ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Variational Deep Embedding for Unsupervised Clustering of Industrial Noise in Steelmaking Plants

Muhammad Waseem Akram, Marco Vannucci, Giorgio Buttazzo, Valentina Colla, Stefano Dettori, Donarella Salvatore *

Scuola Superiore Sant'Anna - TeCIP Institute
Piazza Martiri della Libertá, 33, Pisa - Italy

Abstract.
Industrial plants are major sources of environmental noise, producing complex and high-intensity acoustic emissions that vary across different operational conditions. Automatically characterizing the sources that generate harmful acoustic emissions is crucial to take the necessary actions to reduce them. However, manually labeling these sounds is impractical due to their volume and variability. In this study, we employ an unsupervised deep learning framework for clustering industrial sound emissions in steelmaking plants, focusing on areas such as the hot rolling mill, Electric Arc Furnace, and scrapyard. The approach integrates Variational Autoencoders with Gaussian Mixture Models to learn compact latent representations from Mel-spectrogram features of raw, unlabelled audio data. We compare this approach to traditional clustering techniques such as K-means and GMM, as well as Deep Embedding Clustering. The results demonstrate that the approach significantly outperforms traditional methods, offering reliable and interpretable clustering of industrial acoustic events. This research contributes to the development of automated, efficient, and sustainable noise-monitoring systems for industrial operations, addressing key challenges in environmental noise monitoring.

1 Introduction

Industrial plants are major sources of environmental noise, producing complex acoustic emissions that vary across operational conditions. Activities such as hot rolling, Electric Arc Furnace (EAF) melting, and scrap handling generate overlapping, non-stationary sounds that affect safety and community well-being [1]. Understanding these sounds is key for noise mitigation and regulatory compliance.

The RFCS SILENCE project develops a framework for monitoring noise emissions in steelmaking plants. Microphones are deployed in production zones

*The work described in the present paper was developed within the project *Real-time acoustic sensorS and artificial Intelligence appLications for the rEduction of local eNvironmental impaCt due to noise Emissions (SILENCE)* (Grant agreement No. 101112516) - that has received funding from the RFCS programme, which is gratefully acknowledged. The sole responsibility of the issues treated in the present paper lies with the authors; the Commission is not responsible for any use that may be made of the information contained therein.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

such as the hot rolling mill, coiler, EAF, and scrapyard to capture distinct acoustic signatures and identify activities contributing to elevated noise levels. The volume and variability of data make manual labeling impractical, necessitating unsupervised learning for grouping acoustic events and identifying the processes behind high-noise reports from the surrounding community. Identifying these processes allows us to optimize or modify the corresponding plant activities.

Deep generative models, such as Variational Autoencoders (VAEs) [2] and Generative Adversarial Networks (GANs) [3], have emerged as effective tools for unsupervised learning in complex, unlabeled datasets. While VAEs offer a stable and continuous latent space ideal for clustering, GANs are prone to instability and mode collapse [4], limiting their use in continuous industrial monitoring.

Additionally, self-supervised learning (SSL) has made strides in audio processing [5, 6], but its computational demands and large training corpora requirements make it impractical for industrial applications. Pretrained models like AudioMAE, CLAP, and wav2vec 2.0 offer general audio representations but are insufficient for capturing the specific acoustic characteristics of steelmaking environments. Fine-tuning these models for industrial tasks often yields limited improvements.

To overcome these limitations, we propose a variational autoencoder-based clustering approach that balances data efficiency, interpretability, and deployability [7]. VAEs are particularly well-suited for learning compact, structured latent representations from limited, noisy industrial data, enabling continuous monitoring in real-world conditions [8, 9]. While previous research in sound clustering has focused on applications such as hearing devices and fault detection [8, 9, 7], our study is the first to apply the Variational Deep Embedding (VaDE) [10] model to cluster industrial noise in steel plants.

The goal of this research is to assess the effectiveness of VaDE in identifying and clustering industrial noise sources without labeled data. Our results, tested on both our collected data and the publicly available UrbanSound8K dataset, demonstrate that this approach outperforms traditional methods such as K-means and GMM, showcasing the potential of deep generative models to uncover meaningful patterns in raw, unlabelled industrial sound data.

2 Materials and Method

2.1 Data collection

The data collection process captures acoustic events in the steel mill with minimal human intervention, aiming to identify high-noise processes for noise reduction. Audio recordings are made using strategically placed microphones in production zones. A semi-automatic trigger, activated by decibel threshold detection, records when noise exceeds a predefined level, optimizing storage. Recorded events include machinery operations (coilers, rolling mills, EAF), scrap loading/unloading, and truck transit. Threshold values, set based on preliminary acoustic surveys, depend on the distance from the microphone.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Recorded data, including features like peak amplitude, average sound pressure level, and duration, are stored in a relational database for efficient retrieval. The system, managed by a commercial audio monitoring platform, offers configurable parameters for threshold levels, recording duration, and sampling rates. Despite operator review, the data remains unlabeled due to the impracticality of manual annotation. The dataset consists of raw audio recordings with intensity features but without semantic labels for sound sources.

