Marg.* and *Pierc.* classes more uniformly. A larger test set would be required to disentangle these two effects; this is recommended for future work.

Table 10. McNemar paired-significance test of the proposed 1-D CNN against the threshold-rule baseline and the strongest classical-ML baseline. n_{01} = CNN correct, baseline wrong; n_{10} = baseline correct, CNN wrong. Test-set size: 61 trials.

Pair (Proposed CNN vs.)	n_{01}	n_{10}	Two-Sided p	Verdict
Threshold rule (Basile 2025)	30	4	6.2×10^{-6}	significant ($\alpha = 0.05$)
SFFS-SVM (24-feat library)	6	8	0.791	not significant ($\alpha = 0.05$)

4.6. Interpretability via Grad-CAM

Grad-CAM activation maps over the mel-spectrogram input were computed for one representative held-out trial per class on the proposed 1-D CNN. Based on the existing literature, the expected behaviour was that the network would attend to the 8–26 kHz band for the lack-of-fusion/lack-of-connection transition [31], to the 63 kHz keyhole-resonance peak for the sound-connection class [31,32], and to broadband attenuation for piercing [31,33]. Figure 7 summarises this qualitative class-to-physics correspondence. Figure 8 shows the Grad-CAM attribution maps obtained from the trained network. The CAM correctly localises the 60–110 kHz band for the sound and piercing classes (consistent with stable-keyhole resonance and broadband vapour-venting attenuation, respectively), the mid-frequency 8–26 kHz band for LoC and marginal (consistent with unstable melt-pool oscillation), and the very-low-frequency band for LoF (consistent with the absence of keyhole formation).

These observations confirm the keyhole-resonance physics reported in [31–33] without the network having been given any explicit prior knowledge of these bands.

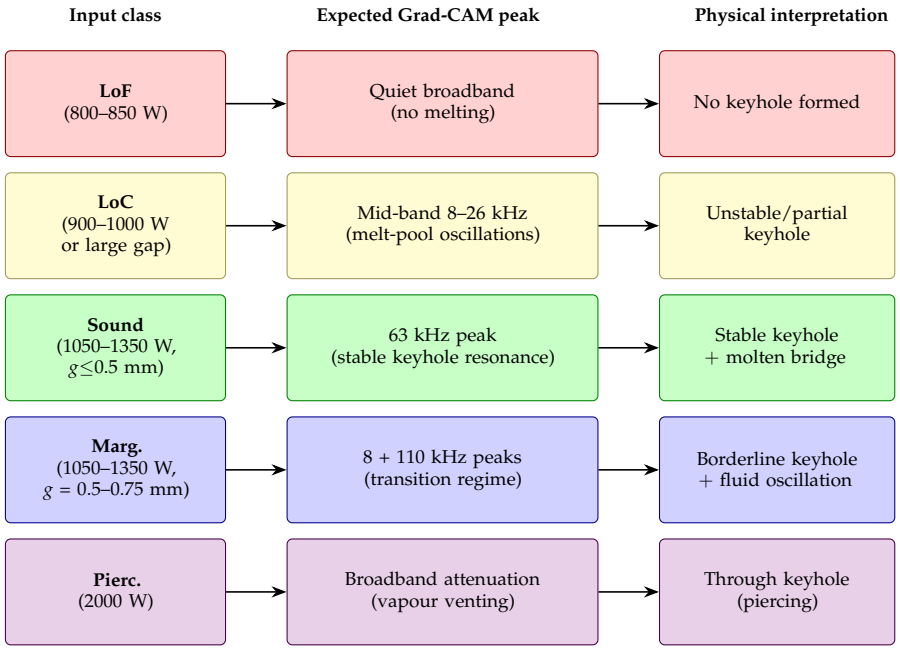

Figure 7. Result synthesis: From input class to expected Grad-CAM spectro-temporal peak to physical interpretation, used as a qualitative cross-check against the empirical Grad-CAM maps.

4.7. Correlation with Process Parameters

Pearson correlation coefficients between the predicted regime score and the two process inputs P_L and g were computed and compared with the values reported by Basile et al. (r = 0.95 between P_L and the total sound power; $r = -0.797$ between g and the band-stop-filtered sound power) [31]. The agreement between the learned regime score and the process inputs provides an external check that the network captured the underlying physical trends rather than trial-specific artefacts. This sanity check follows the Pearson-

coefficient reporting of Sun et al. [10] and Oussaïd et al. [13]. The Pearson correlations between the proposed-model regime score and the two process inputs are $r(\hat{y}, P_L) = 0.751$ and $r(\hat{y}, g) = -0.229$, computed on the 61-trial held-out test set (rep-3-held-out fold). The regime score $\hat{y}$ is the softmax-weighted expected class index $\hat{y}_i = \sum_{c=0}^{4} c\, p_{i,c}$, a continuous value in $[0, 4]$ rather than the discrete argmax prediction.

