International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 7, July, 2026, pp. 2073-2086
Journal homepage: http://iieta.org/journals/jesa

Real-time Noise-Resilient Augmented Reality for Sustainable Industrial Maintenance in Industry 5.0

Mohammed Nabeel Islam[1*], Dhuha Basheer Abdullah[1], Zeena N. Al-Kateeb[2]

[1] Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul 41002, Iraq
[2] Department of Artificial Intelligence, College of Computer Science and Mathematics, University of Mosul, Mosul 41002, Iraq

Corresponding Author Email: mohammed.24csp38@student.uomosul.edu.iq

https://doi.org/10.18280/jesa.590722

Received: 17 May 2026
Revised: 15 July 2026
Accepted: 25 July 2026
Available online: 31 July 2026

Keywords:
augmented reality, real-time, Industry 5.0, speech recognition, digital twin, sustainable maintenance

ABSTRACT

The transition to Industry 5.0 requires human-centric technologies combining efficiency, sustainability and cognitive safety. This paper presents a multi-modal, real-time, noise-resistant augmented reality (AR) system for sustainable industrial maintenance, guiding novice technicians through complex mechanical seal replacement on water pumps. The framework couples a Level 1 Digital Model with an on-device, offline Vosk speech recognition engine and a custom Sound Pressure Level (SPL) adaptation algorithm operating from 50-90 dB. Twenty participants without prior mechanical experience were tested in a controlled environment. AR reduced Task Completion Time (TCT) by 30.11% and lowered performance variability, acting as a behavioral standardizer that removes spatial trial-and-error. NASA Task Load Index (NASA-TLX) workload fell by 55% and execution errors by 51.4%. No critical failures occurred with AR versus two in the manual group (0/10 vs. 2/10), a non-significant observed trend (Fisher's exact p = 0.474) requiring larger-sample confirmation. The System Usability Scale (SUS) score was 84.25 ± 7.17, and voice recognition remained usable across 300 trials (94% at 50-60 dB; 68% at 80-90 dB). By supporting procedural knowledge transfer, reducing downtime and preventing material waste, the framework addresses the environmental, economic and social pillars of Industry 5.0 sustainability, rapidly upskilling novice and cross-trained personnel to strengthen operational resilience.

1. INTRODUCTION

Industry 5.0 is a paradigm shift from Industry 4.0, which is based on automation logic, to a human-centric ecosystem where worker well-being, cognitive safety, and environmental sustainability are treated as first-class design goals, in addition to productivity [1-4]. At the heart of this vision lies the Operator 5.0: a human worker who is not replaced by digital tools, but rather supported by them, which enhance perceptual and cognitive capabilities and retain human agency, adaptability, and tacit knowledge [5, 6]. One important area for this vision to be realized is industrial maintenance, especially of critical fluid-handling equipment such as centrifugal water pumps. Maintenance mistakes result in substantial economic losses due to unscheduled downtime, premature component failure, irreparable damage to high-precision components, and measurable environmental damage through material waste, industrial scrap, and hazardous fluid leaks [7, 8].

The traditional static paper-based maintenance documentation puts a significant cognitive burden on novice technicians. The technician needs to refer to 2D diagrams, use tools in 3D space, and have an awareness of the sequence of procedures, all of which are often beyond the capacity of working memory and increase the likelihood of error [9, 10]. These traditional practices are in direct opposition to the goals of Industry 5.0, which focus on the operator's well-being and the complete avoidance of material waste caused by errors [11]. To solve this conflict, augmented reality (AR) has become a promising candidate technology. AR systems can be used as a digital prosthetic, which overlays procedural guidance to the physical workspace, thereby offloading extraneous cognitive processing and focusing attentional resources on the perceptual-motor aspects of the task that are relevant to the task at hand [12, 13]. This real-time visual feedback is essential for making decisions and preventing errors in real-time. Empirical studies have shown that AR-based guidance can reduce the assembly time by 21% and the perceived workload by 26.8% compared to paper instructions [14], and some configurations can reduce the assembly time by about 49.5% [15].

However, a critical research gap persists at the intersection of multi-modal interaction design and sustainable industrial pedagogy within Industry 5.0 contexts. Most of the current AR maintenance systems are passive visualization systems, which do not take into account the "hands-busy" situation of industrial maintenance, and cannot cope with the high ambient noise (60-90 dB) in pump rooms and mechanical service areas [16, 17]. Moreover, although the Industry 5.0 sustainability discourse has progressed strategically [18, 19], empirical studies quantifying AR's contribution using the Triple Bottom Line (TBL) framework are still limited. In particular, there is a lack of validation in the Environmental Pillar (material waste prevention through error reduction), Economic Pillar (cost mitigation through time efficiency and critical-failure prevention), and Social Pillar (cognitive workload reduction and technical knowledge democratization for Operator 5.0).

This paper aims to fill these gaps by proposing a new multi-modal AR system that combines real-time visual guidance with an automatic speech recognition (ASR) system that is robust to noise for sustainable water pump maintenance. A custom Sound Pressure Level (SPL) Adaptation Algorithm keeps voice command accuracy high in different industrial noise levels, and an interactive Exploded 3D View allows novice users to visualize complex pump assemblies spatially through touch-based visualization. This research makes an empirical contribution to operationalizing the Operator 5.0 construct in industrial maintenance by thoroughly assessing the system's effects in all three TBL dimensions in a controlled between-subjects experiment (N = 20). The fact that non-technical participants have been deliberately included, rigorously testing the intuitiveness of the system and its ability to convey complex procedural knowledge to users with no prior domain knowledge, shows the broad applicability and workforce upskilling potential of the system.

The remainder of this paper is organized as follows: Section 2 synthesizes related literature; Section 3 identifies research gaps; Section 4 presents the main contributions; Section 5 describes the methodology and experimental design; Section 6 reports empirical results; Section 7 discusses findings concerning sustainability and the TBL; and Section 8 concludes with future research directions.

2. RELATED WORK

AR research in industrial maintenance has moved from spatially registered visual guidance to multi-modal interaction for hands-busy work, to Industry 5.0 evaluation in which cognitive load and material waste join time and accuracy as criteria; spatial registration is the common substrate, since it is what makes hands-free control meaningful and error prevention physically measurable. This leaves three tensions that must be resolved together: interaction, as the modality that frees the hands is the most exposed to the environment; robustness, as maintenance is most needed where noise is high and connectivity absent; and evaluation, as sustainability claims require a physically executing operator rather than simulation.

2.1 Theme 1: Augmented reality in pumping and fluid-handling systems-contrasting spatial guidance with cloud-dependent digital twins

Fluid-handling and pumping system maintenance is a new and growing research field for the application of AR. An AR-assisted Product Service System for engineered-to-order manufacturing equipment, similar to industrial pump systems, showed an energy saving of 11% and an inspection time reduction of 50%, highlighting the operational efficiency that can be achieved with spatially registered digital guidance [9]. Moreover, recent case studies on mobile platforms like the assembly of TurtleBots have shown that AR scaffolding can substantially decrease assembly time and increase the accuracy of the procedures in robotic systems [20, 21]. Mourtzis et al. also demonstrated the feasibility of combining digital twins with mixed reality interfaces for remote manipulation of industrial machinery, laying the groundwork for Level 1 digital twin visualization in maintenance scenarios [22]. Begout et al. [23] furthered this research by creating an AR authoring tool that allows operators to co-locate and manipulate digital twin representations on physical workstations, and validated the use in reconfigurable factories. The effectiveness of hierarchical context-aware mobile AR systems for guiding complex assembly tasks, such as large-scale cable products, with adaptive spatial instructions is further demonstrated [24]. Likewise, Lavric et al. [25] created ATOFIS, an AR training system that showed a 13% increase in the speed of task completion and the absence of assembly errors when compared with the current commercial guides. In a similar method, Hasan and Alkan developed a gesture-controlled spatial AR framework called Gest-SAR, which reduced the error rate by seven times compared to traditional paper manuals, further proving the effectiveness of AR in manual assembly [26].

However, a common constraint in these studies is the need for continuous network connectivity or external Internet of Things (IoT) infrastructure to update the digital twin state data in real-time. Spatial awareness can be achieved through methods such as Begout et al. [23] and Fang et al. [24], but at the cost of high latency and a point of failure in the shop floor due to active data synchronization. Network connectivity is often unavailable or unreliable in pump rooms, underground service areas, and remote locations, where maintenance is most often needed. The current framework overcomes this by using a completely offline architecture: the Vosk speech recognition engine, the Unity 6 AR rendering pipeline, and the Level 1 Digital Model visualization (the first maturity level of the digital-twin classification) are all on-device and do not need a network connection to perform the task. This is in contrast to connectivity-based AR metaverse methods [7] that require active IoT infrastructure, but are unable to operate in network-dead industrial areas. Saidi et al. [10] also found hands-free, spatially contextual AR to be an important feature for maintenance technicians, and that head-worn AR assistants could reduce extraneous cognitive load and improve performance; our work builds on this by quantifying sustainability outcomes for pump maintenance. The fact that they depend on external network infrastructure and real-time synchronization also indicates a major weakness, which directly translates to the first research gap (Gap 1): fully offline, hands-free AR solutions in network-dead industrial zones.