The audio data (8941 files) is resampled to 22,050 Hz, converted to mono, and padded or truncated to 7.5 seconds. Mel-spectrograms are computed using 128 Mel frequency bins ($n_{\text{mels}} = 128$), a window size of 2048 samples ($n_{\text{fft}} = 2048$), and a hop length of 512 samples (hop length = 512). The amplitude values are transformed to a logarithmic scale (in decibels) to align with human auditory perception. To standardize the data, global normalization is applied by calculating the mean and standard deviation of the Mel-spectrograms across the entire dataset, ensuring consistent statistical properties and enabling the model to focus on relevant features without being influenced by signal strength variations.

2.2 Methodology

We employ VaDE, which integrates Variational Autoencoders with a Gaussian Mixture Model prior for clustering. The weights of the Variational Autoencoder are initialized using a pretrained Convolutional Stacked Autoencoder (Conv-SDAE). This prior-guided setup enables unsupervised clustering of Mel-spectrogram (the Mel scale is logarithmic, mimicking the way humans perceive pitch) features extracted from audio signals, effectively handling the challenges of high-dimensional industrial noise data by learning compact and meaningful latent representations for precise clustering.

The Conv-SDAE model, central to our feature extraction process, consists of an encoder-decoder structure. The encoder, made up of three convolutional layers and a fully connected layer, maps Mel-spectrograms to a low-dimensional latent space. The decoder reconstructs the original spectrogram from this latent representation. Pre-training is performed using Mean Squared Error (MSE) loss to minimize reconstruction error. The Adam optimizer with a learning rate scheduler ensures effective feature learning.

After pretraining, the Audio VaDE model performs clustering in the latent space. The VaDE model extends the Variational Autoencoder (VAE) framework by incorporating a Gaussian Mixture Model (GMM) prior to model the latent space distribution. The encoder generates the mean (μ) and log-variance ($\log \sigma^2$) of the latent distribution, from which latent variables (z) are sampled using the reparameterization trick. Clusters are assigned based on these latent variables, with cluster centers (μ_c) learned during training. The total loss (reported in equation 1) is a weighted sum of reconstruction loss, KL divergence, and cluster regularization:

$$\text{Total Loss} = \text{Recon Loss} + \beta \times \text{KL Divergence} + \gamma \times \text{KL}(P||Q) \qquad (1)$$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

where β=0.1 and γ=0.5 are hyperparameters controlling the importance of each term and Q is the model's predicted soft cluster distribution, while P is a sharpened target distribution derived from Q to enforce clearer and more separated clusters. We mannually tested differenct values of β and γ and the best result are reported below.

VaDE model weights are initialized using the pre-trained Conv-SDAE model, with cluster centers initialized from the latent code means of Conv-SDAE. The model undergoes two stages of training: first, Conv-SDAE is pre-trained with MSE loss to learn the latent space representation. In the second stage, the VaDE model is trained with the latent representations learned by Conv-SDAE, optimizing cluster assignments and model parameters over multiple runs with varying initializations to ensure robustness.

The model's performance is evaluated by computing clustering metrics such as Silhouette Score (measuring coherence and separation), Calinski-Harabasz Index (CHI) (evaluating cluster dispersion), and Davies-Bouldin Index (DBI) (quantifying cluster similarity). The evaluation is repeated across ten runs and the metrics are averaged for a final assessment.

3 Results and Discussion

The performance of the VaDE clustering model was compared with several other clustering techniques, including K-means, Gaussian Mixture Model (GMM), and Deep Embedding Clustering (DEC) [11]. DEC is evaluated in two configurations: (1) as a feature extractor, followed by K-means clustering on the extracted features, and (2) combined with K-means for final clustering, with simultaneous updates of both K-means and the neural network parameters. Table 1 and 2 summarize the clustering performance across all methods. Results indicate that the latent feature space structure is the dominant factor influencing clustering quality.

Classical algorithms such as K-means and GMM show limited ability to partition Mel-spectrogram representations of industrial audio signals. Their low Silhouette scores (0.2962 for K-means and 0.1600 for GMM) and reduced Calinski–Harabasz Index (3.63×10^3 and 2.32×10^3, respectively) reflect poor cohesion and insufficient separation between operational acoustic regimes. These results are consistent with the spectral properties of industrial soundscapes, which exhibit non-linear, non-Gaussian, and highly multi-modal distributions, violating the assumptions underlying centroid-based and Gaussian-parametric clustering. On the UrbanSound8K [12] dataset, similar trends are observed, where K-means achieves a Silhouette Score of 0.1712 and a CHI of 2.648×10^3, indicating poor cluster separation, with GMM performing even worse (Silhouette Score of 0.1245 and CHI of 1.846×10^3).

