PulmoScope: A Comparative Study of Hybrid TCN-SNN and Temporal Deep Learning Models for Multi-Class Respiratory Disease Detection from Lung Auscultation Sounds

Genheylou Felisilda, Nicole Menorias, Kobe Marco Olaguir, and Joanna Reyda Santos

1 Background of the Study

1.1 Chronic Respiratory Diseases: Burden, Misdiagnosis, and Local Context

Chronic respiratory diseases (CRDs) remain a major cause of morbidity and mortality worldwide, affecting the lungs and airways and significantly impairing breathing and oxygen exchange. According to the Global Burden of Disease (GBD) Study 2021, CRDs accounted for approximately 4.4 million deaths globally, with chronic obstructive pulmonary disease (COPD) and pneumonia among the leading contributors (Momtazmanesh et al., 2023; Cao et al., 2023). Despite gradual improvements in age-standardized mortality rates, CRDs continue to impose a substantial health and economic burden, particularly in low- and middle-income countries where access to specialized diagnostic tools is limited.

In the Philippines, respiratory diseases represent a persistent public health concern. A population-based study conducted in Nueva Ecija reported that 20.8% of adults aged 40 years and older had COPD, with strong associations to smoking and long-term exposure to biomass fuel (Idolor et al., 2011). Pneumonia likewise remains among the top causes of hospitalization and death, especially among older adults and vulnerable populations (Department of Health, Philippines; World Health Organization, 2023). These diseases often present with overlapping clinical symptoms such as chronic cough, dyspnea, wheezing, fever, and chest discomfort, making accurate differentiation challenging in routine clinical settings.

A critical issue in respiratory care is misdiagnosis and delayed diagnosis. COPD is frequently misdiagnosed as asthma, bronchitis, or recurrent pneumonia, particularly in primary care settings where spirometry is underutilized (Ho, T. et al. 2019; Lusuardi et al., 2005). Similarly, pneumonia may be under- or over-diagnosed due to reliance on auscultation findings that are subtle, transient, or masked by background noise. Studies have shown that lung auscultation alone has low sensitivity (≈37%) for detecting acute pulmonary pathologies, even when performed by experienced clinicians (Arts et al., 2020).

These diagnostic challenges are exacerbated in the Philippine healthcare context, where limited access to pulmonary specialists, advanced imaging, and diagnostic equipment often necessitates reliance on basic tools such as the stethoscope. As a result, early signs of COPD exacerbations or pneumonia may go undetected, leading to delayed treatment, increased hospitalizations, and higher out-of-pocket costs for patients (Ang & Fernandez, 2024). This highlights the need for objective, assistive diagnostic tools that can support clinicians in identifying disease-related lung sound patterns more consistently and accurately.

1.2 Conventional Diagnostic Approaches and Real-World Challenges

In contemporary clinical practice, physicians still rely heavily on traditional physical examination techniques—such as auscultation, percussion, palpation, and vocal resonance—as primary and accessible tools for assessing lung function. Despite their widespread use, these methods have important limitations that can reduce diagnostic accuracy, even when performed by experienced clinicians.

A major limitation of auscultation is its low sensitivity. A meta-analysis of 34 studies involving adult patients with acute pulmonary conditions found that lung auscultation had a pooled sensitivity of only 37% and a specificity of 89% (Arts et al., 2020). This indicates that auscultation may fail to detect a substantial number of true respiratory pathologies, limiting its reliability as a stand-alone diagnostic tool.

Auscultation is also less accurate in mechanically ventilated patients. In a study of 200 post–cardiac surgery patients, two independent examiners (blinded to mechanical measurements) performed chest auscultation. They correctly identified decreased or absent breath sounds or crackles in only 34 % of cases for examiner A and 42 % for examiner B. Sensitivities were 25.1% and 36.4%, respectively, while specificities were moderately higher at 68.3% and 63.4% (Xavier, Melo‑Silva, Santos, & Amado, 2019). These findings demonstrate that auscultation may not reliably reflect underlying lung function in such patients.

Interobserver variability further limits reliability. In a longitudinal study of patients with fibrotic interstitial lung disease, nine respiratory physicians independently assessed crackles at baseline and 12 months. Agreement on the presence of crackles yielded a Fleiss' κ of 0.57 (95% CI: 0.55–0.58), and agreement on changes in crackle intensity over time was lower (κ = 0.42, 95% CI: 0.41–0.43) (Sgalla et al., 2024). Although individual physicians were more consistent over time (intra-rater κ = 0.79–0.87), the moderate agreement between different physicians highlights persistent subjectivity in interpretation.

Terminology inconsistencies also contribute to diagnostic challenges. A survey of staff physicians, residents, and medical students found that only approximately 63% of staff physicians and 69% of residents correctly identified crackles, while many used incorrect terms. The study concluded that insufficient auscultation skill, rather than personal preference, was a major factor (Vasquez & Ruiz, 2020). Lack of standardized terminology can lead to miscommunication and misinterpretation among clinicians.

Other physical examination signs also have limitations. A review of patients presenting with dyspnea found that features such as asymmetric chest expansion, diminished breath sounds, egophony, bronchophony, and tactile fremitus may assist in diagnosing pneumonia or pleural effusion. However, for early-stage chronic obstructive pulmonary disease (COPD), no single physical sign demonstrated high accuracy (Shellenberger et. al, 2017). Many signs are particularly insensitive in early or mild disease.

Spirometry remains an essential tool in the traditional assessment of lung function. It provides objective measurements of airflow, including forced expiratory volume in one second (FEV_1), forced vital capacity (FVC), and the FEV_1/FVC ratio, which are critical for diagnosing and staging obstructive lung diseases such as asthma and COPD (Singh et. al, 2025; Agusti et. al, 2023). While spirometry provides reproducible and quantitative data that physical examination alone cannot deliver, its accuracy depends on proper technique and patient cooperation. Additionally, it may be difficult to perform in acute or critically ill patients and in resource-limited settings where equipment or trained personnel are unavailable.

Overall, traditional respiratory assessment methods face several limitations—such as the low sensitivity of auscultation, high subjectivity, variable clinician interpretation, and the reduced diagnostic value of physical signs in early or subtle disease. These constraints make it difficult to reliably detect faint or transient abnormalities, particularly early-stage crackles and wheezes that may signal evolving pulmonary pathology. As a result, there is increasing motivation to explore automated analysis systems capable of providing more objective, sensitive, and reproducible respiratory sound interpretation.

1.3 Existing Studies on Disease-Centered Respiratory Sound Analysis and Research Gaps

Recent advances in deep learning have enabled automated analysis of respiratory sounds with the goal of supporting disease detection and classification. Unlike early work that focused solely on identifying acoustic events such as crackles and wheezes, more recent studies have increasingly framed respiratory sound analysis around specific diseases, including COPD, pneumonia, asthma, and other obstructive or infectious lung conditions.

Several studies have demonstrated that deep-learning models can identify disease-related patterns from lung sound recordings. Srivastava et al. (2025), for example, developed a deep-learning framework for detecting COPD using respiratory sounds, reporting improved accuracy compared with traditional feature-based approaches. Kim et al. (2025) showed that deep-learning models trained on digital stethoscope recordings could outperform medical trainees in differentiating pathological respiratory conditions, emphasizing the potential of AI-assisted auscultation as a clinical decision-support tool.

Disease-oriented respiratory sound datasets have also emerged to support this line of research. The ICBHI 2017 Respiratory Sound Database remains the most widely used benchmark, containing 6,898 respiratory cycles annotated with clinical information linked to conditions such as COPD and pneumonia (ICBHI Challenge, 2017). Similarly, Huang et al. (2023) introduced a respiratory sound database collected using an intelligent stethoscope, explicitly designed to support disease-related analysis. However, both datasets suffer from significant class imbalance, with normal recordings dominating and disease-related samples underrepresented. This imbalance leads to biased models with poor sensitivity for clinically important conditions, particularly in COPD exacerbations and pneumonia-related cases.