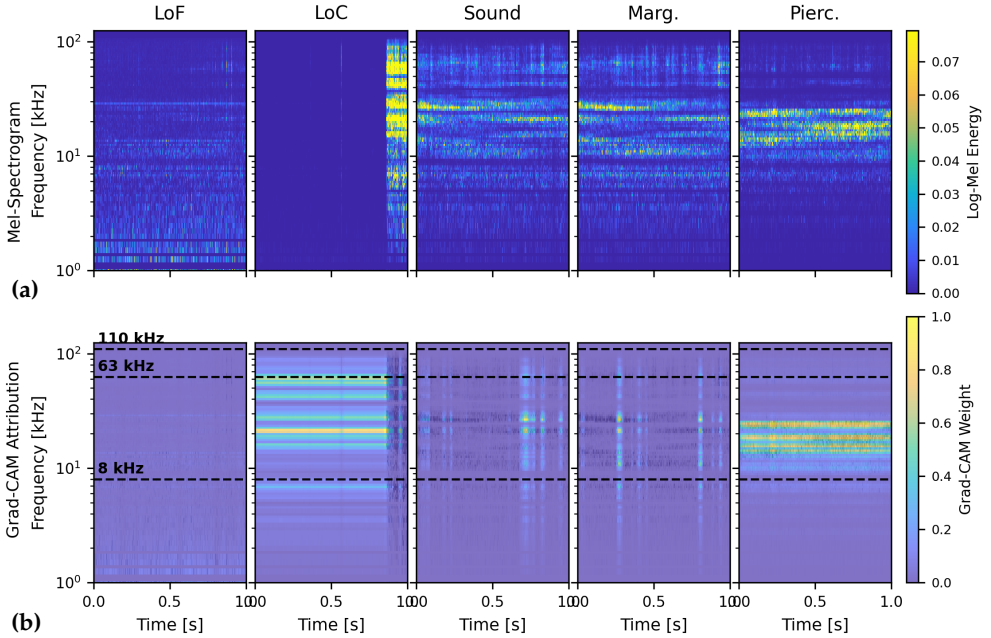

Figure 8. Grad-CAM analysis of the proposed 1-D CNN, one held-out trial per class. (**a**) Log-mel spectrograms and (**b**) Grad-CAM attribution heat-map.

5. Discussion

On the held-out replicate (61 trials, Fold 3), the proposed 1-D CNN achieved 0.721 accuracy and 0.707 macro-F1 with 100% recall on the rare *Pierc.* class (one piercing trial per fold; three piercing trials in total, distributed two train/one test). This is an absolute improvement of +0.43 accuracy (+0.62 macro-F1) over the dual-threshold rule of Basile et al. [31] extended to the five-class taxonomy (0.295/0.092, piercing recall 0). The strongest classical-ML baseline, SFFS-SVM on the 24-feature library, achieved 0.754 accuracy and 0.682 macro-F1; the proposed CNN had slightly lower accuracy (−0.033) but higher macro-F1 (+0.025) because it predicted the rare *Marg.* and *Pierc.* classes more uniformly, and the McNemar test indicates the two are statistically indistinguishable on the 61-trial test set ($p = 0.79$, Table 10). Photodiode-based four-class pipelines such as Caprio et al. [7] (KNN, 100%) and Chianese et al. [14] (DWT+NN, 97.5%) use simpler four-class taxonomies and different alloys (Cu/Ni-steel vs. AA1050) and are not directly comparable; the present results indicate that a single optical-microphone signal supports a five-class classifier of similar quality without the multi-sensor photodiode stacks of [8,15,16,19,20]. The acoustic framework benefits from a much higher sensor bandwidth than the photodiode pipeline of [37] (DC–MHz vs. 50 kHz), capturing the 63 kHz keyhole-resonance peak and broadband piercing attenuation; the Grad-CAM analysis confirmed these are the network's primary discriminators (Figure 8).

On a per-class accuracy basis, the proposed 1-D CNN performs below the vision-based DL state of the art: Knaak et al. [21] report F1 = 0.952 on six classes from dual-IR sequences, Balakrishnan et al. [25] reach 99.55% binary accuracy on radiographs with a Swin–DeiT transformer, and Wang et al. [8] achieve a high F1 on Al-busbar visual data. However, the proposed model operates on a one-dimensional input of 2.5×10^5 samples per trial

after decimation (2.5×10^6 at full bandwidth), one to two orders of magnitude smaller than typical 2-D image inputs. CPU inference latency was 4.1 ms per trial, compared with 1.1 ms on a Jetson AGX Xavier for the CNN–GRU of Knaak et al. [21] on IR images and 20.4 ms for the ST-TCN of Liu et al. [27] on coaxial-vision penetration estimation. The framework is therefore suitable for embedded deployment when camera placement is impractical (closed enclosures, beam-wobble interference, shielding-gas blow-back), such as busbar laser welding [9,10].

