UDC 004.855.5
DOI: https://doi.org/10.36994/2788-5518-2024-01-07-16

ADAPTING AUDIO SPECTROGRAM TRANSFORMERS FOR RESPIRATORY RATE ESTIMATION

***Dmytro Tkachenko**, Ph.D. student, Open International University of Human Development «Ukraine», Kyiv, Ukraine. https://orcid.org/0000-0003-2804-7305, dmitriytkachenko1994@gmail.com*

***Serhii Odribets**, Ph.D., Ass. Prof., Open International University of Human Development «Ukraine», Kyiv, Ukraine. onspo@ukr.net*

***Abstract.** This paper investigates the application of Audio Spectrogram Transformers (AST) for the task of respiratory rate estimation from audio recordings, a critical metric in the screening for sleep apnea. Traditional methods like polysomnography, used for sleep apnea diagnosis, are resource-intensive and not widely accessible, making automated respiratory rate estimation a valuable alternative. Recent advancements in deep learning, particularly transformer-based models, offer promising solutions. This study adapts a pre-trained AST model, originally designed for audio classification tasks, for respiratory rate estimation by incorporating attention mechanisms and a regression layer.*

We fine-tune the AST model on the ICBHI 2017 challenge dataset, utilizing transfer learning to leverage the robust feature representations learned from large-scale audio datasets like AudioSet. Our methodology includes resampling audio files, applying band-pass filters, and converting audio signals into mel-spectrograms. To address the limited size of the ICBHI dataset, we employ data augmentation techniques such as SpecAugment, which introduces variations in the spectrograms to improve the model's generalizability.

We propose two pooling approaches – attention pooling and convolutional pooling – to effectively capture temporal dependencies in respiratory sounds. Experimental results demonstrate that the convolutional pooling variant of our model achieves a Mean Absolute Error (MAE) of 0.8320 and a Mean Relative Error (MRE) of 12.56% on the ICBHI dataset, outperforming the previous model by 16.8% in MAE. These findings highlight the efficacy of transformer-based models in handling complex audio patterns and underscore the potential of AST models for broader applications in respiratory sound analysis.

Our research contributes to the machine learning field by presenting an effective method for respiratory rate estimation using transformer-based models. Future work will explore further refinements in model architecture and the integration of additional physiological signals to enhance predictive accuracy.

***Key words:** audio spectrogram transformer, transfer learning, respiratory rate estimation, sleep apnea.*

АДАПТАЦІЯ АУДІОСПЕКТРОГРАМНИХ ТРАНСФОРМЕРІВ ДЛЯ ВИЗНАЧЕННЯ ЧАСТОТИ ДИХАННЯ

***Ткаченко Д. А.**, аспірант, Відкритий міжнародний університет розвитку людини «Україна», Київ, Україна. https://orcid.org/0000-0003-2804-7305, dmitriytkachenko1994@gmail.com*

***Одрібець С. П.**, к.ф.-м.н., доц., Відкритий міжнародний університет розвитку людини «Україна», Київ, Україна. onspo@ukr.net*

***Анотація.** Ця стаття досліджує застосування аудіоспектрограмних трансформерів (AST) для задачі визначення частоти дихання з аудіозаписів, що є критичним показником під час скринінгу на апное уві сні. Традиційні методи, такі як полісомнографія, що використовуються для діагностики апное уві сні, є ресурсоємними та не широко доступними, що робить автоматизоване визначення частоти дихання цінною альтернативою. Останні досягнення в галузі глибокого навчання, зокрема моделі на основі трансформерів, пропонують перспективні рішення. Це дослідження адаптує попередньо навчену модель AST, первісно розроблену для задач класифікації аудіо, для визначення частоти дихання шляхом додавання механізмів уваги та регресійного шару.*

Ми донавчаємо модель AST на наборі даних ICBHI 2017 challenge, застосовуючи передавальне навчання для використання стійких представлень ознак, отриманих із великих наборів аудіоданих, таких як AudioSet. Наша методологія включає передискретизацію аудіофайлів, застосування смугових фільтрів та перетворення аудіосигналів у мел-спектрограми. Щоб вирішити проблему обмеженого розміру набору даних ICBHI, ми використовуємо методи нарощення даних, такі як SpecAugment, які вносять варіації у спектрограми для покращення узагальнюваності моделі.

