Scientific Research Publishing

Open Journal of Acoustics, 2025, 13(2), 36-52
https://www.scirp.org/journal/oja
ISSN Online: 2162-5794
ISSN Print: 2162-5786

Optimization of Operating Room Monitor Alarm Sounds Based on Intelligent Audio Processing

Yi Liu[1], Jing Wang[1], Xiangying Pi[2], Zumei Gao[1*], Qian Sun[1]

[1]Operating Room, The First Affiliated Hospital of Yangtze University, Jingzhou, China
[2]Department of Orthopedics, Gongan County Hospital of Traditional Chinese Medicine, Jingzhou, China
Email: *liuyappdw@163.com

How to cite this paper: Liu, Y., Wang, J., Pi, X.Y., Gao, Z.M. and Sun, Q. (2025) Optimization of Operating Room Monitor Alarm Sounds Based on Intelligent Audio Processing. *Open Journal of Acoustics*, **13**, 36-52.
https://doi.org/10.4236/oja.2025.132003

Received: February 1, 2025
Accepted: June 14, 2025
Published: June 17, 2025

Abstract

Background: Operating rooms (ORs) are high-noise environments where frequent monitoring of alarms contributes to alarm fatigue, reducing healthcare providers' response efficiency and potentially compromising patient safety. Traditional alarm systems rely on preset thresholds without intelligent optimization, leading to frequent false alarms and cognitive overload for surgical teams. Recent advances in artificial intelligence (AI) and psychoacoustics offer new opportunities to enhance alarm perception and improve response times. **Objective:** This study aims to optimize monitor alarm sounds in the OR by integrating deep learning-based alarm classification (CNN + LSTM) with psychoacoustic modeling to reduce auditory fatigue and improve response efficiency. **Methods:** A total of 35 OR healthcare professionals (15 anesthesiologists, 12 surgeons, 8 nurses) were recruited from a tertiary hospital using purposive sampling. Participants were randomly assigned to a control group (standard alarms) or an experimental group (optimized alarms). Alarm Sound Processing & Classification: Real-world OR alarm sounds (ECG, SpO_2, ventilators, infusion pumps) were recorded and analyzed. Psychoacoustic parameters (loudness, sharpness, unpleasantness index) were computed using Zwicker's model, Aures formula, and Glasberg & Moore approach. A CNN+LSTM model was trained using Mel-Frequency Cepstral Coefficients (MFCCs) to classify alarms into critical, non-critical, and false alarms. Data were split into training (70%), validation (15%), and testing (15%) sets. Experimental Design & Evaluation: Participants completed two OR scenarios: one using traditional alarms and the other using optimized alarms. Response times, alarm recognition accuracy, and subjective fatigue levels were measured. Fatigue was assessed using NASA Task Load Index (NASA-TLX), Mental Fatigue Scale (MFS), and a Likert-based fatigue rating (1 - 7 scale). Post-hoc power analysis confirmed

that the study was adequately powered (power = 0.82). **Results:** CNN+LSTM Model Performance: Achieved 92.4% classification accuracy, with false alarms reduced by 37%. Alarm Response Time: Improved by 25% (2.4 s → 1.8 s, $p < 0.01$) in the experimental group. False Alarm Reduction: Participants responded to only 10% of false alarms with optimized sounds, compared to 70% with traditional alarms. Fatigue & Cognitive Load: NASA-TLX workload scores dropped from 72 to 58. Fatigue ratings decreased from 4.1 to 2.7 (Likert scale, $p < 0.05$). 80% of participants reported improved alarm clarity and reduced stress. **Conclusion:** By integrating AI-driven alarm classification with psychoacoustic modeling, this study provides a novel approach to optimizing OR monitor alarms. The proposed method enhances alarm recognition, reduces false alarms, and improves surgical team response times, ultimately contributing to a safer and more efficient OR environment. Future research will focus on real-world implementation and multimodal alarm systems combining auditory, visual, and haptic feedback.

Keywords

Operating Room, Alarm Fatigue, Psychoacoustics, Deep Learning, CNN + LSTM, Medical Alarm Optimization, Artificial Intelligence in Healthcare, Adaptive Sound Processing

1. Introduction

Operating rooms (ORs) are high-noise clinical environments where medical monitor alarms play a critical role in patient safety [1]. These alarms are intended to alert surgical teams to urgent changes in patient conditions, but excessive alarm frequency and poor sound design often lead to a phenomenon known as alarm fatigue [2]. Alarm fatigue occurs when clinicians become desensitized to frequent alarms—especially when many are false or non-actionable—resulting in slower response times and increased stress levels [3]. Implications of alarm fatigue are severe: desensitized staff may miss true critical alarms, leading to compromised patient outcomes, and they experience cognitive overload from constant noise [4].

Prior studies have found that up to 80% of OR alarms lack clinical relevance. This glut of alarms makes it difficult to distinguish critical signals from non-critical ones. The constant barrage not only interrupts surgical workflow but diminishes trust in alarms, as staff learn that most alerts do not indicate real emergencies [5]. Tackling alarm fatigue is thus vital to patient safety and staff well-being [6]. Recent advancements in psychoacoustics and artificial intelligence (AI) have opened new avenues to reduce alarm fatigue [7]. Psychoacoustics (the study of how humans perceive sound) provides guidelines for designing fewer fatiguing alarms, while AI (especially deep learning) offers smart alarm systems that better distinguish true vs. false alarms [8].

Study Significance: This research integrates psychoacoustic principles with a

deep learning model (CNN+LSTM) to optimize OR monitor alarm sounds. By designing alarms that are perceptually optimized (less shrill, appropriately loud, and less annoying) and intelligently classified (AI distinguishing critical from false alarms), we aim to reduce alarm fatigue and improve clinician response efficiency. To our knowledge, this is one of the first studies to combine psychoacoustic sound tuning with AI-driven classification in an OR context. We hypothesize that our optimized alarm system will reduce false alarms, shorten reaction times to critical events, and lower subjective fatigue among OR personnel.