A central finding is that under three-fold replicate-out CV on the 122-sample training set, every Transformer-hybrid arm stayed well below the proposed CNN. The original 4L/8H/$d_{\text{model}} = 128$ hybrid scored $0.514 \pm 0.028/0.333 \pm 0.036$, and the four light variants clustered between 0.437 and 0.541 accuracy (small1 0.541 ± 0.013, small2 0.514 ± 0.039, small4 0.519 ± 0.020, sinpos2 0.437 ± 0.015); none closed the gap to the proposed CNN's $0.732 \pm 0.041/0.724 \pm 0.034$. The 27-cell hyperparameter grid on Tr-small2 plus the three-cell validation on Tr-original (Table 6) confirms this conclusion is not a tuning artefact: the best-tuned configurations reach macro-F1 $= 0.441 \pm 0.057$ on Tr-small2 and 0.304 ± 0.016 on Tr-original, both well below the proposed CNN's 0.724 ± 0.034. The 90-fold-run search is consistent with the small-data attention findings of recent reviews [1–3] and indicates that the gap reflects an inductive-bias mismatch rather than a missing hyperparameter. ACGAN-style oversampling, evaluated on Fold 3 only, degraded four of the five Transformer arms to accuracy ≤ 0.246 via single-class collapse; only the sinusoidal-position variant was unchanged (0.426/0.279). This contradicts the favourable Transformer-hybrid results of Solovev et al. [28] (mAP 0.85, four classes, IR thermography), Liu et al. [27] (98.96%, three classes, coaxial vision), and Shi et al. [30] (99.39%, ten classes, multi-signal). Two factors explain the difference: those datasets are an order of magnitude larger, consistent with the benchmark-data-scarcity gap noted in [1–3]; and the GAN-style augmentation introduces spurious patterns that the larger attention layers overfit, consistent with Eren's [3] caution that the benefits of synthetic data in laser welding DL are poorly understood. The local-pattern inductive bias of the compact CNN, by contrast, matches the spectro-temporal structure of keyhole-resonance acoustic signatures (Figure 8) and avoids the data-hungry global-attention regime.

The Grad-CAM analysis (Section 4.6) recovered the 8, 63 and 110 kHz keyhole-resonance bands without these being prescribed in the architecture (Figure 8), providing post hoc evidence that the network learned a spectral structure consistent with the keyhole-resonance physics of [31], the capillary-wave melt-pool dynamics of [32], and the laser-ultrasonic attenuation regimes of [33]. The Pearson correlations $r(\hat{y}, P_L) = 0.751$ and $r(\hat{y}, g) = -0.229$ on the 61-trial held-out test set match the sign and order of magnitude of the $r = 0.95$ (P_L vs. s_p) and $r = -0.797$ (g vs. $s_{p,\text{filt}}$) reported by Basile et al. [31], confirming that the network captured physical trends rather than trial-specific artefacts. These observations address the interpretability and physics-grounding gap noted in laser welding DL reviews [1–4].

For battery-pack busbar manufacturing, the framework supports closed-loop laser-power control similar to [13,29,55] and complements pre-process seam-tracking pipelines such as [12]. The 8.1 ms end-to-end latency (4 ms STFT plus 4.1 ms encoder) gives a $\sim$120 Hz update rate, well below the 50–100 ms control-loop budget of the Industry 4.0/5.0 context reviewed in [2,67] but an order of magnitude below the 1 kHz beam-wobble cycle of [10,31], so the framework is suitable for outer-loop quality classification rather than inner-loop wobble control. The framework demonstrates the data-driven, AI-assisted approach to laser-driven manufacturing reviewed in [1–6]: a single high-bandwidth optical-microphone channel replaces multi-sensor stacks, an end-to-end neural classifier replaces hand-tuned thresholds, and a Grad-CAM layer replaces black-box predictions, all at mil-

lisecond latency. The pipeline (single-sensor acquisition → multi-representation front end → compact CNN encoder → Grad-CAM) is transferable in principle to laser cutting, marking, and surface-modification monitoring across metals, polymers, ceramics, and composites; further validation is future work.

