# FL-VetTrans: Privacy-Preserving Translation of Animal Vocalization for Clinical Diagnosis via Federated Learning

**Kausar Ali** [1] **Aasim Zafar** [1]

## Abstract

With the growing integration of artificial intelligence in veterinary healthcare, automated analysis of animal vocalizations offers a promising non-invasive approach for early disease detection. However, traditional centralized learning frameworks raise privacy concerns and struggle with non-IID data distributions across distributed clinical environments. In this study, we propose FL-VetTrans, a privacy-preserving federated learning framework for disease classification using domestic animal vocal signals. The system leverages log Mel-Spectrogram representations of raw audio recordings and employs a convolutional neural network (CNN) for end-to-end acoustic feature learning. Training is performed locally at distributed edge nodes without sharing raw data. To address statistical heterogeneity across clients, FedProx regularization is incorporated to stabilize model convergence under non-IID conditions. Furthermore, differential privacy is implemented using Opacus-based DP-SGD, ensuring formal privacy guarantees through gradient clipping and noise injection. Experimental evaluation on healthy and unhealthy poultry vocal datasets demonstrates that the proposed federated framework achieves 93% global classification accuracy, along with strong precision, recall, F1-score, and ROC-AUC performance. These results highlight the effectiveness of privacy-preserving federated acoustic modeling for veterinary disease diagnosis.

Published at the 6th Muslims in ML (MusIML) Workshop, colocated with ICML 2026, Seoul, South Korea, July 2026. [1] Department of Computer Science, Aligarh Muslim University, Aligarh, India . Correspondence to: Kausar Ali <kausar.cs.amu@gmail.com>.

*Preprint. June 24, 2026.*

## 1. Introduction

Recent advances in speech translation models have reduced reliance on text-based pipelines (Sarim et al., 2025; Jamaluddin, 2026). Animals use nonverbal clues including posture, vocalizations, and behavioral changes to communicate health issues. Effective interpretation of these signals, particularly in companion animals, can facilitate prompt medical intervention. These cues and signals are extremely challenging for humans to decipher, but Machine Learning (ML) has undergone a complete shift how these cues are interpreted for animal medical diagnosis. Even though there are ML algorithms for animal behavior recognition (Kowalczuk et al., 2022; Lencioni et al., 2021) to help diagnose the animals and better interpret these cues, they tend to depend on centralized data collection, which presents logistical, ethical, privacy, and model generalization complexities.A decentralized alternative is Federated Learning (FL) (McMahan et al., 2017), which enables several organizations or hosts to jointly train ML models without exchanging raw data and boost model generalization. In veterinary contexts, where data may be limited, confidential, and species-specific, this framework is notably well-suited. A revolutionary component of precision cattle husbandry, voice-based disease detection uses non-invasive acoustic monitoring to detect respiratory and stress-related disorders. To train models on particular vocalizations like coughs or rales, researchers use a variety of datasets from pigs, cattle, and poultry, frequently obtained from repositories like Zenodo. Machine learning algorithms can identify diseases before clinical symptoms appear by transforming these audio inputs into spectrograms. By providing an affordable diagnostic solution using common microphones or smartphones, this early detection capacity dramatically lowers mortality rates and antibiotic dependence. Mel-Frequency Cepstral Coefficients (MFCC) are used extensively in the field to extract features and input them into sophisticated algorithms like CNNs and LSTMs (Telmem et al., 2025; Abdul & Al-Talabani, 2022). By eliminating extraneous farm noise from feeds and fans, these models attain remarkable accuracy rates, typically exceeding 90% (Rezaul et al., 2024).

Despite these advancements, there are still issues with the scarcity of large-scale, publicly labeled datasets covering a variety of species. Leveraging with our prior studies on

federated learning in biomedical domains, we present FL-VetTrans, an intelligent, scalable, privacy-preserving framework that employs FL to transform animal vocalizations as well as behavioral signals into medical diagnostics that are comprehended by humans. In contrast to previous research, our method combines safe federated optimization, multi-species adaptation, and distributed data (speech/vocals). This study is based on the following main research motivations that will direct the development and assessment of FL-VetTrans, a privacy-preserving structure that supports converting animal speech and behavior into therapeutically significant insights. This naturally raises a key question: can Federated Learning enable accurate veterinary diagnostics without centralized data? If animal vocalizations encode diagnostic cues, how effectively can FL extract them? Moreover, would combining species-specific personalization, multimodal analysis, and privacy-preserving techniques improve accuracy, robustness, and early disease detection over traditional models?