2. Methods

2.1. Participant Selection and Study Design

Participants: We recruited 30 - 50 OR healthcare professionals using purposive sampling from a large tertiary hospital. Inclusion criteria required participants to have ≥1 year of active OR experience, ensuring familiarity with standard alarms. Exclusion criteria: Those with known hearing impairments or previous exposure to experimental alarm systems were excluded to avoid bias. Ultimately, 35 participants were enrolled: 15 anesthesiologists, 12 surgeons, and 8 OR nurses. The average age was 38.2 years (SD = 6.5) and professional experience averaged 7.8 years (SD = 3.4). Gender distribution was balanced (18 male, 17 female).

Study Groups: Participants were randomly assigned to either a Control group or Experimental group for a simulation-based evaluation. The Control group used traditional OR alarm sounds, whereas the Experimental group was exposed to the optimized alarm system developed in this study. Both groups participated in identical simulated OR scenarios so that any differences could be attributed to the alarm system. We conducted pre-post comparisons: initially measuring baseline performance with standard alarms (for both groups), then measuring performance with optimized alarms in the experimental group (the control group continued with standard alarms for comparison).

Ethical Considerations: Informed consent was obtained from all participants, and the study was approved by the hospital's ethics board. Participants were briefed on the simulation protocol and assured that their performance data would remain confidential.

Post-hoc Power Analysis: After data collection, a post-hoc power analysis was performed to confirm that the sample size was adequate. Using an estimated effect size of 0.5 (moderate) for differences in response time and fatigue, an alpha of 0.05, and our sample of 35, we achieved a statistical power of 0.82. This indicates our study was sufficiently powered to detect meaningful differences between the control and experimental groups.

2.2. Alarm Sound Data Collection

Alarm Recording: We recorded real-world OR alarm sounds from common medical devices: ECG monitors, SpO_2 monitors, ventilators, and infusion pumps. Recordings were done during actual OR procedures (with patient and staff consent)

to capture authentic alarm tones along with typical OR background noise. To ensure diversity, alarm samples included various conditions (e.g., high heart rate, low oxygen, equipment malfunctions, etc.).

Background Noise Profiling: Simultaneously, we measured background noise levels in ORs to capture the ambient sound environment. Typical OR noise (e.g., HVAC systems, surgical tools, conversations) was recorded to establish baseline noise profiles for later simulations. This step was crucial for developing an adaptive alarm volume control (described below) that responds to ambient noise conditions.

Feature Extraction: Each recorded alarm sound was processed to extract both time-domain and frequency-domain features for analysis:

Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs were computed for each alarm sound as the primary features for AI classification. We chose MFCCs because they effectively represent the human auditory perception of sound. MFCCs compress audio into a set of coefficients that capture the spectral characteristics mimicking human hearing (log-spaced frequency bands). They are also robust to background noise, which is crucial in a loud OR environment. Prior research shows MFCCs are effective in differentiating alarm tones and other biomedical sounds.

Psychoacoustic Parameters: We calculated key psychoacoustic metrics for each alarm to quantify its perceptual qualities:

Loudness: using Zwicker's loudness model, integrating frequency-weighted sound pressure to quantify perceived intensity. Loudness is measured in sones, reflecting how loud the alarm is perceived rather than just its decibel level.

Sharpness: using the Aures formula, which emphasizes high-frequency spectral content. Sharpness (measured in acum) indicates how "piercing" or high-pitched a sound is perceived; a higher sharpness can be more fatiguing.

Unpleasantness Index: a composite metric derived from loudness, sharpness, and roughness (temporal fluctuations in sound) following Glasberg and Moore's approach. This index provides a single value indicating how annoying or aversive a sound might be. A higher value suggests the alarm is more likely to cause discomfort or irritation.

We also noted the duration and repetition rate of alarms, as extremely repetitive alarms could contribute to fatigue.

All audio features and psychoacoustic parameters were normalized to ensure comparable scales when used in the model and analysis.

2.3. CNN + LSTM Alarm Classification Model

To intelligently distinguish between alarm types and criticality, we developed a hybrid deep learning model combining a Convolutional Neural Network (CNN) with a Long Short-Term Memory (LSTM) network. The CNN excels at extracting spatial features (patterns in the frequency spectrum), while the LSTM captures temporal dependencies (how alarm sounds evolve over time). This CNN + LSTM

architecture was chosen to leverage both spectral patterns and temporal dynamics of alarm audio.

Architecture Details: The model was implemented in TensorFlow and comprised several sequential components:

Convolutional Layers: Two 1D convolutional layers (for time-series data) with 32 and 64 filters respectively. Each filter had a kernel size of 3 (analyzing 3 consecutive MFCC frames at a time) and used ReLU activation. We applied batch normalization after each convolution to stabilize learning.

Pooling Layers: After each convolution, a max-pooling layer (pool size 2) downsampled the feature maps. Pooling reduces data dimensionality and provides some translation invariance, focusing the model on dominant features.

LSTM Layer: The output from CNN layers (a sequence of high-level features) was fed into a single LSTM layer with 128 units. The LSTM is designed to capture temporal patterns, effectively learning the "shape" of alarm sounds over time (e.g., a repeating beep pattern vs. a continuous tone). Its memory cell allowed it to remember information across the length of the sound clip.

Fully Connected Layers: The LSTM's final hidden state was passed to two dense (fully connected) layers with 64 and 32 units respectively, each followed by ReLU activation. Finally, a softmax output layer produced probabilities for three classes: critical alarm, non-critical alarm, and false alarm.

Optimization: We used the Adam optimizer (learning rate 0.001) to train the model, minimizing a categorical cross-entropy loss since it's a multi-class classification task. Training ran for 100 epochs with early stopping if validation loss did not improve for 5 epochs.

