Methodological Paper

Isfahan Artificial Intelligence Event 2024, Challenge I: Respiratory Depression Detection

Abstract

Background: The use of sedative drugs during various medical procedures is on the rise, necessitating close monitoring of respiratory function throughout the administration process. Continuous auscultation of tracheal sounds is an effective method for monitoring respiratory status. However, it requires constant attention from the operator, which may not always be feasible. **Methods:** This concept led to the development of a tracheal sound dataset featuring recordings from 16 patients who underwent cataract surgery at Alzahra Hospital, a university hospital in Isfahan, Iran. To ensure accuracy, the dataset was carefully examined with the assistance of an anesthesiology team, providing precise ground truth annotations for respiratory depression (RD) intervals at a resolution of one second. The Isfahan National Elite Foundation hosted the Isfahan artificial intelligence (AI) 2024 events to advance AI-based detection technologies and offered financial support for five challenges, including the competition for detecting RD from tracheal sounds. Twelve teams from various provinces across Iran participated, utilizing a shared dataset for their evaluations. **Results:** The teams that achieved the first through third places were Houshmandsazan, Houshava, and Hoopad, with F1-Scores of 65.18%, 50.44%, and 21.73%, respectively. All participating teams utilized deep learning techniques to detect RD intervals, achieving notable performance, yet opportunities for further improvement remain. **Conclusion:** This paper summarizes the performance of these teams, detailing the metrics used to assess their results and the methodologies employed by the top three competitors.

Keywords: *Apnea, artificial intelligence, deep learning, Isfahan artificial intelligence event 2024, respiratory depression, tracheal sound*

Submitted: 10-Apr-2025 **Revised:** 02-Jun-2025 **Accepted:** 30-Jun-2025 **Published:** 02-Jan-2026

Azra Rasouli Kenari[1], Neda Esmaeili[2], Mahnoosh Tajmirriahi[1,3], Mehdi Khashei[4], Morteza Ebrahimpour[4], Pariya Alinezhad[4], Ehsan Sheikhi[4], Ali Loghmani[5], Mohammad Reza Torabi[5], Mehdi Abruee[5], Mohamad Kiani[6], Farzad Nekouei[6], Mohamad Yasin Fakhar[6], Mahmoud Saghaei[7], Mohammad Hassan Moradi[8], Hossein Rabbani[1,3]

[1]Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [2]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA, [3]Department of Bioelectrics and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [4]Department of Industrial and Systems Engineering, Isfahan University of Technologies, Isfahan, Iran, [5]Department of Mechanical Engineering, Isfahan University of Technology, Isfahan, Iran, [6]Department of Computer Engineering, Isfahan University, Isfahan, Iran, [7]Anesthesiology and Critical Care Research Center, Isfahan University of Medical Sciences, Isfahan, Iran, [8]Department of Biomedical Engineering, Amirkabir University of Technology, Tehran, Iran

Introduction

Increasingly, both surgical and nonsurgical procedures – such as dental work,[1] endoscopy,[2] cosmetic surgery,[3] and cataract surgery[4] – are performed under sedation analgesia using a combination of sedative and narcotic drugs.[5] These agents must be accurately titrated to meet individual patient needs, with close monitoring of their effects on respiratory function.[6]

Common monitoring techniques, such as pulse oximetry, often have a considerable delay in detecting respiratory complications.[7] In addition, side-stream capnography has limited effectiveness for detecting respiratory depression (RD) due to challenges such as sampling errors, lumen obstruction by airway secretions, and frequent detachment from the patient's airway.[8] Therefore, direct monitoring of airway patency using auscultatory techniques is crucial during sedation analgesia.[9]

Continuous monitoring of tracheal sounds with traditional or electronic stethoscopes can reliably and rapidly detect airway complications before they lead to serious issues. However, the overall efficacy of tracheal stethoscopes as airway monitors depends on the continuous listening to respiratory sounds by the anesthesia team. Continuous operator listening may

Address for correspondence:
Dr. Hossein Rabbani,
Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran.
E-mail: rabbani.h@ieee.org

How to cite this article: Kenari AR, Esmaeili N, Tajmirriahi M, Khashei M, Ebrahimpour M, Alinezhad P, *et al.* Isfahan artificial intelligence event 2024, challenge I: Respiratory depression detection. J Med Sign Sens 2026;16:2.

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.jmss_32_25

Quick Response Code:

not be practical, and interrupted auscultation can be associated with significant airway problems. Therefore, the development of real-time, automatic, and continuous techniques for tracheal sound monitoring and analysis is essential for effective airway monitoring during sedation analgesia.[9,10] Several previous studies have investigated tracheal sound data to detect apnea and intervals of RD using tracheal sound data.[9,11-15]

Recent advances in artificial intelligence (AI) have significantly advanced the field of medical signal analysis, enabling improvements in tasks such as disease staging, event detection, and delineation of pathological abnormalities. For example,[16] utilized deep convolutional neural networks (CNNs) with transfer learning and evaluated their model by employing tracheal sound data for separating obstructive and central respiratory events. They reported 87% accuracy for obstructive and central apnea classification and 83% accuracy for hypopneas.

