Journal of Manufacturing Processes 176 (2026) 179–192

Contents lists available at ScienceDirect

Journal of Manufacturing Processes

journal homepage: www.elsevier.com/locate/manpro

High-resolution CTWD classification in wire arc additive manufacturing using time–frequency acoustic representation and deep convolutional neural networks

Min Seok Jang [a,b], Kyung Soon Jung [c], Kyeung Jun Park [c], Seung Beom Noh [a], Young Hwan Cha [a], Hyub Lee [a], Yong Son [a], Seung-Jun Shin [b], Young Won Kim [a,*]

[a] Korea Additive Manufacturing Technology Innovation Center, Korea Institute of Industrial Technology, Republic of Korea
[b] School of Interdisciplinary Industrial Studies, Hanyang University, Republic of Korea
[c] AHHA Labs Inc., Seoul, Republic of Korea

ARTICLE INFO

Keywords:
WAAM
Contact Tip to Work Distance (CTWD)
Acoustic sensing
Time–frequency analysis
STFT spectrogram
Deep convolutional neural network (Deep CNN)
High-resolution classification
Process monitoring

ABSTRACT

This study proposes a data-driven time–frequency learning framework for classifying Contact Tip to Work Distance (CTWD) at a 1 mm resolution using arc acoustic signals generated in Wire Arc Additive Manufacturing (WAAM). In WAAM processes, CTWD variations directly influence arc length, heat input, and metal transfer behavior, thereby affecting deposition quality and process stability. However, precise in-situ measurement of CTWD remains challenging due to harsh process conditions, including intense arc luminosity, spatter, and high temperatures. To address this limitation, this study formulates the time–frequency structure of acoustic signals as a learnable representation and integrates STFT-based spectrograms with a 2D CNN to directly capture structural pattern differences induced by CTWD variations.

Under a single-session dataset (Set 1), the proposed approach achieved an Accuracy of 0.9955 and a Macro F1-score of 0.9952 for the 11-class (10–20 mm) classification task, demonstrating reliable discrimination at a 1 mm scale. However, when process variability was introduced through randomized CTWD sequences and deposition positions (combined dataset of Set 1–3), the Simple CNN trained under conditions identical to those of the proposed model reached an Accuracy of only 0.6476 and an F1-score of 0.6386, indicating limitations in model capacity under increased data complexity. To address this, a DeepCNN architecture based on stacked Conv_-Blocks composed of Conv2d, BatchNorm2d, and ReLU was developed, achieving an Accuracy of 0.9654 and a Macro F1-score of 0.9662 on the extended dataset. Under a leave-one-session-out evaluation on entirely unseen sessions, however, accuracy dropped to 0.128 ± 0.015, with only weak ordinal CTWD information transferring across sessions; session-robust fine-grained estimation is identified as an open challenge. Furthermore, when image-pretrained backbones (EfficientNetB2, ResNet18) were trained under matched conditions with model-specific learning-rate tuning, they reached accuracies of 0.9834 (ResNet18) and 0.9729 (EfficientNetB2) against 0.9654 for the proposed model, a difference that is statistically significant for ResNet18 (McNemar $p = 0.029$) but not for EfficientNetB2 ($p = 0.442$); the proposed architecture attains this level with roughly one third of the parameters of ResNet18 and without ImageNet pretraining.

Overall, this study quantitatively demonstrates that acoustic signals contain sufficient information for CTWD estimation and that learning time–frequency structures enables stable classification at a 1 mm resolution. These results suggest that acoustic sensing can serve as a low-cost, non-contact alternative to optical methods for improving monitoring resolution in WAAM processes, and further indicate its potential for real-time CTWD estimation and closed-loop process control.

* Corresponding author.
E-mail address: snails2@kitech.re.kr (Y.W. Kim).

https://doi.org/10.1016/j.jmapro.2026.09.024
Received 1 April 2026; Received in revised form 9 August 2026; Accepted 9 September 2026
Available online 14 September 2026
1526-6125/© 2026 The Authors. Published by Elsevier Ltd on behalf of The Society of Manufacturing Engineers. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Wire Arc Additive Manufacturing (WAAM) is a type of Directed Energy Deposition (DED) process in which metallic components are fabricated layer-by-layer using metal wire as feedstock and an electric arc as the heat source [1]. WAAM has attracted significant attention in industries such as aerospace, shipbuilding, and defense due to its ability to produce large-scale components with high deposition rates while substantially reducing material waste and production costs compared to conventional subtractive manufacturing [2–7].

To mitigate the intrinsic instability of the WAAM process, particular attention must be given to the Contact Tip to Work Distance (CTWD), defined as the distance between the contact tip and the workpiece. CTWD directly influences the electrical and thermal behavior of the arc through changes in free wire length and electrical resistance, thereby affecting heat input and molten metal transfer behavior [8]. It has also been reported as a key process parameter governing arc stability, metal transfer mode, bead geometry, and overall deposition quality, where excessive CTWD may lead to arc instability and droplet flow-down phenomena [9]. From a physical standpoint, variations in CTWD alter the resistive heating of the wire, resulting in fluctuations in welding current and heat input [10,11]. These variations can induce changes in arc length, irregular metal transfer cycles, and dynamic behavior of the molten pool, ultimately degrading process stability and leading to geometric inconsistency and reduced deposition quality.

In WAAM processes, geometric irregularities in the deposited bead can cause variations in CTWD, and such errors tend to accumulate in the vertical direction as the number of deposited layers increases. This accumulation of geometric deviation has been identified as a major factor contributing to unstable deposition during multi-layer fabrication [12]. When the torch-to-workpiece distance increases, turbulent mixing and diffusion of shielding gas promote atmospheric contamination, leading to defects such as porosity and reduced arc stability [13]. Conversely, excessively short CTWD can cause abrupt changes in short-circuit frequency and increase the risk of collision between the torch and the workpiece [14]. Therefore, maintaining and monitoring CTWD with high precision in real time is essential for ensuring stable deposition and high-quality builds.

Existing studies on CTWD in WAAM and DED processes have largely focused on indirect approaches, where CTWD variations are mitigated by monitoring bead geometry and adjusting process parameters accordingly. This approach arises from the practical difficulty of directly measuring CTWD during the process. The WAAM environment involves high-temperature arc plasma, intense light emission, spatter generation, and continuously evolving molten pool geometry, all of which degrade the reliability of optical or infrared (IR)-based sensing. In particular, issues such as image saturation due to arc brightness, limited field of view, sensor contamination, and thermal damage impose structural limitations on accurate in-situ measurement of CTWD. As a result, most prior studies have treated CTWD not as a directly measurable variable but as an indirectly controlled parameter inferred from observable outputs such as bead geometry. While effective in circumventing measurement challenges, this approach has limited capability in capturing the fine dynamic variations in arc behavior induced by CTWD changes.

For example, Chabot et al. proposed a geometric closed-loop control strategy in which layer height was measured using a laser profilometer, and process parameters were adjusted to suppress stick-out variation [15]. Similarly, Heralić et al. employed a camera and projected laser line to measure bead width and height, maintaining a consistent relative position between the wire tip and molten pool [16]. Xiong et al. also utilized a CCD camera positioned behind the torch to monitor bead width for process control in GMAW-based deposition [17]. These approaches primarily focus on stabilizing the process through geometric feedback rather than directly estimating CTWD. More recently, Baldauf et al. proposed a vision-based method to directly estimate CTWD by detecting the gas nozzle and wire position in welding images [18]. Although this approach explicitly treats CTWD as a measurable state variable, it still relies on expensive sensing systems and remains sensitive to arc brightness, field-of-view constraints, sensor placement, and environmental variations.

These optical and geometry-based monitoring approaches share common limitations, including high system complexity, cost, and reduced reliability under harsh process conditions. In large-scale WAAM applications or long-duration builds, issues such as limited visibility due to arc brightness and spatter, restricted sensor placement, and degradation of measurement stability further constrain their practical applicability. As an alternative, acoustic sensing has recently emerged as a promising non-contact and cost-effective monitoring technique [19]. Acoustic signals generated during welding are closely related to arc length, current waveform, metal transfer mode (e.g., short-circuit, globular, spray), and molten pool dynamics, all of which are also linked to defect formation [20–22], Since these phenomena are strongly influenced by CTWD, changes in CTWD induce corresponding variations in the frequency content and temporal patterns of acoustic signals. This makes acoustic sensing a viable candidate for real-time, non-contact monitoring of CTWD.

Indeed, in laser additive manufacturing and conventional arc welding processes, acoustic emission (AE) and microphone-based sensing have been widely used to detect porosity, cracking, droplet transfer modes, and molten pool behavior [23,24]. However, the application of acoustic-based monitoring in WAAM remains relatively limited and underexplored [21].

In this context, Chabot et al. proposed an acoustic-based CTWD monitoring framework for arc-based DED processes, where frequency-domain analysis of arc sound was used to identify spectral components associated with short-circuit transfer cycles [25]. Their model achieved an estimation accuracy of approximately ±0.5 mm, demonstrating that acoustic signals contain meaningful CTWD-related information. Nevertheless, their approach relied primarily on a limited set of frequency features, thereby underutilizing the rich time–frequency structure inherent in acoustic signals. While FFT-based analysis effectively captures dominant transfer frequencies, it averages out temporal information, making it difficult to distinguish transient or non-stationary patterns and subtle differences between adjacent CTWD conditions. The reported ±0.5 mm accuracy is sufficient for detecting general trends but insufficient for resolving fine-scale CTWD variations at the 1 mm level required for precise process control. In parallel, Hebert et al. demonstrated microphone-based classification of discrete CTWD levels in WAAM using random-forest models trained on hand-crafted acoustic features [26], further confirming the feasibility of airborne acoustic CTWD monitoring. Accordingly, the use of acoustic signals for CTWD monitoring per se is not claimed as a contribution of the present work. Rather, prior approaches addressed coarse discrete levels or ±0.5 mm-scale estimation based on low-dimensional hand-crafted features, leaving fine-grained discrimination at a 1 mm scale unaddressed.

In summary, previous acoustic-based CTWD studies have successfully established the physical correlation between acoustic signals and CTWD, demonstrating the feasibility of non-contact monitoring. However, they have largely relied on low-dimensional or statistically summarized features, which are insufficient to fully capture the complex and nonlinear time–frequency patterns associated with CTWD variations.

To address these limitations, this study introduces a time–frequency learning framework based on Short-Time Fourier Transform (STFT) representations of arc acoustic signals. By combining STFT spectrograms with machine learning models, the problem is formulated as a high-resolution classification task that directly identifies structural differences in acoustic patterns corresponding to CTWD variations. Unlike previous approaches that rely on averaged or dominant frequency features, the proposed method focuses on learning localized and non-stationary patterns in the time–frequency domain to distinguish CTWD differences at 1 mm intervals.

In WAAM processes, maintaining layer height within ±0.5–1.0 mm

is generally considered necessary for achieving geometric accuracy and process stability [14]. Since CTWD directly affects arc length, heat input, and metal transfer behavior, a 1 mm variation is not a negligible disturbance but a physically meaningful change that can impact process quality. Therefore, the 1 mm interval CTWD classification addressed in this study represents a high-resolution discrimination problem that evaluates whether adjacent CTWD conditions can be reliably distinguished based on acoustic patterns.