From a methodological perspective, most disease-centered studies rely on handcrafted or time–frequency features, such as MFCCs or mel-spectrograms, combined with convolutional or recurrent neural networks (Rocha et al., 2020; Wang et al., 2024). While these approaches achieve promising overall accuracy, several limitations persist. First, many studies evaluate a single disease (e.g., COPD vs. non-COPD), limiting applicability in real-world settings where multiple respiratory diseases coexist. Second, models are often trained and evaluated under controlled conditions, with limited assessment of performance variability across patients, recording environments, and device characteristics.

Furthermore, misclassification between diseases with overlapping acoustic signatures—such as COPD and pneumonia—is rarely analyzed in depth. Many studies emphasize aggregate accuracy metrics without examining class-wise errors, which are crucial for understanding potential clinical risks. As a result, despite technological progress, existing work has not fully addressed the challenges of disease-level differentiation, generalizability, and interpretability in real-world respiratory care.

Table 1.1. Related Works and Research Gaps in Deep-Learning Respiratory Sound Classification

RESEARCH GAPS	DESCRIPTION
Class imbalance and limited generalizability	Public respiratory sound datasets (e.g., ICBHI 2017; Huang et al., 2023) are dominated by normal recordings, while disease-related samples for COPD and pneumonia are underrepresented. This leads to biased models with low sensitivity to clinically significant diseases.
Limited disease-level differentiation	Many studies focus on detecting acoustic events (e.g., crackles or wheezes) rather than directly classifying respiratory diseases. Overlapping sound characteristics between COPD, pneumonia, and other conditions remain insufficiently addressed.
Limited architectural comparison for disease classification	Most works evaluate only one model architecture, restricting insights into how different deep-learning models perform in disease-centered respiratory sound classification tasks.
Insufficient analysis of disease misclassification	Many studies emphasize overall accuracy without examining class-wise errors or confusion between diseases, limiting their clinical interpretability and usefulness.

Sources: Abduh et al. (2018); Perna et al. (2018); Kochetov et al. (2018); Fernando et al. (2021); Fernando et al. (2022); Kim et al. (2025); Rocha et al. (2020); Wang et al. (2024);

Huang et al. (2023); ICBHI Challenge (2017); Srivastava et al. (2025); Tzeng et al. (2025); Yu et al. (2025); Tsai et al. (2023).

2 Objectives

This study aims to develop and evaluate a deep-learning–based respiratory sound analysis system to support disease-centered classification of lung conditions, with particular focus on COPD and pneumonia, under realistic data constraints.

1. **To develop and compare different deep-learning models** for classifying respiratory diseases from lung sound recordings, using standardized preprocessing, training, and evaluation procedures.

2. **To address class imbalance** by applying balanced dataset strategies and assessing their impact on disease-level classification performance through class-wise accuracy, recall, and misclassification analysis.

3. **To evaluate model performance using a held-out test set**, and examine the potential of deep learning as a clinical decision-support tool for reducing misdiagnosis and delayed detection of respiratory diseases, particularly in resource-limited settings.

3 Methodology

This chapter provides an overview of the entire methodology used in the study as shown in Figure 2.1, including the data processing workflow, feature extraction procedures, and analytical techniques applied to ensure reliable and accurate results.

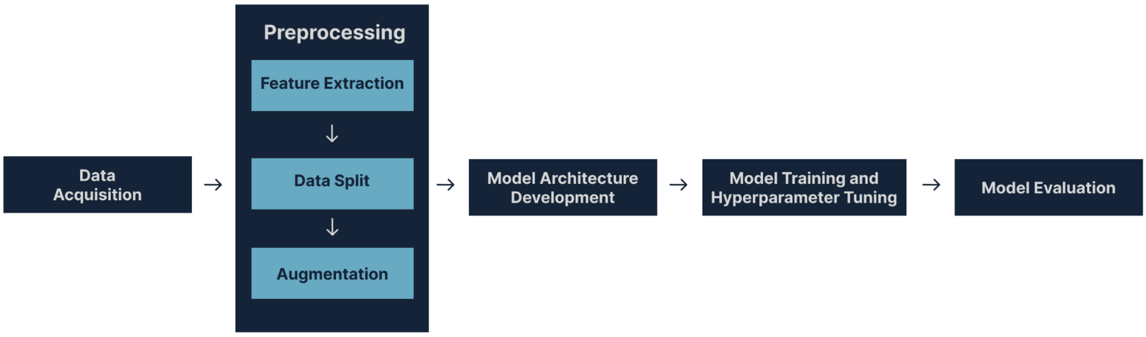


Figure 2.1. Pipeline Overview

3.1 Data Collection

The study uses two complementary respiratory sound datasets: the Respiratory Sound Database from the ICBHI 2017 Challenge.

ICBHI Respiratory Sound Database. This is the primary dataset used, originally released for the 2017 Biomedical Health Informatics Challenge organized by Aristotle University of Thessaloniki and collaborating research teams in Portugal and Greece (Rocha et al., 2020). It contains 920 annotated respiratory sound recordings acquired from 126 patients, with clip durations ranging from 10 to 90 seconds. Each recording is accompanied by annotation files that indicate the presence of adventitious sounds such as wheezes, crackles, or both, alongside demographic and diagnostic information for each subject. This large, heterogeneous collection of real-world respiratory sounds is well suited for developing and evaluating machine learning models for respiratory disease detection.

3.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) phase was performed to understand the clinical–demographic and audio-derived variables, detect inconsistencies, and guide modeling decisions, consistent with recommendations that exploratory analysis is a prerequisite for robust predictive modeling in healthcare and other biomedical domains. In line with Kuhn and Johnson’s framework for applied predictive modeling, EDA here serves to inform later steps such as feature selection, preprocessing, and model tuning rather than being treated as a purely descriptive add-on.

Data Overview and Structure Examination. The analysis began by inspecting the structure of the clinical–demographic dataset and the audio annotation tables, including dimensionality, feature types, and the distribution of diagnostic labels. Summary statistics (minimum, maximum, mean, standard deviation, and number of unique values) were computed for key variables such as age, sex, anthropometric measures, and clinical diagnoses, as well as for the derived audio-based labels. This step allowed verification of schema correctness, identification of missing or inconsistent entries, and an initial view of population characteristics.

Disease Class Mapping. To standardize the diagnostic output on the clinical side, physician-assigned diagnoses were mapped into four target disease classes. This harmonization ensured consistency between multiple diagnostic sources and focused the analysis on major pathologies while grouping less prevalent or co-morbid conditions into a composite category. The final mapping was defined as follows:

- COPD: Includes all confirmed Chronic Obstructive Pulmonary Disease cases.
- Healthy: Includes subjects with no documented history of respiratory pathology.
- Pneumonia: Includes all confirmed pneumonia cases.
- Others: A composite class aggregating Asthma, Bronchiectasis, Bronchiolitis, Upper Respiratory Tract Infections (URTI), and Lower Respiratory Tract Infections (LRTI).

These standardized clinical labels were later aligned with the audio-based labels to enable consistent downstream analysis and model training.

