signals

MDPI

Article

Multi-Output Classification of SMAW Process Parameters from Arc Sound Using MFCC and Deep Audio Embeddings

Luis Viloria [1], Edmanuel Cruz [2,3] and Cesar Pinzon-Acosta [1,2,*]

[1] Facultad de Ingeniería Mecánica, Universidad Tecnológica de Panamá, Panama City 0819-07289, Panama; luis.viloria@utp.ac.pa
[2] Centro de Estudios Multidisciplinarios de Ciencias de la Ingeniería y Tecnología (CEMCIT-AIP), Panama City 0819-07289, Panama; edmanuel.cruz@utp.ac.pa
[3] Centro Regional de Veraguas, Universidad Tecnológica de Panamá, Atalaya 0901, Panama
* Correspondence: cesar.pinzon1@utp.ac.pa

Abstract

Manual arc welding is highly dependent on operator skill, leading to variability in weld quality and an increased risk of defects; therefore, reliable monitoring methods for Shielded Metal Arc Welding (SMAW) are required, particularly in manual environments where process variability and environmental noise are inherent. This study proposes a monitoring approach for classifying SMAW process parameters using airborne acoustic signals generated by the welding arc. Welding experiments were conducted on carbon steel plates of different thicknesses (3, 6, and 12 mm) using E6010, E6011, E6013, and E7018 electrodes under Alternating Current (AC) and Direct Current (DC) configurations; acoustic signals were recorded in real time and processed using Mel-Frequency Cepstral Coefficients (MFCCs) and deep audio embeddings from pre-trained VGGish and YAMNet models as inputs to artificial neural network classifiers for multi-output classification of welding process parameters. Model performance was evaluated using per-target metrics (accuracy and macro F1-score) and joint multi-output metrics (Exact Match and Hamming Accuracy). MFCC-based models significantly outperformed embedding-based approaches, achieving up to 94.51% Exact Match and 97.88% Hamming Accuracy, while reducing computational costs. These results demonstrate the feasibility of SMAW monitoring using arc sound, suggesting that spectral features are an effective solution for welding-process monitoring and a promising foundation for future weld-quality monitoring systems.

Keywords: Shielded Metal Arc Welding; arc sound; MFCC; audio embeddings; multi-output classification; VGGish; YAMNet

Check for updates

Academic Editor: Hua-Liang Wei

Received: 13 April 2026
Revised: 14 May 2026
Accepted: 19 May 2026
Published: 8 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Shielded Metal Arc Welding (SMAW) is one of the most widely used arc welding processes, often performed in manual mode, due to its simplicity, portability, and adaptability to varying operating conditions [1]. Despite the increasing automation of welding processes, SMAW remains prevalent, particularly in Latin America, where it plays a significant role in construction, maintenance, and manufacturing activities [2].

According to the American Welding Society (AWS), weld quality in SMAW is strongly influenced by process parameters, including base metal thickness, welding current type, and covered electrode classification. Improper selection or control of these variables can lead to undesirable weld discontinuities and variability in weld performance [3].

Consequently, reliable methods for monitoring and identifying welding parameters during SMAW operation remain an important challenge in manufacturing practice.

Several sensing techniques have been used for welding process monitoring, including electrical signal monitoring, vision-based sensing, and acoustic emission sensing. Electrical parameters, such as welding current and voltage, are commonly used to evaluate arc stability and process performance, as they are directly related to heat input and metal transfer behavior [4,5]. Signal processing techniques such as the Fourier and Wavelet Transforms have been widely applied to analyze these signals and extract relevant features for welding quality evaluation [5]. Vision-based monitoring systems using cameras, infrared sensors, and optical emission sensors have also been implemented to evaluate weld pool geometry, weld penetration, and weld bead formation [6]. These systems allow direct observation of the welding process; however, they often require complex image processing algorithms and controlled lighting conditions [2]. Acoustic emission and acoustic sensing techniques have also been investigated for welding process monitoring because the generated acoustic signals are directly related to arc behavior, metal transfer, and weld quality [7]. Acoustic emission methods are particularly attractive because they are passive, non-intrusive, and relatively low-cost compared with other sensing techniques [6]. In addition, machine learning techniques such as artificial neural networks and support vector machines have been applied to welding process monitoring to predict weld bead geometry, penetration, and discontinuity formation from sensor data and process parameters [2,8,9].

Acoustic-based monitoring approaches have gained increasing attention as alternative sensing methods for welding process monitoring because acoustic signals are directly related to the physical phenomena occurring during welding. During arc welding processes, acoustic emissions are generated by several sources, including plasma arc fluctuations, metal transfer, droplet detachment, and weld pool dynamics [7,8]. These acoustic signals contain information related to arc stability, metal transfer modes, and process disturbances, making them suitable for welding process monitoring and analysis.

Recent developments in welding monitoring have increasingly explored multi-task learning frameworks, particularly in vision-based monitoring systems for automated welding processes. Xu et al. [10] proposed a multi-output convolutional architecture combined with particle swarm optimization for simultaneous monitoring of multiple state variables in high-speed laser welding from melt-pool imagery. Multi-task and multi-vision frameworks have also been reported for laser beam welding applications, including the welding-segmentation-recognition network proposed by Zhou et al. [11] and the MT-Segformer architecture integrating stereo vision with multi-task decoders [12]. Truong et al. [13] further extended this paradigm to multi-layer multi-pass welding using combined line-scanning and infrared sensing systems. In parallel, multimodal fusion approaches combining heterogeneous sensing modalities have also been investigated from penetration-state estimation in gas-shielded welding processes [14].

In contrast, acoustic-based monitoring has remained largely focused on single-task formulation addressing penetration state, process stability, or discontinuity-related analysis independently. Airborne acoustic sensing provides a non-intrusive and flexible monitoring approach, particularly suitable for manual welding conditions where sensor placement on the workpiece may be impractical.

Acoustic monitoring during welding can be performed using contact acoustic emission sensors attached to the workpiece or airborne acoustic sensors, such as microphones positioned near the welding arc. Acoustic emission sensors capture stress waves propagating through the material, whereas airborne acoustic sensors capture sound waves generated by interactions between the arc and the welding process [6]. Airborne acoustic sensing provides a non-intrusive and flexible monitoring approach, particularly suitable for manual

welding conditions where sensor placement on the workpiece may be impractical. Early studies demonstrated that acoustic signals can be used to monitor welding process conditions and detect process disturbances using signal processing and neural network-based approaches. These studies showed that acoustic signatures contain sufficient information to characterize welding conditions and identify process anomalies [7,8].

Conventional acoustic welding monitoring methods have relied on handcrafted signal descriptors extracted from time-domain and frequency-domain analyses, including statistical features, the Fast Fourier Transform (FFT), spectrograms, and wavelet-based features [5,15]. Among these methods, Mel-Frequency Cepstral Coefficients (MFCC) have been widely used in acoustic signal classification because they provide a compact representation of the spectral envelope of audio signals and have shown good performance in sound classification tasks, including welding acoustic analysis [8,15]. These handcrafted features are commonly used as inputs to machine learning classifiers, such as artificial neural networks, support vector machines, and k-nearest neighbors, for welding condition classification and weld quality prediction [2,7,8]. However, handcrafted feature extraction methods depend heavily on expert knowledge and may not fully capture the complex, non-stationary acoustic patterns in manual welding processes, such as SMAW [9].

Recent advances in deep learning have enabled the use of deep audio embeddings, in which pretrained convolutional neural networks trained on large-scale audio datasets are used to extract high-level representations directly from raw audio signals [9,16,17]. These embeddings can capture complex temporal and spectral patterns and have shown strong performance in general audio classification tasks [16].

Previous studies have explored the use of deep audio embeddings such, as VGGish and YAMNet, for welding defect detection and audio classification tasks, demonstrating the potential of pretrained audio models for welding-related applications [16]. However, the performance of these embeddings compared with traditional acoustic features such as MFCC for welding process parameters identification has not been sufficiently investigated. In addition, the computational cost and the feasibility of real-time implementation of these approaches for welding monitoring have not been widely discussed. Therefore, this study evaluates and compares traditional MFCC-based features and deep audio embedding features (VGGish and YAMNet), combined with artificial neural network classifiers, for the classification of SMAW parameters using arc sound.

In practical welding monitoring applications, particularly for processes such as SMAW performed in field conditions, monitoring systems must provide not only high classification performance but also reliability, low computational cost, and real-time processing capability to enable practical implementation [18]. Manufacturing environments are typically characterized by noise, process variability, and changing operating conditions, which require robust, reliable monitoring systems capable of maintaining stable performance over time [18].

Condition monitoring and fault diagnosis systems used in real-world engineering applications require high diagnostic accuracy and reliability to be useful for process monitoring and decision-making [19]. Therefore, model evaluation should consider not only classification performance but also computational efficiency and implementation feasibility when assessing suitability for practical deployment [18,20]. For this reason, the selection of feature extraction and classification methods for welding monitoring should be based not only on classification performance but also on computational cost and real-time implementation capability.

Accordingly, the objective of this study is to evaluate the feasibility of using airborne acoustic signals (arc sound) for simultaneous identification of multiple SMAW process parameters. Specifically, this work formulates SMAW acoustic monitoring as a multi-output

classification problem in which plate thickness, electrode type, and current type are jointly inferred from the arc sound.

Within this intentionally conventional multi-output framework, selected to isolate the effect of acoustic representation rather than introduce architectural novelty, the study makes three main contributions. First, it establishes a unified and reproducible benchmark for comparing handcrafted spectral descriptors (MFCC) and pre-trained deep audio embeddings (VGGish and YAMNet) under the same acquisition conditions, classifier architectures, and evaluation protocol. Second, it demonstrates that acoustic representation is the dominant performance factor in SMAW parameter classification, with MFCC consistently outperforming embedding-based approaches across all evaluated models. Third, it shows that this performance gap is already evident at the feature-space level prior to classification, providing evidence that the reduced performance of general-purpose embeddings is primarily due to domain mismatch with welding-specific acoustic signals.

These findings provide a practical, domain-specific reference for the design of acoustic monitoring systems for manual SMAW and establish a baseline for future multimodal, domain-adapted welding monitoring approaches.

2. Materials and Methods

This study proposes an acoustic-based methodology for identifying SMAW process parameters using the arc sound generated during the process. The proposed methodology consists of recording arc sound signals during welding experiments, segmenting the recorded audio signals into short time windows, extracting acoustic features using different feature extraction approaches, and classifying welding process parameters using artificial neural network models.

Three acoustic representation approaches were evaluated and compared in this study: Mel-Frequency Cepstral Coefficients (MFCCs) and deep audio embeddings extracted using pre-trained VGGish and YAMNet models. These feature representations were used as inputs to neural network classifiers for multi-output classification of welding process parameters. Model performance was evaluated using both per-target metrics and joint evaluation metrics, as well as computational performance indicators such as training time and inference time.

Figure 1 illustrates the complete data flow from raw audio acquisition to the final multi-output classification output. Raw audio is acquired using the Xiris Automation Inc. (Burlington, ON, Canada) WeldMIC microphone and recorded through the WeldStudio 3 software, which is resampled to 16 kHz in mono format, then segmented into fixed-duration windows, and processed using different feature extraction approaches (MFCC, VGGish, or YAMNet). The resulting feature representations are used as input to the multi-task classifier, which simultaneously predicts Plate Thickness, Electrode Type, and Current Type.

2.1. Experimental Setup

2.1.1. Welding Monitoring Station

The experimental setup was established at a dedicated welding monitoring station designed to ensure repeatability and controlled data acquisition during SMAW experiments. Welding trials were performed using a multipurpose power source capable of operating under both alternating current (AC) and direct current (DC) modes. A welding table equipped with mechanical fixtures was used to maintain consistent plate alignment and joint position throughout all experiments.

A Xiris weld monitoring system was used to support the experimental setup, with the camera mounted on a tripod and oriented toward the welding arc to ensure consistent positioning during the welding process. The monitoring station also included a laptop for

real-time data acquisition, as shown in Figure 2. Acoustic data were acquired using the integrated WeldMIC microphone, positioned at less than 0.8 m from the welded joint. The key specifications of WeldMIC are summarized in Table 1. The distance was selected as a compromise between signal clarity and protection of the sensor from excessive heat and spatter, while maintaining consistent acoustic capture across all trials. The microphone position was kept constant across all experiments to ensure repeatability in acoustic signal acquisition. This configuration was intentionally selected to reduce acquisition variability during the initial feasibility evaluation of the proposed framework and isolate the effect of welding process parameters on the recorded acoustic signals.