Through this approach, the present study aims to benchmark the proposed method against conventional frequency-based approaches and to systematically evaluate the feasibility and limitations of acoustic sensing for high-resolution CTWD monitoring. Ultimately, the key contribution of this work lies not merely in demonstrating the presence of CTWD information in acoustic signals, but in establishing a learning-based framework that captures the time–frequency structure of acoustic patterns and enables precise discrimination of fine CTWD variations from a process control perspective.

2. Methodology

2.1. WAAM system and acoustic sensing setup

The WAAM experiments in this study were conducted using a Gas Metal Arc Welding (GMAW)-based additive manufacturing system equipped with a Fronius TPS 500i power source operating in Cold Metal Transfer (CMT) mode. The CMT process is characterized by the active synchronization of current waveform control and wire feeding motion during the short-circuit phase, enabling stable and controlled metal transfer. This mechanism transforms metal transfer from a stochastic phenomenon into a controllable process, effectively decoupling heat input from deposition rate and significantly reducing spatter generation. As a result, sensitivity to layer geometry variation and arc length or CTWD fluctuation is reduced, allowing for more stable deposition behavior and improved repeatability in WAAM processes. For these reasons, the CMT mode was selected as the baseline process condition in this study [27,28].

To ensure process stability and reliable signal-based analysis, the selection of material with good weldability and low defect sensitivity is critical [28]. In this study, a low-carbon austenitic stainless steel wire, SUS316L, with a diameter of 1.2 mm was used as the feedstock. SUS316L is known for its high-temperature ductility and stable molten pool behavior, which enable consistent metal transfer and repeatable deposition geometry in GMAW- and CMT-based WAAM processes [29–31]. Furthermore, SUS316L is widely used as a benchmark material in WAAM research, facilitating comparison and interpretation of process behavior and signal characteristics. Based on these advantages, SUS316L was selected as the deposition material for systematically analyzing the effects of CTWD on process stability and acoustic signal characteristics.

The detailed WAAM process parameters used in this study are summarized in Table 1, and the chemical composition of the SUS316L wire is listed in Table 2. High-purity argon (Ar) shielding gas was used in all experiments to ensure arc stability and minimize oxidation and environmental disturbances.

Table 1
WAAM process parameters used in this study.

Process parameters	Value
Shielding gas	Ar
Wire feeding speed (mm·s^{-1})	133.33
Travel speed (mm·s^{-1})	6.67
Gas flow rate (L·min^{-1})	20
Welding mode	CMT (synergic control)
Wire diameter (mm)	1.2
CTWD (mm)	10–20 (1 mm interval)
Bead length (mm)	~230

Table 2
Nominal chemical composition of SUS316L stainless steel (wt%).

C	Si	Mn	Cr	Ni	Mo
0.02	0.4	1.73	18.15	11.17	2.22

To systematically investigate the effect of CTWD, the distance was varied from 10 mm to 20 mm in 1 mm increments, resulting in 11 discrete process conditions. The definition of CTWD in the WAAM process is illustrated in Fig. 1. CTWD was controlled using a robot-based torch positioning system, where a baseline CTWD of 15 mm was defined in the global coordinate system. Vertical offsets ranging from −5 mm to +5 mm were applied to achieve the full CTWD range. The horizontal torch position was kept constant to ensure precise and repeatable control of CTWD variation.

To isolate the effect of CTWD and eliminate the influence of cumulative geometric errors, all experiments were conducted using a single-bead deposition strategy. Each bead had a length of approximately 230 mm, with a constant travel speed of 6.67 mm/s and a wire feeding speed of 133.33 mm/s across all CTWD conditions. Welding current and voltage were automatically determined by the synergic control mode of the CMT power source, and no manual adjustments were made between experiments.

Acoustic signals generated during the welding process were collected using a low-cost commercial USB pin-type electret condenser microphone (TB329, Coms, Republic of Korea). The measured quantity is the airborne arc sound propagating through the atmosphere, which is distinct from structure-borne acoustic emission (AE) acquired by contact-type sensors; throughout this paper, the term acoustic signal refers to this airborne sound. According to the manufacturer-declared specifications, the microphone has an omnidirectional polar pattern, a frequency range of 20 Hz–16 kHz, a sensitivity of -34 ± 2 dB (0 dB = 1 V/Pa at 3 V), and a signal-to-noise ratio of $\geq$58 dB; the unit was not individually calibrated. Because the proposed method operates on per-segment standardized time–frequency representations rather than absolute sound pressure levels, calibrated sensitivity is not required for the analysis, and the 0–7 kHz analysis band employed in this study lies well within the specified frequency range of the sensor. The overall acoustic monitoring configuration and sensor placement are illustrated in Fig. 2. The microphone was rigidly mounted beneath the welding camera (WGC400, OTOS, Republic of Korea) and oriented along the welding direction to ensure stable acquisition of arc-generated sound. The microphone–camera assembly was rigidly attached to the torch-side structure; hence the microphone moved together with the torch, and the relative position between the torch and the microphone remained fixed within each experimental session. Because the sensing assembly travels with the torch while the arc remains anchored at the workpiece surface, varying the CTWD slightly changes the microphone-to-arc path length; the influence of this geometric factor on the measured signal is addressed in Section 3.2. The sensing assembly was dismounted and re-installed at the nominally identical mounting position between the Set 1

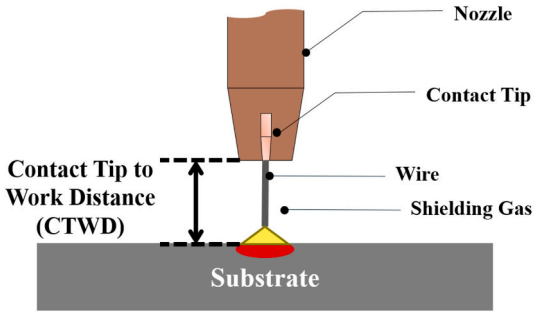

Fig. 1. Schematic illustration defining the Contact Tip to Work Distance (CTWD) in the WAAM process.

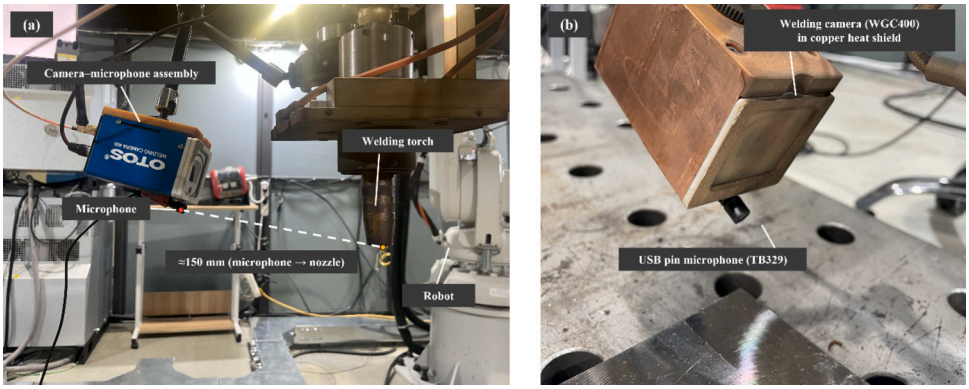

Fig. 2. Acoustic monitoring configuration: (a) overall arrangement of the welding torch, robot, and camera–microphone assembly, photographed before installation of the copper heat shield for visual clarity; the dashed line indicates the microphone-to-nozzle separation (≈150 mm; see Section 3.2); (b) close-up of the as-operated assembly, showing the USB pin-type microphone (TB329) clipped beneath the welding camera (WGC400) enclosed in the copper heat shield.

acquisition and the subsequent campaign in which Sets 2 and 3 were recorded; within that campaign the assembly remained undisturbed. Hence, small deviations in the absolute microphone position between Set 1 and Sets 2–3 cannot be excluded. The acoustic signals were recorded at a sampling rate of 44.1 kHz with a 16-bit resolution.

2.2. Experimental design and data acquisition

In this study, three independent deposition experiment sets (Set 1–Set 3) were designed to capture acoustic characteristics under varying CTWD conditions and to analyze the impact of dataset diversity on classification performance and model generalization. All experimental sets were conducted using the same WAAM system and process parameters; however, differences were introduced in the sequence of CTWD application and deposition location design.

In the first experimental set (Set 1), deposition was performed by sequentially increasing the CTWD from 10 mm to 20 mm, while maintaining a fixed deposition direction (from left to right). This dataset served as a reference under relatively controlled conditions, allowing consistent acquisition of acoustic characteristics associated with CTWD variation.

In contrast, the second and third experimental sets (Set 2 and Set 3) employed a randomized deposition design to incorporate variability representative of real WAAM process conditions. Specifically, CTWD values were applied in a random order rather than sequentially, and deposition locations were also varied without a fixed sequence. This design ensured that non-stationary process factors—such as temporal heat accumulation, location-dependent thermal history, and variations in equipment and environmental conditions—were not systematically correlated with specific CTWD values.

The acoustic signals acquired from each experimental set were segmented into fixed-length time windows of 0.3 s for subsequent analysis. Each segment was assigned the nominal CTWD value (10–20 mm) corresponding to its deposition condition as the ground-truth class label. This formulation enabled the construction of a multi-class classification problem with 1 mm CTWD resolution. Each experimental set contained one continuously recorded bead per CTWD condition, with Set 1 additionally including one repeated recording at 15 mm; this yielded 12, 11, and 11 recordings of approximately 29–30 s for Sets 1–3, respectively. Repetition for each CTWD condition was therefore provided at the session level by the three independent sets rather than by multiple beads within a session. Segmentation used non-overlapping 0.3 s windows, producing approximately 96–100 segments per recording and approximately 3300 segments in total.

All recordings (44.1 kHz, 16-bit, mono) were resampled to 14 kHz prior to analysis. Each 0.3 s segment (4200 samples) was transformed into a time–frequency representation using the short-time Fourier transform with a Hann window of 512 samples (36.6 ms) and a hop length of 256 samples (18.3 ms, 50% frame overlap), yielding 257 frequency bins spanning 0–7000 Hz (approximately 27.3 Hz per bin) and 17 time frames per segment. The linear frequency scale was retained rather than a compressive Mel scale because the informative structure of the CMT arc signature—the short-circuit repetition rate and its harmonics—is distributed across the analysis band rather than concentrated at low frequencies. The resulting power spectrogram was converted to the decibel scale and standardized to zero mean and unit variance on a per-segment basis before being used as the network input.

For the classification experiments reported in Section 3, the segmented data were randomly partitioned into training (80%) and validation (20%) subsets at the individual-segment level with a fixed random seed (8746); the validation subset was also used for the reported performance evaluation, and no additional held-out test set was reserved. The training checkpoint used for evaluation was likewise selected by best accuracy on this same subset, which adds a further optimistic component to the reported figures. Because temporally adjacent segments from the same bead can be assigned to both subsets under this scheme, the resulting figures characterize within-dataset performance and should be interpreted as an optimistic upper bound with respect to unseen deposition sessions. Generalization across independent sessions is therefore assessed separately through a leave-one-session-out evaluation, in which entire recording sessions are held out from training (Section 3).

The three datasets, constructed under different deposition strategies, serve as a key comparative basis in the Results and Discussion section for evaluating the effects of dataset variability and process-induced uncertainty on CTWD classification performance and model design.