Statistical Exploration of Relationships. To explore relationships between demographic variables and audio-derived respiratory patterns, inferential statistics were applied. Age, treated as a continuous predictor, was compared across the four audio-derived lung sound categories (normal, wheeze only, crackle only, both wheezes and crackles) using one-way ANOVA, with the correlation ratio η used as an effect-size measure to complement the F-test and support interpretation of practical significance, consistent with best-practice recommendations for ANOVA reporting in biomedical research (Serdar et al., 2020; Cohen, 2013). Gender, treated as a categorical predictor, was related to respiratory sound patterns and to clinician-assigned diagnosis groups using contingency tables, chi-square tests of independence, and Cramér's V as a scale-free effect-size index, aligning with guidance that significance testing in categorical data analysis should be accompanied by effect-size estimation (McHugh, 2013). The same ANOVA–η framework was planned for age versus clinical diagnosis to assess whether age partitions variance in physician-assigned labels more strongly than in audio-based patterns, in line with epidemiological evidence that respiratory disease distributions (e.g., COPD, asthma, acute infections) show marked age structuring, and the same chi-square–Cramér's V framework was specified for gender versus clinical diagnosis to quantify sex-related differences in disease categories (Ntritsos et al., 2018; Zein & Erzurum, 2015).

Duration and Signal Integrity Analysis. To characterize temporal and spectral consistency across disease groups, audio-level summary features were computed for each recording including mean sample rate, mean, minimum, and maximum duration, mean spectral centroid, spectral bandwidth, spectral rolloff, and spectral flux, and examining their distributions across clinical diagnosis classes (COPD, Healthy, Pneumonia, Other). In addition, box plots were generated per diagnosis to visualize distributional differences and potential outliers in these spectral descriptors, while representative waveforms and their RMS envelopes were plotted for sample recordings from each disease class to qualitatively inspect signal morphology and verify that characteristic respiratory patterns were captured rather than artifacts or noise; the RMS envelope is a robust feature for quantifying energy fluctuations characteristic of abnormal respiratory sounds, as supported by recent work in machine learning–augmented lung sound recognition (Sabry et al., 2024).

These metrics were used to flag abnormally short or long recordings and spectrally implausible signals that might indicate artifacts, environmental noise, or acquisition errors rather than true respiratory activity (Tzeng et al., 2025).

3.3 Data Preprocessing

A rigorous preprocessing pipeline was implemented to transform all respiratory recordings into uniform, diagnostically meaningful inputs suitable for temporal deep-learning models. Given that lung sound recordings inherently vary in duration, amplitude, spectral distribution, and susceptibility to environmental noise, establishing a standardized preprocessing framework is essential for achieving reliable model behavior across heterogeneous clinical conditions. Recent reviews of computerized respiratory sound analysis underscore that preprocessing is not merely a preparatory step, but a decisive factor shaping downstream classification accuracy, particularly because auscultation recordings are prone to interference from stethoscope friction, background noise, and inconsistent sensor placement (Ethala et al., 2025). Furthermore, contemporary evaluations of respiratory sound analytics—including studies grounded in the ICBHI 2017 dataset—demonstrate that effective preprocessing routines involving filtering, segmentation, resampling, and normalization significantly strengthen model robustness and improve generalization when confronted with real-world clinical noise profiles (Yu et al., 2025).

Guided by these insights, the present study employed a preprocessing strategy designed to preserve clinically relevant acoustic cues—such as COPD, pneumonia and transient bursts—while suppressing artifacts that could distort temporal–spectral representations. The pipeline standardized sampling characteristics, minimized extraneous noise, enforced consistent segment lengths, and produced a stable hybrid time–frequency representation through the combined use of MFCCs and Mel-spectrograms. By applying resampling, bandpass filtering, strict segmentation, loop-padding, and Z-score normalization, the dataset attained a uniform structure conducive to fair architectural comparison and robust temporal modeling throughout the experimental phases.

3.3.1 Audio Standardization

Resampling. All recordings were resampled to **16 kHz**, a sampling rate widely adopted in contemporary respiratory sound research. This rate preserves the full diagnostic bandwidth of adventitious pulmonary acoustics—most of which occur below 2.5 kHz—while controlling computational overhead in deep-learning pipelines. Recent reviews emphasize that consistent sampling rates are critical for stabilizing downstream feature extraction and ensuring cross-dataset comparability (Ethala et al., 2025; Yu et al., 2025). These findings align with long-standing evidence that uncontrolled sampling variability introduces spectral distortions that degrade classifier performance in respiratory sound analysis (Pramono et al., 2017).

Fixed-Duration Segmentation (5 Seconds). Each audio file was standardized into 5-second segments, approximating a full respiratory cycle and improving temporal alignment across samples. Studies in computerized auscultation show that fixed window durations reduce temporal drift and provide more consistent cues for models learning pathological events (Yu et al., 2025). Prior work also demonstrates that consistent temporal framing enhances model convergence and interpretability, particularly for architectures processing recurrent cycles of inhalation and exhalation (Li et al., 2016; Ethala et al., 2025). This segmentation strategy therefore ensures that each model receives structurally comparable temporal information.

Loop-Padding Strategy (Signal Tiling). Zero-padding can inject artificial silence, potentially misleading temporal models that rely on energy patterns for distinguishing crackles, wheezes, and breath phases. To avoid this, a loop-padding (tiling) method was applied: the waveform is repeated until it slightly exceeds five seconds and then trimmed precisely. Recent signal-processing and sound-event studies show that tiling preserves physiological continuity and avoids low-energy artifacts that distort spectrotemporal representations (Mulimani et al., 2024; Cheng et al., 2023). This approach aligns with best practices highlighted in respiratory sound reviews, where artifact-free padding is recommended to avoid misleading classifiers that treat silence as meaningful pathological cues (Ethala et al., 2025).

Bandpass Filtering. A 5th-order Butterworth bandpass filter (50–2500 Hz) was applied to isolate diagnostically relevant lung acoustics. The lower cutoff attenuates heart sounds, movement noise, and low-frequency mechanical interference—limitations consistently documented in respiratory sound literature (Pramono et al., 2017; Pramono et al., 2019). The upper cutoff removes high-frequency microphone hiss and environmental artifacts, both of which are common in real-world auscultation data. Filtering within this range is strongly supported by modern deep-learning respiratory analyses that emphasize cleaning high-noise bands to improve feature salience (Ethala et al., 2025; Yu et al., 2025).

Waveform Peak Normalization. Peak normalization was applied by dividing each waveform by its maximum absolute amplitude to ensure dynamic range consistency across recordings obtained from varied microphones and stethoscope types. The respiratory sound literature consistently warns that device-specific amplitude variability can fragment data distributions and reduce model generalization (Yu et al., 2025). Biomedical audio studies further recommend amplitude normalization to minimize inter-device variance and stabilize time–frequency projections (Tzeng et al., 2025; Ethala et al., 2025).

This preprocessing pipeline ensures that every input tensor is saturated with physiologically meaningful acoustic information while actively suppressing “dead-air artifacts”—segments of silence or low-energy noise known to degrade spectrotemporal learning and reduce classifier reliability (Cheng et al., 2023). The resulting uniform structure directly supports the demands of your temporal deep-learning architectures (TCN, TCN-SNN, LSTM) and is fully aligned with the best practices identified in recent reviews of computerized respiratory sound analysis.

3.3.2 Feature Extraction

A **Hybrid Feature Stacking** strategy was employed to merge spectral texture and cepstral dynamics into a unified 2-D representation optimized for temporal deep-learning architectures. This design follows recent methodological progress in computerized respiratory sound analysis, where combining Mel-spectrograms with cepstral features has been shown to enhance robustness, particularly in noisy clinical environments (Fernando et al., 2022; Nguyen & Pernkopf, 2022). Current reviews further emphasize that hybrid feature pipelines improve diagnostic reliability by capturing both high-resolution spectral structure and low-dimensional timbral cues linked to pathological events such as wheezes, crackles, and coarse breathing patterns (Ethala et al., 2025; Yu et al., 2025).