2.2 Theme 2: Voice control challenges in high-noise industrial environments

If offline operation resolves the connectivity dependency identified above, it displaces the problem rather than removing it, since the modality that makes hands-free control possible is itself the one most degraded by the industrial environment. One of the most basic acoustic problems that arises in the integration of voice interaction in industrial AR systems is that factory and pump rooms are often characterized by broadband noise at 65-90 dB SPL, which leads to a significant decrease in the signal-to-noise ratio (SNR) of ASR engines and consequently to a significant drop in accuracy [16, 17]. Bini et al. created a factory-noise dataset and integrated ASR methods for human-robot interaction in industrial settings, showing that standard ASR needs a lot of noise-adaptation before it can be deployed [16]. Their ResNet-based architecture gained accuracy through dynamic data augmentation, but with a low memory footprint (1.76 MB), and performance is dependent on noise profiles [16]. Li described the front-end signal processing pipeline (microphone arrays, sampling optimization, noise reduction filters, echo cancellation, etc.) that is needed to obtain acceptable SNR for industrial ASR [17]. This hardware-centric approach, however, has high engineering overhead and low software-level adaptability, and is not suitable for lightweight commercial AR devices.

Cioflan et al. [27] demonstrated that on-device domain adaptation for noise-robust keyword spotting can boost accuracy by up to 18% on challenging, unseen noise types on Watt-class embedded devices at the algorithmic level. It has been demonstrated to be viable for on-device ASR, but has not been validated in a hands-busy, multi-modal AR context. This discovery directly inspires our SPL Adaptation Algorithm, which is a conceptually similar but architecturally different dynamic threshold for the Vosk engine. Stefaniak et al. [28] validated the use of wearable ASR in underground mining, which is similar to the noise profile of pump rooms, and showed that specialized vocabulary dictionaries and string-matching post-processing can significantly improve accuracy in adverse conditions. But their use of latency-inducing text mining is bad for real-time AR navigation. Valladares-Poncela and Fraga-Lamas [29] set up practical guidelines for on-device ASR in mixed reality (MR) applications, showing that quantized models on HoloLens 2 can reach a Character Error Rate (CER) of less than 6% when noise cancellation is optimized. Their reliance on complex hardware arrays, latency-inducing post-processing, and lack of end-to-end evaluation in active AR workflows directly point to the second research gap (Gap 2): the need for a dynamic, low-latency, noise-adaptive voice architecture.

2.3 Theme 3: Sustainability assessment in Industry 5.0 - the Triple Bottom Line as an evaluative framework

The two preceding tensions concern whether the system functions at all in the target environment; the third concerns the criteria by which its functioning should be judged. The Industry 5.0 sustainability narrative has shifted from efficiency optimization with a single metric to a more comprehensive approach that considers environmental, economic, and social impacts. Gbededo and Liyanage [19] verified a digital-twin simulation framework that evaluates manufacturing processes in all three TBL dimensions, showing that sustainability outcomes can be optimized without compromising productivity when done through simulation. Ojstersek et al. [30] incorporated Environmental, Social, and Governance (ESG) in simulation modelling of collaborative human-robot workplaces, and found improvements in sustainability, productivity, and well-being. Amirkhizi et al. [18] mapped nine Industry 5.0 sustainability dimensions to generative AI and digital twin methods, offering a theoretical architecture to evaluate technology interventions. One of the main limitations of these studies is that they are largely based on simulated environments and theoretical architectures and not on actual execution by human operators.

Bassi et al. [31] conducted a systematic review of 46 cobot collaboration studies, and from a human-centred point of view, they concluded that collaborative automation can lead to a reduction in physical fatigue and an increase in job satisfaction, but can also result in an increase in cognitive workload and psychological stress during complex tasks. This highlights the Social Pillar of the TBL: cognitive burden is a sustainability issue, because a high workload over a long period of time negatively affects the well-being of the workforce, increases the number of mistakes made, and causes environmental costs due to unnecessary damage of components and waste [6, 32]. Calzavara et al. [5] translated this insight into multi-objective task allocation, where human-centric metrics are considered in the operational planning to minimize makespan, energy consumption, and physical effort in human-robot collaborative lines. These studies, however, almost exclusively address collaborative robotics and do not consider the "Operator 5.0" who is only manually performing AR-guided maintenance without the support of robots.

The current study extends this work by using the TBL framework to assess an AR maintenance intervention. This study differs from previous simulation-based studies [19] and [30] and collaborative robotics-only studies [5, 31] by offering empirical evidence that cognitive augmentation (Social Pillar) leads to a decrease in task time (Economic Pillar) and material waste (Environmental Pillar) when physically executing the task. While multi-dimensional sustainability frameworks are well established in theory, the absence of physical maintenance studies evaluating purely manual and AR-augmented tasks directly leads to the third research gap (Gap 3): empirical validation of how cognitive augmentation can simultaneously increase economic efficiency and prevent environmental waste.

Table 1. Comparison of related augmented reality (AR) maintenance studies

Study	Interaction Mode	Noise Adaptation	Digital Twin	Sustainability Metric	Sample	Key Outcome
[9]	Visual AR	None	Yes (Cloud)	Energy, Time	Case study	50% reduction in inspection time, 11% reduction in energy
[14]	Visual AR	None	No	Time, Workload	N = 10 (Experienced)	21% assembly time reduction, 26.8% lower perceived workload
[24]	Mobile AR	None	Yes (Cloud)	Time, Accuracy	N = 12	Improved assembly time and accuracy for cable products
[10]	Head-worn AR	None	No	Cognitive Load	Field Study	Reduced extraneous cognitive load via head-worn AR
[25]	Visual AR	None	No	SUS, NASA-TLX	N = 16	0 errors, 13% faster completion vs. Guides
[26]	Gesture + AR	None	No	NASA-TLX, Errors	Lab Study	7× fewer errors vs. paper manual (0.70 vs. 5.00)
Present Study	Multi-modal (Visual +Voice)	Yes (Custom SPL)	Yes (Level 1, offline)	Triple Bottom Line (TBL)	N = 20	30.11% TCT reduction, 55% workload reduction, 51.4% error reduction, and an observed absence of critical failures in the AR group (0 of 10 versus 2 of 10)

Note: NASA-TLX = NASA Task Load Index; SPL = sound pressure level; SUS = System Usability Scale; TBL = Triple Bottom Line; TCT = task completion time.

Table 1 presents a structured comparison of representative AR maintenance approaches to synthesize the comparative characteristics of the reviewed studies across the three interaction modalities, noise robustness, digital twin integration, and sustainability evaluation. The table shows that the majority of AR systems are visual only, that voice interaction is not widely used and is not noise-adaptive, and that the systems are mostly cloud-based digital twin architectures, which further emphasizes the identified research gaps.

3. RESEARCH GAPS

Based on the synthesis of the literature, three critical gaps that the present study addresses have been identified:

1. Offline Multi-modal AR: While visual AR guidance has been proven effective for maintenance [14, 25], existing systems predominantly rely on cloud connectivity [7, 23] or lack hands-free interaction modalities suitable for "hands-busy" industrial tasks [10]. Furthermore, the real-time processing capabilities of such systems in challenging industrial environments are often not sufficiently addressed.

2. Noise-Adaptive Voice Architecture: Existing industrial ASR studies have not reported end-to-end accuracy profiles across multiple noise bands (50-90 dB) within an active AR maintenance workflow, nor have they described a dynamic SPL threshold adaptation mechanism for the Vosk engine that ensures real-time responsiveness.

3. Empirical TBL Operationalization: The Operator 5.0 construct has been theorized extensively [5, 6] but rarely subjected to controlled experimental validation that simultaneously quantifies social (workload), economic (time, failures), and environmental (error-driven waste) outcomes within a single maintenance study. Specifically, there is a gap in demonstrating the efficacy of such systems when deployed with a workforce lacking prior specialized technical training, thereby proving the system's inherent ability to democratize complex knowledge and tasks.

4. MAIN CONTRIBUTIONS

This research makes the following novel contributions to the field of industrial maintenance and Industry 5.0 sustainability:

1. Integrated multi-modal Noise-Resilient Framework: A novel integration of visual AR guidance with offline, noise-resilient speech recognition (Vosk with a custom SPL Adaptation Algorithm), achieving 94% command accuracy at 50-60 dB and 68% at 80-90 dB, enabling reliable hands-free operation for "Operator 5.0" in high-noise industrial environments without cloud dependency. This framework ensures real-time interaction and guidance. (Addresses Gaps 1 & 2).

2. Engineered SPL-Adaptive Voice Integration: An engineered integration and empirical optimization of energy-based noise-floor tracking for the Vosk engine within a live AR maintenance workflow. The contribution is not a new signal-processing primitive, but the device-specific calibration and system-level coupling of established components: an exponential-moving-average noise floor, a hardware-calibrated SPL mapping, and an SPL-banded acceptance margin relaxed in quiet conditions and tightened as ambient level rises, tuned to sustain real-time responsiveness across 50-90 dB on commodity handheld hardware, and validated end-to-end rather than in isolation. (Addresses Gap 2).

3. Empirical TBL Validation: Rigorous experimental evidence (N = 20) demonstrating simultaneous, large-effect-size improvements across all TBL pillars: Social (55% cognitive workload reduction, d = 2.80), Economic (30.11%-time reduction, d = 2.03; an observed absence of critical failures in the AR group (0 of 10 versus 2 of 10)), and Environmental (51.4% error reduction preventing material waste). Crucially, this validation was achieved with participants specifically chosen for their lack of prior technical experience, thereby providing compelling evidence of the system's capacity to empower a broad spectrum of the workforce and democratize access to complex industrial tasks. (Addresses Gap 3).

4. Sustainable Pedagogy Model: A validated framework positioning AR as a mechanism for operationalizing Green Manufacturing by preventing irreversible material damage and democratizing technical knowledge, thereby fostering a more resilient and sustainable industrial workforce. (Addresses Gap 3).