3. Results and discussion

3.1. Statistical feature-based acoustic analysis

As a baseline—and following the hand-crafted-feature methodology adopted in prior acoustic CTWD studies—a statistical feature-based analysis was first applied to the Set 1 dataset. Seventeen time- and frequency-domain features (e.g., RMS, skewness, kurtosis, spectral centroid, frequency standard deviation, and band peak ratios) were extracted from each acoustic segment to provide a compact description of the overall energy level and its spectral distribution; the mathematical definitions of all features are summarized in Table 3.

Using these features, an MLP classifier was evaluated on 3-class CTWD tasks with progressively narrower class spacings, from (10, 15, 20) to (14, 15, 16) mm. The MLP, illustrated in Fig. 3, comprised three hidden layers of 256, 128, and 64 units, with batch normalization, ReLU activation, and dropout ($p = 0.2$) applied to the first two hidden layers,

Table 3
Statistical features extracted from acoustic signals for CTWD classification.

Domain	Statistic	Expression
Time	Mean	$x_m = \frac{1}{N}\sum_{n=1}^{N} x(n)$
	Standard deviation	$x_{std} = \sqrt{\frac{\sum_{n=1}^{N}(x(n)-x_m)^2}{N-1}}$
	Root Mean Square	$x_{rms} = \sqrt{\frac{\sum_{n=1}^{N}(x(n))^2}{N}}$
	Peak	$x_p = max\|x(n)\|$
	Skewness	$x_{ske} = \frac{\sum_{n=1}^{N}(x(n)-x_m)^3}{(N-1)x_{std}^3}$
	Kurtosis	$x_{Kur} = \frac{\sum_{n=1}^{N}(x(n)-x_m)^4}{(N-1)x_{std}^4}$
	Crest factor	$CF = \frac{x_p}{x_{rms}}$
	Clearance factor	$CLF = \frac{x_p}{\left(\frac{1}{N}\sum_{n=1}^{N}\sqrt{\|x(n)\|}\right)^2}$
	Shape factor	$SF = \frac{x_{rms}}{\frac{1}{N}\sum_{n=1}^{N}\|x(n)\|}$
	Impulse factor	$IF = \frac{x_p}{\frac{1}{N}\sum_{n=1}^{N}\|x(n)\|}$
Frequency	Mean Frequency	$f_{mean} = \frac{\sum_{k=1}^{K} f_k \bullet P(f_k)}{\sum_{k=1}^{K} P(f_k)}$
	Frequency centre	$f_{centre} = \frac{\sum_{k=1}^{K} f_k^2 \bullet P(f_k)}{\sum_{k=1}^{K} P(f_k)}$
	Root means square frequency	$f_{rms} = \sqrt{\frac{\sum_{k=1}^{K} f_k^2 \bullet P(f_k)}{\sum_{k=1}^{K} P(f_k)}}$
	Standard deviation frequency	$f_{std} = \sqrt{\frac{\sum_{k=1}^{K}(f_k - f_{mean})^2 \bullet P(f_k)}{\sum_{k=1}^{K} P(f_k)}}$
	Spectrum peak ratio outer	$SPRO = \frac{\max(s(f_k)\|f_k < f_1)}{\sum_{k=1}^{K} s(f_k)}$
	Spectrum peak ratio inner	$SPRI = \frac{\max(s(f_k)\|f_1 \le f_k < f_2)}{\sum_{k=1}^{K} s(f_k)}$
	Spectrum peak ratio roller	$SPRO = \frac{\max(s(f_k)\|f_k \ge f_2)}{\sum_{k=1}^{K} s(f_k)}$

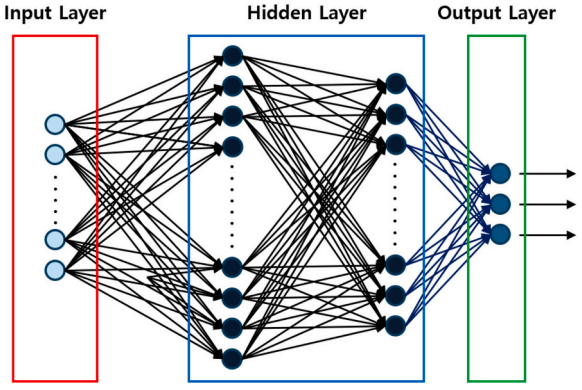

Fig. 3. Architecture of the multilayer perceptron (MLP) used for statistical feature-based CTWD classification.

followed by a softmax output layer over the three classes. Accuracy and F1-score degraded monotonically as the spacing narrowed, reaching approximately 0.82 in the narrowest setting, and replacing the MLP with alternative architectures (1D CNN, RNN, and soft/hard voting ensembles), compared in Fig. 4, yielded no meaningful improvement. The performance ceiling is therefore imposed by the representational capacity of the features rather than by the classifier.

The permutation feature importance analysis presented in Fig. 5 identified the standard deviation of frequency and PeakRatioInner as the dominant descriptors, indicating that CTWD primarily modulates the dispersion and band-wise distribution of acoustic energy—information that low-dimensional statistics summarize only coarsely.

Two implications follow. First, hand-crafted statistical features—although partially informative at coarse (3-class) granularity—cannot resolve 1 mm CTWD differences, consistent with the limitations of prior hand-crafted-feature approaches discussed in Section 1; this motivates the STFT-based time–frequency representation and 2D CNN framework introduced in the following section. Second, because these features are dominated by overall energy-level and coarse spectral summaries, their failure provides direct evidence that absolute signal level alone cannot support fine-grained CTWD discrimination—a point revisited quantitatively in Section 3.2 in the context of the microphone-to-arc distance analysis.

3.2. Time-frequency representation and CNN-based CTWD classification

In this section, the limitations of the statistical feature-based approach are addressed by extending the analysis of arc acoustic signals into the time–frequency domain. As demonstrated in Section 3.1, statistical feature-based models were capable of capturing the existence of energy distribution differences associated with CTWD variations; however, the averaging process within each segment inevitably led to the loss of temporal information, resulting in a structural limitation in distinguishing fine-grained CTWD differences at 1 mm resolution. To overcome this limitation, a Short-Time Fourier Transform (STFT)-based time–frequency representation was introduced to preserve the non-stationary characteristics and localized frequency variations inherent in acoustic signals.

Fig. 6 illustrates the overall architecture of the proposed STFT–2D CNN-based CTWD classification framework. The raw acoustic signal is first segmented into fixed-length windows and then transformed into a two-dimensional spectrogram using STFT. These spectrograms are subsequently used as inputs to a 2D CNN classifier. This design enables simultaneous preservation of temporal evolution and frequency energy distribution, allowing the model to directly learn structural differences associated with CTWD variations.

Importantly, the proposed approach does not rely on absolute sound pressure levels or isolated frequency components. Instead, it focuses on learning the relative energy distribution and localized structural patterns within the time–frequency domain. Such representation inherently provides robustness against variations in signal amplitude caused by external factors such as microphone placement, sensor-to-arc distance, and ambient noise. In other words, the framework emphasizes the identification of recurring structural patterns associated with CTWD changes rather than absolute signal magnitude. This property is enforced by construction in the preprocessing employed here: each dB-scaled spectrogram is standardized to zero mean and unit variance on a per-segment basis, so that any segment-wise offset in absolute level—including the small offset induced by CTWD-dependent changes in the microphone-to-arc distance—is removed from the model input prior to classification.

The magnitude of the distance-induced level change can be bounded explicitly. Because the microphone is rigidly linked to the torch while the arc remains at the workpiece surface, changing the CTWD by an amount δ displaces the arc relative to the microphone by at most δ, irrespective of the mounting geometry. The sensing assembly was installed at a nominal microphone-to-nozzle separation of 150 mm, the microphone being mounted directly on the camera housing so that the two share the same position relative to the torch. Adopting $d = 150$ mm, the free-field sound-level change over the entire 10 mm CTWD range is at most $20\log10((d + 10)/d) \approx 0.6$ dB, i.e., below 0.06 dB per 1 mm class step, and even under a further conservative lower bound of $d \geq 100$ mm it remains below 0.09 dB per step. Such differences are far smaller than the natural short-term fluctuations of the arc sound level and are, in any case, removed from the network input by the per-

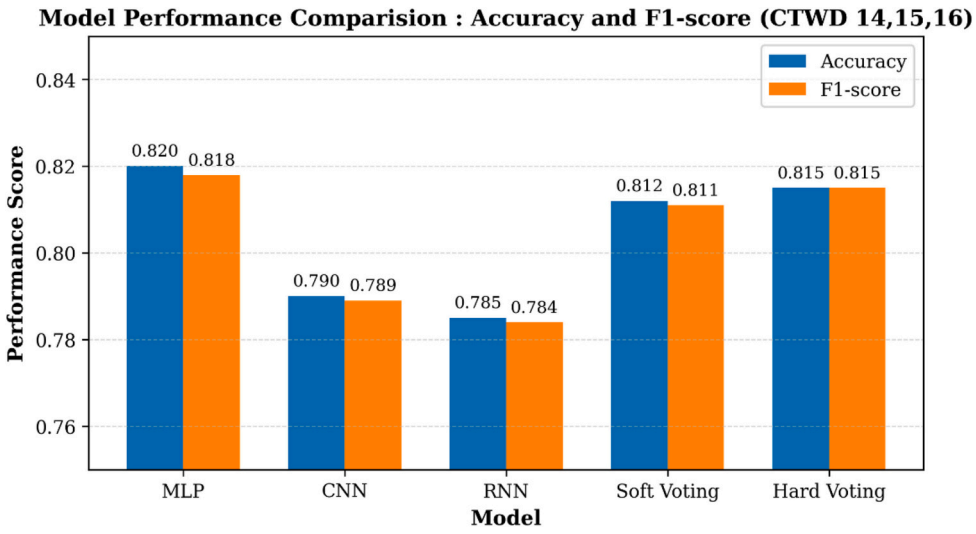

Fig. 4. Performance comparison of statistical feature-based models (MLP, CNN, RNN, Soft Voting, and Hard Voting) for the CTWD (14, 15, 16) 3-class classification task.