The raw 1-D respiratory waveforms were transformed into a stacked 2-D tensor comprising two complementary representations:

High-Resolution Mel Spectrogram. A Mel-spectrogram was computed using 224 Mel bands, an FFT size of 2048, and a hop length of 512. High-resolution Mel projections provide detailed time–frequency distributions of acoustic energy, enabling the model to detect subtle respiratory abnormalities that manifest as transient spectral fluctuations. Reviews of lung-sound analysis note that high-band Mel representations substantially improve sensitivity to adventitious signals embedded within normal breath cycles (Pramono et al., 2017; Tzeng et al., 2025). This is especially important for conditions such as COPD and pneumonia, where harmonic patterns often overlap with background noise.

MFCCs (Mel-Frequency Cepstral Coefficients). A set of 40 MFCCs was extracted to capture cepstral dynamics reflecting timbre, envelope, and airflow characteristics. MFCCs provide a compact representation of the spectral envelope and have been consistently validated in respiratory sound studies for distinguishing between normal and pathological textures, such as differentiating crackles from turbulence-induced broadband noise (Pramono et al., 2019; Ethala et al., 2025). Contemporary research also highlights the importance of MFCCs when generalizing across recording devices and environmental conditions, as their cepstral compression reduces sensitivity to microphone distortions (Yu et al., 2025).

Vertical Feature Stacking and Standardization. The 224-band Mel spectrogram and 40-dimension MFCC matrix were vertically stacked to form a final input tensor of 264 channels, producing a feature representation that jointly encodes spectral locality and cepstral smoothness. This approach aligns with evidence that stacked or multimodal time–frequency features improve classifier robustness in real-world auscultation scenarios, particularly for architectures such as TCNs, SNNs, and CNN-RNN hybrids (Cheng et al., 2023; Mulimani et al., 2024).

Following stacking, Z-score standardization (subtracting the mean and dividing by the standard deviation) was applied to the log-scaled features. Z-scaling was selected because it stabilizes gradient flow and limits the influence of high-energy outliers, allowing temporal models to learn consistent patterns across heterogeneous recordings. Foundational work in deep learning and audio normalization shows that Z-score scaling produces more stable optimization trajectories and mitigates gradient explosions compared to simple min-max scaling (Zhang et al., 2023). This is particularly beneficial for deep temporal architectures, which depend on stable activations when processing long-range acoustic sequences.

Overall, this hybrid feature extraction pipeline ensures that the input tensors provide rich, physiologically grounded representations, enabling the model to capture both fine-grained spectral cues and broader cepstral patterns linked to respiratory pathology.

3.4 Data Splitting

The dataset was divided using a stratified split with 60% for train, 20% for validation and 20% for test to retain proportional class distributions across the training, validation, and test sets, a practice shown to reduce sampling bias and yield more reliable performance estimates in imbalanced classification tasks (scikit-learn, 2024).

3.5 Data Balancing Strategy

Class imbalance is a persistent challenge in respiratory-sound classification, where certain disease categories—such as COPD or chronic obstruction patterns—tend to be overrepresented relative to pneumonia, upper-airway infections, or healthy recordings. This imbalance has been consistently documented across major respiratory datasets and remains one of the primary obstacles to developing clinically reliable models (Ethala et al., 2025; Yu et al., 2025). When left unaddressed, imbalance skews model learning toward majority classes, reducing sensitivity to low-frequency but clinically significant conditions and creating misleading performance metrics. Past systematic analyses of adventitious lung sounds also highlight that minority classes frequently exhibit more acoustic variability, making them disproportionately difficult for models to learn without corrective sampling strategies (Pramono et al., 2017; Pramono et al., 2019).

To mitigate these risks, the present study adopted a **Downsample-then-Augment** strategy. Majority classes were first reduced to limit dominance during gradient updates, preventing the model from exploiting trivial decision boundaries biased toward more abundant disease presentations. After downsampling, targeted augmentation was applied to minority classes to enrich their acoustic diversity without introducing artificial distortions. This approach aligns with best-practice recommendations in recent respiratory-sound literature, which emphasize the importance of balancing sample distributions while preserving the natural acoustic characteristics essential for accurate disease recognition (Ethala et al., 2025; Yu et al., 2025). By correcting imbalance at both the sample-count and feature-diversity levels, the strategy supports a more equitable learning environment across all disease categories and enhances the model's robustness in real-world clinical settings.

3.5.1 Target Class Definition

Given the clinical objective of emphasizing the recognition of high-risk pulmonary infections and chronic obstructive disorders, the class taxonomy was strategically restructured to prioritize the two most clinically consequential categories: Pneumonia and COPD. Contemporary reviews of computerized respiratory sound analysis report that datasets frequently contain numerous low-frequency subclasses—such as asthma, bronchiectasis, or other lower respiratory infections—which introduce substantial sparsity and degrade classification reliability when treated as separate categories (Ethala et al., 2025; Yu et al., 2025). These rare classes often have limited sample availability and exhibit heterogeneous acoustic patterns, making them difficult for deep learning models to learn without overfitting.

To address this imbalance and reduce fragmentation of diagnostically similar but infrequent conditions, all non-target pathologies were consolidated into a unified **“Others”** category. This restructuring aligns with recommendations from respiratory-sound literature, which emphasize that collapsing sparse subclasses enhances model stability and directs representational capacity toward clinically important disease distinctions. By reframing the problem as a focused diagnostic task centered on Pneumonia and COPD, the model is better positioned to achieve reliable performance in scenarios where early detection of these high-risk conditions is critical.

3.5.2 Balancing via Downsample-then-Augment

Following the class consolidation, substantial imbalance persisted, with COPD remaining the dominant class relative to Pneumonia and Others. Such imbalance is a known issue in medical acoustics datasets, where disease prevalence and recording availability rarely distribute uniformly across classes. Recent analyses of medical classification systems emphasize that imbalance can strongly bias model decision boundaries, leading to inflated accuracy for majority classes and suppressed sensitivity for clinically important minority classes (Welvaars et al., 2023; Agyemang et al., 2025). To address this challenge, the study adopted a **Downsample-then-Augment** strategy, which has been shown to produce more stable and equitable model behavior in health-data classification tasks.

Downsampling. The minority-class count ($N_{□i□}$) was used as the reference, and all majority classes were randomly downsampled to achieve balanced representation during training. Downsampling is increasingly recognized as an effective and principled approach in medical datasets, especially where class skew is severe. Comparative evaluations show that downsampling can surpass class-weighted loss or cost-sensitive learning by directly eliminating majority-class dominance and reducing biased gradient updates (Welvaars et al., 2023). This effect is even more pronounced in acoustic and biosignal domains, where imbalanced frequency distributions can cause temporal models to over-specialize on majority-class spectrotemporal patterns. By equalizing class sample counts before augmentation, the training set promotes fair gradient contributions from all disease categories.

Augmentation via Frequency Masking. After downsampling, controlled augmentation was applied to restore training volume while maintaining class parity. Frequency Masking, originally introduced in SpecAugment, was used due to its proven capacity to disrupt local spectral dependencies and force models to learn more distributed, robust representations (Park et al., 2019). This is particularly valuable for respiratory sound analysis, where clinically important cues—such as crackles or wheezes—may be partially masked by noise, microphone placement issues, or patient movement.

Recent research in biomedical acoustics demonstrates that frequency masking enhances generalization by making neural networks more resilient to spectral distortions, a common feature of real-world auscultation conditions (Agyemang et al., 2025). The method is also well suited to spectrogram-based architectures, including TCNs and hybrid TCN-SNN models, which rely on stable high-level spectral patterns for temporal reasoning. By generating diverse yet physiologically consistent spectrogram variations, frequency masking reduces overfitting to narrow spectral features and enhances the robustness of the learned representations.