5. METHODOLOGY

5.1 System architecture

The AR Sustainable Maintenance System is organized into five layers, which enable a holistic and user-friendly experience, as shown in Figure 1. Operator 5.0 represents the User Layer, which communicates with the system through audio input (microphone), spatial data (camera), and touch-based button input (procedural navigation). This interaction is facilitated by the Hardware Layer, which consists of an AR Device, in this case a OnePlus 8T smartphone with Android 11, selected for its powerful processing and camera quality that is suitable for industrial handheld use [33]. This layer also has embedded sensors for spatial tracking and voice input, which can be used to identify a Physical Object, like an industrial water pump, for maintenance. The Software & Engine Layer is the central processing unit, which is based on the Android 11 operating system, Unity 6 development engine, and Vuforia Model Target SDK for marker-less 3D object recognition and spatial registration.

The speech recognition engine is a Vosk offline ASR (vosk-model-small-en-us-0.15). The Vosk engine was chosen for its ultra-low latency and ability to securely process limited grammar vocabularies on-device, as opposed to cloud-based options and heavier offline models. This guarantees real-time interaction and data privacy needs in industrial settings, while avoiding cognitive tunneling delays.

The Logic & Interaction Layer is the intelligence of the system with four integrated modules. The Visual Guidance Module provides step-by-step procedural overlays, with dynamic visual highlighting of the animated target components on the Level 1 Digital Model. The system also uses spatially registered annotated text labels to explicitly indicate the required items, such as the names of the tools in the virtual toolset in Step 0 and the names of the separated parts in the Exploded 3D View, so that the operator can focus on them at a glance without having to search for them. This offers instant feedback and guidance. At the same time, the Exploded 3D View Module animates the pump parts to show the spatial relationships between them, which helps novice users to understand the spatial relationships in real time, as shown in Figure 2. The Speech Recognition Engine is controlled by a Voice Command Controller, which controls a context-aware progression user interface (UI), and a custom SPL Adaptation Algorithm enhances the Speech Recognition Engine. The navigation logic is state-dependent, meaning that the Step 0 interface only has a 'Next' command, Steps 1-14 have both 'Next' and 'Previous' (Back) for procedural review, and the last Step 15 dynamically changes the 'Next' command to an 'Exit' command to safely end the session and log task completion. In addition, a Level 1 Digital Model provides high-fidelity procedural visualization that is synced with the progression of the task. The multi-modal interface design follows the principles of guidance systems for industrial workers [34].

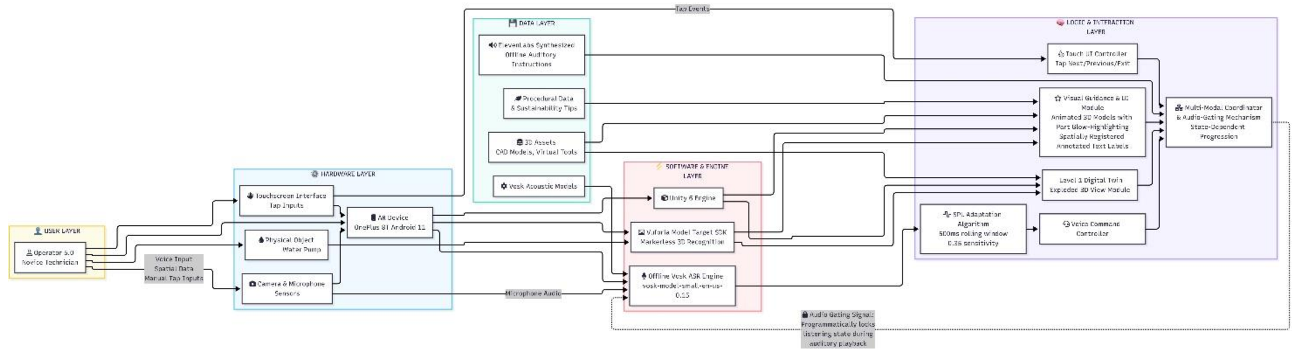

Figure 1. System architecture diagram

Note: AR = augmented reality; ASR = automatic speech recognition; CAD = computer-aided design; SDK = software development kit; SPL = sound pressure level; UI = user interface.

Finally, the Data Layer offers the resources, such as 3D Assets (CAD Models, tools), Procedural Data (e.g., 15-step sequences, safety, sustainability tips), and Acoustic Models (Vosk). The system's completely offline architecture is a conscious design decision, which takes into account the common network constraints in industrial maintenance environments [28, 29] and allows all on-device processing, such as speech recognition and visual guidance, to be performed in real-time.

The interaction layer adopts a sequential audio-gating mechanism to guarantee procedural compliance and reduce cognitive tunneling. Auditory instructions were synthesized beforehand with ElevenLabs text-to-speech (TTS) and kept as audio assets on the device before deployment, thus not needing any network connection during task execution. These high-fidelity auditory instructions are played back while the listening state of the Vosk ASR engine is programmatically locked. Voice commands like 'Next' or 'Back' are only processed once the auditory playback is over. This limitation guarantees that the operator internalizes the instructional material completely before moving on, so that no material is skipped. For added flexibility in various noise environments, the system also includes redundant touch-based navigation buttons (Next, Previous, Exit), which will provide procedural continuity when voice command accuracy is diminished in extreme noise conditions (80 dB or greater).

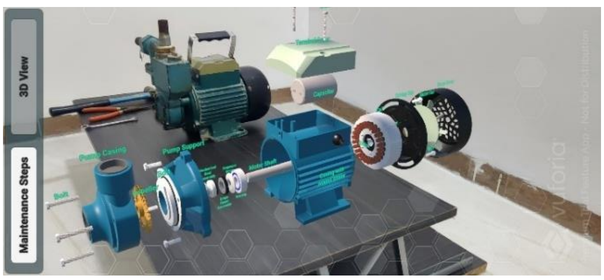

Figure 2. Exploded 3D augmented reality (AR) visualization of the pump assembly for maintenance guidance

5.2 Sound Pressure Level adaptation algorithm

The voice interaction layer employs a fully on-device adaptation mechanism that keeps the Vosk recognizer responsive across the 50-90 dB range without any cloud dependency. Its constituent operations root-mean-square (RMS) energy estimation, exponential-moving-average noise-floor tracking, and threshold-based gating are well-established in speech processing; the contribution here is their calibration to the target hardware and their integration with the procedural state machine, the TTS audio-gating logic and the phonetic-variant grammar, so that recognition remains usable under sustained industrial noise. Rather than thresholding a recognition confidence value, the algorithm operates as an energy-based voice-activity gate built around a continuously tracked ambient noise floor and an adaptive acceptance margin that responds to the estimated SPL. The mechanism comprises four stages: (1) initial noise calibration, (2) continuous noise-floor tracking, (3) SPL-banded margin adaptation, and (4) recognition gating with command validation. The whole pipeline runs as a synchronous per-frame loop, so the system adapts to environmental changes as they happen.

When the microphone stream is opened (16 kHz, mono), the system performs a 1.5 s ambient calibration, averaging the instantaneous input level in decibels relative to full scale (dBFS) over valid samples to establish an initial noise-floor baseline B(0). During operation, the RMS energy of each incoming audio frame is computed, exponentially smoothed, and converted to a digital level L(t) in dBFS. The ambient noise floor is then continuously updated during non-speech (quiet) frames using an exponential moving average (EMA), where the empirically calibrated adaptation rate α = 0.05:

$$B(t) = (1 - \alpha) \cdot B(t-1) + \alpha \cdot L(t), \alpha = 0.05$$

To make the gate aware of the surrounding acoustic environment, L(t) is mapped to an approximate SPL through an empirically calibrated linear transform, clamped to a plausible operating range of 20-110 dB:

$$SPL(t) = M \cdot L(t) + C, M = 2.66, C = 176.6$$

Here M and C are device- and microphone-specific calibration constants determined during pilot testing on the target OnePlus 8T hardware. They provide a practical approximation of ambient level for gating purposes rather than a laboratory-grade SPL measurement. These constants are not portable. They depend on microphone sensitivity, preamplifier gain, enclosure acoustics and the platform audio stack, and must be re-derived for any other device. Recalibration requires only a two-point fit: the device records at two known SPL levels spanning the intended operating range while a reference meter is used, and M and C are obtained from the resulting dBFS-to-SPL regression. The structure of the algorithm, comprising the exponential-moving-average noise floor, the SPL-banded margins and the gating rule, is unchanged by this procedure; only the mapping constants are device-specific. The acceptance threshold for the current frame is then defined as the tracked noise floor plus an SPL-dependent adaptive margin Δ(SPL):

$$Threshold(t) = B(t) + \Delta(SPL)/M$$

where, the adaptive margin Δ(SPL) is selected from three calibrated noise bands:

$$\Delta(SPL) = 5.0\ dB \text{ if } SPL < 70\ dB\ (Quite\ /\ Ambient)$$

$$\Delta(SPL) = 3.5dB \text{ if } 70dB \leq SPL < 85dB\ (\textit{Moderate})$$

$$\Delta(SPL) = 2.0dB \text{ if } SPL \geq 85dB\ (\text{HighNoise})$$

Because the noise floor B(t) itself rises with the ambient level, the absolute acceptance threshold tracks upward in louder environments. The margin is simultaneously tightened so that genuine speech, which, in high-noise conditions, is only modestly above the elevated floor, can still cross the gate, ensuring the system remains responsive rather than becoming unresponsive in noisy rooms. A frame whose level exceeds Threshold(t) is treated as candidate speech and refreshes the most-recent-speech timestamp; otherwise the quiet frame is used to further adapt the noise floor.