Ми пропонуємо два підходи до агрегації: агрегування з механізмом уваги та згорткове агрегування – для ефективного врахування часових залежностей у дихальних звуках. Експериментальні

результати демонструють, що наша модель зі згортковим агрегуванням досягає середньої абсолютної похибки (MAE) 0,8320 та середньої відносної похибки (MRE) 12,56% на наборі даних ICBHI, перевершуючи попередню модель на 16,8% за MAE. Ці результати підкреслюють ефективність моделей на основі трансформерів в обробці складних аудіопатернів і висвітлюють потенціал моделей AST для ширшого застосування в аналізі дихальних звуків.

Наше дослідження робить внесок у галузь машинного навчання, пропонуючи ефективний метод визначення частоти дихання з використанням моделей на основі трансформерів. Подальші дослідження будуть спрямовані на вдосконалення архітектури моделі та інтеграцію додаткових фізіологічних сигналів для підвищення точності.

Ключові слова: *аудіоспектрограмний трансформер, передавальне навчання, визначення частоти дихання, апное уві сні.*

Introduction

Sleep apnea is a prevalent sleep disorder characterized by repeated interruptions in breathing during sleep. These interruptions, or apneas, can lead to various health complications, including cardiovascular disease, cognitive impairment, and daytime fatigue. Early detection and continuous monitoring of sleep apnea are crucial for effective management and treatment. One critical metric in screening for sleep apnea is the respiratory rate, as irregular breathing patterns indicate the disorder.

Traditionally, respiratory rate estimation and sleep apnea diagnosis have relied on polysomnography (PSG), a comprehensive recording of the bio-physiological signals that occur during sleep. However, PSG is expensive, labor-intensive, and requires specialized equipment, making it inaccessible for widespread screening. In contrast, using audio recordings for respiratory rate estimation offers a cost-effective and accessible alternative. Audio-based methods can leverage readily available recording devices and simplify the screening process.

Recent studies have explored various machine learning models for respiratory rate estimation using the ICBHI dataset [1]. Notably, a BiLSTM (Bidirectional Long Short-Term Memory) model [2] achieved a mean absolute error (MAE) of 1.0, demonstrating the potential of deep learning techniques in this domain. However, given the advancements in transformer-based models, we hypothesize that Audio Spectrogram Transformers (AST) [3] could outperform existing approaches due to their superior ability to capture and model temporal dependencies in audio data.

In a related study, the authors of [4] used an AST-based model to achieve state-of-the-art results on the ICBHI dataset for respiratory sound classification. Their success underscores the efficacy of AST models in capturing complex audio patterns and highlights their potential for broader applications, including respiratory rate estimation.

In this study, we investigate an adaptation of the AST model for respiratory rate estimation using audio recordings, specifically incorporating attention mechanisms and a regression layer. This paper details our methodology, the architecture of the proposed model, and the experimental setup, followed by an evaluation of the results.

Methodology and results

Audio Spectrogram Transformer

The Audio Spectrogram Transformer (AST) is a deep learning model designed for audio classification tasks. It converts raw audio waveforms into mel-filter bank features, which are refined representations capturing essential frequency components over time. These mel-filter bank features are treated as images and fed into a vision transformer (ViT) [5]. The vision transformer leverages self-attention mechanisms to model the complex temporal and frequency dependencies in the audio data, allowing it to learn hierarchical representations of the audio signals. The transformer consists of multiple layers of attention heads and feed-forward networks, enabling the model to capture intricate patterns within the spectrogram images.

This study employed a pre-trained Audio Spectrogram Transformer (AST) model [6]. The AST model achieved a mean average precision (mAP) score of 0.4593 on the AudioSet evaluation. Initially, the Vision Transformer (ViT) component of the model was pre-trained on the ImageNet dataset. Subsequently, the entire AST was fine-tuned on the AudioSet dataset, enabling it to learn robust audio and visual features from these extensive datasets.