Deep Embedding Clustering (DEC) shows a clear performance advantage with a higher CHI (45.89×10^3) and a lower Davies-Bouldin Index (0.3873), indicating better intra-cluster compactness and inter-cluster separation. These results suggest that the DEC encoder effectively suppresses nuisance factors

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 1: Clustering Performance Comparison on Industrial Noise Dataset

Method	**Silhouette Score**	**CHI** $\times 10^3$	**DBI**
k-means	0.2962	3.63	1.1634
GMM	0.1600	2.32	1.6287
DEC + k-means	0.2226	1.98	1.4955
DEC	0.7510	45.89	0.3873
VaDE	**0.7900**	**63.55**	**0.2812**

Table 2: Clustering Performance Comparison on UrbanSound8K Dataset

Method	**Silhouette Score**	**CHI** $\times 10^3$	**DBI**
k-means	0.1712	2.648	1.6213
GMM	0.1245	1.846	1.9278
DEC + k-means	0.4221	15.600	0.7602
DEC	0.5466	5.124	1.0149
VaDE	**0.8309**	**131.62**	**0.3334**

such as ambient noise, mechanical reverberation, and non-stationary distortions while preserving the dominant spectral–temporal characteristics of industrial processes. On the UrbanSound8K dataset, DEC + K-means attains a Silhouette Score of 0.4221, which is an improvement but still notably lower than DEC alone. This shows that K-means as a final clustering step does not fully exploit the rich feature representations learned by DEC. The drop in performance when applying K-means to DEC embeddings further confirms that the latent geometry produced by DEC is not well-suited to isotropic Euclidean clustering; imposing centroid-based partitioning disrupts the cluster-consistent structure optimized during joint embedding and clustering.

The VaDE model, combining a Variational Autoencoder (VAE) with a Gaussian Mixture Model (GMM), achieves the strongest clustering performance across all metrics. Its probabilistic latent space yields the highest Silhouette Score (0.7900), the largest CHI (63.55×10^3), and the lowest Davies–Bouldin Index (0.2812), demonstrating superior intra- and inter-cluster separation. This indicates that the VAE effectively captures the multi-modal, non-linear acoustic distributions of steelmaking environments, producing a structured latent representation well aligned with underlying acoustic factors. When paired with GMM, the model forms compact and highly separable clusters, outperforming both traditional and deep representation-learning baselines. On the UrbanSound8K dataset, VaDE again performs best, achieving a Silhouette Score of 0.8309, a CHI of 131.620×10^3, and a DBI of 0.4442, confirming the robustness of the VaDE for complex acoustic modeling.

4 Conlusion and future work

This study demonstrates the use of an unsupervised deep learning framework for clustering industrial noise sources in steelmaking plants, leveraging Varia-

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

tional Autoencoders (VAEs) combined with Gaussian Mixture Models (GMM). The framework learns compact latent representations of Mel-spectrogram features from raw, unlabelled audio data, enabling effective clustering of complex industrial acoustic events. Our results show that this approach outperforms traditional clustering methods such as K-means, GMM, and Deep Embedding Clustering (DEC). These findings highlight the potential of deep generative models, particularly VAEs, to enhance industrial noise monitoring systems, offering an efficient and scalable solution for real-time noise analysis. Future work will focus on refining this framework to address complexities such as varying noise sources, dynamic operational conditions, and real-time deployment, further improving the model's adaptability and robustness in industrial environments.

References

[1] Iulian Petchesi. Analysis of the noise pollution from industrial sources inside agglomerations. Technical Report ETC/ACM Technical Paper 2015/3, ETC/ACM, UAB, Spain, April 2016. Prepared by: Miquel Sáinz de la Maza, Núria Blanes.

[2] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[3] Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4610–4617, 2019.

[4] Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE, 2020.

[5] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12), 2022.

[6] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022.

[7] Jeong-Geun Lee, Kwang Sik Kim, and Jang Hyun Lee. Sound-based unsupervised fault diagnosis of industrial equipment considering environmental noise. *Sensors*, 24(22):7319, 2024.

[8] Luan Vinícius Fiorio, Ivana Nikoloska, Wim van Houtum, and Ronald M Aarts. Clustering of acoustic environments with variational autoencoders for hearing devices. *arXiv e-prints*, pages arXiv–2510, 2025.

[9] Lang Liu, Ying Zheng, and Shaojun Liang. Variable-wise stacked temporal autoencoder for intelligent fault diagnosis of industrial systems. *IEEE Transactions on Industrial Informatics*, 20(5):7545–7555, 2024.

[10] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

[11] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[12] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.