Audio frames are streamed to the Vosk recognizer continuously. To improve robustness against phonetic masking in broadband industrial noise, the recognition grammar is augmented with high-probability acoustic confusions of the three canonical commands (for example, "text", "nest", "nex", "necks" for "Next"; and "pack", "black", "bak", "beck", "sack" for "Back"), which are mapped back to the intended command during post-processing. A recognized token is accepted only when it is associated with recent above-threshold speech activity and is subject to a 1.0 s command cooldown that suppresses duplicate triggers. Validated commands ("Next", "Back", "Exit") then drive the procedural state machine. The complete logic is given in the algorithm below.

Algorithm 1: SPL-Adaptive Voice Recognition

```
INPUT: Microphone audio stream (16 kHz, mono)
OUTPUT: Validated voice command ∈ {Next, Back, Exit}
CONSTANTS: α = 0.05 (EMA rate); M = 2.66, C = 176.6 (SPL map);
        cooldown = 1.0 s; calibration window = 1.5 s
1. INITIALIZE:
  a. Open microphone (16 kHz, mono)
  b. Calibrate for 1.5 s -> B <- mean level (dBFS) of valid samples
  c. Grammar <- {Next, Back, Exit} U {phonetic variants}
2. LOOP (per frame, while NOT playing TTS and NOT in 3D-View):
  a. rms <- RMS(frame); smooth rms
  b. L <- 20 * log10(rms)
  c. SPL <- clamp(M * L + C, 20, 110)
  d. Delta <- 5.0 if SPL<70 ; 3.5 if 70<=SPL<85 ; 2.0 if SPL>=85
  e. Threshold <- B + Delta / M
  f. IF L >= Threshold THEN mark speaking; lastLoudTime <- now
     ELSE B <- (1 - α) * B + α * L
  g. Feed frame to Vosk ASR engine
  h. IF Vosk returns a result AND (now - lastLoudTime < 2.0 s):
        token <- parse(result)
        IF token maps to a command AND (now - lastCommandTime >= cooldown):
            emit command; lastCommandTime <- now
  i. GOTO 2
```

Note: ASR = automatic speech recognition; dBFS = decibels relative to full scale; EMA = exponential moving average; RMS = root mean square; SPL = sound pressure level; TTS = text-to-speech.

5.3 Experimental design

The experiment was a between-subjects controlled experiment with N = 20 novice participants who were not technical, but were recruited as proxies for untrained, entry-level industrial technicians. The use of absolute novices with no mechanical experience provides a stringent evaluation baseline, as the AR system is not dependent on the user's prior technical intuition to transfer procedural knowledge from expert to novice. The participants were randomly divided into two groups: Group A (AR, n = 10) and Group B (Manual, n = 10). The experimental structure is similar to the comparative studies that have recently been conducted to assess the relative effectiveness of AR and traditional paper-based teaching for mechanical tasks [35]. There were 10 male and 10 female participants, equally distributed between the two groups (5 males and 5 females) and the mean age was 22.4 years (SD = 1.8). No one had any experience in mechanical maintenance, AR technology or water pump operation. Based on a similar AR vs. paper manual assembly study [36] and using a priori statistical power analysis with G*Power 3 [37] (one-tailed independent-samples t-test, $\alpha = 0.05$, power = 0.80, effect size d = 1.20), a minimum sample size of N = 18 was determined; N = 20 was used to provide a conservative margin.

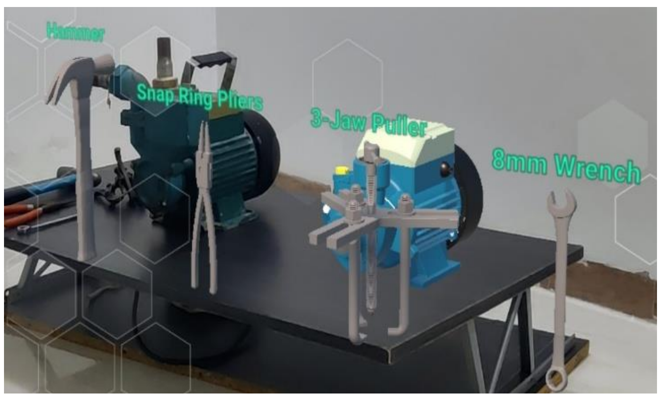

Figure 3. Preparation phase (3D models of the required physical toolset)

In order to standardize the experimental conditions and to avoid the interruption of the tasks, a 'Step 0: Preparation Phase' is required before the 15-step mechanical seal replacement procedure on a peripheral centrifugal water pump (TOTAL TWP105502, 0.75 HP). In this first phase, the AR interface displays 3D models of the necessary physical tools (hammer, snap ring pliers, 3-jaw puller, and 8mm wrench) in a spatially registered manner as illustrated in Figure 3. The system automatically requires participants to confirm the set-up of each tool verbally before unlocking Step 1, thereby reducing material mishandling and preventing artificially increasing the cognitive load by looking for missing materials.

All trials were conducted in a simulated industrial environment with the same physical parameters: 500 lux ambient lighting, 22 °C temperature, and controlled ambient noise levels (between 50 dB and 60 dB). To maintain algorithmic consistency, the custom SPL Adaptation Algorithm used the empirically calibrated parameters described in Section 5.2 (EMA adaptation rate $\alpha = 0.05$, SPL mapping constants $M = 2.66$ and $C = 176.6$, and the three-band acceptance margins) for all AR trials. Identical tool sets were provided to all participants. Group A participants interacted with the AR system by following the procedure using voice commands ("Next," "Back"). Group B participants were given a professionally formatted paper manual with step-by-step written instructions and 2D component diagrams. The comparison is therefore between two complete guidance packages rather than between isolated modalities: Group A received visual AR guidance and hands-free voice navigation as a single integrated intervention, and the reported effects are attributable to that package as a whole. A standard 5-minute familiarization period was provided for both groups before the task.

Immediately after the completion of the task, cognitive workload was measured using the NASA Task Load Index (NASA-TLX). In order to reduce the post-task survey fatigue, the Raw NASA-TLX (RTLX) scoring method was used, which is an unweighted averaging method widely validated in the Human-Computer Interaction (HCI) literature [38]. The System Usability Scale (SUS) [39] was used to measure system usability for the Group A participants only. Task Completion Time (TCT) was captured from the time of tool pick-up to the completion of the final reassembly step, and includes the preparation phase for both groups. Group A viewed the required tools as spatially registered 3D models and Group B viewed the equivalent tool list in the first section of the paper manual; neither condition required per-tool verbal confirmation, so no differential time cost is introduced. Execution errors were recorded by the researcher, who directly observed each participant during the task, and were classified using the error taxonomy in Table 2.

Table 2. Execution error taxonomy and coding scheme

Category	Severity	Definition	Representative Example
Tool-Selection	Minor	Wrong tool chosen but corrected before any damage.	Wrong wrench picked during fastener removal, then corrected.
Procedural-Sequence	Minor	Step done out of order, or a non-critical sub-step omitted, then corrected.	Impeller removal attempted before fastener release.
Component-Handling / Alignment	Minor	Mis-seating or misalignment that is recoverable without replacement.	Snap ring imperfectly seated, re-seated on retry.
Excessive-Force / Over-Torque	Critical	Excessive force causing irreversible damage requiring replacement.	Casing bolt sheared from wrong spanner and over-torque (M-07).
Improper-Extraction / Asymmetric-Load	Critical	Incorrect technique/tool causing asymmetric loading that fractures a precision component.	Ceramic seal fractured by asymmetric extraction (M-03).

Table 3. Sustainable task decomposition and cognitive mapping

Phase	Steps	Cognitive Challenge	Augmented Reality (AR) Solution	Sustainability Tip
Phase 0: Preparation	0	Tool identification and organization	Spatially registered 3D tool models with text labels and Vosk-based “Next” progression lock	Verify all tools to avoid interruptions and material mishandling
Phase 1: Disassembly	1-6	Spatial ambiguity during fastener removal	Dynamic 3D tool animations on the Level 1 Digital Model showing correct tool use and rotation	Collect drained fluids properly and avoid spillage
Phase 2: Inspection & Preparation	7-8	Identifying usable vs. degraded components	Exploded 3D views with labels and highlighted wear points	Separate rubber and metal waste for recycling
Phase 3: Component Replacement	9	Accurate seal alignment	Ghosted 3D models with highlighted insertion depth and orientation	Prevent ceramic seal damage and material waste
Phase 4: Reassembly & Testing	10-15	Remembering reverse assembly steps	Reverse-step animations with virtual tool guidance and voice-based “Exit” command	Apply proper torque to prevent leaks and extend lifespan

5.4 Sustainable task decomposition and cognitive mapping

The 15-step maintenance procedure, which includes a mandatory Step 0: Preparation phase, was broken down into five cognitive-sustainability phases to match the task design with the cognitive needs of novice operators and the Green Manufacturing principles of the Environmental Pillar of the TBL. The overall procedural flow and stepwise AR-guided execution are shown in Figure 4. Sustainability guidance tips were added to each phase in the AR system, using a multi-modal approach of on-screen text labels, animated parts highlighted dynamically, and audio instructions synthesized by ElevenLabs and synced with the animation. These tips are advisory prompts delivered through the interface; the system displays and narrates them but does not detect or enforce compliance, and adherence to them was not measured in the present study. The decomposition is organized by dominant cognitive demand rather than by physical similarity of actions. Phase 0 is a low-interactivity identification task, matching physical objects to named referents. Phases 1 and 4 impose primarily extraneous load: the operator must hold a spatial sequence in working memory, and in Phase 4 must additionally invert it, which is the point at which sequence recall is most fragile. Phase 2 is discriminative, requiring judgment of component condition rather than recall. Phase 3 carries the highest element interactivity, since seal orientation, insertion depth, and applied force must be coordinated simultaneously. The AR treatment for each phase is selected accordingly, externalizing sequence in Phases 1 and 4, supporting discrimination through exploded views and highlighted wear points in Phase 2, and constraining the interacting elements through ghosted geometry in Phase 3, consistent with accounts of AR as a means of reducing extraneous load in maintenance work. The detailed breakdown of this decomposition, outlining the cognitive challenges, AR solutions, and specific sustainability tips for each phase, is presented in Table 3.