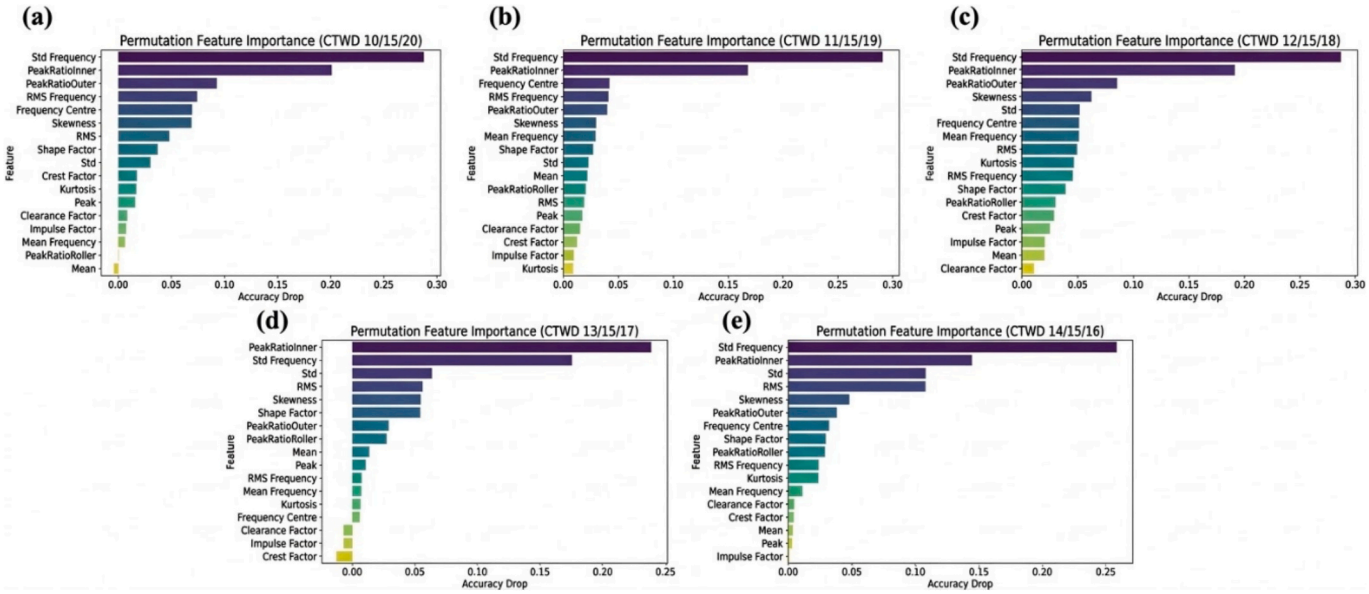

Fig. 5. Permutation feature importance of statistical acoustic features for five 3-class CTWD classification settings: (a) 10/15/20 mm, (b) 11/15/19 mm, (c) 12/15/18 mm, (d) 13/15/17 mm, and (e) 14/15/16 mm.

segment standardization described above. Consequently, absolute-level (distance) cues cannot account for the observed 11-class discrimination, and the classification must rely on the spectral–temporal structure of the signal.

Each acoustic segment, when transformed into a spectrogram, captures the evolution of frequency energy over time. Unlike statistical summaries, this representation retains structural information such as arc transfer periodicity, frequency band energy redistribution, and temporal irregularities in pattern formation. Notably, even when adjacent CTWD conditions exhibit minimal differences in average energy levels, distinct variations can still be observed in the time–frequency plane, including shifts in energy concentration, dispersion patterns, and temporal persistence of spectral features.

Using this time–frequency representation as input, a 2D CNN-based classification model was trained on the Set 1 dataset. Unless otherwise noted, all models in this study were trained with the AdamW optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-3}) and the cross-entropy loss, with a batch size of 16 and up to 200 epochs; the parameters achieving the best validation accuracy were retained for evaluation, and the random seed for data partitioning was fixed at 8746. The model achieved an Accuracy of 0.9955 and a Macro F1-score of 0.9952 for the 11-class (10–20 mm) CTWD classification task, demonstrating stable convergence and high classification performance. This result stands in clear contrast to the statistical feature-based approach, where the same classification problem could not be effectively resolved.

The confusion matrix for the Set 1 dataset is presented in Fig. 7. The majority of samples are correctly classified along the diagonal, indicating a high level of classification consistency across all CTWD classes. This confirms that the STFT–2D CNN framework successfully captures the structural differences in acoustic patterns associated with CTWD variations.

The key factor underlying this performance improvement lies not in increased model complexity, but in the representation of input data. In the statistical feature-based approach, descriptors such as frequency

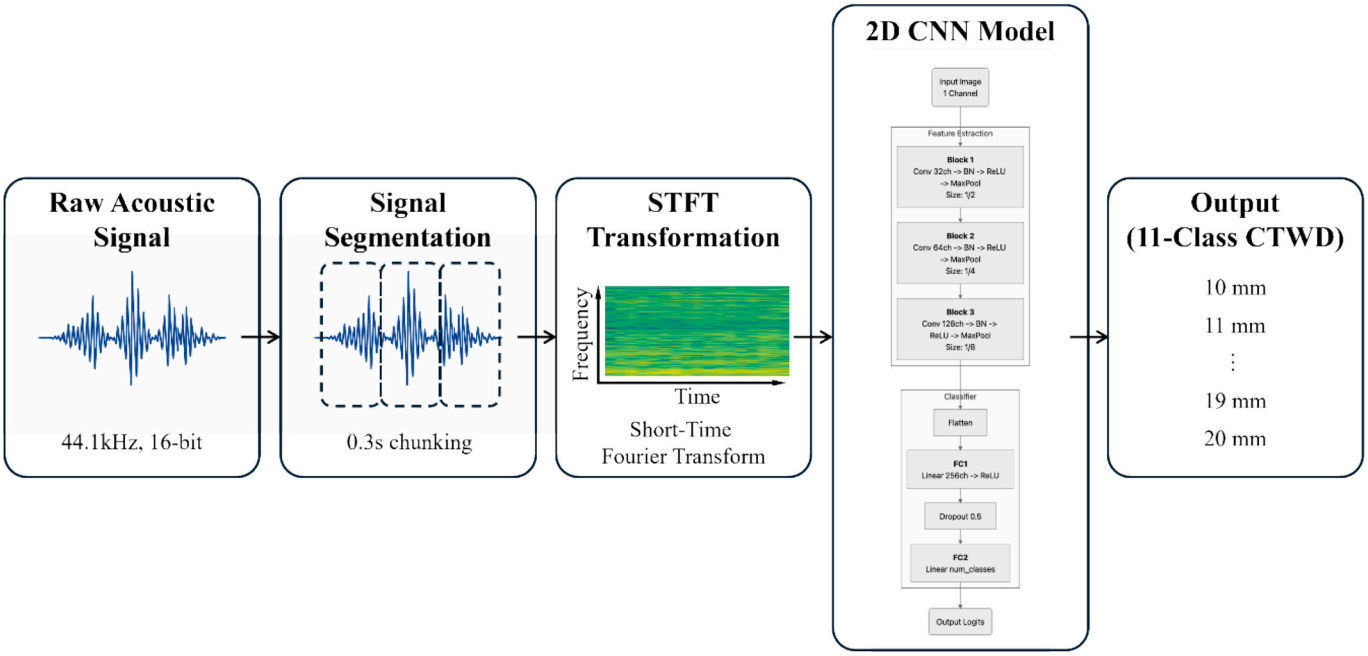

Fig. 6. Overall framework of the proposed STFT–2D CNN based CTWD classification model.

standard deviation and peak ratios capture only partial aspects of CTWD-induced variations and fail to encode temporal sequence and structural relationships. In contrast, STFT-based spectrograms provide a comprehensive representation that includes not only the magnitude of energy distribution but also its temporal evolution and spatial arrangement across frequency bands. The CNN architecture is inherently well-suited to extract such localized patterns in a hierarchical manner, enabling the accumulation of subtle structural differences into distinguishable class boundaries at 1 mm resolution.

Furthermore, the effectiveness of energy distribution-based features observed in the 3-class experiments in Section 3.1 is naturally extended within the STFT framework. Features such as frequency dispersion and peak ratios, which were previously represented as scalar descriptors, are embedded as spatial patterns in the spectrogram and can be automatically learned by the CNN without explicit feature engineering.

In summary, the combination of STFT-based time–frequency representation and 2D CNN significantly outperforms the statistical feature-based approach, demonstrating that preserving and learning the structural characteristics of acoustic signals is essential for fine-grained CTWD classification. These findings confirm that acoustic signals inherently contain sufficient information for CTWD estimation, and that the method of representation and learning plays a decisive role in determining classification performance. In the following section, the effect of dataset expansion and process variability on model performance is further analyzed to evaluate the generalization capability of the proposed approach.

3.3. *Effect of dataset variability and model capacity*

In Section 3.2, it was demonstrated that the STFT-based time–frequency representation combined with a 2D CNN framework can effectively solve the fine-grained CTWD classification problem with 1 mm resolution under a single experimental session (Set 1). For the 11-class (10–20 mm) classification task, the model achieved an Accuracy of 0.9955 and a Macro F1-score of 0.9952, quantitatively confirming that time–frequency representations preserving structural information enable precise discrimination of subtle acoustic differences induced by CTWD variations.

However, in practical WAAM environments, multiple sources of variability—such as repeated experiments, changes in deposition sequence, location-dependent thermal history, and variations in equipment and environmental conditions—simultaneously influence the process. These factors can alter the statistical characteristics and structural patterns of acoustic signals even under identical CTWD conditions. Therefore, this section investigates how model performance changes under increased dataset diversity and process variability, and discusses the necessity of expanding model capacity to address these challenges.

To this end, an extended dataset was constructed by integrating three experimental sets (Set 1–Set 3) acquired under identical process parameters but different deposition strategies. In particular, Set 2 and Set 3 employed a randomized deposition design, where both CTWD values and deposition locations were randomly assigned. This design prevents thermal accumulation and location-dependent thermal history from being systematically correlated with specific CTWD conditions. Consequently, the classification problem shifts from learning relatively fixed pattern distributions within a single session to extracting CTWD-specific structural features that remain consistent across sessions with distribution shift.

When the Simple 2D CNN architecture used in Section 3.2 was applied to the extended dataset under training conditions identical to those of the proposed model, the performance decreased to an Accuracy of 0.6476 and a Macro F1-score of 0.6386. Extending the training budget from 200 to 500 epochs raised these figures only to 0.7018 and 0.6951, indicating that the gap is not attributable to an insufficient epoch count. Misclassifications were observed across the entire CTWD range rather than being confined to specific classes. Although the training loss converged stably, the validation performance remained limited, indicating that the degradation is not due to overfitting but rather to underfitting caused by insufficient model capacity relative to increased data complexity. Notably, the Simple CNN architecture employs a shallow structure in which a single Conv2d layer is immediately followed by MaxPool2d, leading to rapid spatial downsampling of local time–frequency features. While effective under low-variability conditions, this structure fails to preserve shared structural patterns in the presence of inter-session variability. This baseline is also markedly more sensitive to the optimizer setting than the proposed model: at a learning rate of 1×10^{-3} the SimpleCNN reached 0.9864 in one run but collapsed to 0.4880 in a second run under an identical configuration, whereas the