The final training dataset thus consisted of a balanced mixture of original downsampled samples and frequency-masked augmentations, producing a more evenly distributed and acoustically diverse training environment tailored for disease-level classification.

3.6 Model Architecture

The experiment was structured as a Model Family Comparison, enabling a controlled evaluation of how different temporal modeling paradigms interpret respiratory acoustic patterns under identical preprocessing and training conditions. All architectures were implemented with input of 264 mel-frames and a standardized hidden dimension of 192 to ensure a fair comparison and to isolate the effect of temporal modeling strategies rather than parameter count.

TCN-SNN (Hybrid). The TCN-SNN architecture represents a hybrid neuromorphic approach, integrating the sequential modeling strengths of Temporal Convolutional Networks (TCNs) with the event-driven computational characteristics of Spiking Neural Networks (SNNs). The feature extraction backbone is built upon a Multi-Scale TCN designed to capture long-range temporal structure through dilated convolutions—a property shown to outperform traditional recurrent models in sequence modeling (Lea et al., 2016). Specifically, this backbone comprises three stacked blocks employing parallel convolutions with kernel sizes of 3, 5, and 7 to simultaneously capture acoustic patterns across varying temporal scales. To model extended dependencies without downsampling, these blocks utilize exponentially increasing dilation rates (d=1, 2, 4) alongside residual connections. The extracted representations are subsequently processed by a Parametric Leaky Integrate-and-Fire (LIF) node using an arctangent surrogate gradient.

This interface introduces biologically inspired sparse activation patterns (Panda & Roy, 2020), encoding continuous data into discrete spike trains that effectively filter background noise by requiring signal accumulation to cross a firing threshold. Finally, a non-linear attention mechanism aggregates these temporal spikes into a unified feature vector for classification (see Table 2.1), investigating whether combining dense temporal extraction with spike-based decision dynamics offers advantages for interpreting subtle respiratory events.

Table 2.1. TCN-SNN Architecture

Component	Parameters
TCN Layers	3

Kernel	3, 5, 7
Dilation	1, 2, 4
Padding	[1, 2, 3] → [2, 4, 6] → [4, 8, 12]
Channels	128 → 192 → 192
Residual	Yes
Activation	ReLU
Regularization	Dropout (0.2) + BatchNorm(1d) (applied between layers)
SNN Classifier	LIF (Leaky Integrate-and-Fire)
Layer	Linear (192 → 1)
Activation	Tanh → Softmax
Dense	192 → 128 → # of Classes

Pure TCN. The Pure Temporal Convolutional Network (TCN) serves as a conventional deep-learning baseline, explicitly designed to isolate the contribution of the neuromorphic spiking component found in the hybrid model. Structurally, it retains the identical three-layer Multi-Scale TCN configuration as the TCN-SNN, ensuring that feature extraction capabilities driven by parallel kernels (k=3, 5, 7) and dilated convolutions remain constant. This design provides a stable and wide receptive field capable of modeling multi-scale respiratory features while maintaining high temporal resolution—a key advantage of TCNs in acoustic classification tasks (Kang et al., 2024). However, instead of temporal integration via spiking dynamics, this architecture employs a deterministic standard attention mechanism (using Tanh activation followed by Softmax).

This learnable layer assigns importance weights to specific time steps, allowing the model to selectively emphasize diagnostic audio segments, such as transient crackles or wheezes, while suppressing irrelevant periods of silence or ambient noise prior to the final classification stage (see Table 2.2).

Table 2.2. Pure TCN Architecture

Component	Parameters
TCN Layers	3
Kernel	3, 5, 7
Dilation	1 → 2 → 4
Padding	[1, 2, 3] → [2, 4, 6] → [4, 8, 12]
Channels	128 → 192 → 192
Residual	Yes
Activation	ReLU
Regularization	Dropout (0.2) + BatchNorm(1d) (applied between layers)
Attention	Standard Attention
Layer	Linear (192 → 1)
Activation	Tanh → Softmax
Dense	192 → 128 → # of Classes

LSTM. The LSTM network functions as the principal sequential baseline, enabling a direct comparison between classical recurrent memory architectures and the more contemporary convolutional and neuromorphic spiking approaches. LSTMs remain widely utilized in biomedical audio analysis due to their capacity to capture extended temporal dependencies inherent in respiratory cycles, such as evolving airflow patterns, prolonged wheezes, or multi-phase abnormalities distributed across inspiration and expiration. Recent respiratory sound classification research continues to highlight their advantages, demonstrating that LSTM-based models effectively track dynamic acoustic transitions and temporal variability in pathological breath events (Salor-Burdalo & Gallardo-Antolín, 2022).

The architecture implemented in this study consists of **three stacked LSTM layers**, each with a hidden dimension of 192. The Mel-spectrogram is processed sequentially across its full five-second duration, allowing the network to integrate fluctuations in spectral energy, transient adventitious events, and cycle-dependent acoustic signatures. After temporal processing, the hidden state at the final time step (**h**□) is extracted as a condensed representation of the entire respiratory window. This summary vector serves as the input to the downstream classifier, ensuring that the model's final prediction is informed by a temporally holistic view of the input signal, as detailed in Table 2.3.

Table 2.3. LSTM Architecture

Component	Parameters
Recurrent Layers	3 stacked LSTMs
Hidden Dim	192
Regularization	Dropout (0.2) (applied between layers)
Activation	Tanh (hidden) Sigmoid (gates)
Dense	192 → 128 → # of Classes

Vanilla RNN. The Vanilla RNN, implemented as a standard Elman-type recurrent architecture, serves as the minimal sequential baseline for this study. Structured with three stacked recurrent layers of 192 hidden units, it updates its hidden state at each time step solely through simple recurrent transformations based on the current input and the previous state. This absence of gating mechanisms—central to LSTMs—or the dilated hierarchical structure characteristic of TCNs allows the Vanilla RNN to function as a critical control model, isolating whether advanced temporal modeling is required to capture the complex, non-stationary dynamics of respiratory acoustics.

Although no gating or memory stabilization mechanisms are employed, the model still maps learned representations into the same high-dimensional feature space as more advanced architectures, ensuring that performance differences reflect the inherent limitations of simple recurrent dynamics rather than differences in capacity or embedding dimensionality (see Table 2.4). This baseline approach is consistent with prior respiratory-sound literature that evaluates classical recurrent architectures as comparison points for modern deep-learning systems. For instance, Kang et al. (2024) report benchmarking traditional RNN-family models against more sophisticated architectures in respiratory sound classification, highlighting their challenges in modeling long-duration respiratory patterns. Similarly, studies such as Salor-Burdalo & Gallardo-Antolín (2022) use recurrent variants as foundational baselines when assessing attention-equipped or convolution-enhanced models. These works support the role of the Vanilla RNN as an essential reference point for quantifying the benefits of advanced temporal modeling frameworks used in contemporary auscultation analysis.

Table 2.4. Vanilla RNN Architecture

Component	Parameters

Recurrent Layers	3 stacked LSTMs
Hidden Dim	192
Regularization	Dropout (0.2) (applied between layers)
Activation	Tanh
Dense	192 → 128 → # of Classes

3.6.1 Shared Deep Classifier

To ensure a rigorous "apples-to-apples" comparison across all architectures, a unified Deep Classifier Head is employed for final prediction. This module consists of a Multi-Layer Perceptron (MLP) with two dense layers (192 → 128 and 128 → 64), each followed by Batch Normalization, GELU activation, and Dropout (p=0.2) for regularization. The final linear output layer maps the refined features to the four target disease classes (COPD, Healthy, Pneumonia, Other). By keeping this classification logic identical across all experiments, any observed performance variances can be attributed solely to the efficacy of the differing feature extraction backbones (TCN vs. RNN).