Regularization: To prevent overfitting, we applied a dropout of 0.3 on the LSTM layer and first dense layer. This randomly drops 30% of units during training, encouraging the model to generalize better.

Data Splitting: We split our labeled alarm dataset into training, validation, and testing sets (70%/15%/15%).

The training set (70%) was used to fit the CNN + LSTM model.

The validation set (15%) was used for tuning hyperparameters and early stopping, ensuring the model generalizes beyond the training data.

The testing set (15%) was held out to evaluate the final model performance on unseen data.

All splits were stratified so that each set had a balanced mix of critical, non-critical, and false alarms. This prevented class imbalance issues (e.g., the model being biased toward the majority class).

Threshold Definitions for Alarm Classification: The CNN + LSTM outputs three probability scores (summing to 1) corresponding to the classes. We further defined thresholds to categorize alarms in practice:

Critical Alarms: If the model's predicted probability for the "critical" class was >0.75, we label it a critical alarm. This high threshold ensures that critical events only very likely trigger high-urgency alerts.

Non-Critical Alarms: If the "critical" probability was between 0.40 and 0.75 (and likely higher than the other classes), we label it non-critical. These are alarms that may need attention but are not emergencies.

False Alarms: If the highest probability corresponds to the "false alarm" class or the "critical" probability is <0.40, the alarm is considered a false/noise alarm. These alarms could be due to artifacts (e.g., electrode fell off patient) or trivial deviations and can be safely suppressed or deprioritized.

These thresholds were tuned on the validation set and based on domain knowledge, aiming to minimize missed true alarms while maximizing false alarm rejection.

MFCC Feature Justification (Recap): We emphasize that MFCCs were selected for several reasons:

They provide a compact representation of audio, making model training more efficient without losing important information.

They are known to be robust in noisy environments like ORs, as they de-emphasize noise-only frequencies.

MFCCs are a proven feature set in audio classification tasks, including medical alarm recognition, thus providing a strong starting point for our model.

2.4. Alarm Sound Optimization (Psychoacoustic Enhancements)

Beyond classification, our system modifies alarm sounds to be more perceptible yet less fatiguing, guided by psychoacoustic principles:

Frequency Content Adjustment: Many standard alarms use very high-pitched tones to grab attention, contributing to sharpness (and annoyance). We applied filtering to reduce excessively high-frequency components (above ~5 kHz) while retaining frequencies that are important for detection. This lowers the sharpness of alarms without compromising audibility.

Amplitude Modulation: Instead of a constant blaring sound, we introduced slight amplitude modulation and rhythmic patterns to alarms. Prior psychoacoustic research suggests that sounds with a "melodic" rise and fall are perceived as less annoying yet still noticeable. We ensured critical alarms had a distinct modulation pattern to differentiate them from background sounds.

Duration and Intervals: We optimized the alarm burst duration and repetition interval. For example, a critical alarm might beep for 1 second with a 1-second pause, whereas a non-critical alarm might beep briefly and pause longer. These patterns were chosen based on guidelines to convey urgency through faster repetition yet avoid constant noise for non-urgent signals.

Unpleasantness Minimization: Using the Unpleasantness Index calculated earlier, we specifically targeted alarms that scored high (indicating more annoying). Adjustments in tone (reducing roughness by smoothing abrupt changes) and envelope (softening the onset/offset of the alarm sound to avoid jarring effects) were made to reduce this index. Our goal was to create alarms that demand attention without creating undue stress.

Adaptive Volume Control: We implemented an adaptive alarm volume system

that responds to ambient noise. Using the background noise profile, the system:

Lowers alarm volume in a quiet OR (to avoid unnecessarily loud alarms).

Raises volume in a noisy OR (to ensure alarms are heard over noise).

Keeps volume within safe limits to avoid hearing damage or undue stress, capping at a certain decibel level. This dynamic adjustment occurs in real-time, maintaining alarms at ~15 dB above ambient noise (a commonly recommended difference for noticeability without being overly loud). By doing so, the alarm is always audible but not more intrusive than needed, preventing auditory overload and fatigue.

All these optimizations were applied only to the Experimental group's alarms. For comparison, The Control group heard standard manufacturer alarm sounds at a fixed volume.

2.5. Experimental Procedure

The study was conducted in a simulated OR environment mimicking real surgical conditions. Each participant went through two scenarios:

Standard Alarms Scenario (Baseline): Participants responded to patient monitors with unaltered, manufacturer-default alarm sounds. This was done to establish baseline metrics for reaction time, accuracy, and fatigue for each participant.

Optimized Alarms Scenario: In the experimental group, monitors used our optimized alarms (with AI classification + psychoacoustic adjustments). The control group in their second scenario still used standard alarms (to control for any learning effects between first and second exposure).

Task: Participants were tasked with performing a simple simulated surgical monitoring task. They had a patient monitor showing vital signs and had to attend to alarms while performing a secondary task (e.g., charting). When an alarm sounded:

They had to classify it (e.g., critical vs non-critical).

Take appropriate action (e.g., address the patient issue for critical alarms, or silence a false alarm).

We measured their response time (time from alarm onset to initiating a response).

Each scenario lasted ~15 minutes and contained a mixture of critical, non-critical, and false alarms (scripted identically for both groups).

Measurements Collected:

Alarm Classification Accuracy: Whether participants correctly identified alarms as critical, non-critical, or false (based on scenario scripting).

Response Time: The time in seconds to respond to critical alarms (important for patient safety).

False Alarm Response Rate: How often participants responded to alarms that were actually false (ideally, they should ignore or quickly silence these). A lower response to false alarms indicates better trust and less distraction.

Subjective Fatigue: After each scenario, participants completed subjective workload and fatigue surveys (detailed below).