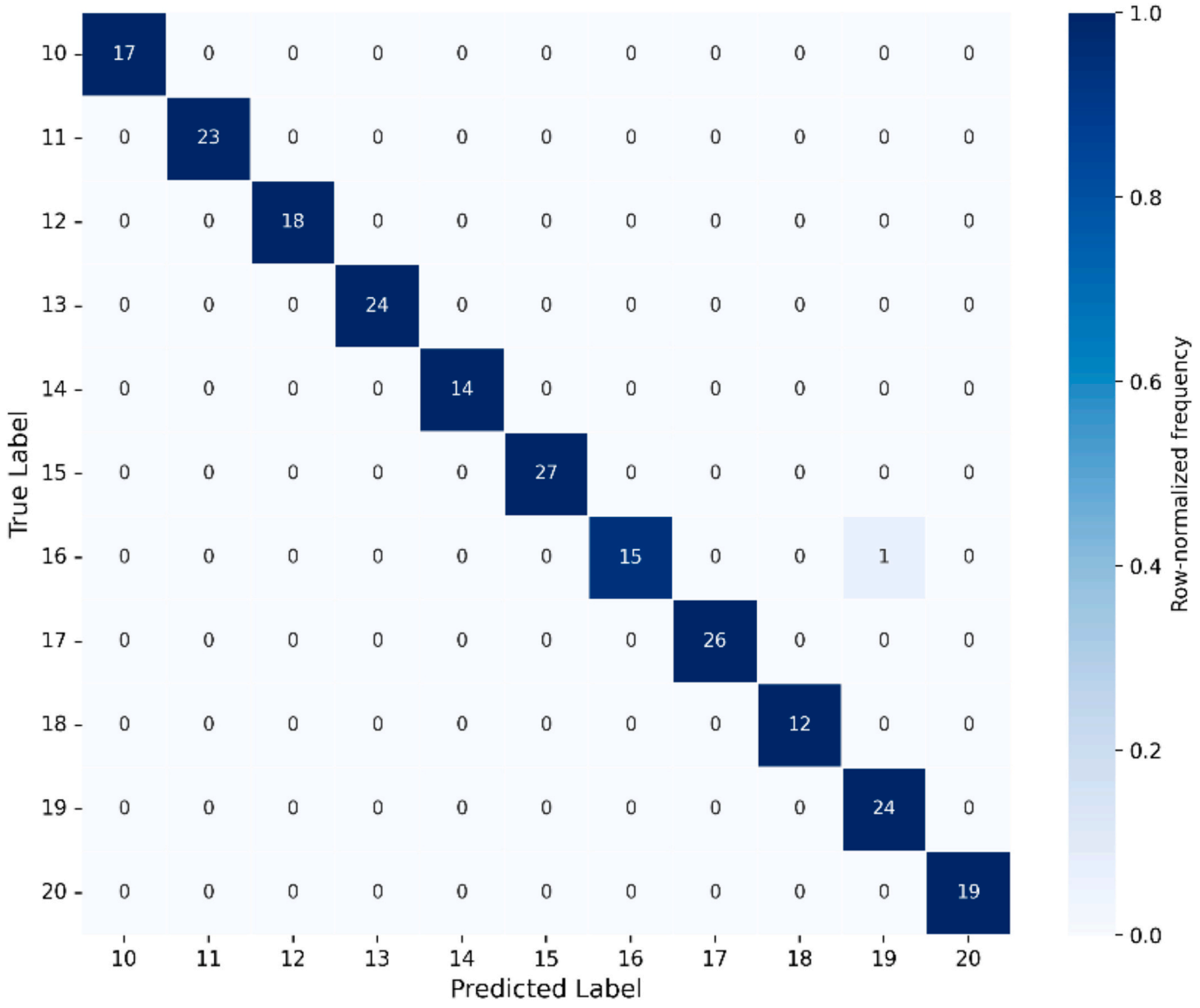

Fig. 7. Confusion matrix of the 11-class CTWD classification on the Set 1 dataset.

proposed DeepCNN attained 0.9654 and 0.9593 at learning rates of 1×10^{-4} and 1×10^{-3}. Stability with respect to the optimization setting is therefore itself a property of the proposed architecture.

To address this limitation, a Deep CNN architecture based on Convolutional Blocks was additionally evaluated; both architectures are illustrated in Fig. 8. The proposed Deep Convolutional Block consists of three consecutive operations of [Conv2d → BatchNorm2d → ReLU], followed by a MaxPool2d layer. Unlike the Simple CNN, which performs immediate spatial reduction after a single convolution, the proposed structure accumulates sufficient nonlinear transformations and normalization steps before pooling, thereby preserving structural information in the time–frequency domain more effectively. Four such Conv_Blocks were stacked, with channel dimensions progressively increasing from 32 to 64, 128, and 256. Each Conv_Block consists of three successive 3×3 convolutions (stride 1, padding 1), each followed by batch normalization and ReLU activation, and ends with a 2×2 max-pooling layer; dropout ($p = 0.3$) is applied for regularization, and the flattened features are passed through a fully connected layer to the 11-way output. This hierarchical design enables the model to extract features ranging from low-level local frequency variations to high-level CTWD-specific patterns, forming robust representations even under dataset variability. The complete layer configuration, including the output feature-map dimensions and the per-stage parameter counts, is summarized in Table 4.

Batch normalization is applied within every convolutional block to stabilize the distribution of time–frequency features during training, and dropout is applied in the classifier for regularization; the sensitivity of the model to the dropout ratio is examined in Section 3.5.

The classification results of the DeepCNN Drop 0.3 model on the extended dataset are presented in the confusion matrix shown in Fig. 9. Of the 664 evaluation segments, 641 were correctly classified, corresponding to an Accuracy of 0.9654 and a Macro F1-score of 0.9662. The confusion matrix is dominated by strong diagonal elements, with per-class recall ranging from 0.912 at 10 mm to 1.000 at 13 mm, the latter being classified without error across all 65 of its segments. The two weakest classes were 10 mm, at 52 of 57 segments, and 17 mm, at 48 of 52; both lie away from the nominal 15 mm working point. Of the 23 misclassified segments, 8 were assigned to a directly adjacent CTWD class, and admitting a ±1 mm tolerance raises the effective accuracy to 0.9774.

These results are physically consistent with the underlying process behavior. Variations in CTWD influence arc length and free wire length,

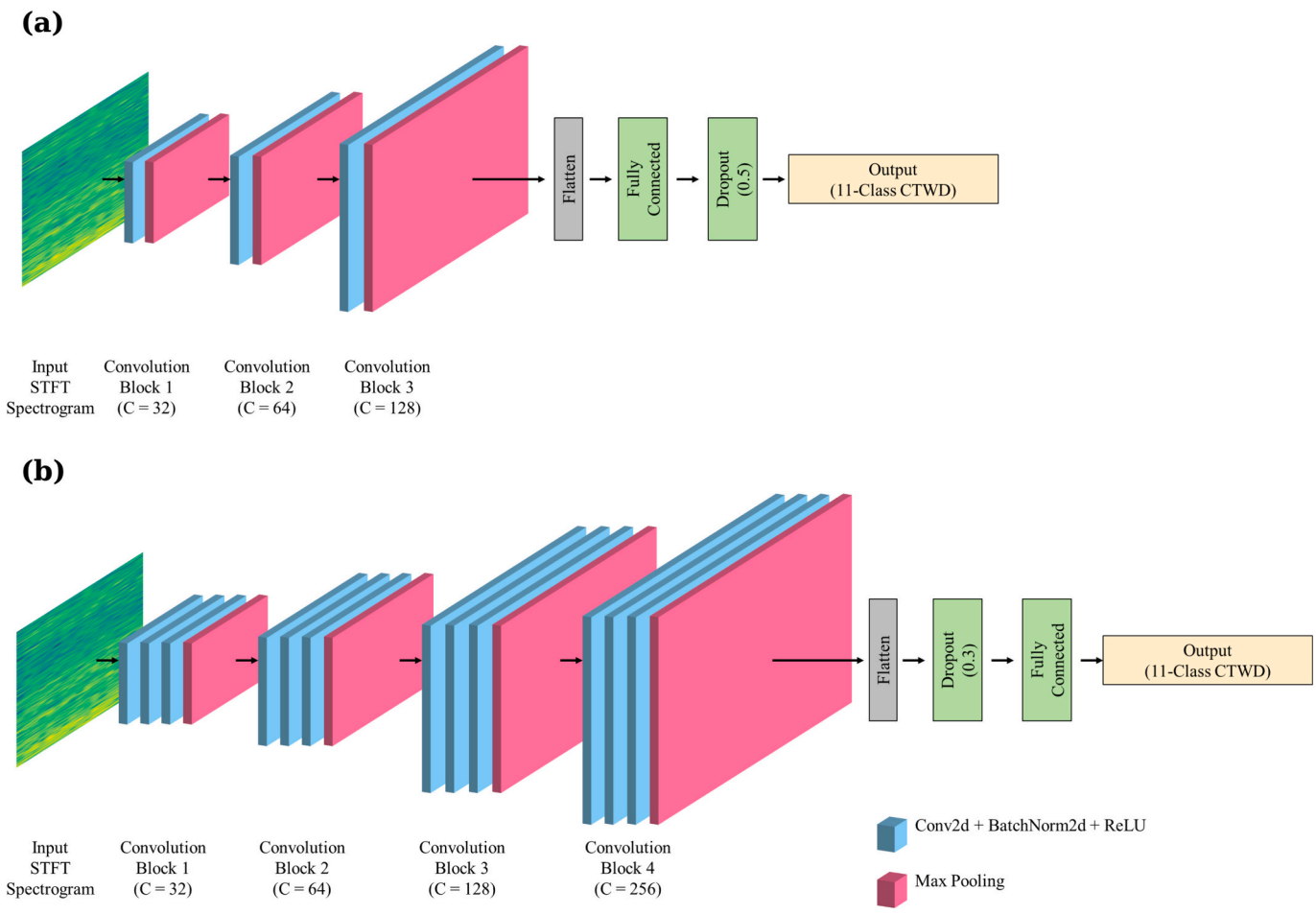

Fig. 8. Architectures of the two CNN models compared in this study: (a) the Simple CNN baseline, with a single convolution per block, and (b) the proposed DeepCNN, with three successive convolutions per block.

Table 4
Layer configuration of the proposed DeepCNN for an input spectrogram of 1 × 257 × 17.

Stage	Configuration	Output (C × F × T)	Parameters
Input	STFT spectrogram	1 × 257 × 17	–
Conv_Block 1	3 × [3 × 3 Conv (s1, p1) + BN + ReLU], 2 × 2 MaxPool	32 × 128 × 8	19,008
Conv_Block 2	3 × [3 × 3 Conv (s1, p1) + BN + ReLU], 2 × 2 MaxPool	64 × 64 × 4	92,736
Conv_Block 3	3 × [3 × 3 Conv (s1, p1) + BN + ReLU], 2 × 2 MaxPool	128 × 32 × 2	369,792
Conv_Block 4	3 × [3 × 3 Conv (s1, p1) + BN + ReLU], 2 × 2 MaxPool	256 × 16 × 1	1,476,864
Classifier	Flatten, Dropout ($p = 0.3$), FC + ReLU, Dropout, FC	512 → 11	2,103,307
Total	–	–	4,061,707

which in turn affect current waveform, short-circuit transfer frequency, and molten pool dynamics in a continuous manner. Therefore, a 1 mm change in CTWD is more likely to manifest as a gradual shift in signal structure than as a discrete transition, which accounts for the confusion observed between adjacent classes. That a substantial share of the residual errors nevertheless span two or more CTWD steps suggests that these errors arise less from the intrinsic similarity of neighbouring conditions than from segment-level variability in arc behavior, in which an individual 0.3 s window occasionally fails to capture a representative transfer pattern. Nevertheless, the ability of the Deep CNN to accurately classify the majority of samples indicates that the proposed hierarchical architecture successfully learns structural differences in time–frequency patterns rather than relying on absolute signal magnitude.

In summary, the two architectures presented in Fig. 8, together with the classification results shown in Fig. 9, demonstrates that the proposed model can robustly distinguish CTWD conditions at 1 mm resolution even under increased dataset variability and process-induced uncertainty within the mixed-data evaluation protocol described in Section 2.2; generalization to entirely unseen sessions is examined in Section 3.4. While the high accuracy observed in a single-session setting reflects the model's representational potential, achieving reliable generalization across varying conditions is essential for practical WAAM applications. The results confirm that a hierarchical Conv_Block-based architecture provides an effective solution for this requirement. In the following section, an ablation study is conducted to systematically analyze the contribution of each architectural component to the overall performance.