Several limitations remain. First, the dataset is restricted to 1.0 mm AA1050 and a single beam-wobble configuration; transferability to AA6063, copper, and copper-to-steel busbars [7,10,14] needs empirical testing through fine-tuning of the trained network on small target-domain sets (20–30 labelled trials per new alloy or thickness), following transfer-learning practice common in welding monitoring. Second, acquisition took place in a controlled laboratory without cross-jet flow or factory-floor noise; production deployment will require noise-robust training via three complementary tracks—(i) noise-injection augmentation using public industrial-acoustic recordings mixed into the training trials at target signal-to-noise ratios; (ii) SpecAugment-style time and frequency masking during training; and (iii) unsupervised domain adaptation (DANN, CORAL) once a small amount of unlabelled factory-floor data is available—along the lines reviewed in [3,4]. Third, a possible label uncertainty remains near weld-regime boundaries. Although the five class labels were adopted from Basile et al. [31] and supported by metallographic inspection and piercing verification, process variability may still affect welds produced under the same nominal conditions. This is especially relevant for the transition between lack of connection, marginal connections, and sound connections, and may explain some confusion between the marginal and sound classes. Future studies should use trial-specific metallographic, radiographic, or CT-based labels [5,6,10,45] to reduce this uncertainty. Fourth, the log-mel front end is a fixed log-frequency compression rather than a physically derived ultrasonic representation, allocating roughly one-sixth of its bands to the unused sub 1 kHz range; a head-to-head three-fold comparison against linear-Hz and log-Hz filterbanks (Table 9) placed all three within 0.022 macro-F1, with the mel scale at the lowest fold variance; a learnable filterbank or a constant-Q transform calibrated to the 8/63/110 kHz keyhole peaks may still extract more signal from the ultrasonic band and is a target for future work.

6. Conclusions

This paper applied an Acoustic Signal Intelligence framework to in-process classification of laser beam welding states from a single optical-microphone signal, developed on an open dataset (183 weld trials of 1.0 mm AA1050 overlap joints, Xarion Eta250 ultra optical microphone, $f_s = 2.5$ MHz) under a five-class taxonomy: lack of fusion, lack of connection, sound, marginal, and piercing.

The contributions are as follows: (i) A compact 1-D CNN encoder on a 128-band log-mel STFT spectrogram (199 237 parameters; three convolutional blocks of 64/128/256 channels and a single Linear($256 \rightarrow 5$) head) trained with class-balanced focal cross-entropy, which under three-fold replicate-out cross-validation achieved 0.732 ± 0.041 accuracy and 0.724 ± 0.034 macro-F1 with a piercing recall of 1.00 in every fold (one piercing trial per fold), the highest macro-F1 across all seventeen models tested (six classical-ML and four deep learning baselines including a five-variant 1-D CNN–Transformer hybrid); a multi-representation fusion variant adding a 64-element Daubechies db-10 wavelet-packet decomposition and a 24-element handcrafted feature library targeting the 8, 63 and 110 kHz keyhole-resonance peaks was evaluated as an ablation arm and not retained in the final model. (ii) A demonstration that Transformer-hybrid models and ACGAN-style oversampling underperformed compared to the proposed network on the 122-trial training set: every Transformer variant fell at least 0.19 accuracy and 0.39 macro-F1 below the CNN, and ACGAN collapsed four of the five attention variants to single-class predictions (accuracy ≤ 0.246). This conclusion is further supported by a 30-configuration

grid search over learning rate, weight decay, and dropout (90 fold-runs in total) under which the best-tuned Transformer-hybrid macro-F1 (0.441 ± 0.057) remains 0.283 below the proposed CNN. (iii) A Grad-CAM analysis that recovered the keyhole-resonance bands (60–110 kHz for *Sound/Pierc.*, 8–26 kHz for *LoC/Marg.*, low-frequency for *LoF*) without explicit prior knowledge.

Three further points follow. First, the dual-threshold rule extended to the five-class taxonomy achieved only 0.295 accuracy/0.092 macro-F1 with zero piercing recall; the proposed network improved on this by $+0.43$ accuracy and $+0.62$ macro-F1 on the held-out replicate (McNemar $p = 6.2 \times 10^{-6}$, $n_{01} = 30$, $n_{10} = 4$). The strongest classical baseline, SFFS-SVM on the 24-feature library, achieved slightly higher accuracy (0.760 ± 0.008) but lower macro-F1 (0.690 ± 0.007); the McNemar test ($p = 0.791$, $n_{01} = 6$, $n_{10} = 8$) indicates the two are statistically indistinguishable on the 61-trial test set. Second, the complementarity of the three signal representations did not survive cross-validation: the full mel + WPT + 24-feature fusion model lifted Fold 3 accuracy by $+0.10$ and macro-F1 by $+0.07$ over the mel-only variant, but under three-fold CV it dropped to $0.628 \pm 0.112/0.531 \pm 0.183$, below the mel-only $0.732 \pm 0.041/0.724 \pm 0.034$ (Table 8); this shows it was a single-fold outlier rather than a real lift, justifying the mel-only branch as the proposed model. Third, the Pearson correlations $r(\hat{y}, P_L) = 0.751$ and $r(\hat{y}, g) = -0.229$ confirmed that the network captured physical trends rather than trial-specific artefacts. CPU inference latency was 4.1 ms per trial on a single x86_64 core; with the 4 ms STFT window, the $\sim$120 Hz update rate fits the 50–100 ms outer-loop budget but lies an order of magnitude below the 1 kHz beam-wobble cycle, so the framework is suitable for outer-loop quality monitoring rather than inner-loop wobble control.

