Generative Modeling of Biomedical Audio Signals for Respiratory Disease Screening Using Reference-Guided Feature Adaptation

Kristopher Carney
Department of Computer Science, University of Houston, Houston, TX, USA.
kristopher1974@uh.edu

Meike Stolze
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
meikestolze691@colostate.edu

Abstract

Respiratory audio offers a low-cost channel for screening conditions such as asthma, chronic obstructive pulmonary disease, pneumonia, and viral respiratory infection, but learned screeners are fragile under acquisition and population shift. Microphone response, room acoustics, breathing effort, body habitus, and subject-specific timbre can dominate the subtle acoustic cues associated with disease. Generative models could improve robustness by learning latent structure and synthesizing underrepresented conditions, yet unconstrained synthesis can reproduce device artifacts, speaker identity, and other nuisance factors rather than pathology. This manuscript develops a reference-guided generative feature adaptation framework for respiratory disease screening. A short, label-free reference recording conditions a scale-shift transformation in a spectral latent space, while a conditional generative model reconstructs and samples respiratory spectro-temporal representations. A screening classifier operates on adapted latents and is regularized by disease classification, domain-adversarial, reference-consistency, and reconstruction objectives. The framework separates what the reference is allowed to explain from what the classifier is allowed to use, and it couples predictions with calibration and uncertainty gates. Analytical assessment identifies the central trade-off: adaptation strong enough to remove device and subject style can also erase disease-related transients, while weak adaptation leaves covariate shift intact. The contribution is a technically specified, testable design that links reference conditioning, generative augmentation, nuisance reduction, and uncertainty-aware decision rules. It provides a foundation for rigorous subject-level and device-held-out evaluation of respiratory audio screening before clinical deployment.

Keywords

respiratory audio; generative modeling; reference-guided adaptation; domain shift; biomedical signal processing; uncertainty-aware screening

1. Introduction

Respiratory disease screening seeks to identify individuals who warrant further clinical assessment using signals that can be collected outside specialized laboratories. Audio is attractive in this context because cough, speech, and breathing can be recorded with consumer devices at low cost and without ionizing radiation. Research on crowdsourced respiratory audio has shown that cough and breathing recordings contain information associated with

COVID-19 status and other respiratory conditions, although reported performance varies substantially across cohorts and recording conditions [1]. Smartphone-based cough classification studies have likewise demonstrated the feasibility of machine-learning pipelines while exposing sensitivity to dataset composition, microphone quality, and participant behavior [2]. Larger breathing, cough, and voice collections have made respiratory audio research more reproducible and have underlined the importance of standardized acquisition protocols [3]. These developments motivate screening tools that can operate under heterogeneous field conditions, but they also reveal a persistent obstacle: the acoustic signal is shaped as much by the recording chain and the person producing the sound as by the disease process.

The central difficulty is not merely limited labeled data. Respiratory audio classifiers are trained on finite populations, devices, and environments, and they are then asked to generalize to new microphones, rooms, breathing maneuvers, ages, body sizes, and disease prevalences. In such settings, a model can learn a shortcut from device frequency response or background noise to the label available in the training cohort. The resulting classifier may appear accurate within a curated dataset and fail when the acquisition condition changes. Domain adaptation methods address related problems in vision and speech by aligning feature distributions or learning domain-invariant representations [9], [10]. Clinical machine learning has increasingly emphasized that dataset shift is not an edge case but a routine property of deployed systems [11], and subgroup analyses have shown that apparently accurate models can systematically underdiagnose underserved populations [12]. Physiological signal models face analogous challenges because subject identity, sensor placement, and acquisition noise are entangled with clinically meaningful variation [13].

Generative modeling offers a complementary route. Generative adversarial networks, variational autoencoders, diffusion models, and autoregressive waveform models can learn complex audio distributions and produce realistic samples [4], [5], [6], [7]. High-fidelity neural vocoders and generative speech systems have demonstrated that spectral and waveform structure can be synthesized with considerable control [8]. For biomedical audio, the promise is that a generative model can represent latent factors of variation, support data augmentation for rare conditions, and provide a feature space in which nuisance factors can be modified without altering disease status. The risk is equally clear: a generative model trained on respiratory recordings may learn to reproduce speaker identity, microphone coloration, and environmental noise more reliably than crackles, wheezes, or other disease-related events. Unconstrained augmentation can therefore amplify exactly the confounders that screening systems must ignore.