The Isfahan National Elite Foundation organized the Isfahan AI Event 2024 (IAI2024) to promote advancements in AI-based detection technologies and provided financial support for five challenges, including tracheal sound RD detection competition with the goal of advancing tracheal sound RD detector technology by impartially assessing a range of AI-based techniques. Participants were given a well-documented dataset of labeled tracheal sound data, referred to as the "training set," to develop their methods. Subsequently, researchers submitted their algorithm outputs for evaluation using a separate test dataset, the labels of which were concealed to ensure an unbiased assessment of performance, free from any influence related to method selection or parameter customization tailored to the data.

In Section II, we will present the dataset and provide the link to access it. In Section III, we will outline the rationale behind the team rankings, the metrics used for evaluation, and a summary of the methodologies employed by the top teams. Section VI will present the statistical results from all twelve teams, detailing how the three leading teams were selected among the competitors. Discussion about the results of the finalists is provided in Section V. Finally, Section VI will provide the conclusion of our study.

Dataset

The dataset was provided by Alzahra Hospital, Isfahan University of Medical Science (Head: M. Saghaei, M. D.) and the School of Advanced Technologies in Medicine (Head: H. Rabbani), Isfahan University of Medical Science. The dataset can be accessed via provided link in.[17]

The dataset was first provided by,[11] and the data labeling structure has changed significantly since then. The study was conducted following approval from the institutional ethical committee and informed consent from the patients. All patients were scheduled for cataract surgery under sedation anesthesia, and those with a history of respiratory diseases were excluded from the study. Upon positioning on the operating table, all patients received supplemental oxygen via a mask. Monitoring included ECG, noninvasive blood pressure, and pulse oximetry. Tracheal sound recording commenced 1 min before the administration of sedative drugs using a C417 omni-directional condenser Lavalier microphone (AKG Acoustics, Vienna, Austria), secured over the suprasternal notch with double-sided adhesive tape. Tracheal sound recording continued throughout the procedure at a sampling rate of 44.1 KHz.

The dataset includes tracheal respiratory sound of 16 adults classified as the American Society of Anesthesiologists I and II. After the study's completion, the recorded sounds were analyzed by the anesthesiologist to identify periods of RD, defined as episodes of apnea, breath-holding, or airway obstructions. The labels were provided in separate. docx files, each containing the onset and offset timings of RD intervals at 1-s resolution.

Figure 1 depicts the plots of two samples from the dataset, showcasing two distinct intervals of tracheal sound data plotted against the time axis, corresponding to apnea and nonapnea intervals.

The dataset was subdivided into training, testing, and hidden sets for the competition. The training data, along with their corresponding labels, was distributed to the teams. As the competition progressed to the second stage, the teams were provided with test data devoid of labels. Subsequently, 12 submissions were received for the test labels, with some submissions showcasing better results, 3 out of the 12 submissions achieving superior scores. Moving on to the final phase of the competition, held live, the three top teams were granted access to the hidden episodes. The number of subjects and RD events in each set is detailed in Table 1.

According to the information presented in Table 1, the dataset comprises 16 recordings, named as S1 to S16. The total duration of these recordings is 21,564 s. Out of this total, only 10.97% is labeled as positive. This indicates that the tracheal sound dataset is a rare case, similar to many medical datasets, highlighting the need for competitors to adapt their methodologies to address this challenge. In addition, it is clear that the tracheal sound data for each of the training, testing, and hidden sets comes from different individuals.

Methods

Twelve competing teams submitted their results for evaluation on the test dataset. The top three winning teams were selected based on the multiple criteria, including the innovation and originality of their proposed methods, the quality and clarity of their preliminary reports, and their

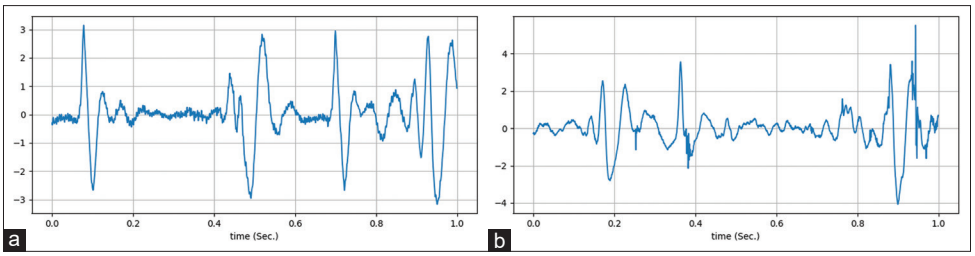

Figure 1: Representative tracheal sound images for 1 s (a) apnea and (b) nonapnea segments

Table 1: Details of tracheal sound dataset

Train		Test		Hidden	
Subject	**Duration***	**Subject**	**Duration***	**Subject**	**Duration***
S1	1929 (223)	S2	1481 (79)	S5	1043 (11)
S3	741 (132)	S6	1397 (56)	S8	2024 (267)
S4	1100 (0)	S12	1868 (0)	Total	3067 (278)
S7**	810 (53)	Total	4746 (135)		
S9	837 (373)				
S10	1007 (312)				
S11**	2036 (66)				
S13	924 (85)				
S14	1591 (219)				
S15	2087 (245)				
S16	689 (134)				
Total	13751 (1842)				

*Duration of tracheal sound data and it's RD events are reported in seconds, RD events are reported in parentheses; **An amount of 110 and 896 seconds should be put aside for the S7 and S11 respectively. RD – Respiratory depression

performance against predefined metrics, which will be detailed for readers in Section III. B.

Ranking of the Teams

The submitted results of the top three competitor teams were approached with caution. In order to rank the three top teams, the final phase of the competition was conducted live and assessed based on the five equally important criteria. These criteria were utilized to rank the top three teams in the competition.