Future work will extend the framework along four directions: (a) validation on AA6063 and copper-to-steel busbars for cross-alloy transferability; (b) integration of cross-section-level porosity and crack labels alongside the parameter-derived taxonomy; (c) edge-hardware deployment for closed-loop laser-power control with a 1 kHz wobble cycle; and (d) self-supervised pre-training and noise-robust domain adaptation for production environments.

Funding: This research received no external funding.

Data Availability Statement: The data utilised in this study were obtained from the work of Basile et al. [31] and are freely available at https://zenodo.org/records/15833960 (accessed on 3 May 2026).

Acknowledgments: The author gratefully acknowledges the Laser Beam Welding Group at WMG, University of Warwick, for the open release of the optical-microphone laser welding dataset used in this study. The author would like to acknowledge the use of ChatGPT (version GPT-5.3, OpenAI, San Francisco, CA, USA) to enhance the clarity and fluency of the manuscript. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACGAN	Auxiliary-Classifier Generative Adversarial Network
AE	Acoustic Emission
ASI	Acoustic Signal Intelligence (the proposed framework)
BEV	Battery Electric Vehicle
CNN	Convolutional Neural Network

CORAL	Correlation Alignment
DANN	Domain-Adversarial Neural Network
DL	Deep Learning
DOE	Design of Experiments
DWT	Discrete Wavelet Transform
FFN	Feed-Forward Network
GAF	Gramian Angular Field
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
LBW	Laser Beam Welding
LoC	Lack of Connection
LoF	Lack of Fusion
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
PSD	Power Spectral Density
RF	Random Forest
RLW	Remote Laser Welding
ROI	Region of Interest
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
WPT	Wavelet-Packet Transform

References

1. Voets, J.; Tercan, H.; Meisen, T.; Esen, C. A Systematic Review and Taxonomy of Machine Learning Methods for Process Optimization and Control in Laser Welding. *Appl. Sci.* **2026**, *16*, 1568. [CrossRef]
2. Gorine, M.E.A.; Nechak, L.; Ichchou, M.; Pichenot, Y. A review of AI and machine learning applications in manufacturing processes: M. El Amine Gorine, L. Nechak, M. Ichchou, Y. Pichenot. *J. Intell. Manuf.* **2026**, 1–89. [CrossRef]
3. Eren, B. Deep learning approaches for weld defect Detection: A comprehensive review of Models, Applications, and future directions. *Comput. Ind. Eng.* **2025**, *212*, 111725. [CrossRef]
4. Zhang, M.; Feng, M.; Chen, C.; Yu, X.; Lian, G. Weld defect detection: Deep learning-based image processing and the mechanisms of defect formation. *Arch. Comput. Methods Eng.* **2025**, *33*, 3525–3563. [CrossRef]
5. Klimpel, A. Review and analysis of modern laser beam welding processes. *Materials* **2024**, *17*, 4657. [CrossRef] [PubMed]
6. Lindsay, E.E.; Botes, A.; Bernard, D.; Gambu, Z. Comprehensive strategies for defect mitigation and process optimisation in laser beam welding of aluminium alloys: A systematic review. *Int. J. Adv. Manuf. Technol.* **2026**, *143*, 3479–3513. [CrossRef]
7. Caprio, L.; Previtali, B.; Demir, A.G. Sensor selection and defect classification via machine learning during the laser welding of busbar connections for high-performance battery pack production. *Lasers Manuf. Mater. Process.* **2024**, *11*, 329–352. [CrossRef]
8. Wang, D.; Zheng, Y.; Dai, W.; Tang, D.; Peng, Y. Deep network-assisted quality inspection of laser welding on power Battery. *Sensors* **2023**, *23*, 8894. [CrossRef]
9. Prijanovič, U.; Prijanovič Tonkovič, M.; Trdan, U.; Pleterski, M.; Jezeršek, M.; Klobčar, D. Remote fibre laser welding of advanced high strength martensitic steel. *Metals* **2020**, *10*, 533. [CrossRef]
10. Sun, T.; Franciosa, P.; Liu, C.; Pierro, F.; Ceglarek, D. Effect of micro solidification crack on mechanical performance of remote laser welded AA6063-T6 fillet lap joint in automotive battery tray construction. *Appl. Sci.* **2021**, *11*, 4522. [CrossRef]
11. Um, J.; Stroud, I.A.; Park, Y.-k. Deep learning approach of energy estimation model of remote laser welding. *Energies* **2019**, *12*, 1799. [CrossRef]
12. Dmitry, R.; Alexander, L.; Valery, P. Development of mechanisms for automatic correction of industrial complex tools in the preprocessing of laser welding for small-scale and piece production using computer vision. *Machines* **2020**, *8*, 86. [CrossRef]
13. Oussaid, K.; Omidi, N.; El Ouafi, A.; Barka, N. Prediction of weld geometry in laser overlap welding of low-carbon galvanized steel. *Metals* **2025**, *15*, 447. [CrossRef]
14. Chianese, G.; Franciosa, P.; Sun, T.; Ceglarek, D.; Patalano, S. Using photodiodes and supervised machine learning for automatic classification of weld defects in laser welding of thin foils copper-to-steel battery tabs. *J. Laser Appl.* **2022**, *34*, 042040. [CrossRef]
15. You, D.; Gao, X.; Katayama, S. Data-driven based analyzing and modeling of MIMO laser welding process by integration of six advanced sensors. *Int. J. Adv. Manuf. Technol.* **2016**, *82*, 1127–1139. [CrossRef]