Reference-guided style transfer provides a useful design idea. In audio generation, a reference exemplar can condition the style or timbre of synthesized output; reference-guided style transfer has been explored for out-of-domain singing voice synthesis, where a short sample informs timbre and rendering [14]. The analogy for respiratory screening is not that breathing is singing, but that a brief reference recording can describe the acquisition and subject-specific rendering of a signal. If a model can use that reference to adapt latent features toward the target domain, then a classifier may rely less on device-specific offsets and more on transient respiratory events. The analogy also exposes a hazard: if the reference is allowed to explain too much, it may absorb disease-related information, especially when the reference contains cough or abnormal breathing.

This manuscript develops a reference-guided generative feature adaptation framework for respiratory disease screening. The framework combines a conditional generative model over spectral representations, a low-capacity reference-conditioned scale-shift adaptation in latent space, a screening classifier, domain-adversarial regularization, reference-consistency constraints, and an uncertainty-aware decision rule. The contribution is methodological and analytical rather than an empirical benchmark. It specifies how generative modeling and reference-guided adaptation can be coupled so that the reference branch is encouraged to describe acquisition and subject style while the classifier is encouraged to use disease-related spectral dynamics. It then analyzes the conditions under which this separation is plausible, the failure modes that arise when it is not, and the evaluation logic required before clinical claims can be made. The intended advance is a precise, testable architecture and a set of design constraints for robust respiratory audio screening under domain shift.

2. Biomedical Audio Generation and Adaptation Landscape

2.1 Respiratory sound mechanisms and screening targets

Respiratory sounds arise from turbulent airflow, airway narrowing, secretions, and tissue vibration. Wheezes are associated with narrowed lower airways and are common in asthma and chronic obstructive pulmonary disease, although they are not specific to either condition. Crackles can reflect opening of collapsed airways or fluid in the small airways and are observed in pneumonia, heart failure, and interstitial lung disease. Stridor suggests upper airway obstruction, while cough is a complex protective reflex with highly variable acoustic morphology. These mechanisms mean that respiratory audio can carry clinically relevant information, but the mapping from acoustic event to disease label is indirect and context dependent [1], [2], [3].

Screening is a more appropriate target than diagnosis for most audio-based systems. A screening model should identify individuals who need spirometry, imaging, microbiological testing, or clinical examination; it should not replace those assessments. This distinction matters for model design because screening emphasizes sensitivity, calibration, and decision thresholds rather than fine-grained disease subtyping. It also changes the cost of errors: a missed high-risk case may be more consequential than a false positive that leads to further evaluation, but excessive false positives can overwhelm clinical pathways and undermine trust. Respiratory audio screening must therefore be evaluated with decision-relevant endpoints, not only with area under the receiver operating characteristic curve.

Acquisition confounders are pervasive. A close-talk microphone captures different high-frequency energy than a smartphone held at arm’s length. Room reverberation smears transient events. Breathing effort changes airflow velocity and spectral tilt. Age, sex, body mass index, smoking history, and nasal versus oral breathing alter the signal. Many of these factors are also associated with disease prevalence, so a model can learn a spurious correlation that appears useful in one cohort and fails in another. The framework developed here treats these factors as nuisance variables to be controlled through reference-guided adaptation and domain-adversarial regularization, while recognizing that complete separation is not guaranteed.

2.2 Generative modeling of audio

Generative models can represent audio at waveform or spectral resolution. Autoregressive models such as WaveNet factorize the waveform into conditional distributions over samples and can produce high-quality speech and music [7]. Generative adversarial networks learn an

implicit distribution through a discriminator and have been used for efficient neural vocoding [4], [8]. Variational autoencoders introduce an explicit latent variable and optimize a tractable lower bound on the data likelihood [5]. Diffusion models define a forward noising process and learn to reverse it, often achieving strong sample quality and stable training [6]. Each family imposes different inductive biases. Autoregressive models capture fine temporal dependencies but are computationally expensive for long recordings. Adversarial models can synthesize realistic textures but may suffer from mode collapse and unstable training. Variational models provide latent structure useful for adaptation but can produce overly smooth reconstructions. Diffusion models offer flexible likelihood-based training but require iterative sampling and careful noise scheduling.