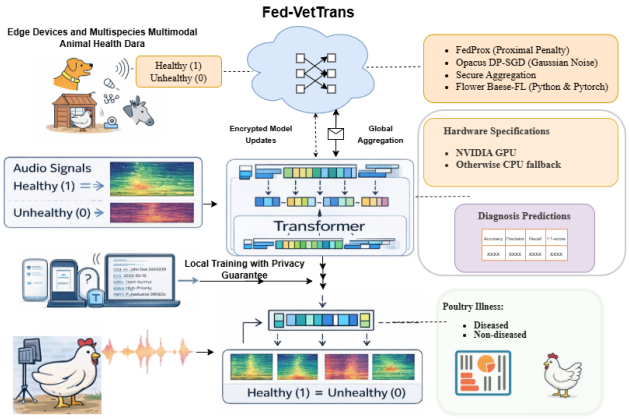


*Figure 1.* FL-VetTrans architecture with edge training and multimodal translation.

These are the primary motivations behind FL-VetTrans to increase accessibility to early diagnostics in rural and underserved veterinary settings, where resources and clinical infrastructure are often limited. Lastly, the framework as depicted in fig 1 is designed to enable cross-species generalization and personalization, ensuring adaptability across diverse animal populations and clinical contexts.

The remaining work is organized as, section 2 has related work to this study and proposed work is presented in the section 3. Results and experimentation of the study is demonstrated in 4, discussion in S.VII, limitations in S.VIII of Supplement file and section 5 concludes the study.

## 2. Related Work

### 2.1. Animal Communication and Health Diagnosis

Animal vocalizations have been linked to emotional or physiological states in plenty of studies (Labra et al., 2013).However, the scope of these studies is often restricted; they are species-specific and lack generalization (Li et al., 2020). The classification of dog barks (Abadi et al., 2016), cat meows (Bonawitz et al., 2017), and horse whinnies has demonstrated the potential of deep learning models; however, they are not integrated with behavioral and physiological data, which is necessary for clinical diagnosis. Recent developments include reinforcement learning techniques for automated behavior correction and CNN-LSTM models for posture-based disease detection (Bonawitz et al., 2017). However, issues related to data scarcity and privacy remain largely unaddressed. Federated Learning (FL), which provides data privacy and model personalization, has been effectively used in human healthcare for ECG classification, COVID-19 diagnosis (Lencioni et al., 2021), and mental health prediction (McMahan et al., 2017). Despite being less studied, veterinary data is just as sensitive because of private clinical information and owner-animal confidentiality. Cutting-edge translation systems use Transformer architectures (Labra et al., 2013) for human languages, but as animal vocalizations lack formal syntactic rules, modifying these models for animal speech necessitates reconsidering embeddings, structure, and contextual representation. Differential Privacy (DP) (Abadi et al., 2016) introduces noise to safeguard individual data, while FedProx (Li et al., 2020) adds a proximal term to stabilize local updates to tackle non-IID data distributions and device heterogeneity. For FL to be implemented in veterinary systems in the real world, both methods are necessary.

### 2.2. Recent Advancement

Animal vocalizations are trustworthy markers of behavior, health, and physiological state, according to recent developments in bioacoustics. An automated audio-surveillance framework for the detection of pig wasting disorders was proposed by Chung et al. (Chung et al., 2013) in one of the first frameworks for acoustic disease detection.The system achieved 94% detection accuracy and 91% classification accuracy using Mel-Frequency Cepstral Coefficients (MFCCs) as acoustic features and a hierarchical structure that combined Support Vector Data Description (SVDD) for anomaly detection and Sparse Representation Classifier (SRC) for disease classification. Cough sounds are early indicators of pathological situations, as demonstrated by the study's focus on respiratory disorders such Porcine Reproductive and Respiratory Syndrome (PRRS), Postweaning Multisystemic Wasting Syndrome (PMWS), and *Mycoplasma Hyopneumoniae* (MH). For the purpose of monitoring respiratory

diseases in poultry, Adebayo et al. (Adebayo et al., 2023) created a publicly available collection of healthy and unhealthy chicken vocalizations. Over the course of 65 days, audio signals were recorded at 24-bit resolution and a sampling frequency of 96 kHz.