16. Liu, G.; Gao, X.; You, D.; Zhang, N. Prediction of high power laser welding status based on PCA and SVM classification of multiple sensors. *J. Intell. Manuf.* **2019**, *30*, 821–832. [CrossRef]
17. Chen, J.; Wang, T.; Gao, X.; Wei, L. Real-time monitoring of high-power disk laser welding based on support vector machine. *Comput. Ind.* **2018**, *94*, 75–81. [CrossRef]
18. Wang, T.; Chen, J.; Gao, X.; Qin, Y. Real-time monitoring for disk laser welding based on feature selection and SVM. *Appl. Sci.* **2017**, *7*, 884. [CrossRef]
19. Fan, K.; Peng, P.; Zhou, H.; Wang, L.; Guo, Z. Real-time high-performance laser welding defect detection by combining ACGAN-based data enhancement and multi-model fusion. *Sensors* **2021**, *21*, 7304. [CrossRef]
20. Zhang, Y.; You, D.; Gao, X.; Wang, C.; Li, Y.; Gao, P.P. Real-time monitoring of high-power disk laser welding statuses based on deep learning framework. *J. Intell. Manuf.* **2020**, *31*, 799–814. [CrossRef]
21. Knaak, C.; von Eßen, J.; Kröger, M.; Schulze, F.; Abels, P.; Gillner, A. A spatio-temporal ensemble deep learning architecture for real-time defect detection during laser welding on low power embedded computing boards. *Sensors* **2021**, *21*, 4205. [CrossRef]
22. Shevchik, S.; Le-Quang, T.; Meylan, B.; Farahani, F.V.; Olbinado, M.P.; Rack, A.; Masinelli, G.; Leinenbach, C.; Wasmer, K. Supervised deep learning for real-time quality monitoring of laser welding with X-ray radiographic guidance. *Sci. Rep.* **2020**, *10*, 3389. [CrossRef]
23. Buongiorno, D.; Prunella, M.; Grossi, S.; Hussain, S.M.; Rennola, A.; Longo, N.; Di Stefano, G.; Bevilacqua, V.; Brunetti, A. Inline defective laser weld identification by processing thermal image sequences with machine and deep learning techniques. *Appl. Sci.* **2022**, *12*, 6455. [CrossRef]
24. Fan, D.; Yu, C.; Sha, L.; Zhang, H.; Liu, X. Failure Detection of Laser Welding Seam for Electric Automotive Brake Joints Based on Image Feature Extraction. *Machines* **2025**, *13*, 616. [CrossRef]
25. Balakrishnan, K.; Narmatha, C.; Rajendran, T.; Arif, M.; Sivaramkrishnan, M.; Kavitha, D. A hybrid Swin–DeiT transformer framework for automated welding defect detection in radiographic images. *Comput. Electr. Eng.* **2026**, *134*, 111135. [CrossRef]
26. Liu, Y.; Yuan, K.; Li, T.; Li, S.; Ren, Y. NDT method for line laser welding based on deep learning and one-dimensional time-series data. *Appl. Sci.* **2022**, *12*, 7837. [CrossRef]
27. Liu, Z.; Ji, S.; Ma, C.; Zhang, C.; Yu, H.; Yin, Y. Penetration state recognition during laser welding process control based on Two-Stage Temporal convolutional networks. *Materials* **2024**, *17*, 4441. [CrossRef]
28. Solovev, G.; Klokov, E.; Krasnov, D.; Sokolov, M. Advancing Defect Detection in Laser Welding: A Machine Learning Approach Based on Spatter Feature Analysis. *Sensors* **2026**, *26*, 1825. [CrossRef] [PubMed]