The five criteria of equal weight were as follows:

1. Innovation of the proposed approach
2. Performance in the initial stage competition (based on the best submission and score)
3. Performance in the final submission (second phase conducted live at Abbasi Hotel, Isfahan, Iran) using the hidden dataset
4. Quality and clarity of the final report
5. Quality and clarity of the presentation.

Each criterion was rated on a scale of 1, 2, or 3. The team that excelled in each criterion received 3 points, while the second-best team received 2 points. The final rankings were determined based on the total points accumulated across all five criteria during the judges' deliberations at the conclusion of the competition.

Performance metrics

In order to assess classification models and provide valuable insights into different aspects of model performance, positive predictive value (PPV), sensitivity, F1-score, and accuracy have been calculated along each second of the tracheal sound data. Intersection over Union (IoU) have also been calculated for each truly detected RD interval and average score along all of the RD events was reported.

PPV, also known as precision, measures the proportion of true positive predictions out of all positive predictions made by the model. It is calculated as (1).

$$PPV\,(\%)=100\frac{TP}{TP+FP} \tag{1}$$

Sensitivity, also known as recall, quantifies the model's ability to correctly identify all actual positive instances in the dataset. It is calculated as (2) and emphasizes the model's capability to capture all positive instances.

$$Sen\,(\%)=100\frac{TP}{TP+FN} \tag{2}$$

The F1-Score combines precision (PPV) and recall (sensitivity) into a single metric, providing a balanced assessment of a model's performance. It is calculated as (3). The F1-score considers both false positives and false negatives (FNs), offering a comprehensive evaluation of the model's precision and recall trade-off.

$$F1\_Score(\%)=100\frac{2\times TP}{2\times TP+FP+FN} \tag{3}$$

The accuracy measure was considered because in the test dataset, there was tracheal sound data of one subject (S12) that did not include any RD event, making accurate detection of true negatives crucial. Accuracy is defined by (4), and considers all the true detected cases, whether it belongs to positive class or the negative class, across all of the classified cases. This metric is crucial in understanding the model's ability to make correct predictions across both classes in the dataset.

$$Acc\,(\%)=100\frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

The TP, TN, FP, and FN are calculated using the output of the network for elementwise classification and label annotations were described by the confusion matrix of Table 2.

Another important measure is the IoU, which was calculated at the event-based level specifically for the intervals of RD that were accurately detected, regardless of the level of overlap. IoU quantifies the overlap between the predicted intervals and the ground truth annotations, providing insight into the model's performance in terms of localization. It is computed as the ratio of the area of overlap between the predicted and ground truth regions to the area of their union.

Summary of winning teams' approach

In this section, we will discuss the methods employed by the top three competitors and the results of their approaches when applied to the hidden dataset described in Table 1.

The first team

The team of Houshmandsazan achieved first place in the competition with an F1-Score of 65.18% for the classification of RD intervals from tracheal sound data.

Their proposed method employs a CNN for effective feature extraction, followed by a long short-term memory (LSTM) network and two dense layers to classify the data into RD and non-RD categories.

The preprocessing steps included segmenting the input data into 1-s intervals, and normalizing the amplitude of the input sound between −1 and 1. To denoise the input signal, they applied discrete wavelet transform (DWT) and soft-thresholding techniques. The noise level was estimated using the mean absolute deviation method, with the threshold determined based on the signal length and noise variance. The resulting signal was then reconstructed to be free of noise.

For feature extraction, the team utilized Mel frequency cepstral coefficients (MFCCs). To address class imbalance in the dataset, they employed the Synthetic Minority Over-sampling Technique (SMOTE) alongside under sampling methods. The CNN architecture included max pooling layers and batch normalization after each of the two convolutional blocks, with filter sizes of 32 and 64, respectively. Following the CNN, two LSTM networks were integrated, featuring recurrent dropout to mitigate overfitting and capture temporal dependencies in the sequential data. The output from the LSTM layers was passed to two fully connected (FC) layers, with the final layer serving as a binary classifier using a sigmoid activation function. To counteract class imbalance, focal loss was implemented, focusing the model's attention on the minority class (RD). The Adam optimization algorithm was employed for model training. Their proposed architecture for the combined CNN-LSTM network is shown in Figure 2.

Table 2: Confusion matrix for elementwise classification

Output of network	0	1
Ground truth annotations		
0	TN	FP
1	FN	TP

TP – True positive; TN – True negative; FN – False negative; FP – False positive

The second team

The team of Houshava secured second place in the competition with an F1-score of 50.44% on the hidden dataset.

Their method consists of two main components: feature extraction and classification. For feature extraction, they employed Short-time Fourier transform (STFT) using a Hanning window, with a window length of 4410 samples and a 50% overlap. They then applied Mel filter banks to compute the logarithm of the spectrogram of the sound data on the Mel frequency scale. To enhance the spectrogram images, they utilized an opening technique, resulting in what they referred to as the processed spectrogram. Subsequently, they implemented a fusion technique to create a fused spectrogram by combining the processed spectrogram with the log Mel spectrogram, which reduced the noise level in the resulting fused spectrogram. Details of the preprocessing block are shown in Figure 3.

For RD interval detection, the team designed CNN architecture comprising four convolutional blocks. Each block utilized a 3 × 3 kernel, progressively increasing the number of input channels from 1 to 16, 32, 64, and finally 128. A ReLU activation function was applied after each CNN block.