For respiratory screening, the generative model is not an end in itself. It serves three functions: it provides a compact latent representation, it can synthesize plausible spectral variants for augmentation, and it can be regularized so that nuisance factors and disease-related factors are represented separately. The choice of representation is consequential. Waveform generation preserves phase and transient detail, but phase is highly sensitive to microphone placement and room acoustics. Spectral magnitude representations discard absolute phase and are more amenable to domain adaptation, but they can obscure temporal fine structure. A practical compromise is to model log-magnitude spectro-temporal patches with a conditional latent generative model and to use a multi-resolution spectral loss to preserve transient events [7], [8].

The generative approach also introduces risks that are specific to biomedical data. A model trained on a small cohort may memorize speaker identity and device characteristics. Synthetic samples can appear realistic while lacking pathological events, producing a classifier that is robust to synthetic noise but not to real disease variability. If the generator is conditioned on disease labels, it may learn label-correlated shortcuts rather than mechanisms. These risks motivate a feature-space generative model with explicit nuisance conditioning and constraints that prevent the reference branch from becoming a disease classifier.

2.3 Domain shift and reference-guided adaptation

Domain shift in respiratory audio can be viewed as a combination of covariate shift, label shift, and concept shift. Covariate shift occurs when the distribution of acoustic features changes across devices or environments while the conditional relationship between disease and features remains approximately stable. Label shift occurs when disease prevalence changes across sites. Concept shift occurs when the same acoustic pattern has different clinical meaning in different populations, for example because of age-related changes in airway mechanics or differences in recording protocol. Domain-adversarial training attempts to learn features that are discriminative for the task but uninformative about the domain [9]. Correlation alignment methods match second-order statistics between source and target representations and offer a simpler alternative when domain labels are available [10]. These methods are useful starting points, but they assume that domain and label are separable in feature space, an assumption that is fragile when disease prevalence varies across domains.

Clinical machine learning research has documented how dataset shift can invalidate models that appear well calibrated in development [11]. Subgroup analyses have shown that models can underdiagnose patients from underserved groups even when overall performance appears acceptable [12]. In physiological signal modeling, contrastive learning across patients has been used to learn representations that generalize across subjects and acquisition

conditions [13]. These findings support the view that respiratory audio screening requires subject-aware and device-aware evaluation, not only random splits.

Reference-guided adaptation adds an explicit mechanism for using a short target-domain recording. In singing voice synthesis, reference-guided style transfer conditions generation on an exemplar so that timbre and style can be transferred to new content [14]. The underlying idea is that a reference encodes aspects of the rendering process that are not captured by the content label. For respiratory audio, a reference could be a brief period of quiet breathing recorded on the same device, in the same room, and by the same subject immediately before the diagnostic segment. The reference is label-free and is not used to infer disease; it is used to estimate the acquisition and subject-specific style. This design is appealing because it turns an unavoidable calibration burden into an explicit input. It is also delicate because the reference may contain disease-related events, especially if it includes cough or labored breathing.

A low-capacity adaptation is preferable to a high-capacity one. Feature-wise linear modulation provides a controlled way to condition neural networks on auxiliary information through scale and shift parameters [15], and style-based generators demonstrate that scale-shift transformations can control hierarchical style attributes [16]. These ideas motivate the reference-conditioned adaptation described in Section 4. The adaptation is deliberately restricted to first-order statistics in a latent space so that it can absorb device gain, spectral tilt, and some subject-specific timbre without redefining the content representation. The analytical question is whether this restriction is sufficient for realistic respiratory audio shift and whether it preserves disease-related transients.

3. Problem Formulation

3.1 Notation and learning problem

Let x denote a raw respiratory audio segment, such as a cough, a sustained vowel, or a sequence of breathing cycles. Let y $\in$ {0, 1} denote the screening label, where y = 1 indicates that the participant meets the clinical criterion for further assessment. Let a denote auxiliary metadata that may include age, sex, body mass index, device type, and protocol variables. Let r denote a short reference recording from the same participant and acquisition session. The reference is available at test time and is not associated with a disease label. The first step is to convert the raw waveform into a spectro-temporal representation suitable for generative modeling and classification.

$$X = S(x) \in \mathbb{R}\hat{}(T \times F)$$

Here S is a short-time spectral operator, such as a log-magnitude short-time Fourier transform or a log-mel transform, X is a matrix with T time frames and F frequency bins, and $\mathbb{R}$ denotes the real-valued domain. The representation should preserve transient events such as crackles and the harmonic structure of voiced sounds while suppressing absolute phase, which is highly sensitive to microphone placement and room acoustics. The choice of window length and hop size controls the trade-off between time and frequency resolution. Short windows localize crackles but blur low-frequency wheezes; longer windows resolve tonal wheezes but smear transient events. The framework can accommodate multiple resolutions by concatenating representations or by using a multi-resolution spectral loss.