29. Weisbrod, N.; Metternich, J. Application of a concept for ML-driven closed-loop quality control in laser beam welding. *Procedia CIRP* **2024**, *126*, 739–744. [CrossRef]
30. Shi, S.; Caluyo, F.; Hernandez, R.; Sarmiento, J.; Rosales, C.A. The utilization of deep learning technology in the detection of laser welding defects. In Proceedings of the 2024 International Conference on Advanced Control Systems and Automation Technologies (ACSAT); IEEE: New York, NY, USA, 2024; pp. 178–181. [CrossRef]
31. Basile, D.; Al Botros, R.; De Maddis, M.; Razza, V.; Franciosa, P. Monitoring part-to-part gap and laser power effects in remote laser welding of 1050 aluminum busbar-to-terminal connections via optical microphone sensing. *Opt. Laser Technol.* **2025**, *192*, 113494. [CrossRef]
32. Volpp, J. Surface tension estimation of steel above boiling temperature. *Appl. Sci.* **2024**, *14*, 3778. [CrossRef]
33. Nomura, K.; Ishifuro, S.; Asai, S. Study on Non-Contact Defect Detection Using the Laser Ultrasonic Method for Friction Stir-Welded Cu–Al Dissimilar Material Joints. *Appl. Sci.* **2026**, *16*, 688. [CrossRef]
34. Sumesh, A.; Rameshkumar, K.; Mohandas, K.; Babu, R.S. Use of machine learning algorithms for weld quality monitoring using acoustic signature. *Procedia Comput. Sci.* **2015**, *50*, 316–322. [CrossRef]
35. Ozkat, E.C.; Abdioglu, M.; Ozturk, U.K. Machine learning driven optimization and parameter selection of multi-surface HTS Maglev. *Phys. C Supercond. Its Appl.* **2024**, *616*, 1354430. [CrossRef]
36. Yilmaz, A.; Ozkat, E.C.; Gul, F. Signal Intelligence: Vibration-Driven Deep Learning for Anomaly Detection of Rotary-Wing UAVs. *Drones* **2026**, *10*, 321. [CrossRef]
37. Ozkat, E.C. Photodiode signal patterns: Unsupervised learning for laser weld defect analysis. *Processes* **2025**, *13*, 121. [CrossRef]
38. Korkmaz Can, N.; Ozkat, E.C.; Ceryan, N.; Ceryan, S. Benchmarking ML Approaches for Earthquake-Induced Soil Liquefaction Classification. *Appl. Sci.* **2025**, *15*, 11512. [CrossRef]
39. Özkat, E.C. A method to classify steel plate faults based on ensemble learning. *J. Mater. Mechatron. A* **2022**, *3*, 240–256. [CrossRef]
40. Kumar, D.; Ganguly, S.; Acherjee, B.; Kuar, A.S. Performance evaluation of TWIST welding using machine learning assisted evolutionary algorithms. *Arab. J. Sci. Eng.* **2024**, *49*, 2411–2441. [CrossRef]
41. Habibkhah, F.; Moallem, M. Application of machine learning for seam profile identification in robotic welding. *Mach. Learn. Appl.* **2025**, *20*, 100633. [CrossRef]
42. Ozkat, E.C.; Franciosa, P.; Ceglarek, D. A framework for physics-driven in-process monitoring of penetration and interface width in laser overlap welding. *Procedia CIRP* **2017**, *60*, 44–49. [CrossRef]