Transfer learning was employed to adapt the pre-trained AST model for respiratory rate estimation. This involves fine-tuning the model on the ICBHI 2017 challenge dataset, enabling it to leverage the rich feature representations learned from the large-scale AudioSet and ImageNet datasets while adapting to the specific characteristics of respiratory sounds. Given the relatively small size of the ICBHI dataset, using a pre-trained model was crucial for achieving the observed performance improvements, as it allowed the model to build on the comprehensive knowledge acquired from diverse audio and visual data.

Dataset and feature extraction

The dataset utilized is the ICBHI 2017 challenge dataset, containing audio recordings annotated with start and end timestamps of respiratory cycles. The ICBHI dataset contains 6898 respiratory cycles, spanning around 5.5 hours (Table 1). It is officially divided into a training set (60%) and a test set (40%), ensuring no patient appears in both sets.

Table 1
Dataset characteristics

	Train set	Test set
Number of recordings	539	381
Number of patients	79	49
Average length, seconds	21.03	19.14
Average number of respiratory cycles	7.68	7.23

The raw audio files are resampled to 16 kHz for consistency. Band-pass filters, focusing on the 50-2500 Hz range – generally accepted [7] as the range where breathing sounds are – are applied to remove noise and enhance the quality of the recordings. Each audio file is segmented into fixed lengths of 10.24 seconds, compatible with the input requirements of the AST model.

The AST feature extractor converts raw audio into mel-filter bank features (or mel spectrograms). Mel-filter bank features transform the spectrogram (the raw frequency content of the signal over time) into a representation that mimics human auditory perception. The extractor generates 128 mel-frequency bins. It pads or truncates the features to a fixed length of 1024 frames and normalizes them using a mean of -4.27 and a standard deviation of 4.57, parameters chosen to match those from AudioSet.

Model adaptation for respiratory rate estimation

To adapt the AST model for respiratory rate estimation, the following modifications were implemented (figure 1):

1. **Pretrained Model and Fine-Tuning:** The AST model was initialized using the pre-trained checkpoint [6]. During the fine-tuning phase, we chose to freeze the first ten layers and only fine-tune the last two layers. This approach leverages the pre-trained knowledge encoded in the initial layers while adapting the final layers to the specific task of respiratory rate estimation. Freezing most layers helps retain the general audio feature extraction capabilities while allowing task-specific features to be learned in the fine-tuned layers.

2. **Pooling Approaches:** Two different pooling methods were explored to aggregate features across the time dimension: attention pooling and convolutional pooling.

• **Attention Pooling:** Attention pooling computes a weighted sum of input features, where the weights are learned during training. This approach allows the model to focus on the most relevant parts of the audio signal for respiratory rate estimation. The following formulas describe the attention mechanism:

- Compute attention scores using a simple dot product: $s = xW_a$, where x is the input feature matrix of the shape $[\text{batch}, \text{time}, \text{features}]$ and W_a are the learnable attention weights of shape $[\text{features}, 1]$.
- Apply softmax to get attention weights: $a = softmax(s)$
- Compute the weighted sum of input features: $y = \sum_i x_i a_i$, where i indexes the time dimension.

• **Convolutional Pooling:** Convolutional pooling involves passing the input features through a series of convolutional layers followed by global pooling. Specifically, the input features are first transposed to shape $[\text{batch}, \text{features}, \text{time}]$. They are then processed by two 1D convolutional layers, each with a kernel size of 5, stride of 1, padding of 2, and dilation of 1. Each convolutional layer is followed by batch normalization and ReLU activation. Dropout with a rate of 0.1 is applied after each batch normalization layer to prevent overfitting. Finally, an adaptive average pooling layer reduces the time dimension to one, producing a fixed-length feature vector for each input sample.

3. **Fully-Connected Layer:** The output of the pooling layer, whether from attention pooling or convolutional pooling, is fed into a fully-connected layer. This layer produces the final prediction of the number of breathing cycles, which can be converted to respiratory rate.

The modified architecture thus combines the AST model's powerful feature extraction capabilities with task-specific adaptations for effective respiratory rate estimation.