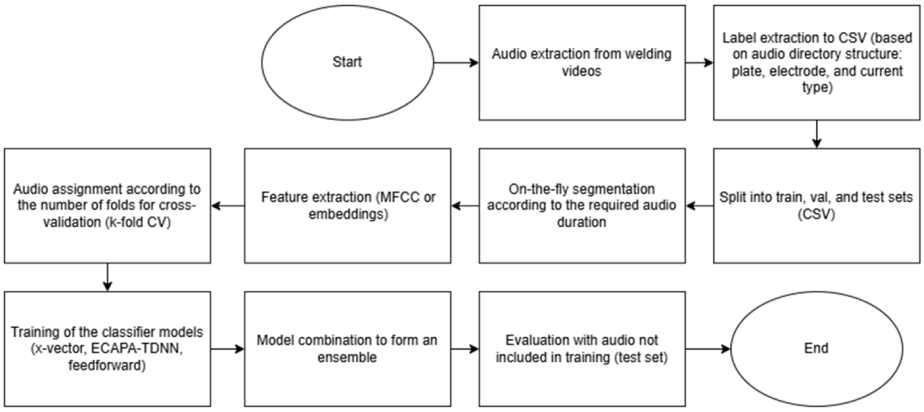


Figure 1. General pipeline for welding audio parameter detection, showing the sequential stages from raw audio acquisition and preprocessing through feature extraction (MFCC, VGGish, or YAMNet) to multi-output classification of plate thickness, electrode type, and current type.

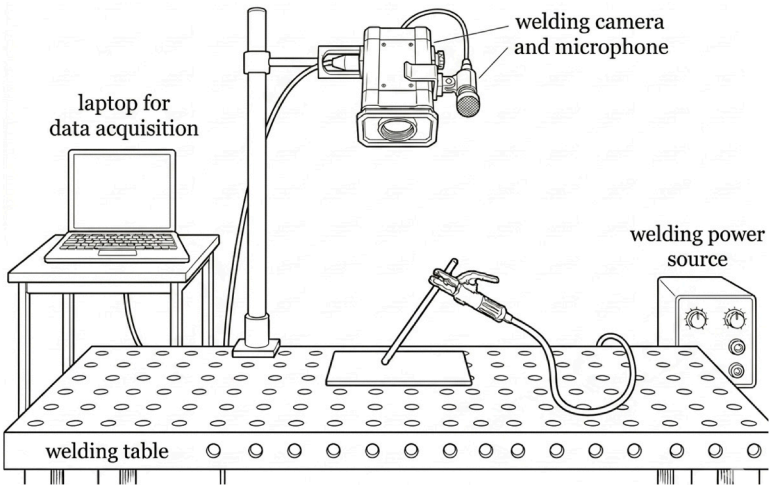


Figure 2. Experimental welding monitoring station. The setup comprises: a welding table with mechanical fixture holes for consistent plate alignment, a multipurpose welding power source with AC/DC capability, a Xiris weld camera and WeldMIC microphone for audiovisual data acquisition, and a laptop for real-time data logging and acquisition.

The recordings were acquired under realistic workshop conditions in the presence of heterogeneous background noise. Occasional impulsive and stationary noise from adjacent workshop activities, including hammering and angle grinding, were intermittently audible at the monitoring location, although these sources were not part of the monitored welding process. Additional noise sources included forced-air ventilation and fume-extraction systems operating continuously in the workshop, nearby welding stations and general workshop activity, as well as low-frequency acoustic coupling from the welding power

source. These sources contributed to the heterogeneous acoustic environment captured in the recordings and are further discussed in Section 4.2.

Table 1. Key specifications of the Xiris WeldMic used for acoustic data acquisition.

Parameter	Specification
Microphone Model	Xiris WeldMIC
Frequency Response	50 Hz–20 kHz
Dynamic Range	>120 dB SPL
Mounting configuration	Clamp/flexible arm
Distance to Arc	<0.8 m
Data format	Audio embedded in video (.WAV, 48 kHz, 16-bit)
Data acquisition software	Xiris WeldStudio 3

Audio was recorded as part of the synchronized system output using the Xiris WeldStudio 3 software and subsequently extracted for analysis. The extracted audio signals were resampled to 16 kHz and converted to mono format for further processing. A dedicated laptop was used for data acquisition, storage, and operation of the WeldStudio 3 software.

2.1.2. Materials

Welding specimens consisted of carbon steel plates measuring 150 mm in length and 50 mm in width with thicknesses of 3 mm, 6 mm, and 12 mm, selected to represent a range of practical welding conditions commonly encountered in structural and fabrication applications. SMAW consumables included E6010, E6011, E6013, and E7018 covered electrodes manufactured by KISWEL Co., Ltd. (Seoul, Republic of Korea), enabling the evaluation of different arc characteristics and metal transfer behaviors.

It should be noted that E6010 and E7018 do not support AC operation and were therefore tested under DC only, while E6011 and E6013 are compatible with both AC and DC and were evaluated under both configurations.

Welding currents were selected according to the manufacturer's specifications for each electrode type to ensure stable arc operation. Both AC and DC welding configurations were included in the experimental design to introduce controlled variability in the acoustic signature associated with different electrical operating conditions.

2.1.3. Welding Protocol

A dedicated Welding Procedure Specification (WPS) was issued for each electrode type prior to all welding trials, in accordance with AWS D1.1 structural welding guidelines. All procedures shared the same base metal specification (ASTM A36, AWS Group I), electrode diameter (3.2 mm, AWS A5.1), and welding position (1 G flat), and no post-weld heat treatment was applied.

Welding was performed as surface beads deposited on a single plate, with no joint preparation or groove geometry applied across all plate thicknesses (3 mm, 6 mm, and 12 mm). Welding was performed under DC or AC conditions, depending on electrode compatibility, following manufacturer-recommended current ranges as shown in Table 2.

Table 2. Recommended current ranges (3.2 mm electrode diameter) by electrode type, plate thickness and current type.

Electrode	Current Type	Plate Thickness		
		3 mm	6 mm	12 mm
E6010	DC	70–95 A	85–115 A	100–125 A
E6011	AC/DC	70–95 A	85–115 A	100–125 A
E6013	AC/DC	80–105 A	90–120 A	110–130 A
E7018	DC	90–110 A	100–125 A	110–140 A

Welding was performed by an experienced but non-certified welder, a condition intentionally selected to introduce natural execution variability and increase the diversity of acoustic patterns representative of realistic manual welding conditions. No corrective feedback was provided during welding to preserve natural process variability.

2.2. Acoustic Dataset

The dataset comprises continuous acoustic recordings collected during SMAW experiments. Each welding trial produced a raw audio signal spanning the full duration of the weld pass, averaging approximately one minute. The complete dataset comprises 373 continuous SMAW sessions, each corresponding to a single weld bead and serving as the atomic unit for dataset partitioning.

Sessions were organized in a session-based directory hierarchy by plate thickness, electrode type, and current type, following the structure: audio/Xmm/EXXXX/{AC,DC}/YYMMDD-HHMMSS_Audio/*.wav. This structure provides full traceability of each sample and consistent labeling for multi-output classification. Figure 3 shows a schematic representation of this directory layout, illustrating how each session folder maps directly to its corresponding combination of plate thickness, electrode type, and current type labels.



Figure 3. Directory structure of the SMAW audio dataset, illustrating the hierarchical organization by plate thickness (Xmm), electrode type (EXXXX), and current type (AC/DC), with individual session folders identified by acquisition timestamp.

2.2.1. Audio Preprocessing

The original recordings were organized as session-based WAV files and stored as mono, 16-bit audio at 16 kHz, and labeled according to the directory structure. For all experiments, audio was processed using a unified preprocessing pipeline consisting of: (i) audio loading, (ii) resampling to 16 kHz when needed, (iii) conversion to a single channel (mono), and (iv) on-the-fly segmentation of each continuous recording into fixed-duration windows with a predefined overlap ratio. This standardized segmentation ensured that MFCC extraction and deep embedding models operated on exactly the same segment boundaries with VGGish and YAMNet subsequently applying their own internal framing within each segment. No filtering, denoising, silence trimming, or amplitude normalization was applied to the welding audio beyond the standard waveform scaling during audio loading.

This minimally processed workflow was intentionally selected to preserve welding-related acoustic variability under realistic workshop conditions.

2.2.2. Audio Segmentation

Continuous welding audio recordings were segmented into fixed-length temporal windows. Seven segment durations were evaluated: 1, 2, 5, 10, 20, 30, and 50 s. Segment

duration directly influences the number of training samples, the amount of temporal information captured, and the computational cost of model training and inference.

A 50% overlap ($p = 0.5$) was used as the baseline segmentation configuration to increase the number of training samples while maintaining sufficient variability between adjacent segments. Table 3 summarizes the total number of segments generated for each segment duration under the baseline overlap configuration, across training, validation, and test sets. The use of overlap also enables data augmentation, as discussed in Section 2.2.3.

Table 3. Total segment counts per duration for training, validation, and test sets at $p = 0.5$ overlap.

Duration (s)	Training Segments	Validation Segments	Test Segments	Total
1	30,176	8006	4988	43,170
2	14,898	3950	2465	21,313
5	5713	1521	951	8185
10	2660	712	447	3819
20	1138	303	199	1640
30	635	170	113	918
50	305	84	59	448

2.2.3. Overlap-Based Data Augmentation

Window overlap was applied as a data augmentation strategy to increase the number of training samples while preserving the original signal characteristics. For a window of duration T and an overlap ratio p, the hop between consecutive windows is defined as $H = T(1 - p)$. This approach increases the number of samples and improves the temporal coverage of transitions without distorting the original signal, thereby stabilizing training for underrepresented classes.

Because overlap creates correlated samples from the same welding session, dataset partitioning was performed at the session level rather than the segment level to prevent data leakage between the training, validation, and test sets.

2.2.4. Labeling

The recorded audio signals were segmented and labeled according to the welding condition associated with each welding session. Labels were assigned based on known welding parameters and post-process inspection of the weld bead, ensuring consistency between acoustic data and process conditions. The resulting label distribution across all sessions is summarized in Table 4.

Table 4. Cross-tabulation of welding sessions across electrode type, current type, and plate thickness.

Electrode	Current	3 mm	6 mm	12 mm	Subtotal
E6010	DC	16	21	20	57
E6011	AC	22	21	20	63
	DC	29	20	20	69
E6013	AC	20	20	20	60
	DC	20	20	20	60
E7018	DC	20	24	20	64
Total		127	126	120	373

The target labels for each session were defined by three variables: plate thickness, electrode type, and current type. Plate thickness was categorized into three levels (3 mm,

6 mm, and 12 mm), while electrode type included E6010, E6011, E6013, and E7018 consumables. Current type was defined as either AC or DC; AC operation was applied only to E6011 and E6013, whereas E6010 and E7018 were evaluated under DC conditions only due to process compatibility constraints. Consequently, the dataset contains structured dependencies between electrode type and current type, reflecting valid SMAW operating conditions rather than independently sampled label combinations.

Each audio segment inherits the labels of the welding session from which it was extracted, resulting in a multi-output classification problem where each input sample is associated with three target labels simultaneously.

Because the proposed framework jointly predicts plate thickness, electrode type, and current type, the joint label distribution across valid operating conditions is relevant for interpreting class balance and label dependency within the dataset. Table 4 summarizes the complete cross-tabulation of the welding sessions for all evaluated parameter combinations.

2.2.5. Dataset Partitioning

Dataset partitioning was performed at the welding session level to prevent data leakage between subsets and to ensure acoustically independent welding trials in each split [21]. A stratified partitioning strategy was applied to preserve the class distribution across plate thickness, electrode type, and current type. A fixed random seed of 42 was used to ensure deterministic, reproducible splits.

A 10% test set was reserved for final model evaluation, while the remaining 90% of the sessions were further split into 80% for training and 20% for validation. This results in 267 training sessions, 68 validation sessions, and 38 test sessions.

Segmentation and overlap-based data augmentation were performed after dataset partitioning to ensure that segments derived from the same welding session were not distributed across different subsets. For the reference segment duration of 5 s, this corresponds to 5713 training segments, 1521 validation segments, and 951 test segments.

2.3. Feature Extraction Approaches

Three approaches were evaluated and compared, one based on spectral analysis (MFCC) and two based on audio embeddings (VGGish, YAMNet) extracted from welding audio recordings.

2.3.1. Spectral Analysis (MFCC)

As a classical spectral baseline, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted. MFCCs model human auditory perception through a non-linear frequency scale (the Mel scale), where perceptually equidistant intervals correspond to equal Mel-scale increments, capturing acoustically relevant patterns concentrated in lower frequency bands [22,23].

MFCC extraction was performed using the librosa audio signal analysis Python 3.10.12 library with consistent parameters across all experiments. Preprocessed audio signals (Section 2.2.1) were used as input to ensure consistency across all feature extraction approaches.

A frame length of 1024 samples (64 ms) and a hop length of 512 samples (32 ms) were used. A Hamming window was applied to each frame prior to the Fourier Transform. MFCCs were computed from a Mel filter bank representation using 40 Mel filters, and a total of 40 MFCCs were extracted per frame. To capture the temporal dynamics, first-order (delta) and second-order (delta-delta) coefficients were computed from the base MFCCs, representing the rate of change and acceleration of the spectral envelope over time, respectively [24,25].

The resulting feature representation consists of three components (static MFCCs, delta, and delta-delta), each of dimension 40, yielding a 120-dimensional frame-level descriptor.