Following the final CNN block, they incorporated a flattening layer leading to two FC layers with 768 and 256 neurons, respectively. A rectified linear unit (ReLU) activation was also applied after the first FC layer, while the last layer employed a sigmoid activation function for binary classification. To mitigate overfitting, dropout techniques were applied across all CNN and FC layers, and

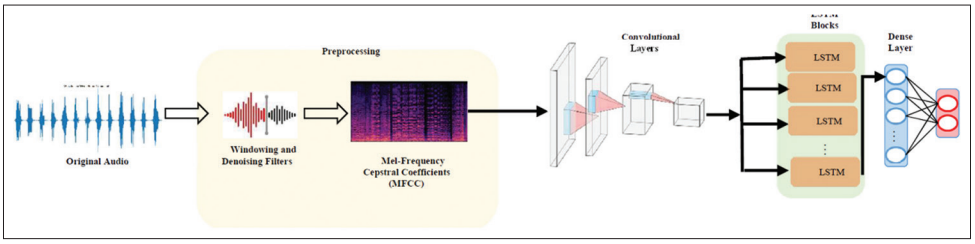

Figure 2: The proposed architecture of the Houshmandsazan team

batch normalization was implemented for each frequency bin. The details of their proposed framework are depicted in Figure 4.

The loss function used was weighted binary cross-entropy to address the challenges posed by the rare class. The Adam optimization algorithm was utilized for weight updates, with a learning rate set at $7e^{-5}$. The team also implemented the two data augmentation techniques: time shifting and time flipping.

The third team

The team of Hoopad achieved an F1-score of 21.73%, securing third place in the competition. Their methodology comprised preprocessing, feature extraction, and classification, leveraging a combination of a pretrained residual network (ResNet) and a transformer model.

Initially, the data were windowed into the segments of 0.5 s in length, with a step size of 0.2 s. The preprocessing phase included downsampling and the application of a bandpass filter, followed by the creation of Mel spectrograms to simultaneously capture time and frequency components. Normalization was applied to the Mel spectrograms to ensure consistent scaling.

For feature extraction, the team utilized a pretrained ResNet-18 model, which extracted local features from the Mel spectrograms. The extracted features were then input into the encoder of a transformer model, which excelled at handling sequences of data and extracting global features for classification. The proposed pipeline of their method is illustrated in Figure 5, showcasing the integration of preprocessing, feature extraction, and classification processes. Figure 6 details thestructure of their designed transformer network.

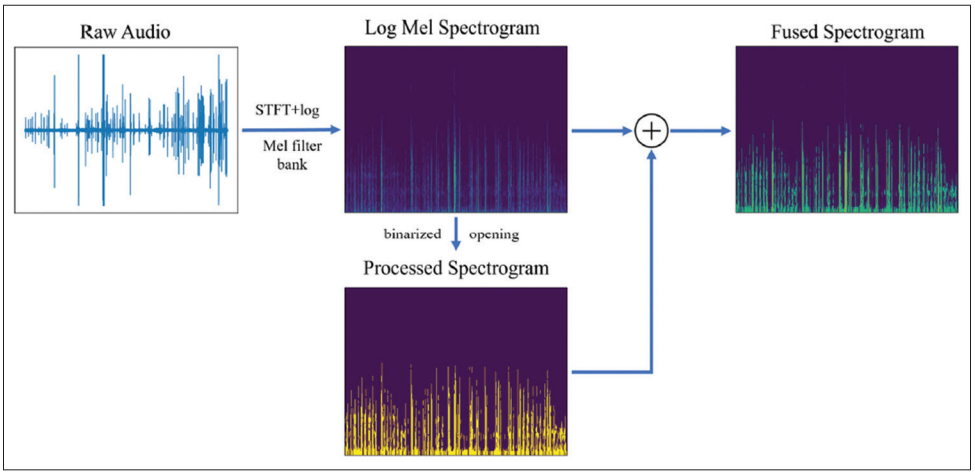

Figure 3: The preprocessing technique of the Houshava team

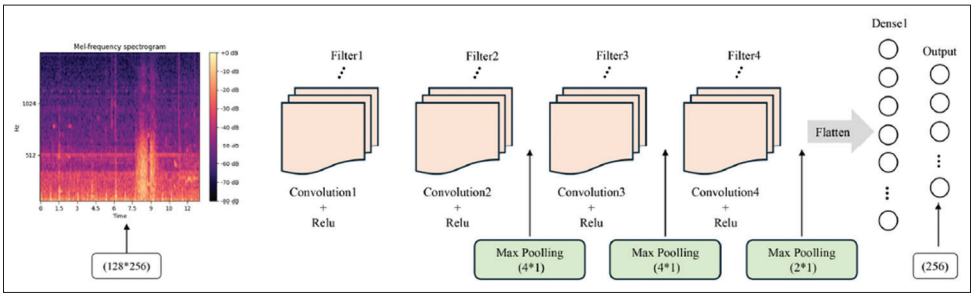

Figure 4: Framework of the Houshava team's respiratory depression detection method. The pipeline illustrates preprocessing with Short-time Fourier transform, Mel filter banks, and spectrogram fusion to reduce noise, followed by a convolutional neural network architecture with four convolutional blocks (3 × 3 kernels, increasing channels from 1 to 128), batch normalization, and dropout. The output is processed through two fully connected layers (768 and 256 neurons) with ReLU and sigmoid activations for the binary classification of respiratory depression (RD) and non-RD intervals