Karaaslan et al. (Karaaslan et al., 2024) proposed a fully autonomous deep learning-based speech analysis system for dogs in companion animal research. The Short-Time Fourier Transform (STFT), MFCC, and Linear-Frequency Cepstral Coefficients (LFCC) were used to extract spectral features after segmentation based on root mean square energy (RMSE). Barking and howling sounds were classified using a variety of convolutional neural network (CNN) architectures, such as AlexNet, DenseNet, EfficientNet, and ResNet variations. The work showed the importance of spectrum representations and CNN-based classifiers in large-scale automated bioacoustic analysis, despite its primary focus on behavioral classification rather than health diagnosis. Lamothe et al. (Lamothe et al., 2025) introduced a large-scale annotated dataset of typical marmoset vocalizations that went beyond domestic species. The collection included over 800,000 segmented recordings that were recorded at 96 kHz.

These studies reveal three key insights: respiratory diseases exhibit acoustic signatures captured by MFCCs; automated segmentation with deep learning improves scalability and accuracy; and annotated datasets remain limited, especially for companion animals. However, reliance on centralized, single-species models highlights the need for federated, multimodal, privacy-preserving diagnostic frameworks.

## 3. Proposed Work

In this section, we demonstrate the proposed work for diagnosing the respiratory disease of poultry birds based on Vocal data using Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. Algorithmic structure is presented in Section S.IV of Supplement file and Feature Engineering and model explanation are discussed in Section S.I and S.II of the Supplement file respectively.

### 3.1. Federated-Aware Feature Optimization

Additional safeguards were included to guarantee robustness along with privacy preservation because training takes place during a decentralized federated setting. FEDPROX REGULARIZATION and DIFFERENTIAL PRIVACY VIA OPACUS is presented in S.IX of the Supplement file.

The proposed framework achieves robustness to non-IID data and robust privacy protection by combining FedProx regularization with Opacus-based DP-SGD. In remote veterinary diagnostic settings, this architecture guarantees stability and privacy protection while enabling end-to-end acoustic representation learning.

### 3.2. Federated Learning Framework

#### 3.2.1. System Architecture

The Flower (FLWR) framework is used to construct a decentralized federated learning system in the proposed FL-VetTrans framework. A central coordination server and several edge clients, such as local monitoring devices or veterinary clinics, make up the system. Without exchanging raw audio data, each client keeps its local dataset and does training. Algorithm S.IV.A of the Supplement file summarizes the entire federated learning workflow, whereas Algorithm S.IV.B of the Supplement file describes the local client training process with differential privacy and Transformer-based classification. Let $K$ represent the total number of clients who are taking part. The global model parameters $\mathbf{w}^{(t)}$ are broadcast to specific clients by the server at communication round $t$. Federated Averaging (FedAvg) is used to aggregate the modified parameters following local optimization:

$$\mathbf{w}^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{\sum_{k=1}^{K} n_k} \mathbf{w}_k^{(t+1)} \tag{1}$$

where $n_k$ denotes the number of local samples at client $k$, and $\mathbf{w}_k^{(t+1)}$ represents the locally updated parameters. To facilitate effective parameter interchange between PyTorch models and the Flower aggregation server, the implementation makes use of the `NumPyClient` interface.

#### 3.2.2. Local Training Procedure

Every client uses mini-batch stochastic optimization for local training. The Adam optimizer with cross-entropy loss is used to optimize the CNN-based classifier previously discussed. FedProx regularization is integrated into the local goal function to lessen the effects of non-IID data distributions between clients:

$$L_k = L_{\text{CE}} + \frac{\mu}{2} \|\mathbf{w}_k - \mathbf{w}^{(t)}\|_2^2 \tag{2}$$

where $L_{\text{CE}}$ denotes the cross-entropy loss, $\mathbf{w}^{(t)}$ represents the global parameters received from the server, and $\mu$ controls the strength of the proximal constraint. Under diverse data conditions, this approach enhances convergence stability and reduces excessive local drift.

Every communication round, local training is carried out for a predetermined number of epochs. Transparency and convergence monitoring are ensured by displaying batch-wise loss values and epoch-level progress during execution.