For the Feedforward classifier, temporal statistics (mean and standard deviation) were computed across all frames within each audio segment, producing a final 240-dimensional feature vector. For the X-Vector [26] and ECAPA-TDNN [27] architectures, the raw MFCC sequence of shape 40 × T was preserved without statistical aggregation, as these architectures incorporate internal temporal pooling mechanisms that operate directly on the frame-level sequence.

2.3.2. VGGish

VGGish is a Convolutional Neural Network (CNN) based on the VGG architecture [28], pre-trained on the large-scale Audio Set database [17]. The model is designed to extract high-level semantic audio features from general-purpose audio signals. As illustrated in Figure 4, the VGGish model processes audio through successive convolutional and pooling layers before producing a compact embedding representation.

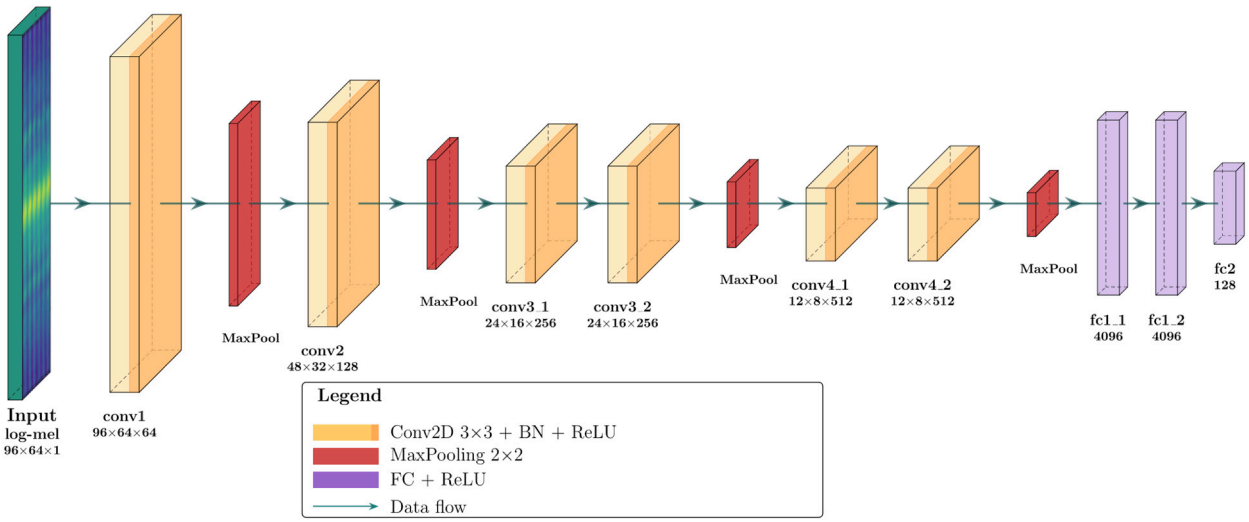


Figure 4. Internal layer structure of the VGGish convolutional neural network, showing the successive convolutional, pooling, and fully connected stages that transform the input log-mel spectrogram into a 128-dimensional embedding vector. Yellow boxes represent conv + ReLU layers and red boxes represent max pooling layers.

VGGish operates on log-Mel spectrogram inputs computed from the preprocessed audio signal (Section 2.2.1). Following the standardized Audio Set pipeline, which includes short-time Fourier Transform, Mel filter bank projection, and logarithmic scaling.

In this study, VGGish was used as a fixed feature extractor. The model processes audio in short-time windows of 0.96 s, generating a 128-dimensional embedding vector per window [28]. For audio segments longer than 0.96 s, a temporal sequence of embeddings is produced. These embeddings were aggregated using temporal average pooling to obtain a fixed-size feature vector suitable for input to the multi-output ANN classifier.

Only audio segments with a duration greater than 1 s were considered to ensure that at least one complete VGGish analysis window could be extracted. This constraint guarantees consistent embedding generation across all samples and enables a fair comparison with MFCC-based and YAMNet-based representations.

2.3.3. YAMNet

YAMNet was evaluated as an alternative to second pre-trained embeddings. It is built on the mobile-optimized MobileNet v1 backbone [29], which is also pre-trained on the large-scale Audio Set dataset [17]. As shown in Figure 5, the network uses depth wise separable convolution layers, which factorize standard convolutions into depth wise and

pointwise operations, significantly reducing the number of parameters and computational cost while maintaining high representational capacity.

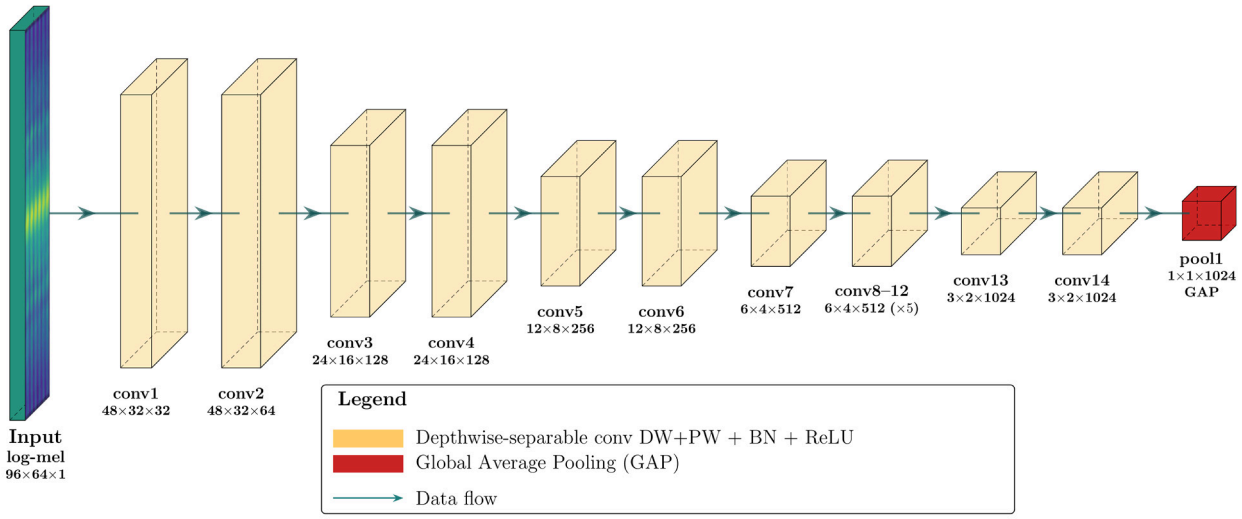


Figure 5. YAMNet architecture, showing the sequence of 14 depth wise separable convolutional layers (conv1 through conv14) that process the 96 × 64 log-mel spectrogram input down to a 1 × 1 × 1024 embedding vector via a final global average pooling stage (pool1).

Similar to VGGish, YAMNet operates on log-Mel spectrogram representations computed from processed audio signals (Section 2.2.1). For each analysis window, the network produces a 1024-dimensional embedding vector providing a higher descriptive capacity compared to the 128-dimensional VGGish embedding at the cost of increased storage requirements, feature extraction time, and classifier training time.

In this work, YAMNet was used as a fixed extractor. For classifiers requiring fixed-size inputs (e.g., feedforward networks), embeddings were aggregated using temporal average pooling. For sequence-based architectures (X-vector and ECAPA-TDNN), the embedding sequence was preserved to enable the model to learn temporal dependencies directly from it. The same experimental protocol, including segment durations, overlap values, and dataset partitioning strategy, was applied to YAMNet to ensure direct comparability with MFCC and VGGish representations.

2.4. Classification Models

The acoustic representations produced by each feature extraction approach (MFCC, VGGish, and YAMNet) were used as inputs to multi-output Artificial Neural Network (ANN) classifiers designed to simultaneously predict the three target variables: plate thickness, electrode type and current type. The problem is formulated as a multi-output (also referred to as multi-target) classification task, in which each input segment is associated with three categorical targets and exactly one class label per target. This formulation is distinct from multi-label classification, where multiple labels may be assigned to a single target, and from single-task classification, where only one target is predicted per model. The shared-backbone, task-specific-head architecture described below is consistent with the multi-task learning paradigm; throughout this paper, the term multi-output is used to describe the prediction task itself, while multi-task is reserved for references to the broader learning literature.

Three classifier architectures were evaluated: a Feedforward Network (FFNN), an X-vector, and an ECAPA-TDNN model. These architectures were selected to compare conventional fixed-vector classifiers with a sequence-based model that captures temporal dependencies in acoustic signals. The Feedforward network operates on fixed-size fea-

ture vectors obtained either directly from aggregated embeddings (mean and standard deviation) or from MFCC-based statistical descriptors. It consists of fully connected layers with non-linear activations and serves as a computationally efficient baseline [30,31]. The X-Vector model employs a Time Delay Neural Network (TDNN) architecture with statistical pooling layers that transform variable-length temporal sequences into fixed-size representations. Originally developed for speaker recognition, this architecture is well suited for extracting discriminative patterns from acoustic signals with temporal variability [26]. The ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation) model extends the X-vector architecture by incorporating channel attention mechanisms and improved feature aggregation, enabling more effective modeling of complex temporal dependencies in audio signals [29].

The classification models architecture details can be found in Table 5. T denotes the number of temporal frames produced by the feature extractor for a given audio segment. For MFCC, frames are computed with a 64 ms analysis window and a 32 ms hop (n_fft = 1024, hop = 512 at 16 kHz), yielding $T = 32$ frames per 1 s segment and $T = 313$ frames for a 10 s segment. For VGGish and YAMNet, embeddings are extracted from 1 s windows with a 500 ms hop, yielding $T = 2$ frames per 1 s segment and $T = 20$ frames per 10 s segment. The temporal dimension T therefore varies with both segment duration and feature extraction strategy, allowing the TDNN-based architecture to process sequential acoustic context at different temporal resolutions depending on the selected representation.

Table 5. Summary of the evaluated classification model architectures and principal configuration parameters.

Parameter	Feedforward	X-Vector	ECAPA-TDNN
Input dimension (MFCC)	240	40 × T	40 × T
Input dimension (VGGish)	128	128 × T	128 × T
Input dimension (YAMNet)	1024	1024 × T	1024 × T
Number of layers	3 Fully Connected layers	5 TDNN layers	3 SE-Res2Net blocks
Hidden dimension	512, 256, 128	512	512
Temporal context (TDNN)	N/A	[−2, +2], [−2, +2], [−3, +3] frames	dilation 2, 3, 4
Attention mechanism	None	None	Channel attention (SE)
Temporal pooling	Pre-computed statistics	Statistics-based temporal pooling	Attention-based statistics pooling
Output heads	3 (parallel FC)	3 (parallel FC)	3 (parallel FC)
Activation	ReLU	ReLU	ReLU
Batch normalization	Yes	Yes	Yes
Dropout rate	0.3	None	0.1
Optimizer	AdamW (weight_decay = 1×10^{-4})		
Initial learning rate	1×10^{-3}		
Batch size	32		
Maximum epochs	100		
Early stopping patience	15 epochs on validation loss		
Loss function (each head)	Cross-entropy (label_smoothing = 0.1)		
Cross-validation folds (k)	10		
Random seed	42		

All models were implemented using a multi-output classification framework, with a shared feature extractor followed by three task-specific output heads for plate thickness, electrode type, and current type. Each output head performs classification over its respective set of classes (3, 4 and 2 classes, respectively).

The same training configuration and evaluation protocol were applied across all model-feature combinations to ensure a fair comparison between feature extraction approaches and classifier architectures.

The objective of this study is not to propose a novel multi-output architecture, loss function, or multi-task optimization strategy. Rather, the proposed framework was intentionally designed to isolate the effect of the acoustic representation itself under a unified classification setting.

2.5. Model Training (K-Fold Ensemble)

Model training was performed using a stratified, group-aware k-fold cross-validation strategy, where groups correspond to individual welding sessions. This ensures that all segments derived from a given session are assigned to the same fold, preventing data leakage and preserving independence between training and validation subsets.

A hold-out test set, comprising 10% of the total welding sessions, was reserved prior to cross-validation and used exclusively for final model evaluation. The remaining 90% of the dataset was used within the k-fold cross-validation framework for model training and validation.

In each fold, a model of the same architecture was trained on a different subset of the training data. The resulting set of k models forms an ensemble, where diversity arises from variations in the training data across folds [32,33]. Several values of k were evaluated, with k = 10 selected as the baseline configuration.

During inference, the final prediction for each input segment was obtained by averaging the output logits of all 10 models in the ensemble (ensemble averaging). The class with the highest averaged logit value was selected as the predicted label for each target variable.

2.6. Evaluation Metrics

Performance was assessed using both per-target classification metrics and joint metrics to fully characterize the classifiers' behavior in the multi-output setting. The evaluation framework is divided into two complementary levels. First, individual metrics are computed independently for each target variable to assess per-target classification performance, following standard machine learning practices [25]. Second, joint evaluation metrics, originally formalized in the multi-label learning literature [32,33] and applicable to multi-output classification with categorical targets, are employed to assess the model's ability to correctly predict all outputs simultaneously. This dual-level evaluation enables a comprehensive analysis of both independent target performance and overall multi-output prediction capability.