3.7 Experimental Framework

A structured three-phase experimental framework was implemented to ensure a systematic, transparent, and empirically sound comparison of all model architectures. This multi-stage design allowed the study to first identify the most promising temporal modeling paradigm, then optimize it for efficiency and accuracy, and finally evaluate it under full training conditions to determine real-world deployability.

3.7.1 Phase 1: Training for Comparative Architecture Search

The dataset was divided using a stratified split with 60% for train, 20% for validation and 20% for test to retain proportional class distributions across the training, validation, and test sets, a practice shown to reduce sampling bias and yield more reliable performance estimates in imbalanced classification tasks (scikit-learn, 2024).

To ensure optimization stability, gradient clipping with a threshold of 1.0 was applied, as recent theoretical and empirical evidence demonstrates that clipping mitigates exploding gradients, improves convergence behavior, and enhances robustness against stochastic gradient noise (Zhang et al., 2019).

A Cosine Annealing Warm Restarts scheduler with an initial restart period of T0=10 was utilized to periodically re-expand the learning rate, encouraging exploration of the loss landscape and preventing premature convergence to sharp local minima (Loshchilov & Hutter, 2016).

To further safeguard against overfitting, an EarlyStopping mechanism monitored validation loss with a patience of 10 epochs, a widely adopted criterion that balances training efficiency with the need to avoid premature termination while supporting more stable generalization (Terry, 2021; Lightning AI, 2024).

All four architectures were trained for 30 epochs using a unified set of Default Hyperparameters. Applying a consistent configuration across all models ensures that observed performance differences arise from architectural design rather than uneven optimization. This approach aligns with best practices for fair neural network benchmarking, see table 2.5.

Table 2.5. Default Training Hyperparameters

Training Hyperparameters	
Learning Rate	1e-3
Batch Size	16
Optimizer	AdamW
Weight Decay	1e-4
Label Smoothing	0.0
Gradient Clipping	1.0

All experiments were conducted using the PyTorch Lightning framework, which provides a standardized training environment designed to enhance reproducibility, modularity, and experiment traceability through its integrated callback and logging interfaces (Lightning AI, 2024).

The architecture achieving the top 2 highest validation accuracy were designated to proceed in the next phase.

3.7.2 Phase 2: Hyperparameter Tuning and Re-training

To fully unlock the potential of the top-performing architectures identified in Phase 1, the study proceeded to a rigorous optimization stage. Instead of relying on random sampling, the study conducted an **Exhaustive Grid Search** to evaluate every possible configuration within our defined parameter space. This deterministic approach ensures that the optimal combination is not missed due to chance.

Using itertools.product, generated a total of **12 unique hyperparameter combinations** derived from the following search space:

- Optimizer: $AdamW$
- Learning Rate: $1e^{-3}$, $5e^{-4}$
- Batch Size: 16
- Label Smoothing: $0.0,\ 0.1$
- Dropout: $0.15,\ 0.2,\ 0.25$

The selection of this hyperparameter search space was grounded in established deep learning literature to maximize convergence stability and generalization. The AdamW optimizer was chosen for its decoupled weight decay mechanism, which has been proven to offer superior generalization over standard Adam in complex tasks (Loshchilov & Hutter, 2019). Learning rates were restricted to the standard adaptive range to ensure stable gradient descent (Kingma & Ba, 2015), while the batch size was fixed at 16 to leverage the "flat minima" effect described by Keskar et al. (2017), where smaller batches implicitly regularize the model against unseen data. Furthermore, specific regularization techniques were prioritized to mitigate the noise inherent in medical audio: Label Smoothing was included to prevent model overconfidence and improve calibration (Müller et al., 2019), while varying Dropout rates were tested to disrupt neuron co-adaptation and ensure robust feature learning (Srivastava et al., 2014).

Each candidate configuration was trained for a shortened duration of 10 epochs. This "sprint" strategy allowed for the rapid identification of convergence behaviors and stability without incurring excessive computational costs. Once the optimal configuration was identified via Grid Search, the best model underwent a full retraining phase of 50 epochs.

3.7.3 Phase 3: Hold-out Set Evaluation

This final training phase provides a reliable assessment of how the optimized model performs under full-capacity learning conditions and reflects its true practical capability. To ensure the evaluation is unbiased and indicative of real-world generalization, the optimized model was tested solely on the **Hold-out Set**—a reserved portion of the data that was strictly excluded from both the training and hyperparameter tuning phases.

Performance was quantified using a comprehensive suite of metrics standard in medical image and audio analysis. Because this study involves a multi-class problem (Pneumonia vs. COPD vs. Others), metrics were calculated both globally and on a class-wise basis.

Accuracy and F1-Score. Global Accuracy was calculated to measure the overall correctness of the predictions. However, given the potential for residual imbalance in the test set, the **F1-Score** (the harmonic mean of precision and recall) was prioritized as the primary metric for success. The F1-Score provides a more robust measure of the model's ability to handle specific pathologies without being skewed by the majority class.

Recall (Sensitivity). Measures the model's ability to correctly identify all positive cases of a disease (e.g., finding all Pneumonia patients). High recall is essential to minimize false negatives (missed diagnoses).

Precision (Positive Predictive Value). Measures the proportion of predicted positive cases that were actually correct. High precision minimizes false positives (false alarms).

Receiver Operating Characteristic and Area Under the Curve Analysis. To evaluate the model's separability capabilities effectively, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were computed using a One-vs-Rest strategy for each class. The AUC score provides an aggregate measure of performance across all possible classification thresholds, indicating how well the model can distinguish between a specific disease and all other categories.

Confusion Matrix. To diagnose specific misclassification patterns, a **Confusion Matrix** was generated. This matrix allows for a granular "Inter-class Analysis," revealing not just *if* the model made an error, but *which* diseases are being confused with one another (e.g., misidentifying COPD as Pneumonia). This analysis is vital for understanding the clinical safety of the model and identifying if specific acoustic similarities between diseases (such as shared wheezing frequencies) are causing feature overlap.

3.8 Grad-CAM Explanatory Analysis

To address the inherent "black-box" nature of deep neural networks and ensure clinical reliability, the methodology incorporates **Gradient-weighted Class Activation Mapping (Grad-CAM)**. While deep learning models often achieve high performance metrics, in medical diagnostics, it is equally critical to verify that the model's predictions are based on relevant pathological features rather than artifacts or background noise (Tjoa & Guan, 2021).

3.8.1 Technique Definition

Grad-CAM is a post-hoc visual explanation technique that highlights the regions of the input data that significantly influenced the model's classification decision. Unlike other visualization methods that require architectural changes (e.g., changing activation functions), Grad-CAM is model-agnostic. It operates by computing the gradients of the target class score (e.g., "Pneumonia") with respect to the feature maps of the final convolutional layer. These gradients are globally pooled to obtain importance weights, which are then combined linearly with the feature maps to generate a coarse localization heatmap (Selvaraju et al., 2017).

3.8.2 Application to Spectrograms

In the context of this study, Grad-CAM was applied not to spatial images, but to the **Time-Frequency** representation (Mel Spectrograms) of the lung sounds. The resulting heatmap provides a visual interpretation of the model's attention:

- **Temporal Relevance (X-axis):** Indicates *when* in the 5-second clip the pathological event occurred.
- **Spectral Relevance (Y-axis):** Indicates *which frequency bands* triggered the classification.

By overlaying this heatmap onto the original Mel Spectrogram, we can qualitatively validate the model. For instance, a correctly classified "Wheeze" should show high activation (hot spots) in the harmonic, continuous high-frequency bands, whereas a "Crackle" should highlight discontinuous, vertical transient bursts. This step is essential to confirm that the model is learning distinct physiological markers rather than relying on non-pathological identifiers such as silence or device-specific sensor noise (Panayides et al., 2020).