5.5 Voice recognition testing protocol

A separate noise-simulation protocol was used as a controlled technical benchmark of voice recognition performance, conducted independently of the participant trials. Three bands of noise stimuli at calibrated SPL levels (50-60 dB, 65-75 dB, 80-90 dB) were created using a calibrated loudspeaker (Brüel & Kjær Type 4292) playing a composite factory noise track (ISO 11690-1 compliant). There were 300 voice command trials (50 trials per command ("Next", "Back") per noise band). The testing protocol only included the high-frequency navigational commands ("Next" and "Back") that are used for the procedural progression, while the "Exit" command is only used once at the end of the last session. Each trial was started by a trained operator who spoke the command at a standard distance of 30 cm from the device microphone. The number of correct commands / total trials was recorded for each condition as a measure of recognition accuracy.

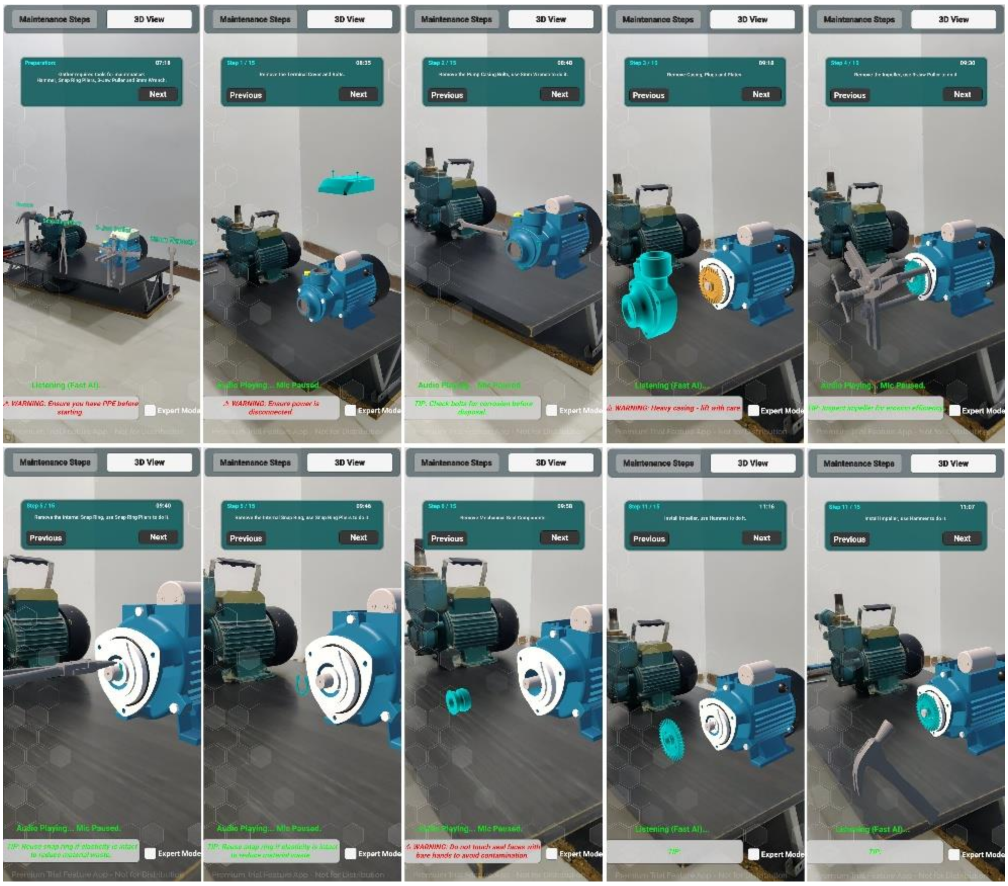

Figure 4. Step-by-step augmented reality (AR) maintenance workflow

6. RESULTS

6.1 Task Completion Time and Usability

The AR-assisted group (Group A) completed the tasks significantly faster than the manual group (Group B) in the controlled ambient condition (50-60 dB). Mean TCT for Group A was 16.85 ± 2.08 minutes, compared to 24.11 ± 4.60 minutes for Group B, representing a 30.11% reduction ($t(18) = 4.55$, $p < 0.001$, Cohen's $d = 2.03$). This is an extremely large effect size, which means that there is a significant practical impact in addition to the statistical impact. Notably, the AR system also decreased the intra-group performance standard deviation by 55% (from SD 4.60 to SD 2.08), acting as a behavioral standardizer that removes the spatial trial-and-error phase that novice technicians would experience when encountering a new pump assembly for the first time. This variability reduction is similar to the results of Wang et al. [13], who found that AR-based maintenance instruction (N = 30) had the highest SUS score, lowest NASA-TLX, shortest TCT, and minimum error rate among three types of instructions, and they believed that the uncertainty of spatial orientation was eliminated.

Table 4. Task Completion Time (TCT) and System Usability Scale (SUS) results

Metric	AR Group (M ± SD)	Manual Group (M ± SD)	Improvement	t-statistic	*p*-value	Cohen's d
TCT (minutes)	16.85 ± 2.08	24.11 ± 4.60	30.11% reduction	t(18) = 4.55	p < 0.001	2.03
SUS Score	84.25 ± 7.17	N/A	Excellent	—	—	—

Note: AR = augmented reality; M = mean; SD = standard deviation; N/A = not applicable.

Table 5. NASA Task Load Index (NASA-TLX) subscale results

Subscale	AR Group (M ± SD)	Manual Group (M ± SD)	Reduction	*p*-value	Cohen's d
Mental Demand	22.5 ± 12.2	75.0 ± 14.5	70%	p < 0.001	3.91
Physical Demand	31.0 ± 9.8	48.5 ± 11.2	36%	p < 0.01	1.66
Temporal Demand	28.0 ± 10.5	58.0 ± 12.8	52%	p < 0.001	2.56
Performance	35.5 ± 13.1	62.0 ± 14.7	43%	p < 0.001	1.90
Effort	33.5 ± 11.4	61.0 ± 13.6	45%	p < 0.001	2.19
Frustration	18.5 ± 10.1	72.0 ± 15.2	74%	p < 0.001	4.14
Total NASA-TLX	28.2 ± 10.9	62.8 ± 13.6	55%	p < 0.001	2.80

Note: AR = augmented reality; M = mean; SD = standard deviation.

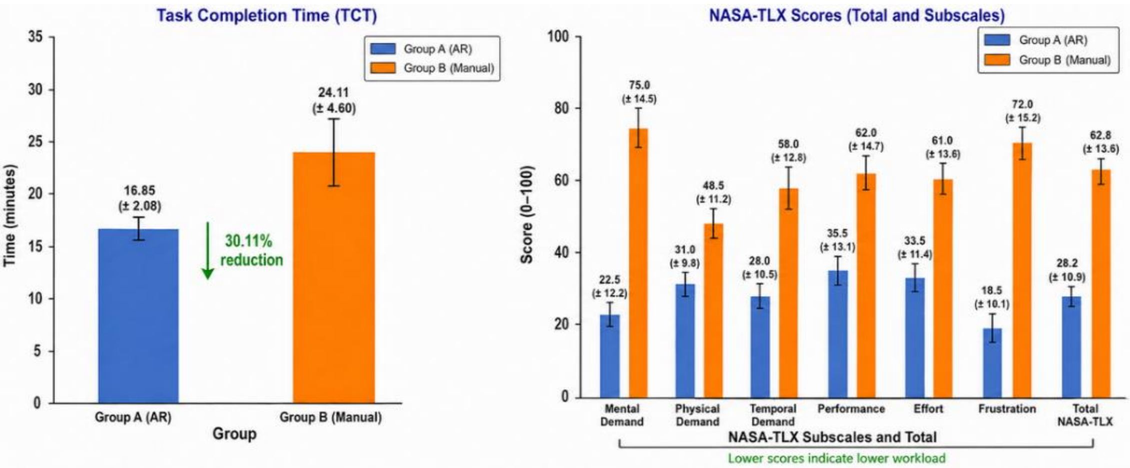

Figure 5. Comparative analysis of Task Completion Time (TCT) and NASA Task Load Index (NASA-TLX) workload scores between augmented reality (AR) and manual groups

The SUS [39] was used to evaluate the system usability for the participants in Group A. Based on the SUS interpretation benchmarks, the AR system was in the "Excellent" range (80-90), with a mean SUS score of 84.25 ± 7.17. The individual SUS scores were between 75.0 and 95.0, with all participants scoring above the 68-acceptability threshold. The quantitative results of TCT and SUS, which show the marked performance gap between the AR-assisted and manual groups, are summarized in detail in Table 4.