*Table 1.* Average Performance Metrics Across Training Rounds

| Metric | Round 1 | Round 2 | Round 3 | Round 4 | Round 5 |
|---|---|---|---|---|---|
| Accuracy | 0.85 | 0.84 | 0.88 | 0.92 | 0.93 |
| Precision | 0.83 | 0.77 | 0.82 | 0.90 | 0.92 |
| Recall | 0.88 | 0.97 | 0.97 | 0.95 | 0.94 |
| F1-score | 0.85 | 0.86 | 0.89 | 0.92 | 0.93 |
| ROC-AUC | 0.92 | 0.94 | 0.95 | 0.95 | 0.95 |

*Table 2.* Global Accuracy Across Training Rounds

| Metric | Round 1 | Round 2 | Round 3 | Round 4 | Round 5 |
|---|---|---|---|---|---|
| Accuracy | 0.85 | 0.85 | 0.89 | 0.92 | 0.93 |

#### 3.2.3. Differential Privacy Mechanism

The Opacus framework in PyTorch is used to implement Differentially Private Stochastic Gradient Descent (DP-SGD) in order to guarantee client-level privacy. Per-sample gradient computation is incorporated into the local optimization procedure via the `PrivacyEngine`. For each mini-batch, individual gradients $\mathbf{g}_i$ are clipped such that $\|\mathbf{g}_i\|_2 \leq C$, where $C$ denotes the maximum gradient norm, thereby bounding the global sensitivity of the update. Gaussian noise drawn from $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ is then added to the aggregated gradient prior to parameter updates.

The method offers dual-layer privacy protection through data localization and stochastic gradient perturbation since only perturbed model parameters are sent to the central server while raw data stays on local devices.

## 4. Experiments, results, and discussion

### 4.1. Dataset Description and Evaluation Matrices

Experiments were conducted on a locally curated audio dataset developed to provide an open, accessible, and high-quality machine learning resource for poultry farm management. The dataset (Aworinde et al., 2023) comprises 346 `.wav` files organized into three categories: healthy (139), noise (86), and unhealthy (121) as mentioned in table S.III.C of the Supplement file. For classification, samples were labeled as Healthy (1) and Unhealthy (0), with noise handled according to the experimental setting. All recordings were resampled to 16 kHz and converted into log Mel-spectrograms with 64 Mel bands and 128 time frames, producing input tensors of size $1 \times 64 \times 128$. The dataset was split across hosts and the number of training epochs for each client is described in Table S.III.A and Table S.III.B of the Supplement file respectively. Evaluation matrices for the model are presented in S.VI of the Supplement file.

### 4.2. Evaluation on Poultry Farm Dataset

Table 2 presents the round-wise performance metrics for comprehensive evaluation of the proposed framework at global model. At round 5, the proposed framework achieved the highest accuracy of 93%, which evidently shows that the increment in rounds leads to accuracy improvements. In table 1, the average of all the performance matrices for each round is recorded. It is clearly seen that all the matrices are getting improve as the training rounds succeed, which serves as an evidence for federated learning environment to make improvements over the rounds of training the models locally and the aggregating the results from all local clients. Accuracy, Pricision, Recall, F1-score and ROC-AUC for individual clients at local model is presented in Table S.V.A of the Supplement file. Table S.V.A of the Supplement file, shows how the individual local models participated in training. In round 5 of Table S.V.A of the Supplement file, the highest accuracy of 96% is achieved by client 2 and client 4 followed by client 5 with accuracy 95%. Client-2 achieves highest precision of 1.00 and second highest presicion is achieved by client-5 of 0.95 followed by client-6 with 0.93. The highest F1 score of 0.96 is achived by client-2 and then client-1, client-4 and client-5 achieve second highest F1 score of 0.95 and client-6 following the same with 0.93. Similarly, client-4 achieved 1.00 of ROC-AUC followed by client-2 with 0.98. Among the participating clients, these results demonstrate the effectiveness of the proposed federated acoustic learning framework across distributed clients under heterogeneous data conditions.