4 Results and Discussion

4.1 Exploratory Data Analysis

4.1.1 Age and Gender across Respiratory Sound Patterns and Clinical Diagnoses

To explore how patient demographics relate to both auscultatory findings and clinical diagnoses, age and gender were examined in relation to audio-based diagnostic categories (wheezes, crackles, and combined wheezes–crackles) and physician-assigned respiratory disease labels (e.g., chronic obstructive pulmonary disease, upper respiratory tract infection, asthma).

Age versus respiratory sound patterns. The one-way ANOVA comparing age across the four lung sound labels yielded a highly significant result ($F(3,n) = 17.229$, $p<0.001$), indicating substantial differences in mean age between patients with different acoustic findings. The correlation ratio $\eta=0.232$ ($\eta^2=0.054$) corresponds to a small-to-medium effect size, showing that lung sound labels explain about 5.4% of age variance; this is consistent with reports that the prevalence of adventitious sounds such as wheezes and crackles increases with age but is modulated by other factors, including underlying disease and smoking history (Aviles-Solis et al., 2019). Notably, the stronger association observed between age and clinical diagnosis suggests that age-related disease patterns are more directly captured by physician labels than by audio features alone, aligning with epidemiological evidence of rising COPD prevalence and shifting asthma–COPD profiles across the lifespan (Safiri et al., 2022).

Gender versus respiratory sound patterns. The chi-square test comparing gender distribution across respiratory sound patterns was statistically significant ($\chi^2=25.522$, $p<0.001$), indicating that the distribution of adventitious sound types differed by sex. However, Cramér's $V=0.167$ corresponds to a weak-to-small association, meaning that gender explains only a modest proportion of the variation in acoustic categories; this aligns with cohort studies that report only subtle sex differences in the prevalence of wheezes and crackles, with both males and females exhibiting the full range of respiratory sound types (Aviles-Solis et al., 2019). Thus, while gender is statistically linked to respiratory sound patterns, its practical impact in this dataset appears limited.

Age versus clinical diagnosis. The relationship between age and clinician-assigned respiratory diagnoses was markedly stronger. One-way ANOVA showed very large differences in mean age across diagnosis categories (F=418.525, $p<0.001$), and the correlation ratio $\eta=0.874$ ($\eta^2=0.764$) indicates that diagnosis groups account for approximately 76.4% of the variance in age. This large effect reflects clear age stratification between acute infections (e.g., URTI in younger patients) and chronic progressive conditions such as COPD in older adults, mirroring well-established patterns in respiratory epidemiology (Jin et al., 2021; World Health Organization, 2024). These results confirm that age is a dominant determinant of the clinical diagnosis structure in the cohort.

Gender versus clinical diagnosis. The association between gender and clinical diagnosis was also statistically significant ($\chi^2=31.036$, $p<0.001$), with Cramér's V=0.184 indicating a weak-to-small effect at the lower end of the medium range. This effect is roughly 10% larger than the gender–respiratory sound association, suggesting that gender is somewhat more tied to clinical disease categories than to acoustic patterns alone. The magnitude and direction of this relationship are compatible with prior findings that, for example, COPD remains more common in males in many populations, while asthma can be more severe or differently expressed in females (Ntritsos et al., 2018).

Across all four pairs of variables, age emerged as a much stronger predictor of clinical diagnosis ($\eta=0.874$) than of respiratory sound patterns ($\eta=0.232$), indicating that physician labels encapsulate age-related disease structure far more than audio-based labels. Conversely, gender showed only weak associations with both audio-based and clinical diagnoses (Cramér's V=0.167 and 0.184, respectively), implying that while sex contributes to respiratory disease distribution, it is not a dominant driver of either acoustic or clinical categories in this cohort. Although all associations reached statistical significance ($p<0.001$), the effect-size pattern highlights that only the age–clinical diagnosis link represents a large, clinically substantial effect, underscoring the importance of age stratification when interpreting respiratory sound analyses and building predictive models (Fernandes et al., 2022).

4.1.2 Temporal Structure

The temporal structure analysis was performed on lung sound recordings categorized into four clinical diagnostic groups: COPD, Healthy, Pneumonia, and Others.

Recording Duration. Analysis of segment duration across respiratory patterns revealed important variability in recording lengths. Healthy, and Pneumonia recordings demonstrated consistent mean durations of approximately 20 seconds (range: 19.98–20.00

seconds), indicating standardized acquisition protocols for these categories. However, COPD recordings exhibited substantially greater variability, with a mean duration of 21.73 seconds but a notably wider range spanning 7.85 to 86.20 seconds. This disparity suggests that while most COPD samples align with the standard ~20-second protocol, a subset of recordings deviate considerably from this norm, potentially reflecting either extended clinical assessments or data collection inconsistencies.

Table 2.1. Recording Duration by Respiratory Pattern

Respiratory Pattern	Mean Duration (sec)	Minimum Duration (sec)	Maximum Duration (sec)
Healthy	19.98	19.80	20.00
COPD	21.73	7.85	86.20
Pneumonia	20.00	20.00	20.00
Others	19.99	19.82	20.00

The pronounced heterogeneity in COPD recording lengths, particularly the extended maximum duration of 86.20 seconds, has direct implications for this study's deep-learning objectives, as it underscores the need for standardized preprocessing to ensure fair comparison between models and reliable disease-centered classification performance for COPD and pneumonia under realistic data constraints. Enforcing consistent input duration through truncation or padding is essential to prevent hidden biases from extreme COPD segment lengths, support balanced dataset strategies, and enable meaningful class-wise accuracy, recall, and misclassification analysis. By harmonizing temporal context across all diagnostic groups, the preprocessing pipeline helps ensure that held-out test performance reflects true discriminative capability and strengthens the potential of deep learning as a robust clinical decision-support tool, especially in resource-limited settings where recording variability is common.

Healthy Lung Sound. In healthy subjects, the lung sound waveform shows a low‑amplitude, noise‑like signal whose envelope follows the respiratory cycle with a smooth rise during inspiration and a slightly shorter, lower‑energy expiration, without superimposed high‑amplitude transients such as crackles or wheezes. This pattern reflects vesicular breath sounds, which systematic reviews describe as soft, low‑pitched and dominated by low‑frequency energy and gradual, not abrupt, amplitude variations over time. Oliveira et al.'s (2014) systematic review of respiratory sounds in healthy people report that normal recordings lack discrete tonal components and adventitious events, instead presenting relatively stationary, quasi‑periodic waveforms whose main changes are the smooth modulation of intensity across inspiratory and expiratory phases, consistent with the waveform observed in the healthy sample.