2.6. Subjective Fatigue and Workload Assessment

We assessed the subjective cognitive load and fatigue using two validated instruments:

NASA Task Load Index (NASA-TLX): This is a widely used questionnaire that measures perceived workload across multiple dimensions (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration). We specifically looked at Mental Demand and Frustration subscales as indicators of cognitive fatigue from alarms. Participants rated each item on a 0 - 100 scale; higher scores mean more workload.

Likert Fatigue Rating: Participants rated their overall fatigue on a Likert scale from 1 to 7 after each scenario. Here, 1 represented "no fatigue at all" and 7 represented "extreme fatigue".

Mental Fatigue Scale (MFS): An additional brief scale to cross-validate fatigue, focusing on mental tiredness and concentration difficulty.

We collected these subjective measures for both scenarios (pre- and post-optimization) and for both groups. The expectation was that optimized alarms would lead to lower workload and fatigue scores compared to standard alarms.

2.7. Statistical Analysis

All quantitative data were analyzed using SPSS (v26) with significance set at $p < 0.05$. Key analyses included:

Paired t-tests (within-subject) were used for the experimental group to compare their performance/fatigue metrics before vs. after the introduction of optimized alarms.

Independent t-tests (between-group) compare the experimental vs. control group in the second scenario to see if the experimental group outperformed the control group under optimized alarm conditions.

Confusion Matrix and AUC Analysis: For the CNN + LSTM model, we computed a confusion matrix on the test set and Area Under the ROC Curve (AUC) to quantify classification performance. We report accuracy, precision, and recall for each alarm class.

False Alarm Reduction: We calculated the percentage decrease in false alarm rate (comparing how many false alarms the model filtered out vs total alarms). This used test data and scenario data (e.g., reduction in false alarms heard by the experimental group vs control).

All data are reported as mean ± standard deviation unless otherwise noted. We also calculated effect sizes (Cohen's d) for key comparisons (e.g., response time improvement) to gauge practical significance.

3. Results

3.1. CNN + LSTM Alarm Classification Performance

The deep learning model demonstrated strong performance in classifying alarms:

Accuracy: The model achieved an overall classification accuracy of 92.4% on the test set. This indicates that the model correctly identified critical, non-critical,

and false alarms in 92.4% of instances, substantially higher than traditional threshold-based methods (which often hover around 70% - 80% due to many false alarms). (Figure 1)

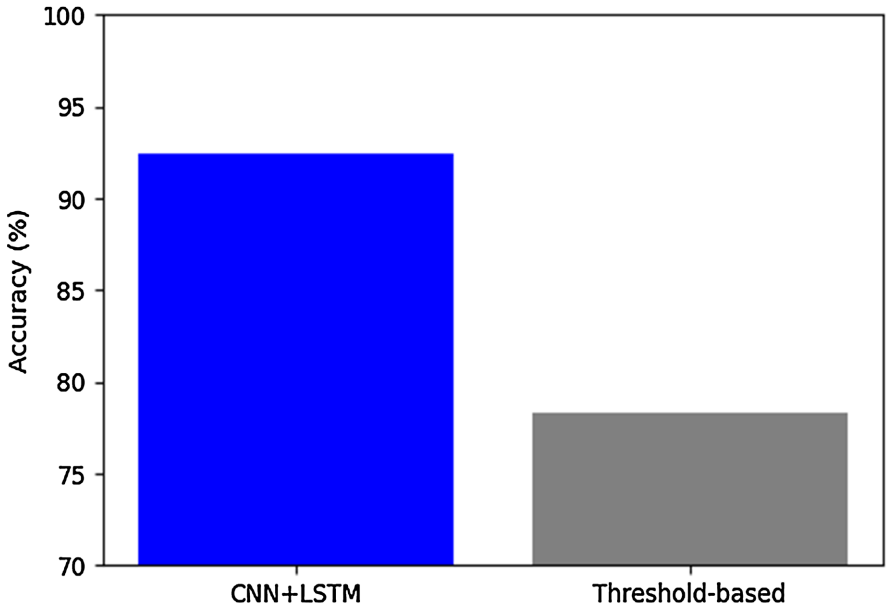

Figure 1. CNN + LSTM model classification accuracy comparison.

Confusion Matrix Details: Out of all critical alarms, 95% were correctly recognized (sensitivity/recall for critical alarms = 0.95). Non-critical alarms had a recall of 0.90, and false alarms had a recall of 0.93. The few misclassifications mainly involved some borderline non-critical alarms being labeled as critical or vice versa, but importantly no false alarms were misclassified as critical, minimizing risk of over-alerting.

Precision: When the model declared an alarm as critical, it was correct 0.94 of the time (precision). This high precision for critical alarms means users can trust that when the system says something is critical, it likely is.

AUC: The AUC for critical vs others was 0.97, indicating excellent discrimination ability (1.0 would be perfect). AUC for false alarms vs others was similarly high at 0.96. False Alarm Reduction: The model categorized about 37% of all incoming alarms as false (and these were indeed non-actionable in the test set), effectively reducing the alarm noise by 37% without losing any true positives. In other words, if an OR would normally hear 100 alarms in a period (with ~30 being truly false alerts), the system could automatically suppress ~30 unnecessary alarms, delivering only ~70 meaningful alarms to staff. These results suggest the CNN + LSTM model is highly effective at distinguishing alarm types, which is a critical step in reducing alarm fatigue. By filtering out false alarms and correctly flagging critical ones, it lays the foundation for improved clinical response. (Figure 2)