2.6.1. Individual Evaluation Metrics

For individual metrics, performance was evaluated independently for each classification target (plate thickness, electrode type, and current type). Let c denote a specific class within a target, C the total number of classes for that target, and N the total number of evaluated samples. TP_c, FP_c, and FN_c represent the true positives, false positives, and false negatives for class C, respectively.

Accuracy is defined as the proportion of correctly classified samples.

$$\text{Accuracy} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} 1(\widehat{y_{i,t}} = y_{i,t})$$

Precision and recall for each class are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c},$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

The Macro F1-score is computed as the unweighted average of the per-class F1-scores, reducing bias toward majority classes in imbalanced datasets [30].

$$F1_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

2.6.2. Joint Evaluation Metrics

Since the model simultaneously predicts multiple welding process variables, joint evaluation metrics were used to assess overall multi-output performance. Let i index individual samples and $l \in \{1, \ldots, \text{L}\}$ index the target variables, where $L = 3$ corresponds to the plate thickness, electrode type, and current type). Let y_i and $\hat{y}_i$ denotes the true and predicted label vector for sample i.

The Exact Match Ratio (also referred to as subset accuracy) measures the proportion of samples for which all target variables are correctly predicted simultaneously:

$$\text{Exact Match} = \frac{1}{N}\sum_{i=1}^{N} 1(\widehat{y_i} = y_i)$$

The Hamming Accuracy evaluates the fraction of correctly predicted labels across all targets:

$$\text{Hamming Accuracy} = \frac{1}{N \cdot L}\sum_{i=1}^{N}\sum_{l=1}^{L} 1(\widehat{y_{il}} = y_{il})$$

While Exact Match provides a strict evaluation criterion requiring all target variables to be predicted correctly simultaneously, Hamming Accuracy offers a more granular measure by accounting for partial prediction correctness across individual targets. Together, these metrics provide complementary perspectives on overall multi-output prediction consistency and per-target classification behavior [34,35].

2.7. Computational Performance Evaluation

Computational performance was evaluated in terms of model training time and inference time for each feature extraction approach and classifier architecture.

Model training and inference were executed on the GPU, and inference time was computed as the average processing time per segment using a fixed batch size of 32. All evaluations were performed using 5 s audio segments with 50% overlap and a 10-fold ensemble configuration.

Experiments were conducted on a desktop workstation equipped with an AMD Ryzen 5 3600 CPU, and NVIDIA GeForce RTX 3070 GPU, 32 GB of DDR4 RAM, and a Samsung 970 EVO 1 TB SSD, running Windows 11 Pro.

3. Results

The experimental evaluation covered a comprehensive set of configurations defined by the combination of three acoustic representation approaches (MFCC, VGGish, and YAMNet), three classifier architectures (Feedforward, X-Vector, and ECAPA-TDNN), seven segment durations (1, 2, 5, 10, 20, 30, and 50 s), four overlap ratios (p = 0, 0.25, 0.5, and 0.75),

and eight ensemble sizes (k = 1, 2, 3, 5, 7, 10, 15, and 20 folds). All configurations were evaluated using the held-out test set.

From this experimental grid, a baseline configuration was selected using a segment duration of 5 s, an overlap ratio of $p = 0.5$, and k = 10 cross-validation folds. This combination provided a balance between classification performance, training stability, temporal resolution, and computational cost. Unless otherwise stated, all comparisons presented in the following sections are based on this baseline configuration.

3.1. Individual Metrics

The following sections analyze individual classification performance in terms of accuracy and Macro F1-score across the three target variables. While all configurations are reported, particular attention is given to models achieving performance above 90% as these represent the most relevant candidates for robust SMAW process characterization.

3.1.1. Accuracy

Figure 6 shows the classification accuracy as a function of segment duration for the three target variables (plate thickness, electrode type, and current type) across all feature extraction approaches and classifier architectures.

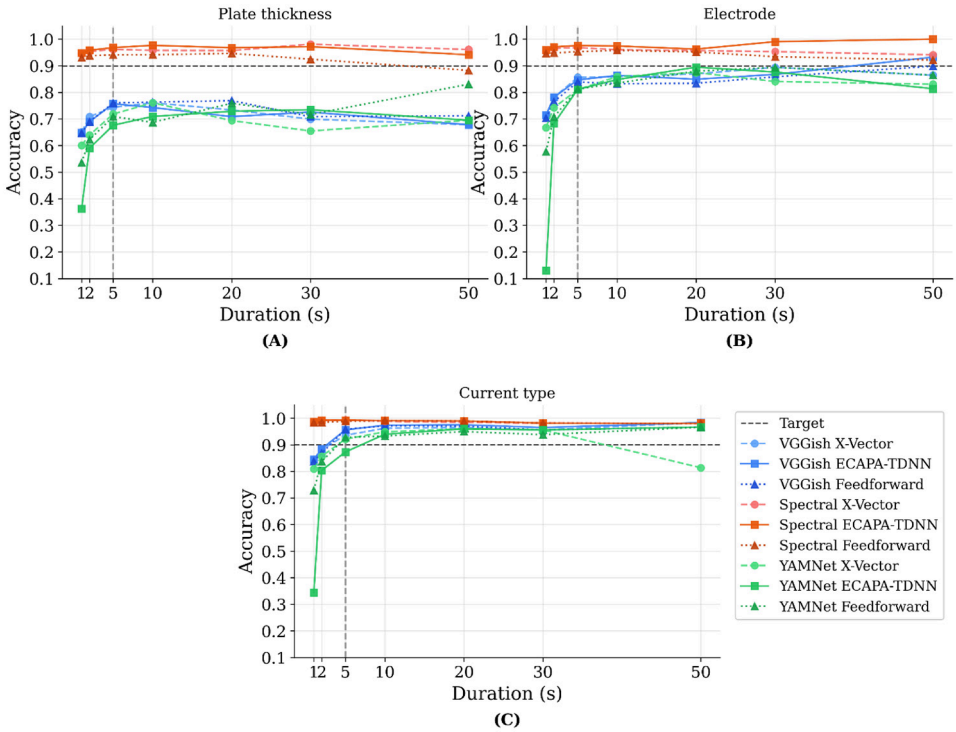


Figure 6. Accuracy as a function of segment duration (1 to 50 s) for all nine approach-architecture combinations at the baseline configuration ($k = 10$, $p = 0.5$). (**A**) Plate Thickness. (**B**) Electrode Type. (**C**) Current Type.

Across all targets, MFCC-based models consistently achieve the highest accuracy, maintaining performance above 90% for most segment durations. The lowest performance is observed at 1 s, indicating that very short segments do not provide sufficient temporal information for reliable classification. At the baseline configuration (5 s), the best performing combination is MFCC + ECAPA-TDNN, which consistently outperforms embedding-based approaches.

In contrast, VGGish-based models exhibit lower accuracy overall, with performance increasing as segment duration increases, suggesting that longer temporal context is re-

quired to stabilize the learned embeddings. YAMNet-based models exhibit the greatest variability in segment durations, suggesting sensitivity to temporal windowing despite their higher-dimensional representation.

Among the three classification targets, the Current Type (AC/DC) task achieves the highest accuracy across all configurations, with several MFCC-based models reaching near-perfect performance at the baseline duration. This behavior reflects the strong, systematic acoustic differences induced by alternating and direct current models.

The Plate Thickness task exhibits intermediate performance, while the Electrode Type task presents the lowest accuracy values and the highest inter-class variability. This is likely due to the similarity in coating composition and arc behavior among certain electrode types, resulting in more subtle acoustic differences that are harder to discriminate.

Overall, these results demonstrate that the MFCC-based representation provides more stable and discriminative features for SMAW acoustic classification than deep audio embeddings, particularly for limited segment durations.

3.1.2. F1-Score

Figure 7 presents the Macro F1-score as a function of segment duration for the three classification targets. In contrast to accuracy, the Macro F1-score provides a more balanced evaluation by accounting for class imbalance, particularly relevant for the electrode type and current type variables.

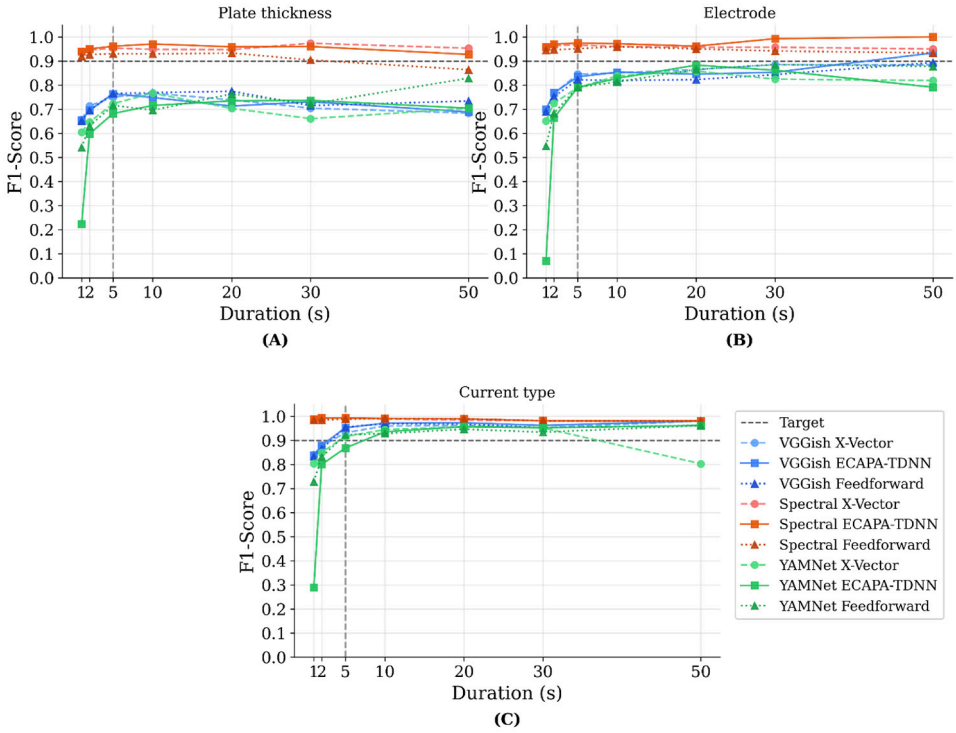


Figure 7. Macro F1-score as a function of segment duration (1 to 50 s) for all nine approach-architecture combinations at the baseline configuration ($k = 10$, $p = 0.5$). (**A**) Plate Thickness. (**B**) Electrode Type. (**C**) Current Type.

Across all targets and segment durations, MFCC-based models consistently achieve the highest F1 scores, maintaining values above 90% in most configurations. At the baseline duration of 5 s, the MFCC + ECAPA-TDNN combination yields the best performance among all evaluated models.

Embedding-based approaches (VGGish and YAMNet) exhibit lower F1 scores overall, with performance improving as segment duration increases. However, even longer

durations remain below the MFCC-based models for most configurations, indicating that handcrafted spectral features provide more robust class discrimination under class imbalance conditions.

Among the three targets, the Current Type task achieves the highest F1 score across all models, with several MFCC-based configurations reaching near-perfect values. The Plate Thickness task shows intermediate performance, while the Electrode Type task presents the lowest F1 scores, reflecting the increased difficulty of distinguishing between electrodes with similar coating characteristics and arc behavior.

These results reinforce the trends observed in the accuracy analysis and demonstrate that MFCC-based representations not only achieve higher accuracy but also provide more balanced and stable classification performance across classes.

3.1.3. Confusion Matrix

To further characterize the classification behavior under the baseline configuration, normalized confusion matrices were computed on the test set for the best-performing model (MFCC + ECAPA-TDNN at 5 s, k = 10, p = 0.5). Figure 8 presents the normalized confusion matrices for Plate Thickness, Electrode Type, and Current Type in panels (A), (B), and (C), respectively.