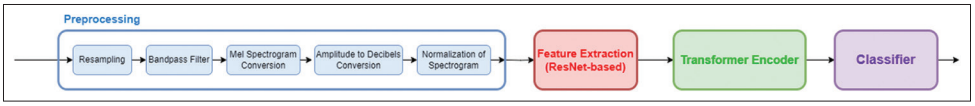

Figure 5: Pipeline of the Hoopad team's respiratory depression detection method. The framework illustrates preprocessing with windowing (0.5 s segments, 0.2 s step), downsampling to 22 kHz, bandpass filtering, and Mel spectrogram generation with normalization. Features are extracted using a pretrained ResNet-18 model, followed by a transformer encoder for global feature extraction and binary classification of respiratory depression (RD) and non-RD intervals, with data augmentation (amplitude scaling and Gaussian noise) integrated

The team also implemented data augmentation techniques. For positive-label signals, the amplitude was scaled randomly within the range of (0.8, 1.2). In addition, white Gaussian noise with a standard deviation of $5e^{-3}$ was added to the data as another augmentation method.

Results

Twelve teams submitted their results for evaluation on the test dataset, with the top three achieving the highest F1-Scores. Figure 7 illustrates the scores assigned to each of the 12 teams. Models are labeled A-L in Figure 7, corresponding to the teams listed in Table 3, which maps team names to their respective model labels. Table 3 ensures clear referencing when comparing performance metrics across teams.

Figure 7 highlights that model "J," is one of the outliers in the bar plot. This model recorded a score of zero for all of the sensitivity, PPV, IoU, and F1-Score metrics. The amount of accuracy of this model is 72.95% which is the lowest reported accuracy among all of the models. These results indicate that their model failed to learn the RD events effectively. Upon reviewing their methodology, it appears that while the team considered data augmentation to address the rare case of RD events, their choice of cross-entropy loss without implementing a weighted loss function, likely contributed to their inability to learn these rare instances.

From Figure 7, we can conclude that the results of six teams are comparable to each other. To further investigate the performance among these six models, Figure 8 compares their precision-recall curves, along with the calculated average precision (AP) for each model.

Overall, the AP values are low, which can be attributed to the challenge of comparing every second of the model outputs with ground truth annotations, making it difficult for any model to succeed. Notably, for models I and B, adjustments were necessary because the output labels were not aligned with the ground truth. For team B, their labels were reformatted into new labels based on 10-s intervals. If at least half of the values in the original interval were 1, the new label for that interval was regarded as positive. For model I, the labels were upsampled by a factor of 20, converting them into 2205 intervals, with each interval containing 10% of ones considered positive. This threshold significantly affected the final results.

In the final phase of the competition, Models H, I, and G were shortlisted for further evaluation. Table 4 presents the performance results of these top three teams on the hidden dataset. The first team significantly outperformed the others across all metrics. While the second and third teams yielded comparable results, the second team demonstrated superior sensitivity but encountered a high number of FNs, resulting in the lowest PPV and accuracy among the three. Nonetheless, the second team secured second place considering the F1-score.

Figure 9 illustrates the annotation diagrams for the three teams, which align closely with the statistics reported in Table 4. The plots clearly show that the first team's detections are more closely aligned with the ground truth

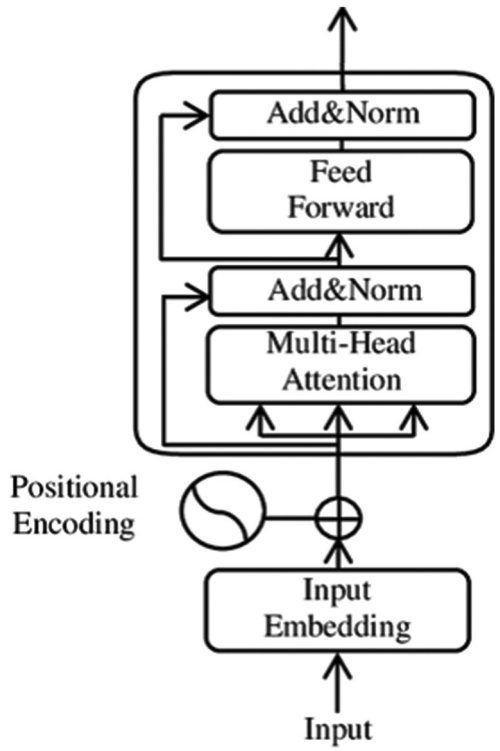

Figure 6: The transformer network utilized by the Hoopad team

Table 3: Mapping of team names to model labels

Team name	Assigned model name
14	A
AI glitch	B
AI medic	C
AIRE	D
Behoush	E
CBRC_Geeks	F
Hoopad	G
Hooshmand Sazan	H
Housh Ava	I
IU team	J
Kimia	K
Sleep Sonic	L

Table 4: Performance evaluation of the three competitors on the hidden dataset

	Sensitivity (%)	PPV (%)	F1-score (%)	Accuracy (%)
1st ranked team (model H)	62.95	67.57	65.18	93.90
2nd ranked team (model I)	92.45	34.68	50.44	83.53
3rd ranked team (model G)	12.23	97.14	21.73	92.01