The generative model defines a conditional distribution over X given metadata a and reference r.

$$p\theta(X|a,r) = \int p\theta(X|z,a,r)\, p(z)\, dz$$

In this expression, z is a latent variable, p(z) is a prior over latent factors, pθ(X|z,a,r) is the decoder distribution parameterized by θ, and the integral marginalizes over latent factors. The conditional dependence on r is the mechanism through which the reference recording informs generation. In practice, the integral is intractable, and training uses a variational lower bound, an adversarial objective, or a diffusion objective. The equation is a specification of the generative assumption rather than a claim that the marginal likelihood can be computed in closed form.

3.2 Causal structure and identifiability assumptions

The framework assumes that the observed spectrogram X is generated by three interacting factors: disease-related physiology, subject-specific and acquisition-related style, and random noise. Disease affects transient events, spectral envelope, and temporal dynamics. Style affects frequency response, gain, reverberation, and timbre. The reference r is assumed to carry information mainly about style, because it is recorded under the same conditions and from the same participant but is not itself a diagnostic segment. This assumption is an approximation. If the reference contains cough or abnormal breathing, it may carry information about disease. If the reference is recorded with a different maneuver than the diagnostic segment, it may not capture the relevant style.

Identifiability is the central theoretical issue. Without constraints, a latent variable can encode any mixture of disease and style, and the generative model can achieve a good reconstruction without learning a clinically meaningful decomposition. The proposed framework imposes four constraints. First, the reference-conditioned transformation is low-capacity and operates on first-order latent statistics. Second, a domain classifier attempts to predict device or site from the adapted latent, and an adversarial term discourages that prediction. Third, a reference-consistency term penalizes changes in the predicted disease probability when the reference is replaced by another reference from the same acquisition condition. Fourth, the reference encoder is not trained with the disease label directly, and gradients from the classifier are stopped or attenuated through the reference branch. These constraints do not guarantee identifiability, but they make the desired decomposition a explicit design target rather than an unconstrained byproduct of training.

The screening classifier estimates the probability that the participant meets the screening criterion.

$$p(y = 1|X,r,a) = \sigma(f\eta(\tilde{z},a))$$

Here σ is the logistic function, fη is a classifier parameterized by η, z̃ is the adapted latent representation, and a is the auxiliary metadata. The classifier is trained on real labeled data and on generated samples whose labels are known by construction. The metadata term allows the model to account for prevalence-related factors such as age or device type, but it also creates a risk of encoding group-specific shortcuts. Metadata should therefore be used with subgroup evaluation and, where appropriate, fairness constraints.

3.3 Screening objective and decision rule

A screening system should output a calibrated probability and a decision. The probability supports threshold selection and communication, while the decision determines whether further assessment is recommended. The decision rule should account for uncertainty because a confident-looking probability from an out-of-distribution device can be misleading. The

uncertainty gate is described in Section 4.5. The important formulation point is that the learning objective should not optimize classification accuracy alone. It should combine reconstruction, classification, domain invariance, reference consistency, and regularization so that the adapted representation is useful for screening while remaining stable across nuisance conditions.

4. Methodological Framework

4.1 Spectral encoder and reference embedding

The framework uses an encoder Eφ that maps the spectrogram X to a latent variable z. The encoder can be a convolutional network with temporal pooling, a recurrent network, or an attention-based sequence model. The choice should reflect the temporal scale of the clinical cue: crackles require fine temporal resolution, while wheezes require spectral resolution. The encoder outputs a mean and log-variance for a variational model, or a deterministic latent for a diffusion or adversarial model. The reference recording r is processed by a separate encoder Rψ to produce a reference embedding e_r. The reference encoder should be small relative to the main encoder to limit its capacity to encode disease information. It can be trained with reconstruction and domain objectives but not with the disease label.