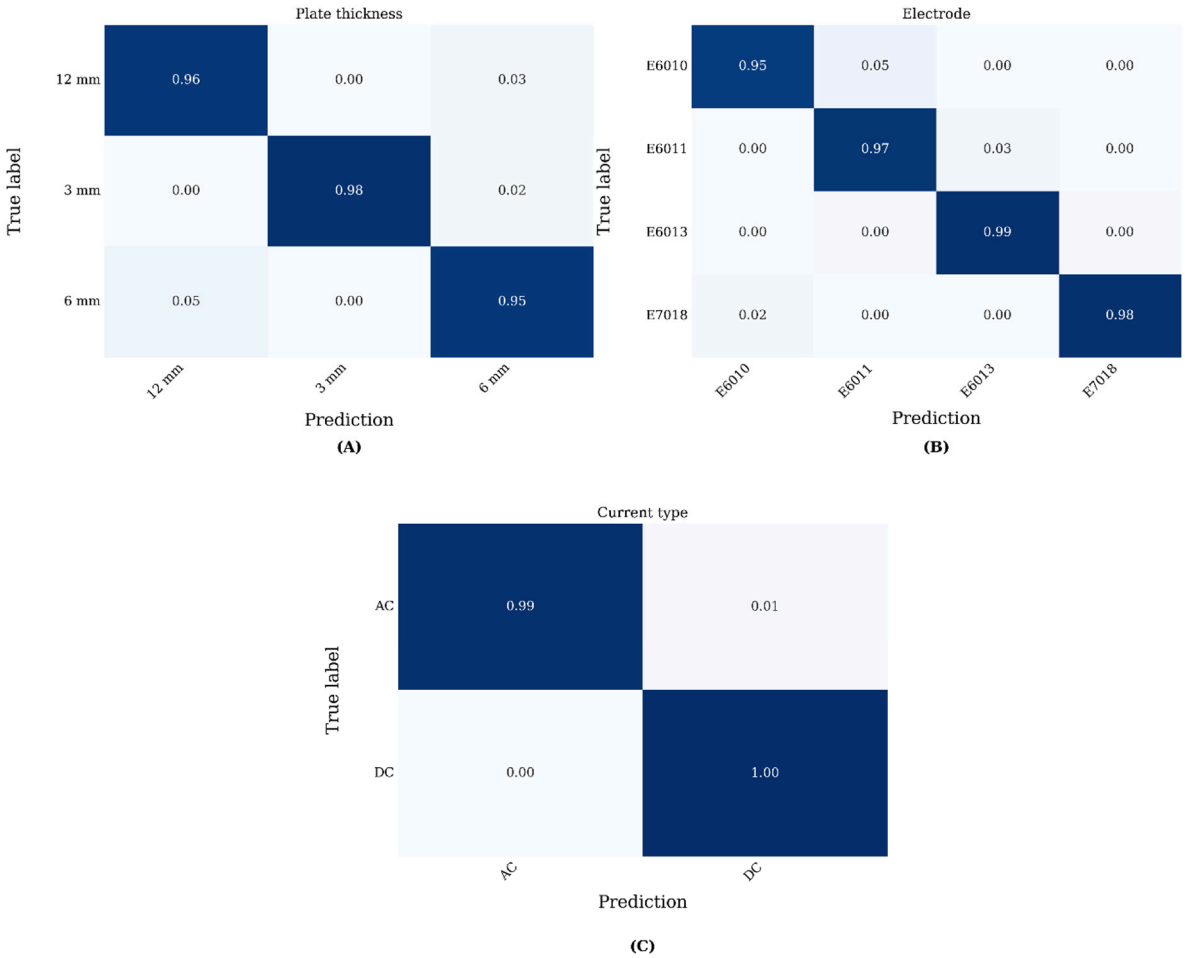


Figure 8. Normalized confusion matrices on the test set for the MFCC + ECAPA-TDNN model at the baseline configuration (5 s, k = 10, p = 0.5). (**A**) Plate Thickness (3 mm, 6 mm, 12 mm). (**B**) Electrode Type (E6010, E6011, E6013, E7018). (**C**) Current Type (AC, DC). The color intensity represents the normalized proportion of samples, with darker shades indicating higher values.

For the plate thickness classification task, high diagonal values across all classes (3 mm, 6 mm, 12 mm) indicate strong discriminative performance. The most frequent misclassifications occur between 3 mm and 6 mm plates, suggesting that these thicknesses produce a more similar acoustic signature compared to the 12 mm conditions.

In the electrode type classification task, the highest confusion is observed between E6010 and E6011, which is consistent with their similar cellulosic coating composition and arc behavior, leading to comparable metal transfer dynamics and acoustic signatures [36]. In contrast, E6013 and E7018 exhibit clearer separation, indicating more distinct acoustic signatures.

For the current type of classification task, near-perfect separation is observed between AC and DC conditions, with minimal misclassification. This result is consistent with the physical characteristics of the SMAW process, where alternating current introduces a periodic oscillatory component associated with arc current reversal (typically around 50–60 Hz), resulting from cyclic arc extinction and re-ignition, and producing a distinctive acoustic pattern, that facilitates reliable discrimination [3,37].

Overall, the confusion matrix analysis confirms that classification performance is strongly influenced by the physical differences between welding conditions, with greater separability observed for variables that induce more pronounced changes in arc dynamics and acoustic emissions.

3.2. Time-Domain Signal Inspection

To provide qualitative insight into the acoustic differences between welding conditions prior to feature extraction, representative short waveform fragments were examined directly in the time domain. Figure 9 shows 200 ms excerpts of raw audio signals under four representative welding conditions (A) E6011-AC-6 mm, (B) E6010-DC-6 mm, (C) E6010-DC-6 mm, (D) E7018-DC-6 mm.

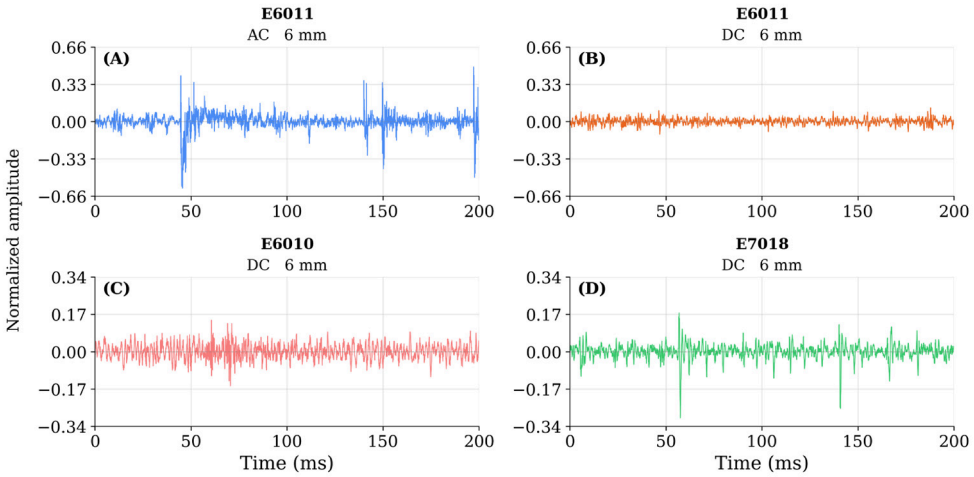


Figure 9. Representative 200 ms waveform segments from welding arcs. (**A**,**B**) E6011/AC vs. E6011/DC: AC exhibits denser activity with amplitude modulation from 50 Hz re-ignition; DC is more uniform and lower amplitude. (**C**,**D**) E6010/DC vs. E7018/DC: E6010 shows sustained high-density crackling (cellulosic), while E7018 produces a sparser, lower-energy signal (stable low-hydrogen arc). Amplitudes normalized to full-scale range; rows share common vertical scale for direct comparison.

Panels (A) and (B) compare the effect of current type under the same electrode and plate thickness conditions. The AC waveform in panel (A) exhibits more pronounced low-frequency amplitude modulation and stronger transient excursions, consistent with the periodic arc extinction and re-ignition behavior characteristic of alternating current

welding. in contrast, the DC waveform in panel (B) appears more continuous and stationary, with reduced amplitude fluctuation over time.

Panel (C) and (D) compare the effect of electrode type under the same current mode and plate thickness conditions. The E6010 waveform in panel (C) exhibits a more irregular and impulsive structure, consistent with its more aggressive arc behavior, whereas the E7018 waveform in panel (D) appears comparatively smoother and more stable.

These waveform-level differences provide qualitative evidence that some welding conditions are already partially distinguishable directly in the time domain. However, although these differences are visually observable, raw waveform inspection alone is not sufficiently robust for reliable classification under realistic workshop conditions. Feature extraction remains necessary to obtain stable and discriminative representations for automated classification.

3.3. Feature Space Analysis

To further examine the intrinsic separability of the evaluated feature representations, a feature space analysis was conducted using t-distributed Stochastic Neighbor Embedding (t-SNE) on the test-set feature representations produced by each approach under the baseline configuration (5 s segments, $p = 0.5$). The t-SNE projections were generated using a perplexity of 30 and 1000 iterations. Figure 10 presents the resulting two-dimensional projections color-coded by electrode type (top row), current type (middle row), and plate thickness (bottom row) for MFCC, VGGish, and YAMNet features, respectively.

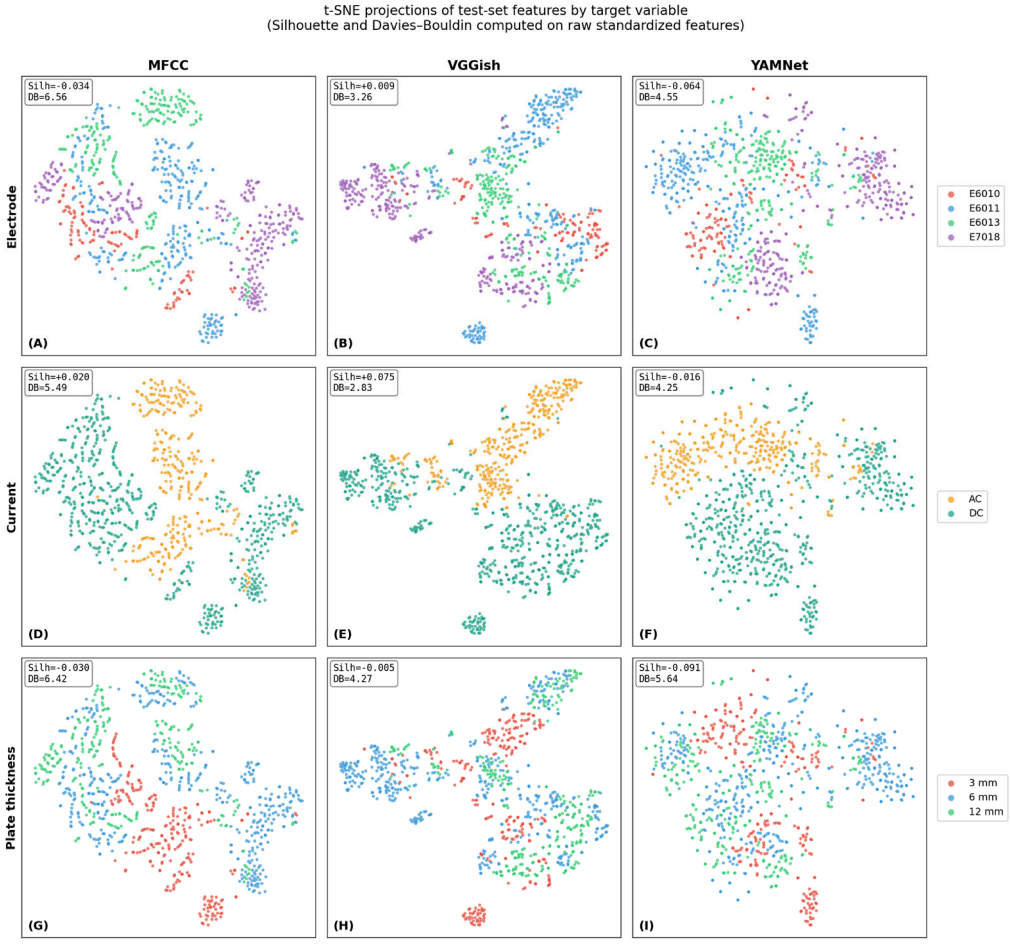


Figure 10. t-SNE visualizations of test-set feature representations for (**A**–**C**) electrode type, (**D**–**F**) current type, and (**G**–**I**) plate thickness using MFCC (**A**,**D**,**G**), VGGish (B,E,H), and YAMNet (**C**,**F**,**I**) features. Each point represents a 5 s audio segment. Compact, well-separated clusters in the MFCC space confirm the inherent discriminability of spectral features for SMAW parameter classification.

MFCC-based features form more structured and compact distributions, particularly for the current type, where a clearer separation between AC and DC samples is observed. In contrast, VGGish and YAMNet embeddings exhibit substantially greater intra-class spread and inter-class overlap, especially for electrode type, indicating weaker intrinsic separability in the raw feature space. This difference in feature organization is evident before classifier inference, indicating that the advantage of MFCC is inherent to the representation and not solely an artifact of the ECAPA-TDNN architecture.

To complement the visual analysis, the silhouette coefficient and Davies-Bouldin index were computed on the original standardized feature spaces and are reported in Figure 10. Although the absolute values remain modest due to the partial overlap and inherent similarity between welding classes, the overall trend is consistent with the visual projections. MFCC yields a more stable, task-aligned feature organization, whereas embedding-based representations produce less compact, more entangled clusters.

To further assess whether this advantage emerges only under the ECAPA-TDNN architecture, additional comparisons were performed using classifiers of increasing complexity on the same test set. As shown in Figure 11, MFCC maintains a measurable advantage over VGGish and YAMNet even with linear and shallow non-linear classifiers, while the performance gap widens substantially larger with ECAPA-TDNN. This indicates that the superiority of MFCC stems from the feature representation and is amplified, rather than created by the classifier.