3.4. Cross-session generalization assessment (leave-one-session-out)

The evaluations reported in Sections 3.1–3.3 partition the segmented data at the individual-segment level and therefore characterize within-dataset performance, as noted in Section 2.2. To assess generalization to unseen deposition sessions—the deployment-relevant scenario in which a trained model must interpret recordings from welds it has never observed—a leave-one-session-out (LOSO) evaluation was conducted: the model was trained on two of the three experimental sets and tested on all segments of the held-out set, rotated across three folds (A: train Sets 2–3, test Set 1; B: train Sets 1 and 3, test Set 2; C: train Sets 1–2, test

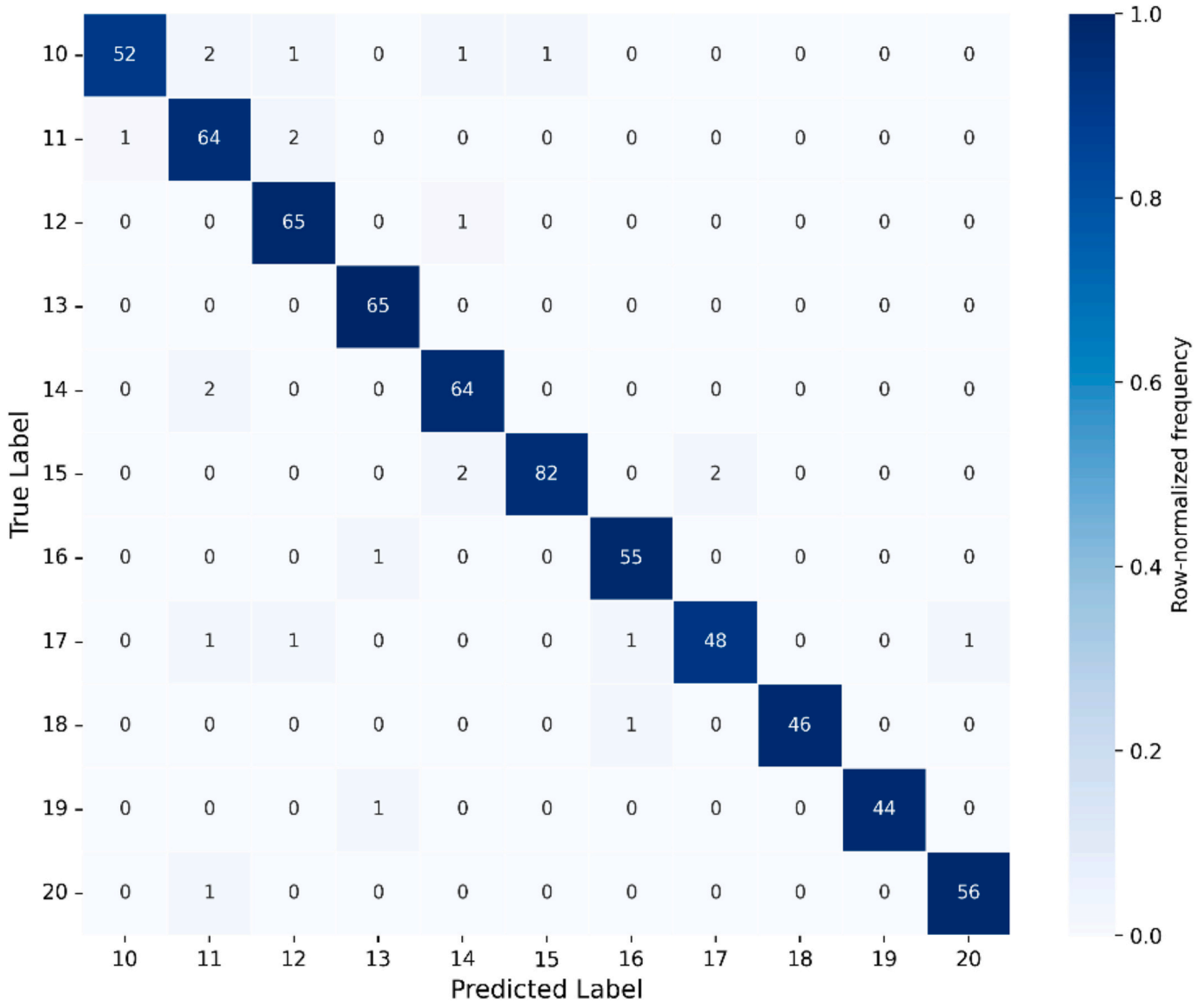

Fig. 9. Confusion matrix of the proposed DeepCNN (Dropout = 0.3) for 11-class CTWD classification on the extended dataset under the segment-level random split.

Set 3). All preprocessing steps, the proposed Conv_Block architecture, and all training settings were identical to those of Section 3.3; only the data partition differed. Two additional single-session folds (D: train Set 2, test Set 3; E: train Set 3, test Set 2) were included to isolate cross-plate transfer within a single recording campaign.

Under LOSO, the 11-class accuracy fell to 0.112, 0.122, and 0.132 for folds A–C, respectively—only marginally above the chance level of 0.091—although validation accuracy on held-in data reached 0.98–0.99 in every fold. Repeating the evaluation over five random seeds yielded 0.128 ± 0.015 (range 0.105–0.159 over 15 runs), confirming that the degradation is systematic rather than seed-dependent. The gap was insensitive to model capacity (channel widths of 32–256 versus 64–256), STFT resolution (mean accuracies of 0.113, 0.122, and 0.126 for n_fft = 256, 512, and 1024, with hop lengths of 128, 256, and 256 respectively), the time–frequency representation (0.118, 0.117, and 0.128 for Mel filterbanks of 32, 64, and 128 bands), session-wise feature standardization, and SpecAugment-style masking; none of these variations narrowed it, and the differences among them lie within the seed-to-seed variation reported above. The fold-wise accuracies for all of these configurations are summarized in Table 5.

The single-session folds D and E yielded accuracies of 0.131 and

Table 5
Leave-one-session-out accuracy under different model, feature, and preprocessing configurations.

Configuration	Fold A	Fold B	Fold C	Mean
Baseline (n_fft 512, linear STFT, 32–256 ch)	0.112	0.122	0.132	0.122
Five seeds × three folds (15 runs)	–	–	–	0.128 ± 0.015
n_fft 256 (hop 128)	0.113	0.123	0.104	0.113
n_fft 1024 (hop 256)	0.113	0.15	0.115	0.126
Mel filterbank, 32 bands	0.093	0.112	0.15	0.118
Mel filterbank, 64 bands	0.071	0.144	0.135	0.117
Mel filterbank, 128 bands	0.092	0.132	0.161	0.128
Channel widths 64–256	0.129	0.147	0.123	0.133
Session-wise standardization	0.101	0.084	0.142	0.109
Session-wise standardization + SpecAugment	0.103	0.101	0.139	0.114
RMS level only (control)	0.088	0.086	0.099	0.091
Single-session folds D and E	0.131 (D)	0.102 (E)	–	0.116
Chance level	–	–	–	0.091

0.102, indistinguishable from folds A–C, even though Sets 2 and 3 originate from the same recording campaign and share identical preprocessing. The failure therefore cannot be attributed to the characteristics of any particular session. A structural explanation follows from the dataset design: because each CTWD condition corresponds to a single bead within a session, class identity and recording identity are confounded within any one session, and the segment-level protocol of Section 3.3 rewards the recognition of recording-specific signatures—which, by construction, do not transfer to new beads.

The transfer failure is nevertheless not total. The ordinal structure of CTWD partially survives the session boundary: over folds B–E, the Spearman rank correlation between true and predicted CTWD reached 0.30–0.47, the ±1 mm tolerance accuracy 0.38–0.46 (random expectation ≈ 0.26), and a coarse three-band re-scoring (low: 10–13 mm; nominal: 14–16 mm; high: 17–20 mm), illustrated for fold C in Fig. 10, 0.40–0.51 against a proportional chance level of 0.34 (majority-band baseline: 0.364). Fold A was the exception, with predictions collapsing toward the extreme classes. Mean absolute errors of 2.6–2.9 mm are consistent with this picture: the monotonic trend of the acoustic signature transfers weakly, whereas the absolute 1 mm scale does not. Because Sets 2 and 3 were recorded within a single campaign without disturbing the sensing assembly, folds D and E further isolate cross-session transfer from any effect of sensor re-installation, indicating that the failure is not attributable to sensor repositioning. The same behavior was observed for pretrained backbones: under LOSO, ResNet18 and EfficientNetB2 reached accuracies of 0.109 and 0.102, respectively, and the ordinal statistics of ResNet18 on fold C (Spearman 0.43; three-band accuracy 0.47) closely matched those of the proposed model, indicating that the residual ordinal signal reflects the information available in the data rather than the representational capacity of a particular architecture.

Finally, a level-only control model trained on the per-frame RMS sequence alone remained at chance accuracy (0.09), and its validation loss did not depart from ln(11) even on held-in data, confirming that the discriminative information exploited by the spectrogram-based models resides in spectral–temporal structure rather than in absolute signal level. Taken together, these results delimit the scope of Sections 3.2–3.3: the proposed framework resolves 1 mm CTWD differences within the acoustic conditions represented in training, whereas transferring the absolute 1 mm scale to unseen sessions remains an open problem, with only weak ordinal information carrying over. Achieving session-robust fine-grained CTWD estimation—for example, through per-session calibration or representation learning that disentangles CTWD-related structure from recording-specific characteristics—is identified as the principal direction for future work.

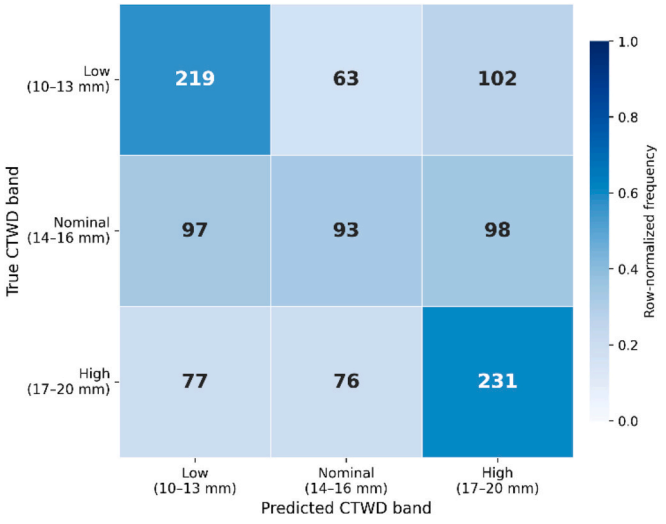

Fig. 10. Confusion matrix of the leave-one-session-out evaluation (fold C) scored on three coarse CTWD bands (low: 10–13 mm; nominal: 14–16 mm; high: 17–20 mm).

3.5. Discussion and ablation study

In this section, the experimental results are comprehensively interpreted by integrating perspectives on data representation, model architecture, regularization, and comparisons with state-of-the-art (SOTA) models. The comparative performance of the proposed model and the image-pretrained backbones under the extended dataset condition is summarized in Table 6.