The auxiliary metadata a is embedded and concatenated with the adapted latent or used to condition the decoder and classifier. Metadata can include device type, sampling rate, and protocol variables. Continuous variables such as age should be normalized; categorical variables should be embedded. If metadata is unavailable at test time, the framework can fall back to a reference-only adaptation, but the classifier should be evaluated under that missingness condition. The reference embedding is computed from the same reference recording that is available at test time, so there is no train-test mismatch in the reference pathway.

4.2 Reference-conditioned latent adaptation

The core adaptation step transforms the source-standardized latent code toward the reference style. Let μ_s and σ_s denote the mean and standard deviation of the source latent representation, estimated from the training cohort. Let Aω be an adaptation network that takes the reference embedding e_r and, optionally, metadata a and outputs scale and shift parameters. The adapted latent is computed as follows.

$$\tilde{z} = \gamma\omega(e_r) \odot ((z - \mu_s) \oslash \sigma_s) + \beta\omega(e_r)$$

In this equation, z is the latent code produced by the encoder, e_r is the reference embedding, γω and βω are the scale and shift vectors produced by Aω, ⊙ denotes element-wise multiplication, and ⊘ denotes element-wise division. The term (z − μ_s) ⊘ σ_s standardizes each latent dimension with respect to the source cohort. The adaptation then re-scales and re-centers the standardized code according to the reference. The transformation is intentionally low-capacity: it can absorb global gain, spectral tilt, and some subject-specific timbre, but it cannot arbitrarily rearrange the latent space. This restriction is a safeguard against removing disease-related information.

The design draws on feature-wise linear modulation, which conditions neural networks through scale and shift parameters [15], and on style-based generators, which use scale-shift transformations to control hierarchical attributes [16]. In those settings, scale-shift modulation is effective because it changes the statistics of feature maps without discarding content. In respiratory audio, the analogous hope is that the reference can explain device and subject style

while the residual latent code retains transient disease cues. The adaptation network should be regularized toward identity when the reference is uninformative or noisy. This can be achieved by penalizing large deviations of γω and βω from one and zero, respectively, and by training with reference dropout so that the classifier remains usable when the reference is missing.

4.3 Conditional generative decoder and classifier

The decoder Gθ reconstructs or generates a spectrogram from the adapted latent and metadata. The decoder can be a convolutional upsampling network, a diffusion model over spectral patches, or a conditional generative adversarial network. The choice affects sample quality and training stability. A diffusion decoder is attractive because it provides stable training and flexible conditional sampling, but it requires iterative generation. A variational decoder is faster but may produce overly smooth spectra. A hybrid design can use a variational encoder for latent inference and a diffusion decoder for high-fidelity reconstruction. The decoder is trained with a reconstruction loss and a perceptual or multi-resolution spectral loss to preserve transient events.

The generative loss is specified as follows.

$$L_gen = E[||X - G\theta(\tilde{z}, a)||_1] + \lambda_p \, L_perc(X, G\theta(\tilde{z}, a))$$

Here E denotes expectation over training examples and sampled latent codes, Gθ(z̃,a) is the generated spectrogram, ‖·‖_1 is the L1 reconstruction loss, L_perc is a multi-resolution spectral loss that compares magnitude spectrograms at several window sizes, and λ_p is a positive weight. The L1 term encourages accurate reconstruction of dominant spectral energy. The perceptual term encourages the preservation of transients across time scales. Neither term guarantees that the generated samples are clinically valid; they are training signals that shape the generative distribution.

The classifier fη operates on the adapted latent z̃ and metadata a. It can be a small multilayer perceptron or a temporal attention network. The classifier should be trained on real labeled examples and on generated examples whose labels are assigned by the generative model's conditioning variable. Generated examples should not be treated as independent clinical observations. They are synthetic regularizers that increase the diversity of nuisance conditions seen by the classifier. To reduce the risk of synthetic artifacts, generated samples can be filtered by a discriminator or by a disagreement criterion among an ensemble of classifiers. The classifier output is the screening probability in Eq. (4).