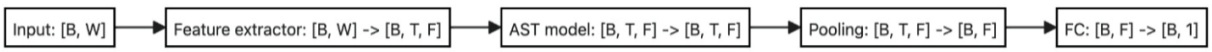

Fig. 1. Overall architecture of AST model adapted for respiratory rate estimation. In square brackets are tensor shapes: B – batch, W – waveform, T – time, F – features. FC denotes the fully connected layer

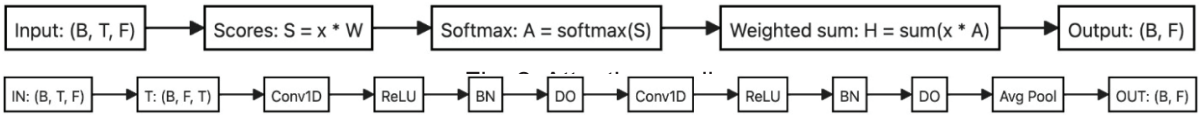

Fig. 3. Convolutional pooling. Layer abbreviations: BN – batch normalization, DO – dropout

Data Augmentation

To improve the robustness and generalizability of the AST model, data augmentation was applied using the SpecAugment technique (introduced in [8], applied to the classification problem on the ICBHI dataset in [4]). SpecAugment is particularly effective for audio data, as it introduces spectrogram variations without altering the underlying audio content. The augmentation process involved both frequency and time masking. Frequency masking was applied with a parameter $F = 48$ (maximum number of masked frequency bins; the actual length is sampled for each batch uniformly from $[0;48)$) and two frequency masks per spectrogram. Time masking was applied with a parameter T=160 (maximum number of masked audio frames) and two time masks per spectrogram. This augmentation strategy helps the model learn more invariant features by simulating the effect of missing frequencies and time segments in the input spectrograms.

Training

The training was conducted with a batch size of 64, and the model was trained for up to 100 epochs. Throughout the training process, we observed that the best performance was typically achieved between epochs 30 and 50. This observation guided the early stopping criteria, ensuring the model training was efficient and effective without overfitting.

The Mean Absolute Error (MAE) was used as the loss function, which is suitable for regression tasks like respiratory rate estimation. The MAE formula is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

where y_i and $\hat{y}_i$ are the true and predicted values, respectively.

Our Audio Spectrogram Transformer (AST) model was optimized using the AdamW optimizer [9]. This optimizer is well-suited for transformer architectures due to its ability to handle weight decay decoupled from the gradient update step. The learning rate was set to 10^{-3}, with betas $(0.9, 0.999)$, an epsilon value of 10^{-6}, and a weight decay of 0.01. To enhance the training process, a cosine learning rate scheduler with warmup was employed. This scheduler included 120 warmup steps and was applied over 2000 training steps, adjusting the learning rate dynamically at each step to improve convergence and model performance.

Results

The results are summarized in Table 2 below and include two variants of our model: one with attention pooling and another with convolutional pooling. The parity plot (figure 4) compares predicted and expected respiratory cycle counts for our best model (convolutional pooling). The result from the prior publication [2] is also included for comparison. Note that its authors used a smaller dataset with 335 samples in the training set (out of 539), and the test set consisted of 84 samples (out of 381). Our results are from a final training run performed on the entire training dataset and evaluated on the whole test set. In addition to the MAE described above, the table includes two other metrics:

- MRE (Mean Relative Error, %):

$$MRE = \frac{100}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{y_i}$$

- R^2 (coefficient of determination)

Table 2
Results

Method	Dataset	MAE	MRE, %	R^2
Prior work – BiLSTM [2]	ICBHI (subset)	1.0	Not specified	Not specified
Our model (AST + attention pooling)	IBCHI (full)	0.9948	15.33	0.7594
Our model (AST + convolutional pooling)		**0.8320**	**12.56**	**0.8192**

Conclusions

In this study, we successfully adapted the Audio Spectrogram Transformer (AST) model for respiratory rate estimation using audio recordings. Our approach leverages the AST's robust feature extraction capabilities and introduces task-specific adaptations, such as attention pooling and convolutional pooling, to effectively capture the temporal dependencies and complex patterns in respiratory sounds.