6.2 Cognitive workload reduction

The NASA-TLX assessment was used to provide empirical evidence of the effectiveness of the AR framework to reduce cognitive load. The AR group exhibited a 55% overall reduction in cognitive workload compared to the manual group (28.2 ± 10.9 vs. 62.8 ± 13.6, $t(18) = 6.26$, $p < 0.001$, Cohen's d = 2.80). The greatest change was in the Frustration subscale (74% reduction, d = 4.14), indicating that the system could reduce the psychological strain of unfamiliar and complex technical tasks. Mental Demand and Temporal Demand also experienced significant decreases of 70% (d = 3.91) and 52%, respectively, demonstrating that the multi-modal annotations and dynamic 3D visual guidance were effective in reducing working memory demands. Table 5 shows a detailed breakdown of the cognitive workload on all subscales.

A comparative bar chart of TCT and NASA-TLX scores for the AR-assisted group (Group A) and manual group (Group B) is shown in Figure 5. As illustrated, the AR-based approach resulted in lower TCTs and consistently lower cognitive workload across all NASA-TLX subscales for Group A, demonstrating the efficiency and usability benefits of the AR-based approach.

6.3 Execution errors and critical failures

Execution errors were classified into two categories, following the taxonomy in Table 2: (1) minor errors, which were procedural errors or errors in the selection of tools that did not cause physical damage to the pump components and could be corrected; and (2) critical failures, which were errors that resulted in irreversible physical damage to the pump components and required replacement. The AR group committed significantly fewer total execution errors (1.70 ± 0.67) compared to the manual group (3.50 ± 0.97), representing a 51.4% reduction ($t(18) = 4.81$, $p < 0.001$, Cohen's d = 2.15).

Most importantly, there were no critical failures (0%, n = 0) in the AR group, while 20% (n = 2) of the manual group had critical failures. Two critical failures in the manual group were:

1. Casing Bolt Shearing (Participant M-07): The participant

in the manual group used the wrong spanner during the casing-bolt tightening step of reassembly (Step 13) and applied excessive mechanical force while tightening, resulting in the bolt shearing. Those bolts are opened with an 8 mm wrench. This kind of error is particularly important to reinforce the procedure during torqueing operations, which has been solved by recent innovations in AR-guided connected torque wrenches that prevent this failure mode [40].

2. Mechanical Seal Fracture (Participant M-03): The participant used the wrong hand tools during the seal extraction process and used an asymmetric extraction technique, resulting in an unequal lateral force on the ceramic seal face and fracturing the mechanical seal assembly. This fault necessitated the replacement of the entire seal cartridge and housing parts.

These critical failures are direct violations of the principles of Green Manufacturing under the Environmental Pillar of the TBL: they used irreplaceable raw materials, produced industrial waste that needed specialist disposal, and required replacement parts to be bought, with the associated energy costs of manufacturing, packaging and supply chain logistics. The AR group has achieved an observed absence of critical failures in the AR group (0 of 10 versus 2 of 10), which shows the effectiveness of the system in preventing material waste and promoting sustainable maintenance practice. It is important to note that the difference in critical failure rates is clinically meaningful and practically significant, but the small sample size (n = 10 per group) precludes statistical significance for rare binary outcomes; Fisher's exact test yielded p = 0.474 (two-tailed). Future studies with larger samples are needed to statistically validate this finding. These execution errors and critical failures for both AR-assisted and manual groups are analyzed and presented in Table 6, which shows the effectiveness of the AR system in preventing costly errors and material waste.

Table 6. Execution errors and critical failures

Error Category	AR Group (M ± SD)	Manual Group (M ± SD)	Reduction	t-Statistic	p-Value	Cohen's d
Total Errors	1.70 ± 0.67	3.50 ± 0.97	51.4%	t(18) = 4.81	p < 0.001	2.15
Critical Failures	0 (0%)	2 (20%)	Observed 100%	—	p = 0.474 (Fisher)	—

Note: AR = augmented reality; M = mean; SD = standard deviation.

Table 7. Voice recognition accuracy by noise band

Noise Band (dB SPL)	Environment Type	"Next" Accuracy	"Back" Accuracy
50-60	Ambient industrial	94% (47/50)	92% (46/50)
65-75	Moderate industrial	80% (40/50)	72% (36/50)
80-90	High industrial	68% (34/50)	56% (28/50)

Note: dB = decibels; SPL = sound pressure level.

6.4 Voice recognition performance across noise levels

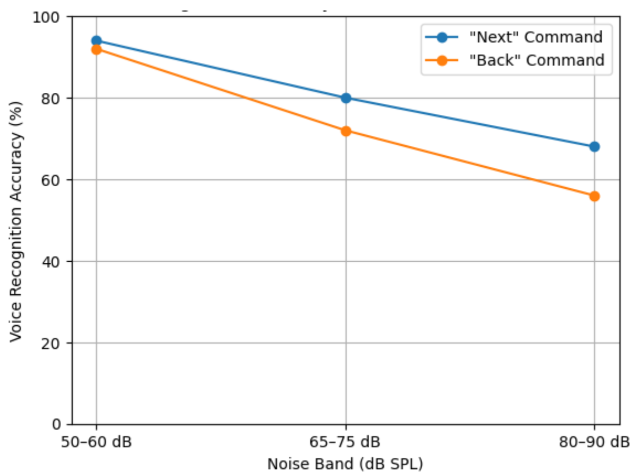

Figure 6. Voice recognition accuracy across noise levels

The accuracy of voice recognition was tested in three bands of industrial noise. The system was tested at 50-60 dB (ambient industrial) and resulted in 94% accuracy for "Next" commands (47/50 trials) and 92% accuracy for "Back" commands (46/50 trials). At 65-75 dB (moderate industrial), accuracy decreased to 80% for "Next" (40/50) and 72% for "Back" (36/50). At 80-90 dB (high industrial), accuracy further decreased to 68% for "Next" (34/50) and 56% for "Back" (28/50). The SPL Adaptation Algorithm was able to keep the recognition accuracy usable even at the highest noise band, while the unmodified Vosk during pilot testing at 80-90 dB had a baseline accuracy of 45%. These figures represent a controlled technical benchmark and an upper bound on in-task accuracy, since commands were issued by a trained speaker at a fixed 30 cm distance; participant trials were conducted only in the 50-60 dB band. Table 7 and Figure 6 illustrate the strong performance of the SPL Adaptation Algorithm under different acoustic conditions, namely the voice recognition accuracy for the three calibrated industrial noise bands.

7. DISCUSSION

7.1 Economic pillar: Efficiency and variance reduction

The AR group showed a TCT reduction of 30.11% (d = 2.03), which is a direct economic benefit due to reduced maintenance downtime and reduced labour costs. Moreover, the performance standard deviation was reduced by 55% (4.60 → 2.08), showing that the multi-modal AR system is a strong behavioral standardizer. The system removes the spatial uncertainty that leads novice technicians to time-consuming trial-and-error activities, resulting in very predictable maintenance times. This predictability is essential for planning the operating time in industrial fluid-handling systems. The Exploded 3D View and spatially registered tool guidance are the most plausible mechanisms for the reduction of tool-handling ambiguity, although the present design cannot isolate their contribution from that of hands-free navigation [9, 26]. The substantial economic gains were made by participants with no prior mechanical experience, further highlighting the system's ability to quickly integrate and upskill new workers, thereby helping to build a more agile and economically resilient workforce.

7.2 Social pillar: Procedural knowledge transfer and cognitive relief

The deep 55% decrease in cognitive workload (d = 2.80) and the 74% decrease in Frustration (d = 4.14) directly operationalize the Operator 5.0 construct [1, 5]. These findings show that AR can serve as a cognitive orthosis, offloading the extraneous cognitive load of consulting documentation from working memory to the perceptual-motor demands of the task. The AR interface allows for the offloading of sequential memory and spatial reasoning, supporting procedural knowledge transfer to users without prior domain experience. Importantly, the participants who were not trained in mechanics were able to complete a 15-step industrial procedure with substantially lower workload than participants performing the same task using conventional documentation. This indicates potential for reducing dependence on specialist expertise for procedural tasks, although establishing this at the level of a workforce would require evidence across a wider range of ages and backgrounds and under real factory conditions [4].

7.3 Environmental pillar: Mitigating waste through error prevention

The principal environmental contribution of the system is the statistically robust 51.4% decrease in execution errors (d = 2.15), accompanied by an observed absence of critical failures in the AR group (0 of 10 versus 2 of 10). The latter is a practically meaningful trend rather than a confirmed effect: with ten participants per group, the comparison is underpowered for a rare binary outcome (Fisher's exact p = 0.474), and confirmation requires a substantially larger sample. The causal chain from guidance to material saving is mechanical rather than exhortative. Spatially registered tool models specify which tool applies at each step, and the animated Level 1 Digital Model specifies the extraction geometry and insertion depth. In the two critical failures observed in the manual group, the proximate cause was in both cases a tool-and-technique error: an incorrect spanner producing over-torque on a casing bolt, and an asymmetric extraction imposing unequal lateral force on a ceramic seal face. Each of these failure modes is addressed by a specific guidance element, and each avoided failure removes a determinate quantity of scrapped material from the process, namely the sheared fastener or the replaced seal cartridge and housing. The environmental benefit therefore follows from preventing the physical damage, not from the advisory content of the interface. A critical failure (e.g., sheared bolt or broken ceramic seal) in a mechanical system directly contradicts Green Manufacturing principles. These failures use up irreplaceable raw materials, create industrial waste that needs special disposal, and require energy-intensive raw material procurement, manufacturing, and transportation of replacement parts [18, 19]. The AR framework operationalizes environmental sustainability by ensuring that the right tools are used and aligned correctly, thereby avoiding material waste caused by errors, which aligns with the findings that digital twin integration in industrial IoT scenarios can lead to tangible energy savings and waste reduction [41]. The fact that there were no critical failures among participants with no technical skills also underscores the system's high level of error-proofing, which can minimize material waste even when used by an inexperienced workforce.