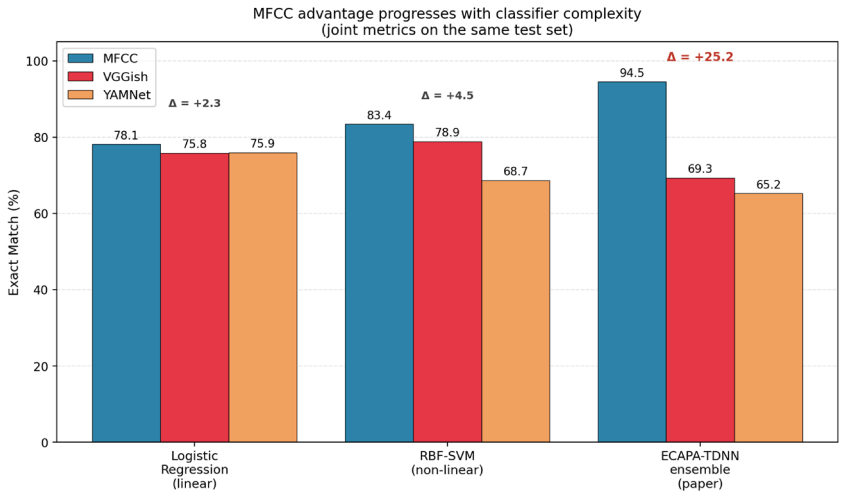


Figure 11. Exact Match comparison across classifiers of increasing complexity using the same test set. MFCC maintains a measurable advantage over VGGish and YAMNet under linear and shallow non-linear classifiers, while the performance gap increases substantially with ECAPA-TDNN, indicating that MFCC superiority originates in the feature representation and is amplified by classifier complexity. Red Δ values indicate the Exact Match performance gap (percentage points) between MFCC and the best embedding-based representation, highlighting the more than 25 percentage point advantage observed for the ECAPA-TDNN classifier.

These results provide additional evidence for the interpretation discussed in Section 4.3, in which the performance gap is examined in terms of domain mismatch between general-purpose audio embeddings and welding-specific acoustic signals.

3.4. Joint Metrics

3.4.1. Multi-Output Performance

The joint evaluation on the test set under the baseline configuration reveals a clear and consistent performance gap in the acoustic representation. Figure 12 presents the Exact Match and Hamming Accuracy results for all feature–architecture combinations.

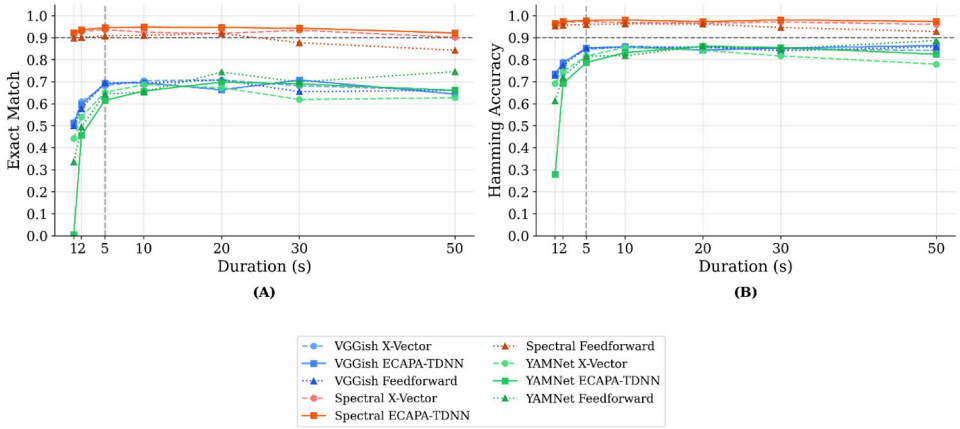


Figure 12. Joint evaluation metrics for all tested approach-architecture combinations at the baseline configuration (5 s, k = 10, p = 0.5). (**A**) Exact Match grouped by acoustic approach (MFCC, VGGish, YAMNet), with individual bars for each classifier architecture (Feedforward, X-Vector, ECAPA-TDNN). (**B**) Hamming Accuracy for the same groupings.

Across all models, MFCC-based representation significantly outperforms deep audio embeddings in both joint metrics. The best-performing configuration (MFCC + ECAPA-TDNN) achieves 94.51% Exact Match and 97.88% Hamming Accuracy, indicating highly consistent simultaneous prediction across all three welding variables.

In comparison, VGGish-based models reach a maximum of 69.30% Exact Match and 85.31% Hamming Accuracy, while YAMNet-based models achieve 65.19% and 81.63%, respectively. This corresponds to a performance gap of more than 25 percentage points in Exact Match between MFCC and the best embedding-based approaches.

Within each feature extraction approach, differences between classifier architectures are relatively small. For example, MFCC-based models vary by only 3.6 percentage points in Exact Match, while VGGish-based models vary by 1.4 points. This indicates that the choice of acoustic representation has a significantly greater impact on performance than the selection of classifier architecture.

The difference between Exact match and Hamming Accuracy further highlights the challenge of multi-output prediction. While Hamming Accuracy remains relatively high across all approaches, Exact Match reveals that embedding-based models struggle to consistently predict all three variables simultaneously.

These results demonstrate that MFCC-based representations provide more reliable and consistent multi-variable classification, making them better suited for practical SMAW monitoring applications where simultaneous identification of welding conditions is required.

3.4.2. Stratified Joint Performance Analysis

To account for the structural dependency between electrode type and current type labels, Table 6 reports Exact Match and Hamming Accuracy separately for DC-only sessions (E6010 and E7018) and dual-mode sessions (E6011 and E6013 evaluated under both current types). Across both subgroups, MFCC consistently outperformed embedding-based features, with ECAPA-TDNN achieving the highest Exact Match in DC-only sessions (95.71%) and dual-mode sessions (93.91%), indicating that this advantage is not solely explained by label dependency.

Embedding-based features, however, showed a pronounced performance drop in DC-only sessions, with VGGish and YAMNet exhibiting reductions of approximately 20 percentage points in Exact Match relative to dual-mode conditions. This behavior suggests that these representations are less effective at capturing the welding-specific acoustic characteristics associated with E6010 and E7018 weld signatures.

Table 6. Stratified classification performance by feature extractor and electrode subgroup. Exact Match and Hamming Accuracy are reported separately for DC-only sessions (E6010 and E7018) and dual-mode sessions (E6011 and E6013 evaluated under both AC and DC conditions).

Approach	Session Subgroup	X-Vector		ECAPA-TDNN		Feed Forward	
		Exact Match (%)	Hamming Acc. (%)	Exact Match (%)	Hamming Acc. (%)	Exact Match (%)	Hamming Acc. (%)
MFCC	DC-only	93.40	96.92	95.71	97.91	90.76	95.71
	Dual-mode	93.91	97.70	93.91	97.86	90.95	96.22
VGGish	DC-only	56.75	75.50	59.00	77.42	56.75	75.92
	Dual-mode	76.41	91.05	76.77	91.05	78.95	92.01
YAMNet	DC-only	53.25	72.17	47.75	65.08	53.75	72.17
	Dual-mode	73.87	88.51	71.51	88.51	72.23	88.45

While the structural constraint between electrode and current type reduces the effective difficulty of joint prediction in DC-only combinations, the stratified results demonstrate that the overall performance hierarchy across feature extractors remains consistent regardless of subgroup composition.

3.4.3. Session-Level Aggregation

Although all dataset partitions were performed at the welding-session level to prevent data leakage, the previous analyses were computed at the segment level to evaluate low-latency inference performance. To further assess operational stability at the session level, majority voting was performed across overlapping segments within the same welding session using the best-performing ECAPA-TDNN classifier for each feature representation. The resulting session-level metrics are summarized in Table 7.

Table 7. Session-level joint classification performance obtained through majority voting across overlapping segments within each welding session. Results are reported for the best-performing ECAPA-TDNN configuration of each feature representation.

Feature Representation	Exact Match (%)	Hamming Acc. (%)	Plate Acc. (%)	Electrode Acc. (%)	Current Acc. (%)
MFCC	100.00	100.00	100.00	100.00	100.00
VGGish	94.74	98.25	94.74	100.00	100.00
YAMNet	89.47	94.74	92.11	97.37	94.74

Majority voting consistently improves joint and per-target metrics relative to segment-level prediction for all three representation approaches. The MFCC + ECAPA-TDNN model achieves 100% session-level Exact Match and 100% Hamming Accuracy, compared to 94.51% and 97.88% at the segment level. This improvement reflects the redundancy provided by multiple overlapping segments per session and demonstrates that the proposed framework effectively leverages temporal consistency across the full welding pass. Session-level accuracy is particularly relevant for practical monitoring applications, where a decision per weld bead, rather than per individual segment, is typically required.

The results indicate that many segment-level prediction inconsistencies are temporally localized and can be effectively mitigated through session-level aggregation, thereby substantially improving operational stability during continuous welding monitoring.

3.5. Computational Performance

In addition to classification performance, computational efficiency was evaluated to assess the practical feasibility of each approach for real-time or near real-time SMAW

monitoring. Training time and inference time were analyzed as functions of segment duration and feature extraction method, considering that both data volume and model complexity directly impact deployment constraints.

The following section analyzes the computational cost of each feature–architecture combination, highlighting the trade-offs between classification performance and computational efficiency.

3.5.1. Training Time

Training time is strongly influenced by the number of training samples, which increases as segment duration decreases. In the evaluated dataset, the number of segments ranges from 448 samples at 50 s to 43,170 samples at 1 s, resulting in a significant increase in computational cost for shorter segment durations. Additionally, the ensemble size k directly scales the training time, as each fold produces an independent model.

Figure 13 shows the training time as a function of segment duration for all feature–architecture combinations. Across all approaches, training time increases sharply for segment durations below 5 s, reflecting the rapid growth in dataset size.

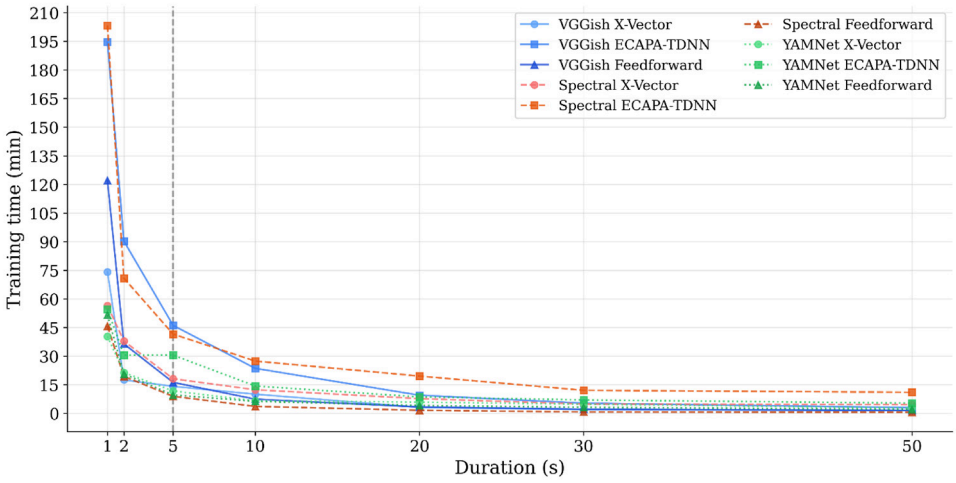


Figure 13. Training time as a function of segment duration (1 to 50 s) for all approach-architecture combinations (MFCC, VGGish, YAMNet, each with Feedforward, X-Vector, and ECAPA-TDNN classifiers). The vertical dashed line marks the baseline configuration at 5 s.

In the baseline configuration (5 s, k = 10), MFCC-based models consistently achieve the lowest training times across all architectures, whereas YAMNet-based models incur the highest computational cost due to their higher-dimensional embeddings. The VGGish-based model exhibits intermediate behavior.

For longer segment durations ($\geq$20 s), training times decrease significantly across all approaches, as the number of training samples is reduced. However, this reduction in computational cost comes at the expense of lower temporal resolution.

These results highlight the strong dependency between segmentation strategy and computational efficiency and demonstrate that MFCC-based representations provide a more favorable trade-off between performance and training cost.

3.5.2. Inference Time

Per-segment inference latency remained low across all configurations and approaches, confirming the feasibility of the proposed system for near real-time acoustic monitoring.

Figure 14 presents the average inference time per segment for each feature–architecture combination under the baseline configuration (5 s segments, k = 10). Ensemble inference was executed on the GPU using a batch size of 32.