The results indicate that our model, particularly the variant with convolutional pooling, outperforms previous approaches, achieving a Mean Absolute Error (MAE) of 0.8320 and a Mean Relative Error (MRE) of 12.56% on the ICBHI dataset. This result represents a 16.8% improvement in MAE over the previous best result. This performance highlights the efficacy of transformer-based models in handling audio data for respiratory rate estimation. Furthermore, the use of data augmentation techniques like

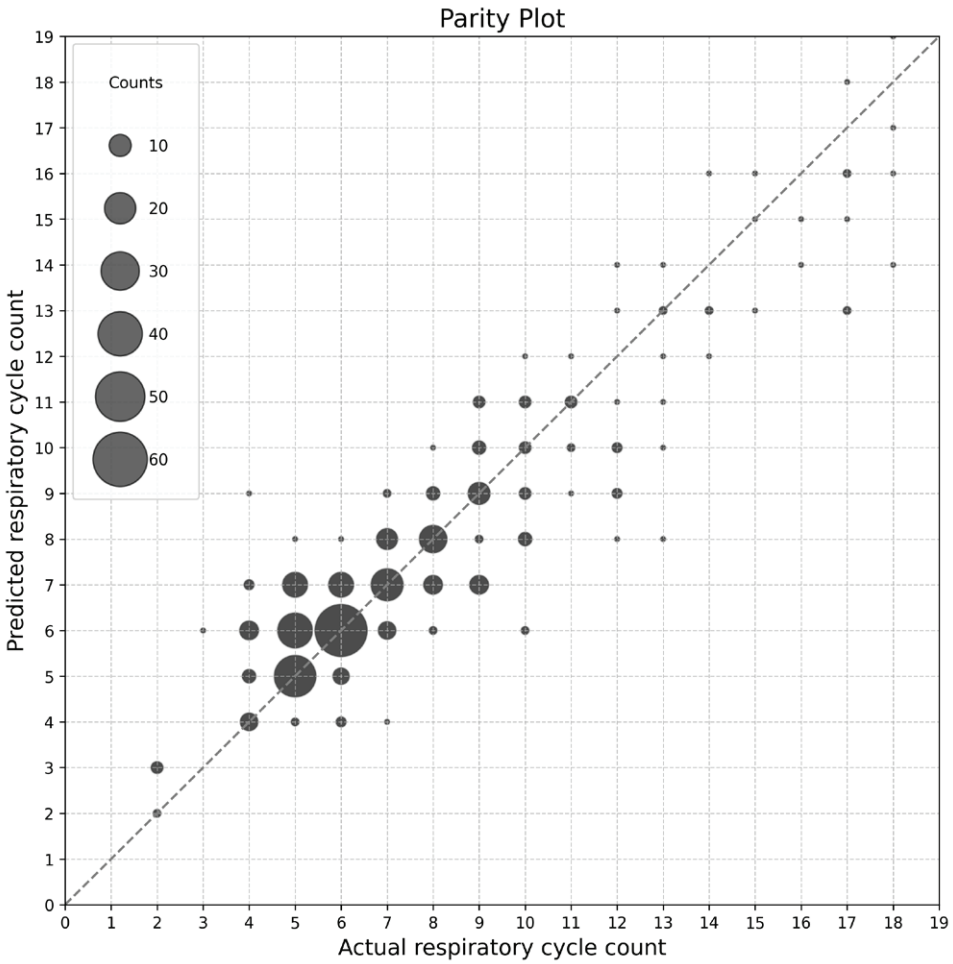

Fig. 4. Parity plot comparing actual respiratory cycle counts to those predicted by our convolutional pooling model variant. The dashed 45-degree line corresponding to perfect predictions is shown for reference

SpecAugment significantly enhances the model's robustness, ensuring reliable predictions across diverse audio samples.

Our findings underscore the potential of utilizing pre-trained AST models for health-related applications, where capturing subtle variations in audio signals is crucial. The approach presented in this paper offers a cost-effective, non-invasive alternative to traditional methods like polysomnography, making continuous monitoring and early detection of sleep apnea more accessible.

Future work could explore further refinements to the model architecture and investigate the integration of additional physiological signals to enhance the predictive accuracy.