7.4 Technical feasibility of noise-resilient offline interaction

The successful deployment of the custom SPL Adaptation Algorithm demonstrated that fully offline, hands-free voice interaction is possible even in the high-noise pump room environment, allowing real-time command processing. The results of 94% accuracy at 50-60 dB and 68% at 80-90 dB are a significant improvement over the baseline accuracy of 45% for unmodified Vosk at the highest noise band. This dynamic threshold mechanism proved to be effective in balancing sensitivity and specificity, as reported in studies [27, 28] that on-device domain adaptation helps in keyword spotting in noise. The asymmetric accuracy between "Next" (68%) and "Back" (56%) at 80-90 dB is due to the phonetic structure of "Back", which ends with a plosive consonant (/k/), which is very vulnerable to masking by broadband industrial noise [17]. Importantly, this offline architecture addresses the cloud-dependency vulnerability found in previous AR literature [7, 23], and the system can be used in network-dead zones.

7.5 Limitations and future work

There are several limitations to the current study. First, the sample size (N = 20) and the fact that the participants were young adult novices, rather than industrial operators, restrict the generalizability of the results to novice and cross-trained personnel. Experienced technicians differ in cognitive profile, tactile intuition, and tolerance of procedural constraint, and the benefits reported here should not be assumed to transfer to them. Step-locked guidance that relieves a novice may instead constrain an expert who already holds the procedure in long-term memory, in which case the overlay could act as a redundant visual load rather than a cognitive aid. It should be emphasized, however, that this was a conscious methodological decision to test the AR system's basic effectiveness in knowledge transfer, without relying on any mechanical intuition that the user might have, and to rigorously isolate the AR system's ability to guide absolute novices. Although this is a controlled laboratory environment, which is not a complete representation of the physical conditions and variability of real industrial deployments, it does give a good baseline for future deployments. Future longitudinal studies should therefore test the framework in the hands of professional technicians in actual factory settings to establish ecological validity. Second, the AR condition combined spatial visual guidance with hands-free voice navigation, so the two modalities are confounded by design. The present data cannot apportion the observed gains between spatial visualization and voice interaction, and the results should be read as the effect of the integrated multi-modal package rather than as evidence for visual AR alone. Third, the Level 1 Digital Model does not currently use real-time sensor data, and integration with vibration, temperature and pressure sensors would take the system to Level 2/3. Fourth, the voice evaluation was a controlled benchmark with one trained speaker at a fixed distance, and therefore an optimistic bound on live in-task accuracy; multi-speaker and cross-accent robustness testing is needed before deployment in various international workforces. Fifth, a further limitation concerns the granularity of the workload measurement. NASA-TLX was administered once after task completion and therefore reports an aggregate impression; it cannot detect momentary peaks during the more information-dense steps, in particular inspection and reassembly, where highlights, labels, virtual tools and animations coincide. It remains possible that transient overload occurred at specific steps without being visible in the global score. Sixth, the SPL mapping constants were calibrated on a single device, the OnePlus 8T, and no cross-device validation or sensitivity analysis was performed. Their transferability to other handhelds or to head-mounted platforms such as HoloLens, whose microphone arrays and audio pipelines differ substantially, is untested, and the recognition accuracies reported here should be regarded as specific to the calibrated configuration. Seventh, the sustainability prompts embedded in the procedure (Table 3) are advisory only. The system does not sense fluid collection or waste segregation and cannot verify that these behaviours occurred, so the environmental claims of this study rest on the measured reduction of damage-causing errors rather than on demonstrated compliance with the prompts. Finally, the five-phase decomposition functions as a design framework for allocating guidance, not as a manipulated experimental variable. Workload was measured once at task level, so the phase-specific demands described in Section 5.4 are a theoretically motivated basis for interface design rather than an empirically validated partition, and confirming them would require phase-resolved measurement.

Three next steps follow directly from the limitations above. First, a modality-isolating ablation comparing visual guidance with touch navigation against the full visual-plus-voice configuration, to apportion the observed effects between modalities. Second, cross-device validation of the SPL mapping across handheld and head-mounted hardware, with a sensitivity analysis of the calibration constants. Third, a field study with professional technicians under authentic noise and time pressure, incorporating step-resolved measures such as per-step completion times, eye-tracking and in-task recognition logging, together with confirmation gating for sustainability actions so that compliance is verified rather than assumed. Longer-term extensions, including physiological workload measurement, progression toward higher digital-twin maturity, gesture as a fallback modality, and application to other fluid-handling procedures, lie beyond the scope of the present validation.

8. CONCLUSION

The research introduces and empirically verifies a multi-modal AR system that is resilient to noise for sustainable industrial maintenance in Industry 5.0 water pump systems. The system brings the Operator 5.0 construct to life through visual AR guidance, interactive Exploded 3D Views, and a custom SPL Adaptation Algorithm for offline noise-robust real-time voice control. The framework acts as a cognitive orthosis for inexperienced operators, enabling them to carry out a complex maintenance procedure safely, efficiently and sustainably. The very high effect sizes reported here were obtained with participants deliberately selected for their lack of prior technical experience, and the findings should therefore be bounded accordingly: they establish the system's value for novice and cross-trained personnel entering an unfamiliar procedure, and do not extend to experienced maintenance technicians, whose established procedural routines and tactile expertise may interact differently with guided AR overlays. This helps to promote all three aspects of the TBL. The Social Pillar was promoted by reducing the cognitive workload by 55% ($d = 2.80$) and the frustration by 74% ($d = 4.14$), indicating potential for reducing procedural strain on the operator and for transferring procedural knowledge to inexperienced users. The Economic Pillar was promoted with a 30.11% reduction in TCT ($d = 2.03$), together with an observed absence of critical failures in the AR group (0 of 10 versus 2 of 10) that, while not statistically confirmed at this sample size, points to reduced costs from unplanned downtime and component replacement. The Environmental Pillar was promoted by a statistically robust decrease of 51.4% in execution errors ($d = 2.15$), which prevented the component damage that generates industrial scrap. In a controlled benchmark of 300 trials, the SPL adaptation mechanism-maintained recognition accuracy at 94% at 50-60 dB and 68% at 80-90 dB, supporting hands-free operation in realistic industrial noise without cloud connectivity. This constitutes an end-to-end validation of offline noise-adaptive voice control within a complete AR maintenance procedure. Immediate next steps are to isolate the contribution of each interaction modality and to validate the system with professional technicians under authentic industrial conditions. This study concludes that multi-modal AR is a key enabling technology for Industry 5.0, offering a scalable mechanism to reconcile industrial productivity with a holistic sustainability objective.

ACKNOWLEDGMENT

The authors are grateful to the 20 volunteers who participated in the empirical study. This research received no external funding.

REFERENCES

[1] Dourado, I.L., Moreira, L.C., Kauffman, M., Horan, B. (2024). A Wearable AR system for maintenance in Industry 5.0: Assessing mental workload and usability. In 2024 29th International Conference on Automation and Computing (ICAC), Sunderland, United Kingdom, pp. 1-6. https://doi.org/10.1109/ICAC61394.2024.10718838

[2] Oliveri, L., Chiacchio, F., Lo Iacono, N., Vitti, M., Facchini, F. (2025). Enhancing maintenance operations in Industry 5.0: A conceptual user interface design for task assignment. Procedia Computer Science, 253: 3015-3024. https://doi.org/10.1016/j.procs.2025.02.026

[3] Hasan, B.T., Abdullah, D.B. (2022). Real-time resource monitoring framework in a heterogeneous Kubernetes cluster. In 2022 Muthanna International Conference Engineering Science Technology, Samawah, Iraq, pp. 184-189. https://doi.org/10.1109/MICEST54286.2022.9790264

[4] Raffak, H., Ghatous, H., Mansouri, M., Lakhouilii, A. (2025). A data-driven predictive maintenance approach for Industry 4.0 using LSTM with cross-validation and the IDAIC framework. Journal Européen des Systèmes Automatisés, 58(1): 21-29. https://doi.org/10.18280/jesa.580103

[5] Calzavara, M., Faccio, M., Granata, I. (2023). Multi-objective task allocation for collaborative robot systems with an Industry 5.0 human-centered perspective. The