43. Ozkat, E.C.; Franciosa, P.; Ceglarek, D. Laser dimpling process parameters selection and optimization using surrogate-driven process capability space. *Opt. Laser Technol.* **2017**, *93*, 149–164. [CrossRef]
44. Velazquez de la Hoz, J.L.; Cheng, K. Development of an intelligent quality management system for micro laser welding: An innovative framework and its implementation perspectives. *Machines* **2021**, *9*, 252. [CrossRef]
45. Darwish, A.; Persson, M.; Ericson, S.; Ghasemi, R.; Salomonsson, K. Weld Defect Detection in Laser Beam Welding Using Multispectral Emission Sensor Features and Machine Learning. *Sensors* **2025**, *25*, 5120. [CrossRef] [PubMed]
46. Cai, W.; Jiang, P.; Shu, L.; Geng, S.; Zhou, Q. Real-time laser keyhole welding penetration state monitoring based on adaptive fusion images using convolutional neural networks. *J. Intell. Manuf.* **2023**, *34*, 1259–1273. [CrossRef]
47. Asif, K.; Zhang, L.; Derrible, S.; Indacochea, J.E.; Ozevin, D.; Ziebart, B. Machine learning model to predict welding quality using air-coupled acoustic emission and weld inputs. *J. Intell. Manuf.* **2022**, *33*, 881–895. [CrossRef]
48. Ozturk, U.K.; Abdioglu, M.; Ozkat, E.C.; Mollahasanoglu, H. Extended 2-D magnetic field modeling of linear motor to investigate the magnetic force parameters of high-speed superconducting maglev. *IEEE Trans. Appl. Supercond.* **2023**, *33*, 1–8. [CrossRef]
49. 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. [CrossRef]
50. Kumar, P. Hybrid Multi-Scale CNN and Transformer Model for Motor Fault Detection. *Machines* **2026**, *14*, 113. [CrossRef]
51. Liu, C.; Zou, W.; Hu, Z.; Li, H.; Sui, X.; Ma, X.; Yang, F.; Guo, N. Bearing health state detection based on informer and CNN+ Swin transformer. *Machines* **2024**, *12*, 456. [CrossRef]
52. Pang, B.; Liang, J.; Liu, H.; Dong, J.; Xu, Z.; Zhao, X. Intelligent bearing fault diagnosis based on multivariate symmetrized dot pattern and LEG transformer. *Machines* **2022**, *10*, 550. [CrossRef]
53. Wu, Z.; He, L.; Wang, W.; Ju, Y.; Guo, Q. A fault prediction method for CNC machine tools based on SE-ResNet-transformer. *Machines* **2024**, *12*, 418. [CrossRef]
54. Li, B.; Zhang, Y.; Ren, R.; Liu, W.; Xu, G. Time-Frequency Conditional Enhanced Transformer-TimeGAN for Motor Fault Data Augmentation. *Machines* **2025**, *13*, 969. [CrossRef]
55. Božič, A.; Kos, M.; Jezeršek, M. Power control during remote laser welding using a convolutional neural network. *Sensors* **2020**, *20*, 6658. [CrossRef] [PubMed]
56. Kim, B.J.; Kim, Y.M.; Kim, C. Transfer learning-based multi-sensor approach for predicting keyhole depth in laser welding of 780DP Steel. *Materials* **2025**, *18*, 3961. [CrossRef]
57. Murua, O.; Arrizubieta, J.I.; Lamikiz, A.; Schneider, H.I. A case study of a laser beam welding model for the welding of inconel 718 sheets of a dissimilar thickness. *Metals* **2024**, *14*, 829. [CrossRef]
58. Yang, P.; Li, F.; Zhu, Z.; Chen, H. Comparison of Interfaces Between In Situ Laser Beam Deposition Forming and Electron Beam Welding for Thick-Walled Titanium Alloy Structures. *Micromachines* **2024**, *15*, 1383. [CrossRef]
59. Zhang, X.; Hu, X.; Li, H.; Zhang, Z.; Chen, H.; Sun, H. Research on predicting welding deformation in automated laser welding processes with an enhanced DEWOA-BP algorithm. *Machines* **2024**, *12*, 307. [CrossRef]
60. Laureto, J.J.; Dessiatoun, S.V.; Ohadi, M.M.; Pearce, J.M. Open Source Laser Polymer Welding System: Design and Characterization of Linear Low-Density Polyethylene Multilayer Welds. *Machines* **2016**, *4*, 14. [CrossRef]
61. Luo, Z.; Wu, D.; Zhang, P.; Ye, X.; Shi, H.; Cai, X.; Tian, Y. Laser welding penetration monitoring based on time-frequency characterization of acoustic emission and CNN-LSTM hybrid network. *Materials* **2023**, *16*, 1614. [CrossRef]
62. Zhang, Z.; Qin, R.; Li, G.; Du, Z.; Li, Z.; Lin, Y.; He, W. Deep learning-based monitoring of surface residual stress and efficient sensing of AE for laser shock peening. *J. Mater. Process. Technol.* **2022**, *303*, 117515. [CrossRef]
63. Zhang, H.; Lai, Y.H. A Lightweight Audio Spectrogram Transformer for Robust Pump Anomaly Detection. *Machines* **2026**, *14*, 114. [CrossRef]
64. Özkat, G.Y.; Aasim, M.; Bakhsh, A.; Ali, S.A.; Özcan, S. Machine learning models for optimization, validation, and prediction of light emitting diodes with kinetin based basal medium for in vitro regeneration of upland cotton (*Gossypium hirsutum* L.). *J. Cotton Res.* **2025**, *8*, 19. [CrossRef]
65. Altunkaya, A.N.; Ozkat, E.C.; Avci, M. Analytical-to-AI pipeline: Modeling and optimization of entropy generation in pulsating non-Newtonian heat flow. *Comput. Math. Appl.* **2026**, *205*, 195–211. [CrossRef]
66. Yalçin-Özkat, G. Computational studies with flavonoids and terpenoids as BRPF1 inhibitors: In silico biological activity prediction, molecular docking, molecular dynamics simulations, MM/PBSA calculations. *SAR QSAR Environ. Res.* **2022**, *33*, 533–550. [CrossRef]
67. Hasan, N.; Alkan, B. Gest-SAR: A Gesture-Controlled Spatial AR System for Interactive Manual Assembly Guidance with Real-Time Operational Feedback. *Machines* **2025**, *13*, 658. [CrossRef]