EfficientNetB2 and ResNet18 were trained under conditions matched to those of the proposed model—single-channel spectrograms replicated to three channels, bilinear resizing to 224 × 224, ImageNet normalization, ImageNet-pretrained initialization, a learning-rate sweep over 1×10^{-4} and 1×10^{-3}, and checkpoint selection by best validation accuracy—with each model reported at its own optimum. Under these conditions the backbones reached accuracies of 0.9834 (ResNet18) and 0.9729 (EfficientNetB2), compared with 0.9654 for the proposed DeepCNN evaluated in the same environment and in the same training run. Paired McNemar tests on the same evaluation segments indicate that the difference is significant for ResNet18 ($p = 0.029$) but not for EfficientNetB2 ($p = 0.442$). The proposed architecture therefore does not surpass these backbones in absolute accuracy; rather, it attains comparable performance with roughly one third of the parameters of ResNet18, without ImageNet pretraining, and without resampling the spectrogram to a natural-image input size. The structural differences between spectrogram-based and natural-image classification nevertheless remain relevant to the design of efficient models. EfficientNet and ResNet architectures are optimized for extracting spatial features from RGB images, relying on strong translation invariance and hierarchical spatial abstraction. However, STFT-based spectrograms exhibit anisotropic structures along time and frequency axes, where the preservation of relative energy distribution and repetitive temporal patterns is critical. The inherent inductive bias of image-based models, particularly their emphasis on spatial invariance, may suppress the relative arrangement of time–frequency structures or lead to learning distribution shifts as noise. It should be emphasized that the accuracies of 0.96–0.98 discussed above are all obtained under the segment-level random partition described in Section 3.3. When the same three models are evaluated under the leave-one-session-out protocol of Section 3.4, their accuracies fall to 0.128 (proposed DeepCNN), 0.109 (ResNet18), and 0.102 (EfficientNetB2), close to the 0.091 chance level, and the ranking is inverted with respect to the random-partition results. Greater capacity and image-domain pretraining therefore improve the recognition of recording-specific signatures without improving cross-session transfer. The comparison therefore positions the proposed

Table 6
Comparison of the proposed model and image-pretrained backbones trained under matched conditions. Random 80:20 values are taken from a single training run per model, namely the runs shown in Figs. 9–11; a repeat of the same configurations gave 0.9744, 0.9729, and 0.9639 respectively, indicating a run-to-run variation of about ±0.01 attributable to non-deterministic GPU computation. McNemar *p*-values are for paired predictions against the proposed model on the same evaluation segments.

Model	Parameters	Optimal lr	Random 80:20	LOSO (mean)	McNemar p
DeepCNN (proposed)	4.06 M	1×10^{-4}	0.9654	0.128	–
ResNet18 (ImageNet)	11.18 M	1×10^{-3}	0.9834	0.109	0.029
EfficientNet-B2 (ImageNet)	7.72 M	1×10^{-3}	0.9729	0.102	0.442
Chance level	–	–	0.091	0.091	–

model as a compact alternative that matches, rather than exceeds, the accuracy of substantially larger image-pretrained networks on this task.

In contrast, the SimpleCNN baseline, trained under conditions identical to those of the proposed model, reached an F1-score of only 0.6386. Under the same settings the proposed DeepCNN with a Dropout ratio of 0.3 attained an Accuracy of 0.9654, Precision of 0.9676, Recall of 0.9654, and F1-score of 0.9662, as shown in Table 7. This corresponds to an improvement of more than 32 percentage points in F1-score over the shallow baseline under matched training conditions. Importantly, this improvement is not merely a result of increased network depth but is attributed to the structural design of the model. The DeepCNN consists of convolutional blocks composed of repeated Conv2d, BatchNorm2d, and ReLU operations followed by MaxPooling, stacked across four stages with progressively increasing channel dimensions from 32 to 64, 128, and 256. Unlike the SimpleCNN, which performs immediate spatial reduction after a single convolution, the proposed architecture preserves structural information by applying sufficient nonlinear transformations prior to pooling. This enables hierarchical feature extraction from low-level local frequency variations to high-level CTWD-specific patterns.

The convergence behavior of the three models is presented in Fig. 11 and Fig. 12. Under the matched training conditions all three reach a comparable accuracy plateau but differ markedly in stability. ResNet18 rises fastest, exceeding 0.90 validation accuracy by epoch 8, followed by the proposed DeepCNN at epoch 22 and EfficientNetB2 at epoch 33, so that the proposed model does not converge more rapidly than the pretrained backbones. Once the plateau is reached, however, the validation loss of the proposed DeepCNN is substantially smoother, as shown in Fig. 11: over the final 50 epochs its standard deviation is 0.033, against 0.112 for ResNet18 and 0.066 for EfficientNetB2. The corresponding validation accuracy curves are given in Fig. 12. The pretrained backbones exhibit repeated loss excursions throughout training, so that the epoch at which the best checkpoint is obtained, and hence the accuracy reported for them, depends appreciably on where in the schedule such an excursion happens to fall. These observations indicate that BatchNorm-based normalization within a compact architecture yields a more predictable training trajectory, which is the practically relevant property when a model must be retrained on new process data without an extensive validation budget.

The ablation study results related to regularization are also summarized in Table 7. Accuracies of 0.9684, 0.9654, and 0.9669 were obtained for ratios of 0.2, 0.3, and 0.4 respectively, with corresponding F1-scores of 0.9678, 0.9662, and 0.9673. These differences are not statistically significant: paired McNemar tests against the 0.3 configuration give $p = 0.832$ for 0.2 and $p = 1.000$ for 0.4, and the spread is comparable to the run-to-run variation observed for a fixed configuration. The model is therefore insensitive to the dropout ratio over this range, and a value of 0.3 was adopted for all reported experiments without a claim of optimality.

Overall, the results demonstrate that CTWD classification is fundamentally a problem of identifying structural differences in time–frequency patterns rather than comparing absolute energy magnitudes.

Table 7
Quantitative performance comparison (Accuracy, Precision, Recall, F1-score) between the Simple CNN baseline and the proposed DeepCNN with different Dropout ratios for 11-class CTWD classification on the extended dataset (664 evaluation segments). All models share identical preprocessing, optimizer, learning rate (1×10^{-4}), batch size, epoch budget, checkpoint criterion, and data partition, and all values are taken from the same training run as Figs. 9–11. Differences among the three dropout ratios are not statistically significant (McNemar $p = 0.832$ for 0.2 and $p = 1.000$ for 0.4, both against 0.3).

Model	Accuracy	Precision	Recall	F1-Score
SimpleCNN (Baseline)	0.6476	0.6412	0.642	0.6386
DeepCNN (Drop = 0.2)	0.9684	0.9692	0.9669	0.9678
DeepCNN (Drop = 0.3)	0.9654	0.9676	0.9654	0.9662
DeepCNN (Drop = 0.4)	0.9669	0.9697	0.9656	0.9673

A moderately deep architecture composed of convolutional blocks that aligns with the characteristics of the data matches the accuracy of substantially larger image-pretrained models at a fraction of their parameter count. The proposed DeepCNN, combined with STFT-based representation, achieves a robust F1-score of 0.9662 under extended dataset conditions, quantitatively confirming the feasibility of acoustic-based approaches for fine-grained CTWD classification at 1 mm resolution.

4. Conclusion

This study proposes a data-driven analysis framework for classifying Contact Tip to Work Distance (CTWD) at a 1 mm resolution using arc acoustic signals generated in Wire Arc Additive Manufacturing (WAAM). Unlike conventional approaches that interpret CTWD variations through bead geometry-based indirect control or limited frequency-domain analysis, this study formulates the time–frequency structure of acoustic patterns induced by CTWD variations as a learnable representation. Statistical feature-based analysis confirmed that CTWD variations influence the overall energy distribution and frequency dispersion characteristics of acoustic signals. While certain frequency-domain statistical features exhibited discriminative capability under limited conditions, the feature distributions increasingly overlapped as the CTWD interval decreased, revealing the inherent limitation of statistical representations in resolving fine-grained CTWD differences. These findings are consistent with prior frequency-based interpretations while experimentally demonstrating that CTWD variations inherently involve non-stationary and localized acoustic pattern changes. To address this limitation, an STFT-based time–frequency representation was introduced to preserve non-stationary signal characteristics and localized frequency patterns. Under a single-session dataset, the proposed approach achieved an Accuracy of 0.9955 and a Macro F1-score of 0.9952 for the 11-class CTWD classification task, indicating that CTWD variations are reflected as consistent structural differences in the time–frequency domain. However, under extended dataset conditions incorporating process variability such as repeated experiments, sensor re-installation, and randomized deposition sequences, the SimpleCNN model trained under matched conditions reached an Accuracy of only 0.6476 and an F1-score of 0.6386, highlighting the importance of model capacity in handling increased data complexity. Based on these observations, a Conv_Block-based Deep CNN architecture was designed, achieving an Accuracy of 0.9654 and an F1-score of 0.9662 under extended dataset conditions. This represents a substantial improvement over the shallow baseline under matched training conditions, while matching the accuracy of ImageNet-pretrained backbones at roughly one third of their parameter count. A leave-one-session-out evaluation, however, quantified the remaining challenge: when entire sessions were held out from training, accuracy fell to 0.128 ± 0.015, with only weak ordinal information transferring across sessions, underscoring session-robust generalization as the principal open problem for practical deployment. The present findings are established for SUS316L single-bead single-layer deposition under CMT-based WAAM. Because the periodic short-circuit transfer characteristic of CMT produces a strongly structured acoustic signature, the extent to which the proposed framework transfers to other transfer modes, materials, or energy sources remains an open question. Extending the approach to less periodic transfer modes and to additional materials, and identifying the acoustic conditions under which spectrogram-based CTWD estimation remains effective, are therefore planned as future work, together with the session-robust generalization identified above. The ablation study further indicated that performance is insensitive to the dropout ratio over the range 0.2–0.4, and that an appropriately deep hierarchical structure aligned with time–frequency data characteristics is more critical than parameter count alone.

In summary, the main contribution of this study is not merely demonstrating that acoustic signals contain CTWD information, but

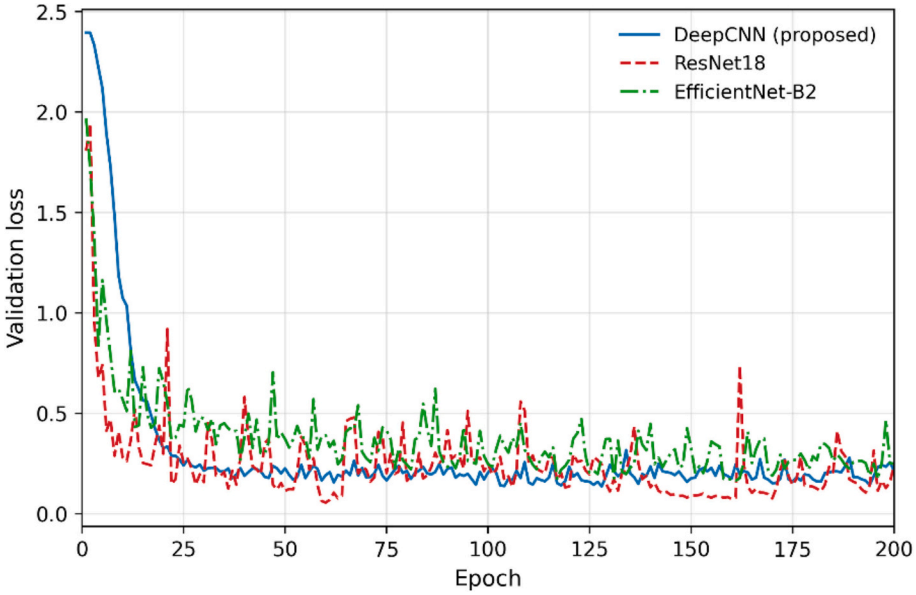

Fig. 11. Validation loss convergence of the proposed DeepCNN and the image-pretrained backbones under matched training conditions.