International Journal of Advanced Manufacturing Technology, 128(1): 297-314. https://doi.org/10.1007/s00170-023-11673-x
[6] Karbouj, B., Tobias Schuster, P.S., Blumhagen, M., Krüger, J. (2024). Optimizing human-robot collaboration in Industry 5.0: A comparative study of communication mediums and their impact on worker well-being and productivity. In 2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI), Washington, DC, USA, pp. 182-188. https://doi.org/10.1109/CogMI62246.2024.00032
[7] Liu, C., Tang, D., Wang, Z. (2023). AR-driven industrial metaverse for the auxiliary maintenance of machine tools in IoT-enabled manufacturing workshop. In 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), Auckland, New Zealand, pp. 1-6. https://doi.org/10.1109/CASE56687.2023.10260530
[8] Valentini, L., Grandi, F., Peruzzini, M., Pellicciari, M. (2025). A novel human-centric framework for maintenance digitisation using augmented reality. International Journal of Agile Systems and Management, 18(2): 239-260. https://doi.org/10.1504/IJASM.2025.145449
[9] Angelopoulos, J., Mourtzis, D. (2022). An intelligent product service system for adaptive maintenance of engineered-to-order manufacturing equipment assisted by augmented reality. Applied Sciences, 12(11): 5349. https://doi.org/10.3390/app12115349
[10] Saidi, H., Carreteros, L., Rey, S., Truscello, L., Miloudi, Y. (2022). Bl. MIXEDR: Augmenting traditional maintenance procedures to better exploit the capabilities of head-worn AR. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece, pp. 176-184. https://doi.org/10.1145/3529190.3529210
[11] Abdullah, D.B., Hasan, B.T. (2023). HRRMLQ: Container scheduling algorithm on edge nodes cluster. AIP Conference Proceedings, 2834(1): 50005. https://doi.org/10.1063/5.0171070
[12] Nagy, A., Spyridis, Y., Mills, G.J., Argyriou, V. (2024). User experience evaluation of AR assisted industrial maintenance and support applications. In 2024 10th International Conference on Virtual Reality (ICVR), Bournemouth, United Kingdom, pp. 270-277. https://doi.org/10.1109/ICVR62393.2024.10869099
[13] Wang, C., Lo, W., Wang, M.J. (2022). Usability evaluation of augmented reality-based maintenance instruction system. Human Factors and Ergonomics in Manufacturing and Service Industries, 32(3): 239-255. https://doi.org/10.1002/hfm.20942
[14] Eversberg, L., Lambrecht, J. (2023). Evaluating digital work instructions with augmented reality versus paper-based documents for manual, object-specific repair tasks in a case study with experienced workers. The International Journal of Advanced Manufacturing Technology, 127(3): 1859-1871. https://doi.org/10.1007/s00170-023-11313-4
[15] Hou, Z., Zhang, Q., Wang, S., He, W., Zhang, S. (2024). A blind area information perception and AR assembly guidance method based on RGBD data for dynamic environments and user study. The International Journal of Advanced Manufacturing Technology, 133(1): 115-128. https://doi.org/10.1007/s00170-024-13599-4
[16] Bini, S., Saggese, A., Vento, M. (2024). Enhancing noise robustness of speech-based human-robot interaction in industry. In European Robotics Forum, Cham: Springer Nature Switzerland, pp. 29-33. https://doi.org/10.1007/978-3-031-76428-8_6
[17] Dong, M., Peng, L., Nie, Q., Li, W. (2023). Speech signal processing of industrial speech recognition. Journal of Physics Conference Series, 2508(1): 012039. https://doi.org/10.1088/1742-6596/2508/1/012039
[18] Jourabchi Amirkhizi, P., Pedrammehr, S., Pakzad, S., Shahhoseini, A. (2025). Generative artificial intelligence in adaptive social manufacturing: A Pathway to achieving Industry 5.0 sustainability goals. Processes, 13(4): 1174. https://doi.org/10.3390/pr13041174
[19] Gbededo, M., Liyanage, K. (2024). Advancing sustainable manufacturing through digital twin-based simulation: Assessing and optimising processes for social, environmental and economic impacts. In 21st International Conference on Manufacturing Research (ICMR2024), Strathclyde, Glasgow, pp. 1-6. https://doi.org/10.1051/matecconf/202440105003
[20] Wu, M., Koh, Y.S., Yeong, C.F., et al. (2024). Optimizing assembly processes with augmented reality: A case study on TurtleBots. Indonesian Journal of Electrical Engineering and Computer Science, 35(3): 1547-1555. https://doi.org/10.11591/ijeecs.v35.i3.pp1547-1555
[21] Fernando, J., Oktavia, T. (2025). Evaluation of the utilization of augmented reality technology to increase interaction satisfaction in remote assistance. Journal Européen des Systèmes Automatisés, 58(4): 755-761. https://doi.org/10.18280/jesa.580409
[22] Mourtzis, D., Angelopoulos, J., Panopoulos, N. (2022). Closed-loop robotic arm manipulation based on mixed reality. Applied Sciences, 12(6): 2972. https://doi.org/10.3390/app12062972
[23] Begout, P., Kubicki, S., Bricard, E., Duval, T. (2022). Augmented reality authoring of digital twins: Design, implementation and evaluation in an industry 4.0 context. Frontiers in Virtual Reality, 3: 1-14. https://doi.org/10.3389/frvir.2022.918685
[24] Fang, W., Zhang, T., Chen, L., Cao, Y., Hu, H., Bi, J. (2026). HiMAR: A hierarchical context-aware mobile AR assembly for large-scale cable products. International Journal of Production Research, 64(12): 4808-4829. https://doi.org/10.1080/00207543.2025.2568740
[25] Lavric, T., Bricard, E., Preda, M., Zaharia, T. (2022). ATOFIS, an AR training system for manual assembly: A full comparative evaluation against guides. In 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Singapore, Singapore, pp. 558-567. https://doi.org/10.1109/ISMAR55827.2022.00072
[26] Hasan, N., Alkan, B. (2025). Gest-SAR: A gesture-controlled spatial AR system for interactive manual assembly guidance with real-time operational feedback. Machines, 13(8): 658ttps://doi.org/10.3390/machines13080658
[27] Cioflan, C., Cavigelli, L., Rusci, M., De Prado, M., Benini, L. (2022). Towards on-device domain adaptation for noise-robust keyword spotting. In 2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS), Incheon, Korea, pp. 82-85. https://doi.org/10.1109/AICAS54282.2022.9869990

[28] Stefaniak, P., Stachowiak, M., Koperska, W., Skoczylas, A., Śliwiński, P. (2022). Application of wearable computer and ASR technology in an underground mine to support mine supervision of the heavy machinery chamber. Sensors, 22(19): 7628. https://doi.org/10.3390/s22197628

[29] Valladares-Poncela, A., Fraga-Lamas, P., Fernández-Caramés, T.M. (2025). On-device automatic speech recognition for low-resource languages in mixed reality industrial metaverse applications: Practical guidelines and evaluation of a shipbuilding application in Galician. IEEE Access, 13: 77017-77038. https://doi.org/10.1109/ACCESS.2025.3564137

[30] Ojstersek, B., Javernik, R., Buchmeister, A. (2024). Integrating simulation modelling for sustainable, human-centred Industry 5.0: ESG-based evaluation in collaborative workplaces. Advances in Production Engineering & Management, 19(4): 527-538. https://doi.org/10.14743/apem2024.4.522

[31] Bassi, G., Orso, V., Salcuni, S., Gamberini, L. (2025). Understanding workers' well-being and cognitive load in human-cobot collaboration: Systematic review. Journal of Medical Internet Research, 27: e75658. https://doi.org/10.2196/75658

[32] Jongi, L., Vasdeki, V., Baalsrud Hauge, J. (2024). Towards sustainable and technology-enabled engineering psychology in production. In Sustainable Production through Advanced Manufacturing, Intelligent Automation and Work Integrated Learning: Proceedings of the 11th Swedish Production Symposium (SPS2024), pp. 346-357. https://doi.org/10.3233/ATDE240179

[33] Leins, N., Gonnermann-Müller, J., Teichmann, M. (2024). Comparing head-mounted and handheld augmented reality for guided assembly. Journal on Multimodal User Interfaces, 18(4): 313-328. https://doi.org/10.1007/s12193-024-00440-1

[34] Khalil, H., Jubran, O.F., Muntini, L., Wagner, A., Lachmann, T. (2025). Comparing AR interaction techniques and user interfaces for worker guidance systems for industrial applications. ECCE '25: Proceedings of the 36th Annual Conference of the European Association of Cognitive Ergonomics. https://doi.org/10.1145/3746175.3746216

[35] Abdullah, M., Vibhav, N., Rahman, M.M. (2024). Comparing the efficacy of virtual reality training, augmented reality instruction, and traditional paper-based instruction methods for manual assembly tasks. Human Factors in Design, Engineering, and Computing, 159: 1038-1047. https://doi.org/10.54941/ahfe1005670

[36] Morales Méndez, G., del Cerro Velázquez, F. (2024). Impact of augmented reality on assistance and training in Industry 4.0: Qualitative evaluation and meta-analysis. Applied Sciences, 14(11): 4564. https://doi.org/10.3390/app14114564

[37] Faul, F., Erdfelder, E., Lang, A.G., Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2): 175-191. https://doi.org/10.3758/BF03193146

[38] Hart, S.G. (2006). NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Sage CA: Los Angeles, CA: Sage Publications, 50(9): 904-908. https://doi.org/10.1177/154193120605000909

[39] Brooke, J. (1996). SUS-a quick and dirty usability scale. Usability Evaluation in Industry, 189: 4-7.

[40] Fau, A., Ghobrial, M., Seitier, P., et al. (2025). Enhancing performance in bolt torque tightening using a connected torque wrench and augmented reality. arXiv preprint arXiv:2505.10047. https://doi.org/10.48550/arXiv.2505.10047

[41] Agarwal, A., Sinha, I., Bhattacharyya, S., Mamodiya, U. (2025). Leveraging digital twin technology in industrial IoT for energy optimization and waste reduction. In Accelerating Product Development Cycles with Digital Twins and IoT Integration, IGI Global Scientific Publishing, pp. 301-322. https://doi.org/10.4018/979-8-3373-2028-1.ch015