PPV – Positive predictive value

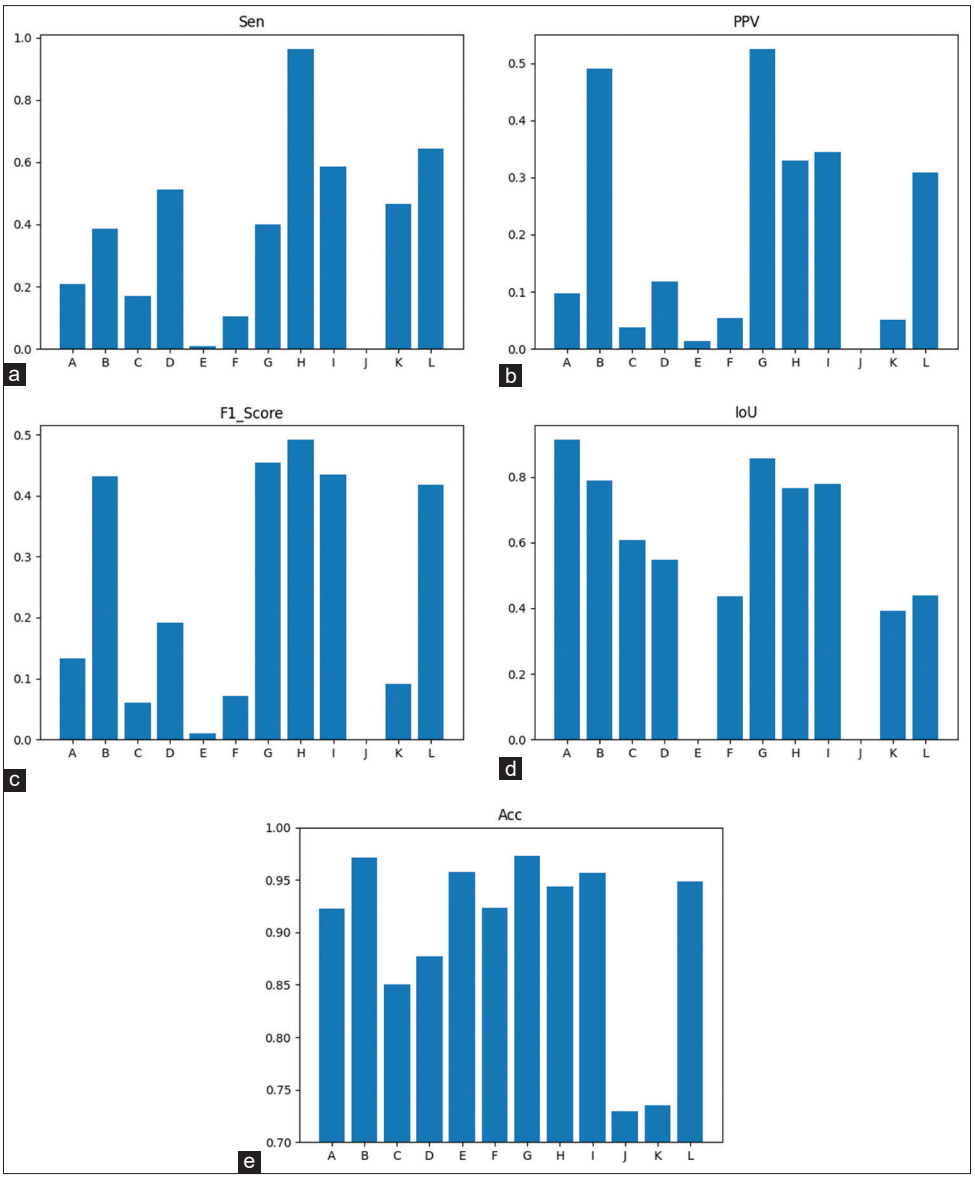

Figure 7: Performance evaluation during the initial phase of the competition. A bar plot showcasing various statistics for the 12 participating teams is presented. Panel (a) displays sensitivity, panel (b) represents positive predictive value, panel (c) depicts F1-score, panel (d) represents Intersection over Union for only those events that were detected true, panel (e) shows accuracy, in relation to the detection of respiratory depression intervals from the test tracheal sound data. Models are labeled as A-L, corresponding to the teams listed in Table 3

annotations compared to those of the second and third teams. The second team shows a substantial number of FPs, while the third team achieves good precision but also suffers from a high FN rate.

Discussion

The Houshmandsazan team, emerged as the top performer, achieving an F1-score of 65.18%, outperforming the other models across various metrics. The model's ability to correctly identify both positive and negative cases contributed to its high accuracy and F1-score. The key advantage of the first team might lie in a good tailoring of their method to the tracheal sound data especially in the preprocessing scheme. They used entropy of wavelet transform which provides a good insight of the respirational flow, and was a talent design that could effectively handle the problem of environmental and speech noise present in the dataset. They also used normalization of data at the first step of preprocessing. The first team did not utilize any augmentation techniques, data augmentation can be a valuable tool in medical data contexts, but its application should be approached with caution and careful consideration of the specific circumstances. It is crucial to ensure that any augmented data remains clinically relevant and realistic.