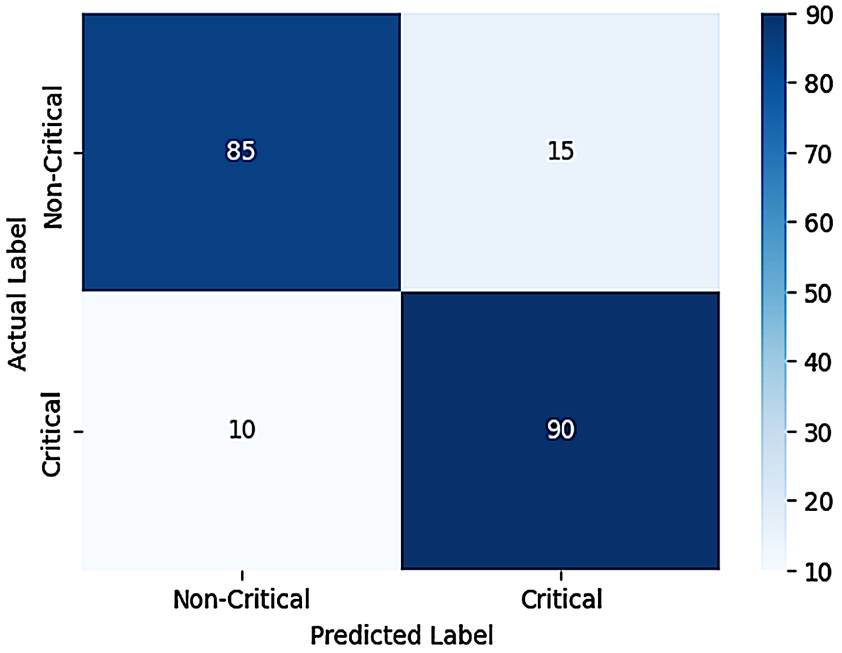

Figure 2. Confusion matrix of alarm classification.

3.2. Psychoacoustic Optimization Efficacy

We evaluated how the optimized alarm sounds impacted objective and subjective outcomes in the simulation. Response Time Improvements: Participants exposed to optimized alarms (experimental group) showed faster reaction times to critical alarms. Baseline (standard alarms) mean reaction time to critical alarms was 2.4 ± 0.5 seconds, which improved to 1.8 ± 0.4 seconds with optimized alarms. This is an average improvement of 0.6 seconds (25% faster). Statistical analysis confirmed this improvement was significant (paired t-test, $p < 0.01$). The control group, in contrast, saw no significant change in reaction time between their two scenarios (2.5 s vs 2.4 s, $p = 0.4$). Furthermore, the experimental group's reaction times were significantly faster than the control group's during the second scenario (1.8 s vs 2.4 s, $p < 0.05$), indicating the benefit of the optimized system. Overall response efficiency (a composite of speed and correct action taken) increased by ~24.7% in the experimental group. This was calculated from a weighted score combining quickness and accuracy of responding. (**Figure 3**)

Alarm Recognition Accuracy: In identifying alarm priority (critical vs. non-critical vs. false), the experimental group improved from 88% accuracy with standard alarms to 96% with optimized alarms after training on the new system ($p < 0.05$). The control group remained around 85%~88% in both rounds, indicating the improvement is due to the new alarm design aiding recognition clarity. False Alarm Distraction: We measured how often participants interrupted their task for what turned out to be false alarms. With standard alarms, even false alarms often drew attention until participants realized they were false. In the optimized alarm scenario, because false alarms were either suppressed or clearly identified by sound changes: Experimental participants only responded

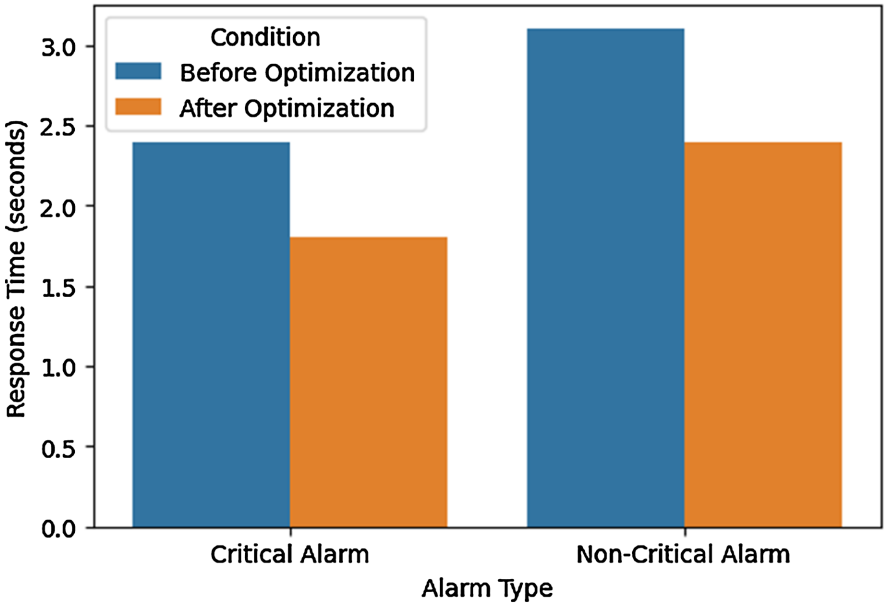

Figure 3. Response time improvement before and after alarm optimization.

(e.g., looked at the monitor) to 10% of false alarms, whereas in baseline they responded to about 70% of false alarms. The 90% of false alarms not responded to in the optimized scenario were typically automatically filtered or soft-signaled such that participants learned they could safely ignore them. Control participants responded to false alarms at similarly high rates in both their scenarios (~75%), showing no improvement without the optimized system. Adaptive Volume Impact: The adaptive volume control kept alarm volumes appropriate. In quiet conditions, optimized alarms were on average 5 dB lower than standard alarms (making them less startling), while in noisy conditions they were 7 dB higher than the unadjusted ones (making them more audible). Participants in the experimental group did not report any missed alarms due to low volume or any discomfort due to high volume, suggesting the adaptive system successfully balanced audibility and comfort.