References

1. Rocha, B. M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y. P., Jakovljevic, N., Turukalo, T. L., Vogiatzis, I. M., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Maglaveras, N., Pedro Paiva, R., Chouvarda, I., & de Carvalho, P. (2019). An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement, 40(3)*. https://doi.org/10.1088/1361-6579/ab03ea

2. Harvill, J., Wani, Y., Alam, M., Ahuja, N., Hasegawa-Johnsor, M., Chestek, D., & Beiser, D. G. (2022). Estimation of Respiratory Rate from Breathing Audio. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2022, 4599–4603. https://doi.org/10.1109/EMBC48229.2022.9871897

3. Gong, Y., Chung, Y., & Glass, J.R. (2021). AST: Audio Spectrogram Transformer. *Interspeech*. https://doi.org/10.21437/interspeech.2021-698

4. Bae, S., Kim, J., Cho, W., Baek, H., Son, S., Lee, B.H., Ha, C.W., Tae, K., Kim, S., & Yun, S. (2023). Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification. *Interspeech*. https://doi.org/10.21437/interspeech.2023-1426

5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2010.11929

6. Gong, Y., Chung, Y.-A., & Glass, J. (2021). *Checkpoint of Audio Spectrogram Transformer (AST) model fine-tuned on AudioSet (0.4593 mAP)* [Model checkpoint]. Hugging Face. https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593

7. Reichert, S., Gass, R., Brandt, C., & Andrès, E. (2008). Analysis of respiratory sounds: state of the art. *Clinical medicine. Circulatory, respiratory and pulmonary medicine, 2*, 45–58. https://doi.org/10.4137/ccrpm.s530

8. Park, D.S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E.D., & Le, Q.V. (2019). SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Interspeech*. https://doi.org/10.21437/interspeech.2019-2680

9. Loshchilov, I., Hutter, F. (2017). Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1711.05101

Література

1. An open access database for the evaluation of respiratory sound classification algorithms / B. M. Rocha et al. *Physiological measurement*. 2019. Т. 40, № 3. URL: https://doi.org/10.1088/1361-6579/ab03ea (дата звернення: 27.05.2024).

2. Estimation of respiratory rate from breathing audio / J. Harvill et al. *44th annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, м. Glasgow, Scotland, United Kingdom, 11–15 лип. 2022 р. 2022. С. 4599–4603. URL: https://doi.org/10.1109/embc48229.2022.9871897 (дата звернення: 27.05.2024).

3. Gong Y., Chung Y.-A., Glass J. AST: audio spectrogram transformer. *Interspeech*. 2021. URL: https://doi.org/10.21437/interspeech.2021-698 (дата звернення: 27.05.2024).

4. Patch-Mix contrastive learning with audio spectrogram transformer on respiratory sound classification / S. Bae et al. *Interspeech*. 2023. URL: https://doi.org/10.21437/interspeech.2023-1426 (дата звернення: 27.05.2024).

5. An image is worth 16x16 words: transformers for image recognition at scale / A. Dosovitskiy et al. *International conference on learning representations (ICLR)*. 2020. URL: https://doi.org/10.48550/arXiv.2010.11929 (дата звернення: 27.05.2024).

6. Gong Y., Chung Y.-A., Glass J. Checkpoint of audio spectrogram transformer (AST) model fine-tuned on audioset (0.4593 mAP). URL: https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593 (дата звернення: 27.05.2024).

7. Analysis of respiratory sounds: state of the art / S. Reichert et al. *Clinical medicine. Circulatory, respiratory and pulmonary medicine*. 2008. Т. 2. P. 45–58. URL: https://doi.org/10.4137/ccrpm.s530 (дата звернення: 27.05.2024).

8. SpecAugment: a simple data augmentation method for automatic speech recognition / D. S. Park et al. *Interspeech*. 2019. URL: https://doi.org/10.21437/interspeech.2019-2680 (дата звернення: 27.05.2024).

9. Loshchilov I., Hutter F. Decoupled weight decay regularization. *International conference on learning representations (ICLR)*. 2019. URL: https://doi.org/10.48550/arXiv.1711.05101 (дата звернення: 27.05.2024).