The Houshava team secured second place with an F1-score of 50.44%. Their model exhibited strength in sensitivity. However, faced significant challenges, due to a high

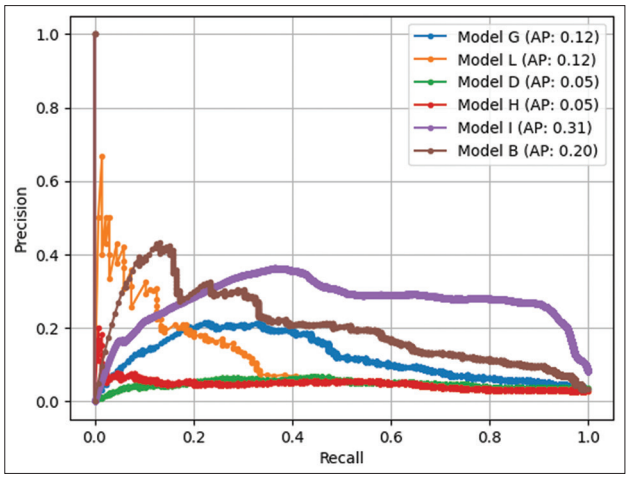

Figure 8: The precision-recall curves for 6 of the 12 teams are shown, with each team's average precision indicated in the legend. Models are labeled A-L, corresponding to the teams listed in Table 3

number of false positives, which adversely affected its PPV and overall accuracy.

The Hoopad team, placed third with an F1-score of 21.73%. The architecture of their model was innovative, particularly the decision to utilize a pretrained network to address the challenge of limited data availability. However, the high number of parameters in their network posed a challenge for maintaining good sensitivity. The data were processed through augmentation, down-sampling to a frequency of 22 kHz, and segmentation into 0.5-s intervals, likely designed to adapt the data for the extensive parameter set. Despite these efforts, the techniques employed were insufficient to provide the necessary information for effective training. The model demonstrated good precision in its detections but was also hindered by a considerable number of FNs. This impacted its reliability in identifying true events, ultimately affecting its overall evaluation metrics. While the model's precision is commendable, the substantial FN rate indicates a need for enhanced sensitivity, potentially achievable through a larger dataset.

Conclusion and Outlook

In this study, we evaluated three models through their performance on the test and hidden dataset, each exhibiting distinct strengths and weaknesses in addressing the task of detecting RD intervals from tracheal sound data. All of the three competitors utilized deep learning models, with good architecture and smart design.

In summary, The Houshmandsazan team achieved the first place in the competition with an F1-score of 65.18%. Their method utilized a CNN for feature extraction, followed by LSTM networks, and included effective preprocessing techniques such as DWT Transform for denoising and MFCC for feature extraction, addressing class imbalance with SMOTE and focal loss.

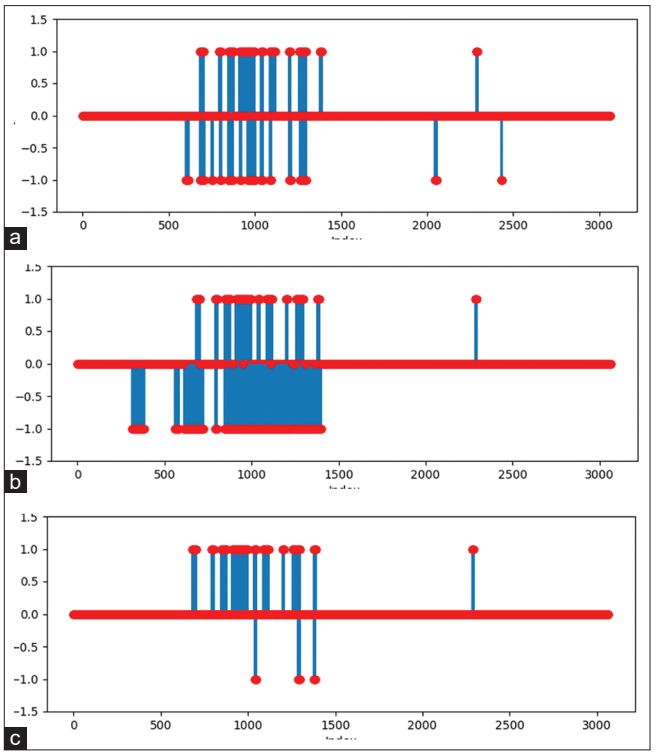

Figure 9: Annotation status plot for the two patients (S8 and S5) from the hidden dataset. For each panel, patient data are shown successively along the time axis–top stem plots depict ground truth annotations, and bottom stem plots depict each team's model output. For each second, an annotation of 1 (+1) indicates the presence of an event (ground truth), while −1 reflects a detection by the model. (a-c) Display the outputs of the first, second, and third-ranked teams, respectively. This figure illustrates how closely each team's detections align with the true annotations over time, providing a visual assessment of performance

The Houshava team secured second place in the competition with an F1-score of 50.44%. Their approach involved feature extraction using STFT with Mel filter banks, followed by CNN architecture with four convolutional blocks for classification. Despite demonstrating strong sensitivity, the model faced challenges with a high number of false positives, impacting its overall accuracy and PPV.

The Hoopad team achieved an F1-score of 21.73%, securing third place in the competition. Their methodology involved preprocessing, feature extraction using a pretrained ResNet-18 model, and classification with a transformer model, which handled sequences effectively. Despite employing data augmentation techniques, the model struggled with a high number of FNs, impacting its reliability in identifying true events and highlighting the need for improved sensitivity.

While all the three models exhibited notable performance characteristics, there are clear opportunities for improvement. Houshmandsazan's model leads in accuracy and overall detection capability, while Houshava and Hoopad models underscore the ongoing challenge of balancing sensitivity and precision in the

medical data applications. Future work should focus on optimizing preprocessing techniques, exploring data augmentation strategies, and refining model architectures to enhance the performance across all metrics.