4.4 Training objectives and constraints

The full objective combines reconstruction, classification, domain invariance, reference consistency, and parameter regularization.

$$L = L_gen + \alpha \, L_cls + \beta \, L_dom + \gamma \, L_ref + \lambda||\theta||^2$$

In this expression, L_gen is the generative loss in Eq. (5); L_cls is the binary cross-entropy between the true label y and the predicted probability in Eq. (4); L_dom is a domain-classifier loss that predicts a device or site label d from the adapted latent z̃; L_ref is a reference-consistency loss; α, β, and γ are positive weighting coefficients; λ is a regularization coefficient; and ‖θ‖² is the squared norm of trainable parameters. The domain loss is implemented with a gradient-reversal layer or an alternating optimization scheme so that the encoder and adaptation network are encouraged to remove domain information while the domain classifier attempts to retain it. The reference-consistency loss penalizes changes in the

predicted disease probability when the reference is replaced by another reference from the same acquisition condition. This term encourages the classifier to use the reference for style adaptation rather than as a label cue.

The objective is a design specification, not a claim that a particular weighting will work in all settings. The relative weights determine the strength of adaptation. A large β increases domain invariance but may remove disease-related variation. A large γ enforces reference invariance but may prevent useful calibration. A large α improves classification but may encourage shortcut learning. These trade-offs should be tuned on a validation set that includes device-held-out and subject-held-out conditions, not on a random split. The regularization term controls the capacity of the adaptation network and the classifier. The reference encoder should be regularized more strongly or trained with a stop-gradient from the classifier to limit disease leakage.

4.5 Uncertainty-aware decision rule

Screening decisions should not be based on a probability alone when the input may be out of distribution. The framework uses predictive uncertainty to gate the decision.

$$screen(X) = 1 \; if \; p(y = 1|X, r, a) \geq \tau \; and \; u(X) \leq \kappa$$

In this rule, screen(X) = 1 indicates that the participant should be referred for further assessment, τ is the screening threshold, u(X) is an uncertainty score, and κ is the maximum uncertainty allowed for an automatic decision. The uncertainty score can be the predictive entropy, the variance across an ensemble, or the mutual information between predictions and model parameters. Monte Carlo dropout provides a practical approximation to Bayesian uncertainty in deep networks [17]. Ensembles are often more reliable but more expensive. Calibration is essential because the threshold τ is meaningful only if the predicted probabilities are calibrated. Modern neural networks can be poorly calibrated, and temperature scaling or other post hoc calibration methods should be applied and evaluated on held-out data [18].

The uncertainty gate serves two purposes. First, it prevents confident predictions on inputs that are far from the training distribution, such as a new microphone or a noisy recording. Second, it supports a triage workflow in which uncertain cases are routed to manual review or repeat recording. The gate does not eliminate domain shift; it makes the system's limitations visible at the decision point. The choice of τ and κ should be based on clinical utility and subgroup error rates, not only on overall accuracy.

5. Analytical Assessment and Failure Modes

5.1 Why reference guidance can reduce first-order domain effects

In the log-spectral domain, a linear time-invariant microphone response and a stationary room response can be approximated as an additive offset across frequency. If the encoder is approximately equivariant to such offsets, then a scale-shift transformation in latent space can absorb a substantial part of the device effect. The reference recording provides an estimate of that offset because it is recorded through the same chain and in the same environment. The low-capacity adaptation in Eq. (3) is well matched to this first-order mechanism. It can correct global gain, spectral tilt, and some subject-specific timbre without changing the temporal structure that carries crackles and wheezes.

This argument also explains the limits of the approach. If the device response is nonlinear, time-varying, or dependent on signal level, a single reference may not capture it. If the

reference is recorded during a different breathing maneuver than the diagnostic segment, the estimated style may be inappropriate. If the reference contains disease-related events, the adaptation may remove the very transients that the classifier needs. These are not implementation details; they are structural properties of reference-guided adaptation. The design should therefore include a reference-quality check, a fallback to unadapted features when the reference is unreliable, and an evaluation that separates the effect of reference quality from the effect of the classifier.

5.2 Separation of disease and style

The central analytical trade-off is between nuisance removal and disease preservation. A transformation that is strong enough to remove all reference-correlated variation will also remove any disease-related variation that happens to correlate with the reference. A transformation that is too weak will leave device and subject effects in the classifier. The low-capacity scale-shift is a compromise: it targets global statistics while leaving higher-order temporal structure intact. The domain-adversarial term further discourages the adapted latent from encoding device identity, but it can also remove disease information if disease prevalence differs across devices. The reference-consistency term helps only when the reference swap is label-preserving; if the reference itself carries label information, consistency can enforce the wrong invariance.