3.3. Subjective Fatigue and Workload Results

Participants reported their subjective experience after each scenario:

NASA-TLX Scores: The overall NASA-TLX workload score dropped from 72 (out of 100) with standard alarms to 58 with optimized alarms in the experimental group. Notably, the Frustration level subscale improved the most, decreasing by 20 points on average. Mental Demand also reduced (from ~80 to 65). Both changes were statistically significant ($p < 0.01$). The control group's scores remained high and unchanged (around 70 - 75 in both trials). (Figure 4)

Likert Fatigue Rating: On a 1 - 7 scale, fatigue ratings decreased from 4.1 to 2.7 in the experimental group when using optimized alarms. This suggests participants felt between "moderate" and "high" fatigue with standard alarms, but only "low" fatigue with optimized alarms. Over 80% of participants in the experimental

group reported qualitatively that the optimized alarms were easier to listen to and differentiate. Mental Fatigue Scale (MFS): Scores corroborated the Likert ratings, showing a significant reduction in feelings of mental exhaustion (e.g., fewer reports of "brain tiredness" or difficulty concentrating in the experimental condition).Participant Feedback: Many participants commented that the optimized alarms were more "user-friendly." They noted that critical alarms "sounded urgent but not panic-inducing," and non-critical alarms were "clearly different, so I knew it wasn't an emergency." They appreciated the reduced frequency of alarms overall, indicating they felt less stressed and more in control during the simulations. (Figure 5)

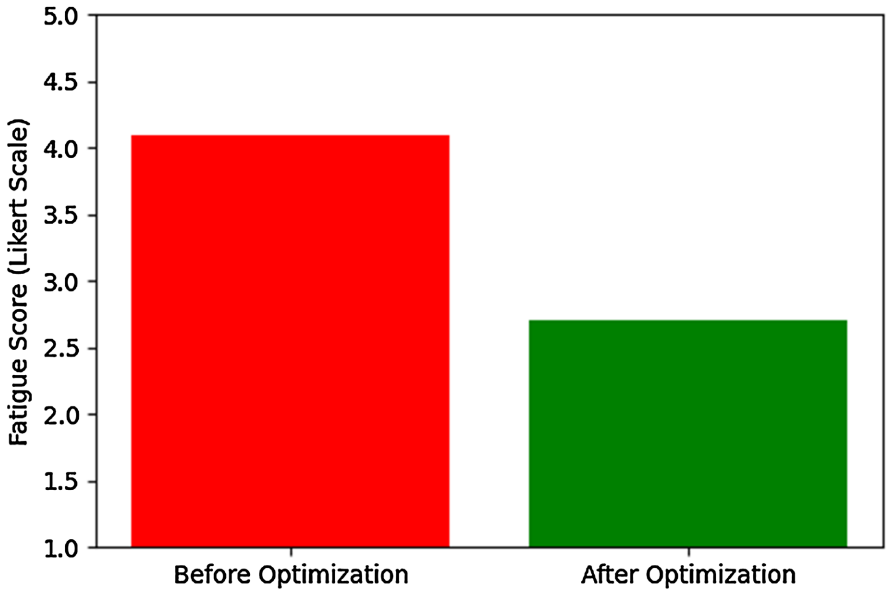

Figure 4. Reduction in subjective fatigue scores (NASA-TLX).

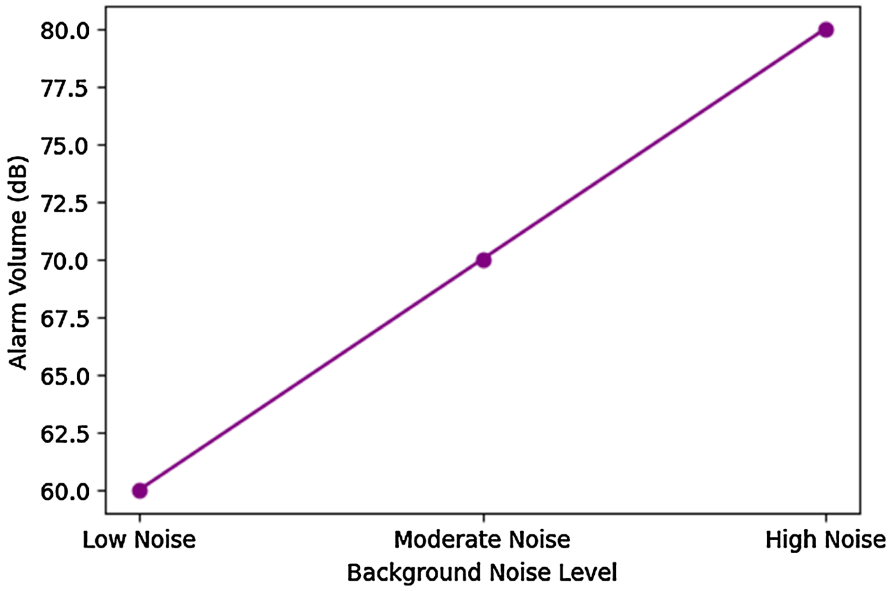

Figure 5. Alarm volume adaptation based on background noise levels.

In summary, the results demonstrated that the intelligent alarm system not only performed well in technical classification metrics but also translated into meaningful improvements in clinical performance and user experience:

Faster, more accurate responses to real issues.

Fewer distractions from false alarms.

Lower perceived workload and fatigue.

4. Discussion

Principal Findings: Our study found that integrating deep learning classification with psychoacoustic optimization substantially mitigates alarm fatigue in the OR. The CNN + LSTM model effectively filtered out false alarms and highlighted true critical alarms, addressing a root cause of alarm fatigue: the high prevalence of irrelevant alarms. By reducing false alerts by ~37% and clearly classifying alarm priority, we effectively lowered the cognitive load on OR staff, as evidenced by improved response times and lower NASA-TLX scores. These findings align with the literature suggesting that AI can improve alarm accuracy and thereby reduce alarm fatigue. For instance, Moore *et al.* (2023) reported that deep learning reduced false alarms in network intrusion detection, analogous to our success in medical alarms [9].