Acknowledgments

We express our gratitude to the policy council members: Prof. Behrouz Minaei Bidgoli from Iran University of Science and Technology, Prof. Mohammad Hassan Moradi from Amirkabir University of Technology, Prof. Heshaam Faili from the University of Tehran, Prof. Emad Fatemizadeh from Sharif University of Technology, Prof. Arash Amini from Sharif University of Technology, Prof. Hamid Soltanian Zadeh from Tehran University of Technology, Prof. Shohre Kasaei from Sharif University of Technology, and Prof. Hossein Rabbani from Isfahan University of Medical Sciences, for their roles in making the final decisions regarding the winners of all five challenges.

Finally, we thank the anonymous reviewers for their valuable feedback and contributions that improved this paper.

Financial support and sponsorship

This work received support from the IEF, which sponsored IAI2024. The IEF organized the event and provided financial backing for five challenges, including Challenge I: RD detection. Several winners were awarded prizes by the IEF.

Conflicts of interest

The authors declare the following potential conflicts of interest:

MTR was an organizer of the IAI 2024 competitions on behalf of IEF, which included five challenges.

MHM, ARK, MS, and NE served as members of the scientific committee for Challenge I: RD Detection, responsible for evaluating the methodologies and results of all participating teams.

MK, ME, PA, ES, MRT, MA, AL, FN, MK, and MYF are members of the winning teams in this challenge. None of the organizers or scientific committee members (MTR, MHM, ARK, MS, and NE) contributed to the methods developed by the participating teams. The final decision regarding winners was made by the policy council members based on the following criteria:

- Technical contributions of the developed models by the teams
- Results on both initial and final test data of each team
- Submitted reports and team presentations.

The authors have disclosed these relationships to maintain transparency and uphold the integrity of the research.

References

1. Nelson TM, Xu Z. Pediatric dental sedation: Challenges and opportunities. Clin Cosmet Investig Dent 2015;7:97-106.
2. Sidhu R, Turnbull D, Haboubi H, Leeds JS, Healey C, Hebbar S, *et al*. British society of gastroenterology guidelines on sedation in gastrointestinal endoscopy. Gut 2024;73:219-45.
3. Butz DR, Gill KK, Randle J, Kampf N, Few JW. Facial aesthetic surgery: The safe use of oral sedation in an office-based facility. Aesthet Surg J 2016;36:127-31.
4. Kumar CM, Seet E, Eke T, Irwin MG, Joshi GP. Peri-operative considerations for sedation-analgesia during cataract surgery: A narrative review. Anaesthesia 2019;74:1601-10.
5. Schweickert WD, Kress JP. Strategies to optimize analgesia and sedation. Crit Care 2008;12 Suppl 3:S6.
6. Carrasco G. Instruments for monitoring intensive care unit sedation. Crit Care 2000;4:217-25.
7. Maurer WG, Walsh M, Viazis N. Basic requirements for monitoring sedated patients: Blood pressure, pulse oximetry, and EKG. Digestion 2010;82:87-9.
8. Kodali BS. "Capnography: The science, logistics, applications, and limitations for procedural sedation. In: Pediatric Sedation Outside of the Operating Room. A Multispecialty International Collaboration. Cham: Springer International Publishing; 2021. p. 155-69.
9. Yu L, Ting CK, Hill BE, Orr JA, Brewer LM, Johnson KB, *et al*. Using the entropy of tracheal sounds to detect apnea during sedation in healthy nonobese volunteers. Anesthesiology 2013;118:1341-9.
10. Liu J, Ai C, Zhang B, Wang Y, Brewer LM, Ting CK, *et al*. Tracheal sounds accurately detect apnea in patients recovering from anesthesia. J Clin Monit Comput 2019;33:437-44.
11. Esmaeili N, Rabbani H, Makaremi S, Golabbakhsh M, Saghaei M, Parviz M, *et al*. Tracheal sound analysis for automatic detection of respiratory depression in adult patients during cataract surgery under sedation. J Med Signals Sens 2018;8:140-6.
12. Yadollahi A, Giannouli E, Moussavi Z. Sleep apnea monitoring and diagnosis based on pulse oximetry and tracheal sound signals. Med Biol Eng Comput 2010;48:1087-97.
13. Cummiskey J, Williams TC, Krumpe PE, Guilleminault C. The detection and quantification of sleep apnea by tracheal sound recordings. Am Rev Respir Dis 1982;126:221-4.
14. Nakano H, Hayashi M, Ohshima E, Nishikata N, Shinohara T. Validation of a new system of tracheal sound analysis for the diagnosis of sleep apnea-hypopnea syndrome. Sleep 2004;27:951-7.
15. Nakano H, Furukawa T, Tanigawa T. Tracheal sound analysis using a deep neural network to detect sleep apnea. J Clin Sleep Med 2019;15:1125-33.
16. Saha S, Ghahjaverestan NM, Yadollahi A. Separating obstructive and central respiratory events during sleep using breathing sounds: Utilizing transfer learning on deep convolutional networks. Sleep Med 2025;131:106485.
17. Tracheal Sound Dataset. Available from: https://misp.mui.ac.ir/en/dataset-description-tracheal-sound-dataset-detection-respiratory-depressions-rd-intervals. [Last accessed 2025 July 26].