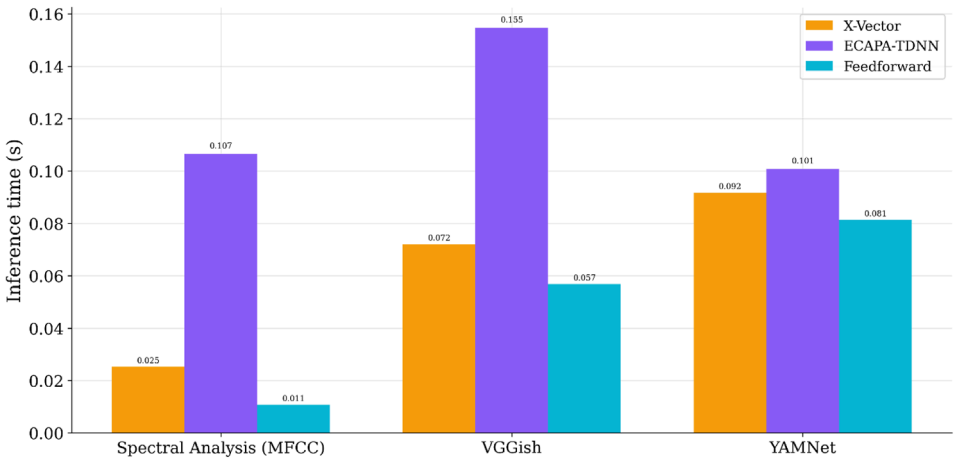


Figure 14. Per-segment inference latency at the baseline configuration (5 s, k = 10, p = 0.5), for MFCC, VGGish, and YAMNet.

Across all approaches, Feedforward models consistently exhibit the lowest inference latency, followed by X-vector architectures, while ECAPA-TDNN models incur the highest computational cost due to their more complex temporal modeling. The MFCC-based model achieves the lowest overall inference times overall, with latencies of approximately the order of 10 ms per segment for Feedforward networks and approximately 25 ms for X-vector models. In contrast, embedding-based approaches introduce additional computational overhead, with VGGish and YAMNet requiring substantially longer inference times across all architectures.

Among embedding-based methods, YAMNet generally exhibits higher inference latency than VGGish due to its higher-dimensional representations, although ECAPA-TDNN shows comparable behavior across both embeddings.

These results reinforce that MFCC-based representations not only provide superior classification performance but also enable faster inference, making them more suitable for real-time SMAW monitoring systems where both accuracy and latency are critical.

4. Discussion

4.1. Confusion Matrix Results

The confusion matrix analysis provides additional insight into the separability of the welding parameters and helps explain the trends observed in the classification metrics.

Among the evaluated targets, the current type exhibits the highest separability, as reflected in the near-perfect classification performance across all models. This behavior can be attributed to the distinct acoustic signature introduced by alternating current operation, where periodic arc extinction and re-ignition generate a characteristic oscillatory component in the emitted sound.

In contrast, electrode type presents the greatest classification difficulty. The observed confusion between E6010 and E6011 is consistent with their similar cellulosic coating compositions and arc characteristics, leading to comparable metal transfer dynamics and acoustic emissions. This reduced inter-class variability limits the discriminative capacity of both handcrafted and learned features.

Plate thickness shows intermediate behavior, with occasional confusion between adjacent thickness levels. This suggests that the corresponding acoustic differences are more subtle and likely influenced by secondary effects such as heat input and weld pool dynamics.

Overall, these observations indicate that classification performance is strongly governed by the physical distinctiveness of the welding conditions. Variables that induce more pronounced changes in arc behavior and acoustic emission are more easily separable,

highlighting the importance of incorporating process knowledge when designing acoustic monitoring systems.

4.2. Welding Audio Noise

The acoustic recordings were acquired in real workshop environments characterized by non-stationary background noise from surrounding machinery, ventilation systems, and human activity. Although no controlled external disturbance was intentionally introduced during the welding trials, occasional stationary and impulsive noise from adjacent workshop operations, including angle grinding and hammering, was intermittently present and could be perceived at the monitoring station despite occurring at a distance from the monitored welding setup. Rather than applying aggressive noise suppression, which risks filtering acoustically relevant components of the welding process, the methodology intentionally preserved the raw signal structure and relied on robust session-level partitioning to prevent data leakage.

In welding applications, the distinction between signal and noise is not always well defined. Acoustic emissions generated by arc instability, metal transfer, and spatter are inherently stochastic and can resemble background noise, yet they carry important information about process conditions. Previous studies have shown that welding acoustic signals arise from multiple superimposed sources, including arc ignition and extinction, metal transfer dynamics, and process-induced disturbances, resulting in complex, non-stationary acoustic behavior [38]. Although such studies are often conducted in gas-shielded processes, similar acoustic complexity is present in SMAW due to flux decomposition and arc-plasma interactions. Consequently, filtering strategies designed to remove noise may inadvertently suppress discriminative features relevant for classification.

The acoustic environment during data acquisition exhibited a heterogeneous mixture of noise contributions, which can be broadly grouped into three categories: (i) approximately stationary broadband noise produced by ventilation and fume-extraction systems; (ii) non-stationary transient noise from adjacent metalworking activities such as grinding, cutting, and hammering; and (iii) process-intrinsic stochastic emissions generated by the welding arc itself, including spatter, droplet detachment, and arc instabilities. The third category is particularly relevant for welding acoustic analysis, since it simultaneously represents a noise component from a signal-processing perspective and an informative component from a process-monitoring perspective.

Consequently, the conventional definition of signal-to-noise ratio (SNR) becomes ambiguous in this context. A physically meaningful SNR estimate would require a calibrated arc-silent reference channel recorded synchronously at the same microphone position, which was not available within the present acquisition protocol. For this reason, the robustness of the evaluated representations is reported here implicitly through classification stability across sessions acquired under the heterogeneous workshop conditions described in Section 2.1.1, rather than through a quantitative SNR distribution. A controlled SNR-stratified evaluation, in which canonical noise types are injected at known levels, is identified as an important direction for future work and is further discussed in Section 4.5.

The experimental results indicate that all three approaches can operate under these conditions without explicit filtering, although with different levels of robustness. MFCC-based representation demonstrates greater stability across configurations, suggesting that their spectral compression effectively captures relevant acoustic patterns while mitigating the impact of environmental variability. In contrast, embedding-based approaches such as VGGish and YAMNet, which are trained on the general-purpose AudioSet, appear more sensitive to non-stationary noise and domain mismatch in welding environments.

These findings support the use of minimally processed acoustic signals in welding monitoring applications, where preserving the full signal structure may be more beneficial than applying aggressive denoising techniques. Under the tested workshop conditions, the proposed approach remained stable despite intermittent environmental disturbances, indicating that relevant welding information can still be extracted without explicit noise suppression. However, stronger or sustained external disturbances were not systematically evaluated in this study and should be assessed in future work to quantify the operational robustness limits of the proposed approach. Therefore, the robustness demonstrated in this study should be interpreted within the acquisition conditions represented in the present dataset rather than universal robustness across arbitrary industrial environments.

4.3. Embeddings vs. Spectral Analysis for Welding

The results show that the choice of acoustic representation has a dominant impact on classification performance. Despite being trained on large-scale datasets, the pre-trained embedding approaches (VGGish and YAMNet) are consistently outperformed by the spectral representation (MFCC) across all evaluation metrics. MFCC-based models exceed embedding-based approaches by more than 25 percentage points in Exact Match on the held-out test set. This large performance gap suggests that the observed differences are primarily driven by the suitability of the feature representation to the welding domain rather than by classifier complexity alone.

The observed performance gap between MFCC and deep audio embeddings can be attributed to several factors that compound. VGGish and YAMNet exhibit a strong domain mismatch with SMAW arc sound, since both models were optimized for general-purpose audio classification across thousands of heterogeneous AudioSet categories rather than for the quasi-periodic and quasi-stationary spectral patterns associated with welding arc dynamics. Notably, both embedding approaches yielded remarkably similar classification performance across all evaluated tasks, suggesting that the dominant limitation arises primarily from representation-domain mismatch rather than from architectural differences between the two models. On the other hand, MFCC explicitly preserves the signal's low-frequency spectral envelope via Mel-scale filtering, making it inherently sensitive to arc modulation, metal transfer dynamics, and the oscillatory components introduced by AC operation. In contrast, general-purpose deep audio embeddings may suppress these narrow-band spectral structures as non-salient variability relative to the semantic audio categories for which they were originally optimized.

The relatively small size of the welding-specific dataset (373 sessions) further limits the feasibility of reliably fine-tuning large embedding backbones such as VGGish and YAMNet without substantial overfitting risk. Consequently, both embedding models were used as fixed feature extractors, restricting adaptation to welding-specific acoustic characteristics. Together, these factors explain why MFCC-based representations, despite their lower dimensionality, provide substantially more informative inputs for SMAW parameter classification under the evaluated experimental conditions.

Furthermore, the results indicate that the acoustic representation constitutes the primary performance bottleneck. Within each feature-extraction approach, differences in classifier architecture are comparatively small, whereas switching from MFCC to embedding-based representations results in substantial performance degradation. This interpretation is further supported by the feature space analysis in Section 3.3, where MFCC exhibits more structured feature organization prior to classification and maintains a measurable advantage even under simpler classifiers.

This trend also remains consistent under the subgroup analysis presented in Section 3.4.2, where MFCC-based models maintain superior Exact Match performance

in both DC-only and dual-mode operating conditions, indicating that the observed performance advantage is not solely explained by structured label dependency. Nevertheless, the session-level aggregation results presented in Section 3.4.3 also demonstrate that embedding-based representation substantially improves under majority-vote temporal aggregation, indicating that many prediction inconsistencies are locally transient rather than systematically incorrect.

These findings suggest that, for domain-specific applications involving manual SMAW processes in real-world environments, handcrafted spectral features may provide more effective and data-efficient representations than general-purpose deep audio embeddings. Beyond the quantitative performance gap itself, the present analysis indicates that the observed differences are structural rather than purely architectural. The superiority of MFCC emerges prior to classifier training, remains consistent across model complexity, persists under subgroup stratification, and reflects a domain-level mismatch between general-purpose audio representations and welding-specific acoustic patterns.

It is important to emphasize that the inferior performance of VGGish and YAMNet in this study does not imply a general limitation of deep audio embeddings for acoustic monitoring tasks. Rather, the results reflect a specific domain mismatch: both models were pre-trained on AudioSet, a corpus of general-purpose audio events with statistical properties substantially different from those of SMAW arc sounds. Under conditions where deep embeddings are either pre-trained on welding-specific data or fine-tuned on a sufficiently large in-domain dataset, performance comparable to or exceeding MFCC may be achievable. The development of welding-specific deep audio embedding models is therefore identified as an important direction for future work.

4.4. Process Parameters and Acoustic Monitoring

The classification results reflect the inherent physical separability of the evaluated welding parameters. The current Type (AC/DC) is consistently identified with the highest accuracy across all models, as expected given its distinct acoustic signature. This distinction is also qualitatively visible in the raw waveform fragments shown in Section 3.2, where AC signals exhibit stronger low-frequency amplitude modulation and more pronounced transient excursions than DC signals under comparable welding conditions. Alternating current introduces a systematic oscillatory component associated with periodic arc extinction and re-ignition, which is absent in direct current operation. In addition, the current type labels are partially structured by SMAW process compatibility constraints, since some electrode classes (E6010 and E7018) were evaluated only under DC operation due to their coating composition and arc stability requirements. Consequently, the observed separability reflects both the intrinsic acoustic differences between AC and DC welding and the physically valid operating conditions represented in the dataset. This dependency also affects the interpretation of the join evaluation metrics. Exact Match requires simultaneous correct prediction of all output labels and therefore benefits more from structured label correlations, whereas Hamming Accuracy evaluates each target independently and is less sensitive to inter-label dependency. However, the subgroup analysis presented in Section 3.4.2 shows that the overall performance hierarchy across feature extractors remains consistent under both DC-only and dual-mode operating conditions.

The session-level aggregation analysis presented in Section 3.4.3 further indicates that many segment-level prediction inconsistencies are temporally localized and can be mitigated through majority-vote aggregation across complete welding sessions. While segment-level inference remains more appropriate for low-latency process monitoring and real-time feedback, the session-level results demonstrate that the proposed framework maintains strong temporal consistency throughout the welding pass.

In contrast, electrode type and plate thickness were more difficult to classify. In particular, the difference between chemically similar consumables (e.g., E6011 and E6013) was more challenging, likely due to their comparable coating composition and arc behavior, which produce similar acoustic signatures. At the waveform level, this contrast is also reflected in the more irregular and impulsive behavior observed for E6010 relative to the smoother and more stable waveform of E7018, as shown in Section 3.2. These results suggest that arc sound is particularly effective for identifying coarse process conditions, whereas finer distinctions between consumables may require more discriminative or domain-adapted representation.

Although the present study focused exclusively on process-parameter classification, the observed sensitivity of acoustic signals to changes in welding conditions supports the broader potential of acoustic monitoring for process-state assessment. Variations in arc stability, metal transfer behavior, and transient disturbances are known to influence acoustic emissions and may provide discriminative information for identifying unstable operating conditions.

However, weld discontinuity detection was not addressed in this work and should not be inferred directly from the present results. Detecting discontinuities such as porosity, lack of fusion, or slag inclusion would require dedicated annotations at the defect-event level and a task-specific modeling framework designed for event localization or anomaly detection. In this context, the present study should be interpreted as a process-monitoring foundation that may support future work on acoustic-based discontinuity detection, rather than as a direct defect detection system.

4.5. Limitations

While the proposed methodology demonstrates strong performance for SMAW acoustic classification, several limitations should be acknowledged.