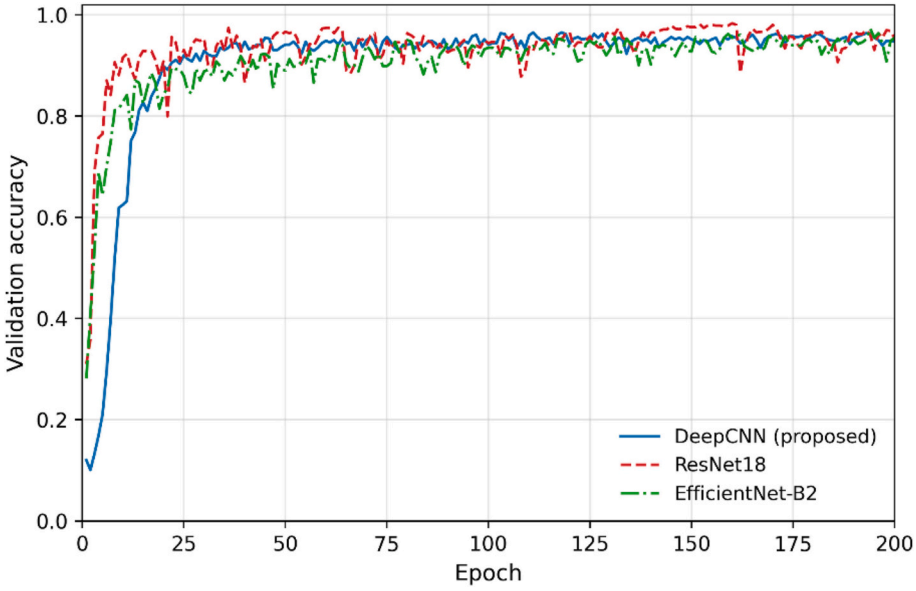

Fig. 12. Validation accuracy convergence of the proposed DeepCNN and the image-pretrained backbones under matched training conditions.

rather establishing that the time–frequency structure of acoustic patterns can be effectively learned to enable fine-grained CTWD classification at a 1 mm scale. This suggests that acoustic sensing can serve as a low-cost, non-contact alternative to optical systems for improving monitoring resolution in WAAM processes. Nevertheless, this study is limited to single-bead, single-layer deposition conditions under the CMT mode with a specific material system. Therefore, it should be regarded as an initial step toward acoustic-based CTWD monitoring rather than a complete solution for general WAAM processes. Future work should extend this framework to multi-layer deposition environments, considering cumulative geometric errors, thermal history effects, and the interaction between CTWD and bead geometry.

Furthermore, the proposed classification framework can be naturally extended to regression models for continuous CTWD estimation in real time. By combining STFT-based representations with lightweight CNN or temporal models, it is expected that real-time CTWD estimation with sufficient temporal resolution and low latency can be achieved. With respect to control latency, the effective sensing delay of the present formulation is dominated by the 0.3 s analysis window, as the computational cost of the STFT and a single forward pass of the compact CNN is negligible in comparison. CTWD deviations in practice develop through the accumulation of bead-height errors and thermal distortion over successive layers, so the quantity being tracked evolves on the timescale of layer deposition rather than of individual segments; the present update interval is therefore short relative to the dynamics it is intended to follow, and overlapping analysis windows can further shorten the update interval without modifying the architecture. Integrating such predictions with robotic torch positioning or wire feeding control could enable the development of a closed-loop WAAM process control system based on acoustic sensing.

Beyond WAAM, whether the proposed framework transfers to other manufacturing processes remains an open question rather than a demonstrated capability. In particular, the CMT mode employed in this study produces a highly structured and periodic acoustic signature

owing to its mechanically synchronized short-circuit transfer, and this regularity is likely a favorable precondition for the fine-grained separability observed here. Processes with more stochastic transfer behavior (e.g., globular or spray transfer GMAW) or with fundamentally different acoustic generation mechanisms (e.g., laser-based processes) may not exhibit comparably separable time–frequency structures. Systematic verification across transfer modes, materials, and multi-layer conditions is therefore required before general applicability can be claimed.

CRediT authorship contribution statement

Min Seok Jang: Writing – original draft. **Kyung Soon Jung:** Data curation. **Kyeung Jun Park:** Data curation. **Seung Beom Noh:** Validation. **Young Hwan Cha:** Software. **Hyub Lee:** Project administration. **Yong Son:** Supervision. **Seung-Jun Shin:** Supervision. **Young Won Kim:** Writing – review & editing, Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Claude Opus 4.8 (Anthropic) in order to correct grammar and improve the language of the manuscript. The author(s) reviewed and edited the output as needed and take full responsibility for the content of the published article.

Declaration of competing interest

This research was supported by the Ministry of Trade, Industry and Energy of Korea through the project “Development of 3D Printing Digital Transformation Platform Technology Based on Process Metadata” (KM260157), and by the Korea Institute of Industrial Technology (KITECH) through its in-house research project Development of Autonomous Diagnosis and Compensation Technology Based on Laser Output Characteristics in Metal Additive Manufacturing” (UR260093).

References

[1] Williams SW, et al. Wire+ arc additive manufacturing. Mater Sci Technol 2016;32:641–7.
[2] Bekker ACM, Verlinden JC. Life cycle assessment of wire + arc additive manufacturing compared to green sand casting and CNC milling in stainless steel. J Clean Prod 2018;177:438–47.
[3] Treutler K, Wesling V. The current state of research of wire arc additive manufacturing (WAAM): a review. Appl Sci 2021;11:8619.
[4] Srivastava M, Rathee S, Tiwari A, Dongre M. Wire arc additive manufacturing of metals: a review on processes, materials and their behaviour. Mater Chem Phys 2023;294:126988.
[5] Zhang J, Wang X, Paddea S, Zhang X. Fatigue crack propagation behaviour in wire + arc additive manufactured Ti-6Al-4V: effects of microstructure and residual stress. Mater Des 2016;90:551–61.
[6] Hauser T, et al. Porosity in wire arc additive manufacturing of aluminium alloys. Addit Manuf 2021;41:101993.
[7] Wu B, et al. A review of the wire arc additive manufacturing of metals: properties, defects and quality improvement. J Manuf Process 2018;35:127–39.
[8] Henckell P, Gierth M, Ali Y, Reimann J, Bergmann JP. Reduction of energy input in wire arc additive manufacturing (WAAM) with gas metal arc welding (GMAW). Materials (Basel, Switzerland) 2020;13:2491.
[9] Han S-W, et al. Parameter optimization of WAAM with pulsed GMAW for manufacturing propeller-shaped blade. Int J Precis Eng Manuf 2023;24:1103–10.
[10] Han JH, Lee JS, Kim S-H, Yun JO. A study on the improvement of a contact tip for the wire melting rate enhancement. Weld World 2017;61:1181–7.
[11] e Silva RHG, dos Santos Paes LE, Barbosa RC, Sartori F, Schwedersky MB. Assessing the effects of solid wire electrode extension (stick out) increase in MIG/MAG welding. J Braz Soc Mech Sci Eng 2018;40:31.
[12] Ding D, Pan Z, Cuiuri D, Li H. Wire-feed additive manufacturing of metal components: technologies, developments and future interests. Int J Adv Manuf Technol 2015;81:465–81.
[13] Dreher M, et al. Methods and results concerning the shielding gas flow in GMAW. Weld World 2013;57:391–410.
[14] Xiong J, Zhang G. Adaptive control of deposited height in GMAW-based layer additive manufacturing. J Mater Process Technol 2014;214:962–8.
[15] Chabot A, Rauch M, Hascoët J-Y. Towards a multi-sensor monitoring methodology for AM metallic processes. Weld World 2019;63:759–69.
[16] Heralić A, Christiansson A-K, Ottosson M, Lennartson B. Increased stability in laser metal wire deposition through feedback from optical measurements. Opt Lasers Eng 2010;48:478–85.
[17] Xiong J, Zhang G, Qiu Z, Li Y. Vision-sensing and bead width control of a single-bead multi-layer part: material and energy savings in GMAW-based rapid manufacturing. J Clean Prod 2013;41:82–8.
[18] Baldauf M, Lohrer P, Hauser T, Jauer L, Schleifenbaum JH. Camera-based measurement and control of the contact tip to work distance in wire arc additive manufacturing. Prog Addit Manuf 2024;9:565–74.
[19] Ansari MJ, et al. Optimized analytical approach for the detection of process-induced defects using acoustic emission during directed energy deposition process. Addit Manuf 2024;86:104218.
[20] Bhattacharya S, Pal K, Pal SK. Multi-sensor based prediction of metal deposition in pulsed gas metal arc welding using various soft computing models. Appl Soft Comput 2012;12:498–505.
[21] Xia C, et al. A review on wire arc additive manufacturing: monitoring, control and a framework of automated system. J Manuf Syst 2020;57:31–45.
[22] Pal K, Bhattacharya S, Pal SK. Prediction of metal deposition from arc sound and weld temperature signatures in pulsed MIG welding. Int J Adv Manuf Technol 2009;45:1113.
[23] Wang F, Mao H, Zhang D, Zhao X, Shen Y. Online study of cracks during laser cladding process based on acoustic emission technique and finite element analysis. Appl Surf Sci 2008;255:3267–75.
[24] Ye D, Hong GS, Zhang Y, Zhu K, Fuh JYH. Defect detection in selective laser melting technology by acoustic signals with deep belief networks. Int J Adv Manuf Technol 2018;96:2791–801.
[25] Chabot A, Rauch M, Hascoët J-Y. Novel control model of Contact-Tip-to-Work Distance (CTWD) for sound monitoring of arc-based DED processes based on spectral analysis. Int J Adv Manuf Technol 2021;116:3463–72.
[26] Hebert D, Thien A, Saldana C. Acoustic process monitoring of contact tip to work piece distance in wire arc additive manufacturing by random forest algorithms. In: International manufacturing science and engineering conference. vol. 85802. American Society of Mechanical Engineers; 2022. V001T01A020.
[27] Selvi S, Vishvaksenan A, Rajasekar E. Cold metal transfer (CMT) technology-An overview. Defence Technol 2018;14:28–44.
[28] Srinivasan D, Sevvel P, Solomon IJ, Tanushkumaar P. A review on cold metal transfer (CMT) technology of welding. Mater Today Proc 2022;64:108–15.
[29] Lee SH. CMT-based wire arc additive manufacturing using 316L stainless steel: effect of heat accumulation on the multi-layer deposits. Metals (Basel) 2020;10:278.
[30] Xie B, Xue J, Ren X. Wire arc deposition additive manufacturing and experimental study of 316L stainless steel by CMT+ P process. Metals 2020;10:1419.
[31] Jin W, et al. Wire arc additive manufacturing of stainless steels: a review. Appl Sci 2020;10:1563.