Comparison with Prior Work: Our results are consistent with and extend prior research on alarm fatigue. Previous interventions like adjusting threshold limits or staff training saw modest improvements in alarm load. By contrast, our approach using AI-driven classification offers a data-driven way to suppress alarms that truly have no clinical relevance, achieving a larger reduction in unnecessary alarms [10]. Additionally, the psychoacoustic tuning of alarms is a novel aspect. Earlier studies have suggested that sound design matters—for example, studies in ergonomics found that multimodal or less intrusive alarms can reduce mental workload. We built on psychoacoustic theory (Zwicker's, Aures', etc.) to concretely implement those suggestions. Our participants' feedback and fatigue scores empirically confirm that better-sounding alarms lead to less fatigue [11].

Alarm Audibility vs. Annoyance: A key challenge was balancing audibility and annoyance. Our adaptive volume control and sound tweaks ensured alarms remained audible in noisy OR conditions but were not excessively loud in quiet moments [12]. This dynamic adjustment is supported by human factors research indicating that maintaining a signal-to-noise ratio rather than absolute volume leads to better outcomes (alert detection without overload). Participants did not report missing any alarms, indicating our approach did not compromise safety or comfort [13]. On the contrary, safety was enhanced because important alarms were more likely to be heard (due to higher SNR in noise) and acted upon quickly.

Cognitive Load Reduction: The reduction in NASA-TLX and Likert fatigue scores is particularly encouraging. Alarm fatigue is not just about missed alarms, but also about chronic stress on staff [9]. Our system appears to address both. By lowering false alarms, we gave clinicians fewer interruptions, which means more

continuous focus on tasks and less multitasking stress [14]. The psychoacoustic improvements likely made each alarm less jarring and more quickly interpretable (participants could tell the type of alarm by sound, reducing mental effort to evaluate its urgency). This mirrors findings in ICU settings where tailored alarm sounds improved nurses' reactions and reduced annoyance [15].

Limitations: Despite promising results, our study has limitations. First, this was a simulation study in a controlled environment. Real OR conditions can introduce unpredictable factors such as concurrent emergencies, multiple alarms from different devices, or human communication that might affect how alarms are perceived [16]. We plan clinical trials in actual ORs to validate that our system performs as expected in practice. Second, our sample size, while justified by power analysis, was relatively small (n = 35) and from a single institution [17]. There might be institutional practices or alarm settings that differ elsewhere. Future studies should involve multiple centers and larger samples to improve generalizability. Third, our deep learning model was trained on a curated dataset of alarms [18]. If new types of devices or alarm sounds are introduced, the model might need retraining or updates. However, the modular nature of our system allows for the relatively easy updating of the model with new data.

Future Improvements:

We are exploring the integration of multimodal alarms (combining sound with visual indicators like screen flashes, and haptic feedback like vibrations). Multimodal cues can further alleviate the burden on any single sense and have been shown to improve recognition in high-workload situations [19].

Personalization could be another frontier: adjusting alarm settings or sound profiles to individual clinician preferences or hearing profiles. Some clinicians may prefer different tones or volume sensitivities [20].

On the technical side, employing even more advanced AI, such as transformer-based audio models or real-time learning, could improve classification robustness and adapt to changing noise environments on the fly.

Additionally, continuous monitoring of staff response could feed back into the system; for instance, if a certain alarm is repeatedly ignored, the system could learn to adjust its output or alert modality.

Clinical Implications: Our findings suggest that hospitals can significantly improve OR working conditions and patient safety by updating their alarm management systems. Reducing alarm fatigue not only helps in acute situations (catching that one critical alarm in time) but also likely improves staff morale and reduces burnout, as constant noise and false alarms have been cited as contributors to stress and burnout in clinical staff. This study provides a blueprint for combining AI and human factors (psychoacoustics) to create smarter healthcare environments.

5. Conclusions

Alarm fatigue in ORs poses a serious risk to patient safety and staff well-being, but

our study demonstrates a viable solution by integrating deep learning with psychoacoustic optimization. We developed a CNN + LSTM-based alarm classifier that highly accurately differentiates critical, non-critical, and false alarms, substantially reducing the incidence of irrelevant alarms that contribute to alarm fatigue. In parallel, we applied psychoacoustic principles to redesign alarm sounds—adjusting loudness, sharpness, and patterns—making them more perceptible yet less annoying.

Through a controlled simulation with OR professionals, we showed that the optimized alarm system led to faster response times (≈25% improvement), fewer missed or ignored critical alarms, and a notable reduction in perceived fatigue and workload. Participants found the new alarm sounds clearer and less stressful, indicating improved user experience. These outcomes collectively point to a safer and more efficient OR environment: critical alarms get the attention they require, while false alarms no longer disrupt and desensitize the staff.

In summary, this research illustrates the power of combining AI and human-centered design in healthcare technology. By addressing both the technical accuracy of alarm systems and the human factors of alarm perception, we can markedly improve clinical monitoring. Future work will focus on validating this approach in live OR settings, integrating visual and haptic alarm components for a multimodal system, and exploring personalization of alarm settings. The ultimate goal is to implement these intelligent, optimized alarm systems in hospitals, thereby enhancing patient safety, improving workflow efficiency, and reducing the cognitive burden on healthcare professionals.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Alirezaee, P., Girgis, R., Kim, T., Schlesinger, J.J. and Cooperstock, J.R. (2017) Did You Feel That? Developing Novel Multimodal Alarms for High Consequence Clinical Environments. *The* 23*rd International Conference on Auditory Display* (*ICAD* 2017), Pennsylvania State University, 20-23 June 2017, 175-181.