A useful diagnostic is counterfactual reference substitution. For a fixed diagnostic segment, the predicted disease probability should be stable when the reference is replaced by another reference from the same acquisition condition. If the probability changes substantially, the classifier is using the reference as a label cue. Conversely, if the probability is unchanged when the reference is replaced by a reference from a different device, the adaptation may be too weak. These two tests bound the desired behavior. They can be evaluated without claiming clinical efficacy, and they provide a direct check on the separation assumption.

5.3 Generative augmentation and synthetic artifacts

Generative augmentation can improve classifier robustness by exposing it to a wider range of nuisance conditions, but it can also introduce artifacts that the classifier learns to exploit. A generator trained on a small cohort may reproduce speaker identity more faithfully than pathological events. Synthetic samples may have unrealistic noise floors, spectral holes, or temporal discontinuities. If the classifier is trained only on real and synthetic data from the same generator, it may become sensitive to the generator's idiosyncrasies. The framework mitigates this risk in three ways. First, generation occurs in a spectral latent space rather than in the raw waveform, which reduces the model's ability to memorize speaker-specific phase and fine timing. Second, the generative loss includes a multi-resolution spectral term that encourages realistic transient structure. Third, generated samples are used as regularizers rather than as replacements for real data, and their influence is controlled by the classification weight α.

Even with these safeguards, synthetic augmentation cannot create clinical information that is absent from the training data. If the training cohort contains no examples of a particular disease or acquisition condition, the generator may interpolate in ways that are not clinically meaningful. The framework should therefore be evaluated with a clear distinction between interpolation and extrapolation. Augmentation is most defensible when it varies nuisance factors within the support of the training distribution. It is least defensible when it is used to claim coverage of unseen devices, populations, or disease stages.

5.4 Privacy and identity leakage

Respiratory audio can contain speaker-identifying information. A generative model that synthesizes realistic audio may reproduce voice characteristics, breathing patterns, and background speech. Feature-space generation reduces some of this risk because the model does not directly synthesize raw waveforms, but it does not eliminate it. Latent representations can still encode identity, and generated spectrograms can preserve formant structure. The framework should be paired with privacy auditing, including speaker identification tests on generated samples and membership inference tests on the generative model. Differential privacy can provide formal guarantees, but it may degrade sample quality and classification performance. The appropriate privacy mechanism depends on the intended use and the sensitivity of the data. These considerations are part of the design space, not an afterthought.

5.5 Evaluation logic

A rigorous evaluation of this framework would require subject-level splits, device-held-out tests, site-held-out tests, and temporal validation. Random splits are insufficient because the same participant or device can appear in both training and test sets. The evaluation should report calibration, sensitivity at clinically relevant thresholds, subgroup performance, and decision-curve analysis. It should also include ablations that remove the reference pathway, the generative loss, the domain-adversarial term, and the uncertainty gate. These ablations test whether the proposed mechanisms contribute to robustness rather than merely adding complexity. The broader machine-learning literature on COVID-19 diagnosis has emphasized the dangers of biased datasets, inadequate external validation, and overoptimistic performance estimates [19]. Respiratory audio screening is vulnerable to the same pitfalls, and the evaluation protocol should be specified before model development.

The analysis in this section does not report empirical results. It identifies the conditions under which the framework is expected to help, the conditions under which it may fail, and the experiments that would distinguish between the two. The value of the framework lies in making these conditions explicit and testable.

6. Clinical Translation and Boundary Conditions

6.1 Intended use and clinical endpoints

The framework is intended for screening and triage, not for diagnosis. A positive screen should trigger further evaluation, such as spirometry, imaging, or clinical examination. The screening label must be defined by a clinically meaningful endpoint, such as a confirmed diagnosis, an exacerbation, or a need for urgent care. The choice of endpoint affects prevalence, label noise, and the consequences of false positives and false negatives. A model trained to predict a proxy label, such as a positive PCR test, may not generalize to a different endpoint, such as disease severity. The framework should therefore be evaluated separately for each intended use.