First, the study is restricted to the SMAW process under workshop conditions. Although the recordings naturally incorporate the ambient and non-stationary background noise sources detailed in Sections 2.1.1 and 4.2, the experimental setup employed a fixed microphone position near the welding arc. Variations in sensor placement, distance, and orientation may affect the acoustic signal characteristics and should be evaluated in future studies. These acquisition constraints were intentionally maintained during the present feasibility study to isolate the contribution of the acoustic representations under controlled acoustic acquisition conditions.

In addition, although intermittent background disturbances from adjacent workshop activity were present during acquisition, the robustness of the proposed approach against stronger or sustained external noise sources was not systematically evaluated.

A quantitative SNR characterization was not performed because a physically meaningful estimate would require a calibrated arc-silent reference channel recorded synchronously at the same microphone position, which was not available within the present acquisition protocol. Consequently, the robustness of the evaluated representations was not assessed under controlled interference levels or under systematically isolated industrial noise conditions. Future work should therefore include controlled noise-injection experiments, systematic evaluation across varying microphone-to-arc distances and orientations, and cross-site validation in workshops with distinct acoustic environments.

Second, the dataset reflects variability introduced by a single operator, which, while representative of manual welding conditions, does not fully capture the range of skill levels, electrode manipulation styles, arc lengths, and execution variability encountered in practice. Extending the dataset to include multiple welders and broader operational variability would improve the generalizability of the proposed approach.

Third, the evaluated deep audio embeddings (VGGish and YAMNet) are pre-trained on general-purpose AudioSet and are not optimized for welding-specific acoustic phenomena, as reflected in the performance differences observed in this study. This domain mismatch limits their effectiveness, suggesting that domain-adapted or task-specific embedding models may yield improved performance. In this regard, another direction of this work is the development of a domain-specific audio embedding model trained on welding arc sound datasets. This approach could address limitations in general-purpose embeddings, which are not optimized for welding-specific acoustic phenomena and potentially improve classification performance in multi-output prediction tasks.

In addition, the proposed multi-output framework relies on standard shared-backbone classification architectures and does not incorporate task-specific loss balancing, cross-task attention, or other specialized multi-task learning strategies. While this design was intentionally selected to isolate the effect of acoustic representation, more advanced multi-task optimization strategies may further improve joint prediction performance and should be explored in future work.

Finally, although this study focuses on acoustic-based classification, welding processes are inherently multimodal. Future work should explore integrating acoustic signals with electrical measurements and visual monitoring to improve robustness and enable more comprehensive characterization of welding conditions.

5. Conclusions

This study evaluated the feasibility of multi-output classification of SMAW process parameters using airborne acoustic signals generated during welding. Three acoustic representation approaches (MFCC, VGGish, and YAMNet) were evaluated under a unified experimental framework using ANN-, X-Vector-, and ECAPA-TDNN-based classifiers to simultaneously predict plate thickness, electrode type, and current type from a single airborne acoustic sensing channel. The results demonstrated that MFCC-based representation consistently outperformed deep audio embeddings across both classification and computational performance metrics, with the MFCC + ECAPA-TDNN configuration achieving the highest performance, reaching 94.51% Exact Match and 97.88% Hamming Accuracy at the segment level. In addition to the classification advantage of more than 25 percentage points in Exact Match relative to the best embedding-based approaches, the MFCC-based model also exhibited lower training and inference times. Strong classification performance was maintained even under short analysis windows, supporting the feasibility of low-latency acoustic monitoring for SMAW applications.

The analysis indicates that the observed superiority of MFCC is primarily due to the representation's suitability for welding-specific acoustic phenomena rather than to an inherent limitation of deep audio embeddings. Feature-space visualization revealed that MFCC representations exhibited substantially greater separability prior to classifier training, while VGGish and YAMNet showed higher intra-class dispersion and inter-class overlap, consistent with a representation-domain mismatch between AudioSet pre-training and SMAW arc sound. Among the evaluated targets, Current Type (AC/DC) exhibited the highest classification separability across all feature representations, consistent with the distinct oscillatory acoustic behavior associated with alternating current operation. The session-level aggregation analysis further demonstrated strong temporal consistency across welding passes, with majority-vote inference substantially improving prediction stability across all feature representations.

Overall, the results support the feasibility of acoustic-based monitoring for multi-output characterization of SMAW process conditions under realistic workshop environments. However, the present work should be interpreted as a process-monitoring frame-

work rather than a direct weld discontinuity detection system. Future work should therefore focus on developing welding-specific audio-embedding models, evaluating them under broader operational variability involving multiple operators and workshop conditions, and extending the proposed framework toward multimodal, defect-event-aware welding monitoring systems.

Author Contributions: Conceptualization, L.V., E.C. and C.P.-A.; methodology, L.V. and C.P.-A.; software, L.V.; validation, L.V., E.C. and C.P.-A.; formal analysis, L.V.; investigation, L.V. and C.P.-A.; resources, C.P.-A.; data curation, L.V. and E.C.; writing—original draft preparation, L.V.; writing—review and editing, E.C. and C.P.-A.; visualization, L.V.; supervision, E.C. and C.P.-A.; project administration, C.P.-A.; funding acquisition, C.P.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Secretaria Nacional de Ciencia Tecnologia e Innovacion (SENACYT), through contract No. 178-2023 (FID23-056).

Data Availability Statement: The acoustic signal data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to acknowledge Alejandra Avila, for the support during the development of this work. The authors also thank all welders who contributed to the construction of the experimental dataset, with special recognition to Nacor Martinez for his contribution. Special thanks are extended to Olmedo Caceres for his valuable guidance. During the preparation of this manuscript/study, the author(s) used ChatGPT (OpenAI, GPT-5.3) for the purposes of language edition and improvement of readability. The authors have reviewed and edited the output and taken full responsibility for the content of this publication. We would also like to thank M.A.P.G for his kind contribution.

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

References

1. Bohnart, E.R. *Welding: Principles and Practices*, 5th ed.; McGraw-Hill Education: New York, NY, USA, 2017.
2. Cocota, J.A.N.; Garcia, G.C.; da Costa, A.R.; de Lima, M.S.F.; Rocha, F.A.S.; Freitas, G.M. Discontinuity Detection in the Shield Metal Arc Welding Process. *Sensors* **2017**, *17*, 1082. [CrossRef] [PubMed]
3. American Welding Society. *Welding Handbook: Welding Science and Technology*, 9th ed.; American Welding Society: Doral, FL, USA, 2001; Volume 1.
4. Kumar, V.; Parida, M.K.; Verma, O.P. Evaluation of Power Sources and the Effect of Varying Current in SMAW Process. *Int. J. Syst. Assur. Eng. Manag.* **2020**, *11*, 455–465. [CrossRef]
5. Kumar, V.; Ghosh, S.; Parida, M.K.; Albert, S.K. Application of Continuous Wavelet Transform Based on Fast Fourier Transform for the Quality Analysis of Arc Welding Process. *Int. J. Syst. Assur. Eng. Manag.* **2024**, *15*, 917–930. [CrossRef]
6. Zhang, L.; Basantes-Defaz, A.C.; Ozevin, D.; Indacochea, E. Real-Time Monitoring of Welding Process Using Air-Coupled Ultrasonics and Acoustic Emission. *Int. J. Adv. Manuf. Technol.* **2019**, *101*, 1623–1634. [CrossRef]
7. McCardle, J.R.; Taylor-Burge, K.L.; Sc, B.; Stroud, R.R.; Tech, M.; Harris, T.J. The Real Time Analysis of Acoustic Weld Emissions Using Neural Networks. In *Proceedings of the 4th International Conference on Condition Monitoring and Diagnostics Engineering Management (COMADEM)*; Brunel University: Egham, UK, 1997.
8. Sumesh, A.; Rameshkumar, K.; Mohandas, K.; Babu, R.S. Use of Machine Learning Algorithms for Weld Quality Monitoring Using Acoustic Signature. In *Proceedings of the Procedia Computer Science*; Elsevier B.V.: Amsterdam, The Netherlands, 2015; Volume 50, pp. 316–322.
9. Stemmer, G.; Lopez, J.A.; Ontiveros, J.A.D.H.; Raju, A.; Thimmanaik, T.; Biswas, S. Unsupervised Welding Defect Detection Using Audio And Video. *arXiv* **2024**, arXiv:2409.02290. [CrossRef]
10. Xue, B.; Chang, B.; Du, D. Multi-Output Monitoring of High-Speed Laser Welding State Based on Deep Learning. *Sensors* **2021**, *21*, 1626. [CrossRef]
11. Zhou, Y.; Xue, B.; Guan, K.; Yue, H.; Zou, H.; Chang, B.; Du, D. An Online Monitoring Approach for Laser Beam Welding Process Based on Multi-Vision-Task Learning. *J. Intell. Manuf.* **2025**, *2025*, 1–17. [CrossRef]

12. Zhou, Y.; Guan, K.; Wang, L.; Chang, B.; Du, D. In-Situ Vision Monitoring for Laser Beam Welding of Large-Scale Thin-Walled Aerospace Structural Components Based on Multi-Vision-Task Learning and Stereo Vision. *J. Manuf. Process.* **2026**, *164*, 279–298. [CrossRef]
13. Truong, V.D.; Wang, Y.; Won, C.; Yoon, J. A Deep Learning-Based Machine Vision System for Online Monitoring and Quality Evaluation During Multi-Layer Multi-Pass Welding. *Sensors* **2025**, *25*, 4997. [CrossRef]
14. Shua, Y.; Gao, P.; Han, J.; Bai, L.; Zhao, Z. Online Welding Quality Monitoring via Vision-Acoustic Fusion with Future-State Prediction. *Weld. World* **2026**, *2026*, 1–17. [CrossRef]
15. Günaydin, L.; Cetin Akinci, T.; Kocyigit, G.; Gokmen, G.; Akgun, O. Analysis of Sounds Generated from Welding Machine While in Idle and Operational Mode by Acoustic Method. *Appl. Res. Tech. Technol. Educ.* **2015**, *3*, 214–223.
16. Avila, A.; Pinzon-Acosta, C. Comparative Analysis of VGGish and YAMNet Models for Welding Defect Detection. In *International Scientific-Technical Conference Manufacturing*; Springer Nature: Cham, Switzerland, 2024.
17. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2017; pp. 776–780.
18. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A.K. Applications of Machine Learning to Machine Fault Diagnosis: A Review and Roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
19. Jardine, A.K.S.; Lin, D.; Banjevic, D. A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510. [CrossRef]
20. Kusiak, A. Smart Manufacturing. *Int. J. Prod. Res.* **2018**, *56*, 508–517. [CrossRef]
21. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1995; Volume 2, pp. 1137–1143.
22. Davis, S.; Mermelstein, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoust.* **1980**, *28*, 357–366. [CrossRef]
23. Abdul, Z.K.; Al-Talabani, A.K. Mel Frequency Cepstral Coefficient and Its Applications: A Review. *IEEE Access* **2022**, *10*, 122136–122158. [CrossRef]
24. Gourisaria, M.K.; Agrawal, R.; Sahni, M.; Singh, P.K. Comparative Analysis of Audio Classification with MFCC and STFT Features Using Machine Learning Techniques. *Discov. Internet Things* **2024**, *4*, 1. [CrossRef]
25. Rezaul, K.M.; Jewel, M.; Islam, M.S.; Siddiquee, K.N.E.A.; Barua, N.; Rahman, M.A.; Sulaiman, R.B.; Shaikh, M.S.I.; Hamim, M.A.; Tanmoy, F.M.; et al. Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 37–53. [CrossRef]
26. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2018; pp. 5329–5333.
27. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proceedings of the Interspeech 2020*; ISCA: Singapore, 2020; pp. 3830–3834.
28. Hershey, S.; Chaudhuri, S.; Ellis, D.P.W.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN Architectures for Large-Scale Audio Classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
29. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
30. Zaman, K.; Sah, M.; Direkoglu, C.; Unoki, M. A Survey of Audio Classification Using Deep Learning. *IEEE Access* **2023**, *11*, 106620–106649. [CrossRef]
31. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2017.
32. Dietterich, T.G. Ensemble Methods in Machine Learning. In *Proceedings of the Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
33. Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
34. Zhang, M.-L.; Zhou, Z.-H. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837. [CrossRef]
35. Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. *Classifier Chains for Multi-Label Classification*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 254–269.
36. AWS. *AWS A5.1 Specification for Carbon Steel Electrode for Shielded Metal Arc Welding*; AWS: Hong Kong, China, 2012.

37. Rodríguez, A.G.; Pérez, C.R.G.; Borroto, O.M.R.; Oria, J.V.M. Behaviour of Short-Circuit Frequency and Duration Time and Electrical Conductivity during Arc Re-Ignition in SMAW (AC) with E6013 Electrodes. *Weld. Int.* **2010**, *24*, 673–679. [CrossRef]
38. Szłapa, P.; Marczak, W. Arc Welding Noise Assessment from the Measured Ultrasound Pressure Levels. Part III: Modern Welding Techniques. *Ultrasonics* **2020**, *108*, 106220. [CrossRef] [PubMed]

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