## 5. Conclusions

This study presents FL-VetTrans, a federated and privacy-preserving framework for animal disease diagnosis using vocalization data. By integrating CNN-based acoustic feature extraction with a Transformer-based classifier, the proposed approach effectively captures both local spectral patterns and long-range temporal dependencies associated with disease-related vocal anomalies. The adoption of federated learning enables decentralized model training across multiple clients without sharing raw audio data, thereby addressing privacy concerns and supporting real-world deployment in distributed veterinary environments with enhanced generalization. The incorporation of FedProx enhances robustness under non-IID data distributions, while differential privacy mechanisms ensure secure model updates. Experimental results demonstrate that the proposed framework achieves reliable performance across multiple evaluation metrics, including accuracy, F1-score, and ROC-AUC, indicating its effectiveness for non-invasive and scalable animal health monitoring. Moreover, the use of log Mel-spectrogram features proves to be highly informative for distinguishing healthy and diseased vocal patterns.

Despite these promising outcomes, challenges such as data heterogeneity, environmental noise, and limited annotated datasets remain. Future work will focus on extending the

framework to multi-species datasets, incorporating additional modalities such as behavioral and physiological signals, and optimizing privacy–utility trade-offs.

## References


Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Abdul, Z. K. and Al-Talabani, A. K. Mel frequency cepstral coefficient and its applications: A review. *Ieee Access*, 10:122136–122158, 2022.

Adebayo, S., Aworinde, H. O., Akinwunmi, A. O., Alabi, O. M., Ayandiji, A., Sakpere, A. B., Adeyemo, A., Oyebamiji, A. K., Olaide, O., and Kizito, E. Enhancing poultry health management through machine learning-based analysis of vocalization signals dataset. *Data in Brief*, 50, 10 2023. ISSN 23523409. doi: 10.1016/j.dib.2023.109528.

Aworinde, H., Adebayo, S., Akinwunmi, A., Alabi, O., Ayandiji, A., Oke, O., Oyebamiji, A., Adeyemo, A., Sakpere, A., and Echetama, K. Poultry vocalization signal dataset for early disease detection, 2023. URL https://doi.org/10.17632/zp4nf2dxbh.1.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Chung, Y., Oh, S., Lee, J., Park, D., Chang, H. H., and Kim, S. Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. *Sensors (Switzerland)*, 13:12929–12942, 9 2013. ISSN 14248220. doi: 10.3390/s131012929.

Jamaluddin. Thesis proposal: Development of end-to-end speech translation models for Indian languages. In Baez Santamaria, S., Somayajula, S. A., and Yamaguchi, A. (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 535–543, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-383-8. doi: 10.18653/v1/2026.eacl-srw.41. URL https://aclanthology.org/2026.eacl-srw.41/.

Karaaslan, M., Turkoglu, B., Kaya, E., and Asuroglu, T. Voice analysis in dogs with deep learning: Development of a fully automatic voice analysis system for bioacoustics studies. *Sensors*, 24, 12 2024. ISSN 14248220. doi: 10.3390/s24247978.

Kowalczuk, Z., Czubenko, M., and Żmuda-Trzebiatowska, W. Categorization of emotions in dog behavior based on the deep neural network. *Computational Intelligence*, 38 (6):2116–2133, 2022.

Labra, A., Silva, G., Norambuena, F., Velásquez, N., and Penna, M. Acoustic features of the weeping lizard's distress call. *Copeia*, 2013(2):206–212, 2013.

Lamothe, C., Obliger-Debouche, M., Best, P., Trapeau, R., Ravel, S., Artières, T., Marxer, R., and Belin, P. A large annotated dataset of vocalizations by common marmosets. *Scientific Data*, 12, 12 2025. ISSN 20524463. doi: 10.1038/s41597-025-04951-8.

Lencioni, G. C., de Sousa, R. V., de Souza Sardinha, E. J., Corrêa, R. R., and Zanella, A. J. Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. *PloS one*, 16(10): e0258672, 2021.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Rezaul, K. M., Jewel, M., Islam, M. S., Siddiquee, K., Barua, N., Rahman, M., Sulaiman, R., Shaikh, M., Hamim, M., Tanmoy, F., et al. Enhancing audio classification through mfcc feature extraction and data augmentation with cnn and rnn models. *International Journal of Advanced Computer Science and Applications*, 15(7):37–53, 2024.

Sarim, M., Shakeel, S., Javed, L., Nadeem, M., et al. Direct speech to speech translation: A review. *arXiv preprint arXiv:2503.04799*, 2025.

Telmem, M., Laaidi, N., and Satori, H. The impact of mfcc, spectrogram, and mel-spectrogram on deep learning models for amazigh speech recognition system. *International Journal of Speech Technology*, 28(1):299–312, 2025.