Clinical confounding is substantial. Age, sex, smoking history, body mass index, medication use, and comorbidities affect both respiratory sounds and disease prevalence. Some of these variables are available in metadata, but many are not. The classifier should be evaluated within clinically relevant subgroups, and performance disparities should be reported. If a subgroup has a higher false-negative rate, the decision threshold or the uncertainty gate

may need to be adjusted. These adjustments should be justified by clinical utility, not by overall accuracy alone.

6.2 Deployment and monitoring

Deployment introduces conditions that are difficult to simulate in development. Devices change, users record in noisy environments, and disease prevalence varies over time. A reference-guided system can partially adapt to these changes, but it requires a reference recording that is representative of the diagnostic segment. The deployment protocol should specify how the reference is collected, how its quality is checked, and what happens when it is missing or unusable. The system should log reference quality, prediction uncertainty, and decision outcomes so that drift can be detected. Periodic recalibration and re-evaluation on local data are necessary. These are standard requirements for clinical machine learning, but they are especially important for audio because the signal is sensitive to small changes in hardware and environment.

6.3 Boundary conditions

The framework has several boundary conditions. It assumes that a short reference recording is available and that it captures the relevant acquisition and subject style. It assumes that the generative model is trained on data that cover the nuisance conditions encountered at test time. It assumes that the disease signal is not entirely explained by the reference. It assumes that the classifier's uncertainty estimate is meaningful under distribution shift. These assumptions may hold in a controlled research setting and fail in a new population. The framework should therefore be treated as a design hypothesis rather than a validated clinical tool. Its claims are about the internal consistency of the architecture and the plausibility of its mechanisms, not about clinical effectiveness.

7. Conclusion

This manuscript has developed a reference-guided generative feature adaptation framework for respiratory disease screening. The framework addresses a specific problem: respiratory audio classifiers are vulnerable to acquisition and subject-specific style, and naive generative augmentation can amplify nuisance factors rather than pathology. The proposed design uses a short, label-free reference recording to condition a low-capacity scale-shift transformation in a spectral latent space. A conditional generative model reconstructs and samples respiratory spectrograms, while a screening classifier operates on adapted latents and is regularized by reconstruction, classification, domain-adversarial, and reference-consistency objectives. An uncertainty gate supports decisions by withholding automatic screening when the input is unreliable.

The analytical contribution is the explicit treatment of the trade-off between nuisance removal and disease preservation. Reference guidance can reduce first-order domain effects, but strong adaptation can erase clinically meaningful transients. The framework addresses this trade-off through a restricted transformation, a separate reference encoder, gradient control, and counterfactual reference tests. It also clarifies why generative augmentation should be used as a regularizer rather than as a substitute for clinical evidence. The resulting design is technically specified and testable. It does not establish clinical efficacy, but it provides a coherent foundation for subject-level, device-held-out, and site-held-out evaluation of generative modeling and reference-guided adaptation in respiratory audio screening.

References

1. Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., & Mascolo, C. (2020). Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3474–3484). Association for Computing Machinery. https://doi.org/10.1145/3394486.3412865

2. Pahar, M., Klopper, M., Warren, R., & Niesler, T. (2021). COVID-19 cough classification using machine learning and global smartphone recordings. Computers in Biology and Medicine, 135, 104572. https://doi.org/10.1016/j.compbiomed.2021.104572

3. Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Ghosh, P. K., & Ganapathy, S. (2020). Coswara—A database of breathing, cough, and voice sounds for COVID-19 diagnosis. arXiv. https://arxiv.org/abs/2005.10548

4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (Vol. 27, pp. 2672–2680). Curran Associates.

5. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In International Conference on Learning Representations.

6. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851). Curran Associates.

7. Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv. https://arxiv.org/abs/1609.03499

8. Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Advances in Neural Information Processing Systems (Vol. 33, pp. 17022–17033). Curran Associates.

9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

10. Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1, pp. 2058–2065).

11. Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283–286. https://doi.org/10.1056/NEJMc2104626

12. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

13. Kiyasseh, D., Zhu, T., & Clifton, D. A. (2021). CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 5606–5615). PMLR.

14. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

15. Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1, pp. 3942–3951).

16. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401–4410).

17. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050–1059). PMLR.

18. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR.

19. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., Rudd, J. H. F., Sala, E., & Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199–217. https://doi.org/10.1038/s42256-021-00307-1