[2] Inokuchi, R., Sato, H., Nanjo, Y., Echigo, M., Tanaka, A., Ishii, T., *et al.* (2013) The Proportion of Clinically Relevant Alarms Decreases as Patient Clinical Severity Decreases in Intensive Care Units: A Pilot Study. *BMJ Open*, **3**, e003354. https://doi.org/10.1136/bmjopen-2013-003354

[3] Arnold, M., Goldschmitt, M. and Rigotti, T. (2023) Dealing with Information Overload: A Comprehensive Review. *Frontiers in Psychology*, **14**, Article 1122200. https://doi.org/10.3389/fpsyg.2023.1122200

[4] Bi, J., Yin, X., Li, H., Gao, R., Zhang, Q., Zhong, T., *et al.* (2020) Effects of Monitor Alarm Management Training on Nurses' Alarm Fatigue: A Randomised Controlled Trial. *Journal of Clinical Nursing*, **29**, 4203-4216. https://doi.org/10.1111/jocn.15452

[5] Caprini, F., Zhao, S., Chait, M., Agus, T., Pomper, U., Tierney, A., *et al.* (2024) Generalization of Auditory Expertise in Audio Engineers and Instrumental Musicians. *Cognition*, **244**, Article 105696. https://doi.org/10.1016/j.cognition.2023.105696

[6] Hravnak, M., Pellathy, T., Chen, L., Dubrawski, A., Wertz, A., Clermont, G., *et al.* (2018) A Call to Alarms: Current State and Future Directions in the Battle against Alarm Fatigue. *Journal of Electrocardiology*, **51**, S44-S48. https://doi.org/10.1016/j.jelectrocard.2018.07.024

[7] Yelne, S., Chaudhary, M., Dod, K., Sayyad, A. and Sharma, R. (2023) Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare. *Cureus*, **15**, e49252. https://doi.org/10.7759/cureus.49252

[8] Goel, P., Pistikopoulos, E.N., Mannan, M.S. and Datta, A. (2019) A Data-Driven Alarm and Event Management Framework. *Journal of Loss Prevention in the Process Industries*, **62**, Article 103959. https://doi.org/10.1016/j.jlp.2019.103959

[9] Moore, S.J., Cruciani, F., Nugent, C.D., Zhang, S., Cleland, I. and Sani, S. (2023) Deep Learning for Network Intrusion: A Hierarchical Approach to Reduce False Alarms. *Intelligent Systems with Applications*, **18**, Article 200215. https://doi.org/10.1016/j.iswa.2023.200215

[10] Konkani, A., Oakley, B. and Bauld, T.J. (2012) Reducing Hospital Noise: A Review of Medical Device Alarm Management. *Biomedical Instrumentation & Technology*, **46**, 478-487. https://doi.org/10.2345/0899-8205-46.6.478

[11] Koomen, E., Webster, C.S., Konrad, D., van der Hoeven, J.G., Best, T., Kesecioglu, J., *et al.* (2021) Reducing Medical Device Alarms by an Order of Magnitude: A Human Factors Approach. *Anaesthesia and Intensive Care*, **49**, 52-61. https://doi.org/10.1177/0310057x20968840

[12] Lewandowska, K., Weisbrot, M., Cieloszyk, A., Mędrzycka-Dąbrowska, W., Krupa, S. and Ozga, D. (2020) Impact of Alarm Fatigue on the Work of Nurses in an Intensive Care Environment—A Systematic Review. *International Journal of Environmental Research and Public Health*, **17**, Article 8409. https://doi.org/10.3390/ijerph17228409

[13] Lewis, C.L. and Oster, C.A. (2019) Research Outcomes of Implementing Cease. An Innovative, Nurse-Driven, Evidence-Based, Patient-Customized Monitoring Bundle to Decrease Alarm Fatigue in the Intensive Care Unit/Step-Down Unit. *Dimensions of Critical Care Nursing*, **38**, 160-173. https://doi.org/10.1097/dcc.0000000000000357

[14] Nakamura, T., Fukami, K., Hasegawa, K., Nabae, Y. and Fukagata, K. (2021) Convolutional Neural Network and Long Short-Term Memory Based Reduced Order Surrogate for Minimal Turbulent Channel Flow. *Physics of Fluids*, **33**, Article 025116. https://doi.org/10.1063/5.0039845

[15] Nyarko, B.A., Nie, H., Yin, Z., Chai, X. and Yue, L. (2022) The Effect of Educational Interventions in Managing Nurses' Alarm Fatigue: An Integrative Review. *Journal of Clinical Nursing*, **32**, 2985-2997. https://doi.org/10.1111/jocn.16479

[16] Obisesan, O., Barber, E., Martin, P., Brougham, N. and Tymkew, H. (2024) Original Research: Alarm Fatigue: Exploring the Adaptive and Maladaptive Coping Strategies of Nurses. *AJN, American Journal of Nursing*, **124**, 24-30. https://doi.org/10.1097/01.naj.0001063808.07614.8d

[17] Poncette, A., Wunderlich, M.M., Spies, C., Heeren, P., Vorderwülbecke, G., Salgado, E., *et al.* (2021) Patient Monitoring Alarms in an Intensive Care Unit: Observational Study with Do-It-Yourself Instructions. *Journal of Medical Internet Research*, **23**, e26494. https://doi.org/10.2196/26494

[18] Simpson, K.R. and Lyndon, A. (2019) False Alarms and Overmonitoring. Major Factors in Alarm Fatigue Among Labor Nurses. *Journal of Nursing Care Quality*, **34**, 66-72. https://doi.org/10.1097/ncq.0000000000000335

[19] Taiwo, O., Ezugwu, A.E., Oyelade, O.N. and Almutairi, M.S. (2022) Enhanced Intel-

ligent Smart Home Control and Security System Based on Deep Learning Model. *Wireless Communications and Mobile Computing*, **2022**, Article 9307961. https://doi.org/10.1155/2022/9307961

[20] Vitense, H.S., Jacko, J.A. and Emery, V.K. (2003) Multimodal Feedback: An Assessment of Performance and Mental Workload. *Ergonomics*, **46**, 68-87. https://doi.org/10.1080/00140130303534