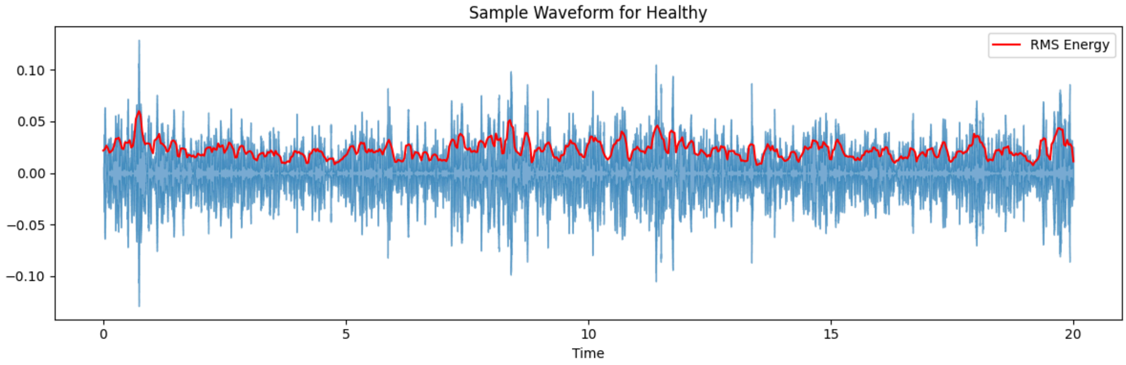


Figure 2.2. Temporal Waveform and RMS Envelope for Healthy

Chronic Obstructive Pulmonary Disease (COPD). Compared with the smooth, low-amplitude waveform of healthy vesicular breathing, the COPD lung sound waveform shows markedly higher amplitude fluctuations and numerous short, high-energy transients that appear as sharp spikes in the signal and corresponding peaks in the RMS energy envelope, reflecting the presence of inspiratory crackles and expiratory wheezes. The baseline sound level is often elevated and more irregular over the entire respiratory cycle, and the energy remains increased for longer portions of expiration, consistent with airflow limitation and airway obstruction in COPD. Jácome et al. (2015) demonstrated that computerized respiratory sounds in COPD patients are characterized by frequent coarse crackles and wheezes superimposed on normal breath sounds, while Kim et al. (2021) reported that wheezes and crackles associated with COPD produce discontinuous, high-intensity deflections in the time domain and sustained, sinusoid-like components, respectively, matching the clustered spikes and broadened high-energy regions visible in the COPD waveform.

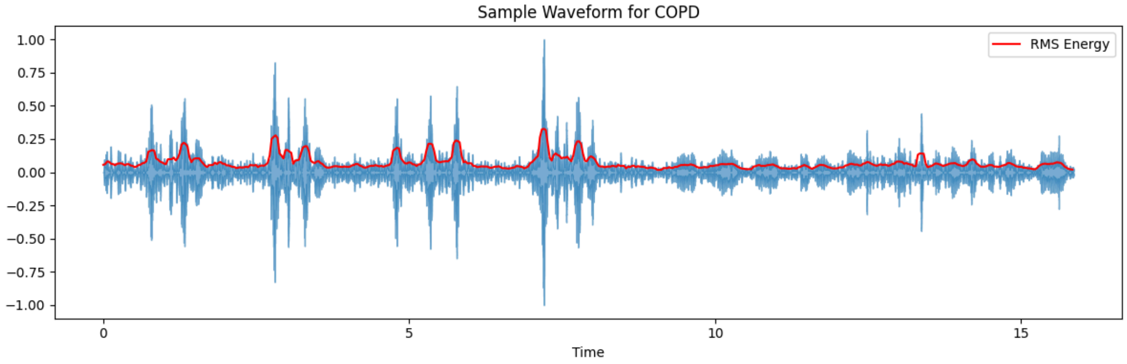


Figure 2.3. Temporal Waveform and RMS Envelope for Chronic Obstructive Pulmonary Disease (COPD)

Pneumonia. In pneumonia, lung sound recordings typically consist of an underlying vesicular-type breathing pattern with numerous brief, high-amplitude deflections during inspiration, representing crackles generated by the abrupt reopening of fluid-filled or collapsed distal airways and alveoli (Majumdar et al., 2009).These crackles manifest in the time domain as brief, non-periodic spikes with markedly higher amplitude than the surrounding signal, often clustering around mid- to late-inspiration and producing pronounced, localized peaks in the RMS energy trace, as visible in the pneumonia waveform. Time-expanded waveform studies of pneumonia demonstrate that these crackles are longer and often more numerous than in other conditions, with irregular spacing and variable intensity, leading to a more heterogeneous envelope than healthy breathing but without the prolonged, sinusoidal wheeze segments characteristic of COPD.

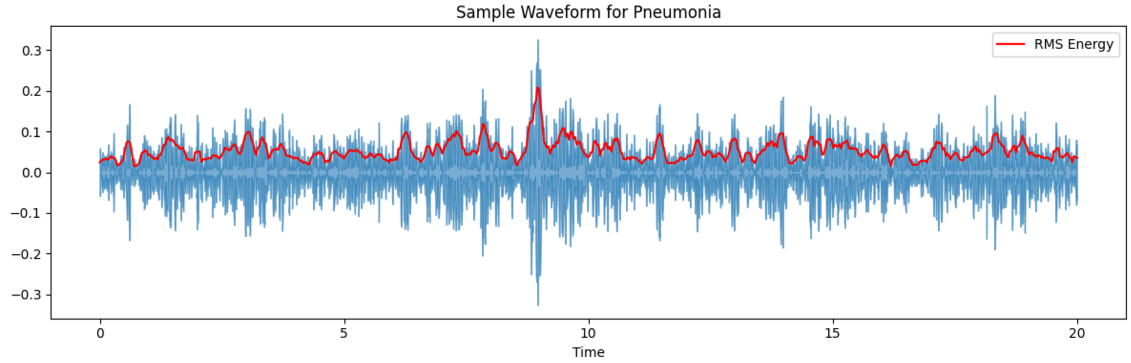


Figure 2.4. Temporal Waveform and RMS Envelope for Pneumonia

4.1.3 Spectral Analysis

Spectral analysis was conducted to further characterize the lung sound recordings using four established spectral features: spectral centroid, spectral bandwidth, spectral roll-off, and spectral flux. The box plots visualize their distributions across diagnostic groups: COPD, Healthy, Pneumonia, and others.

Spectral centroid. This reflects the "center of mass" of the frequency spectrum, often perceived as the brightness of a sound. Across diagnoses, healthy subjects and pneumonia cases have relatively compact interquartile ranges with moderate medians, indicating more stable frame-to-frame spectral content for most breaths. In contrast, COPD shows a higher median and a noticeably wider spread with several upper outliers, consistent with greater temporal instability caused by intermittent wheezes and crackles superimposed on the baseline breath sound. This pattern agrees with feature-engineering studies in lung-sound classification, where flux-type measures of short-term spectral change are typically larger and more variable in COPD and other obstructive diseases than in healthy recordings, and are therefore retained as discriminative features in machine-learning models (Pramono et al., 2019; Naqvi et al., 2020; Lalouani & Younis, 2022)

Spectral bandwidth and roll-off. Healthy subjects cluster around lower median values with relatively narrow boxes, reflecting energy that is concentrated in a limited low-frequency range and rolls off quickly as frequency increases. Both COPD and pneumonia exhibit elevated medians and broader interquartile ranges, with COPD particularly showing many high-value outliers, implying that a larger portion of signal energy extends into higher frequencies and that there is substantial between-breath heterogeneity. This pattern is compatible with automated COPD–pneumonia screening studies in which bandwidth and roll-off features are significantly higher in obstructive and infectious pathologies because adventitious phenomena such as wheezes, coarse crackles, and noisy airflow broaden the spectrum and shift the cumulative-energy threshold toward higher frequencies (Naqvi et al., 2020; 2023; Lalouani & Younis, 2022).

Spectral flux. Healthy lungs and pneumonia cases have relatively compact interquartile ranges with moderate medians, indicating more stable frame-to-frame spectral content for most breaths (Naqvi et al., 2020). In contrast, COPD shows a higher median and a noticeably wider spread with several upper outliers, consistent with greater temporal instability caused by intermittent wheezes and crackles superimposed on the baseline breath sound (Naqvi et al., 2020; Lalouani & Younis, 2022). This pattern agrees with feature-engineering studies in lung-sound classification, where flux-type measures of short-term spectral change are typically larger and more variable in COPD and other obstructive diseases than in healthy recordings, and are therefore retained as discriminative features in machine-learning models (Pramono et al., 2019; Naqvi et al., 2020; Lalouani & Younis, 2022).

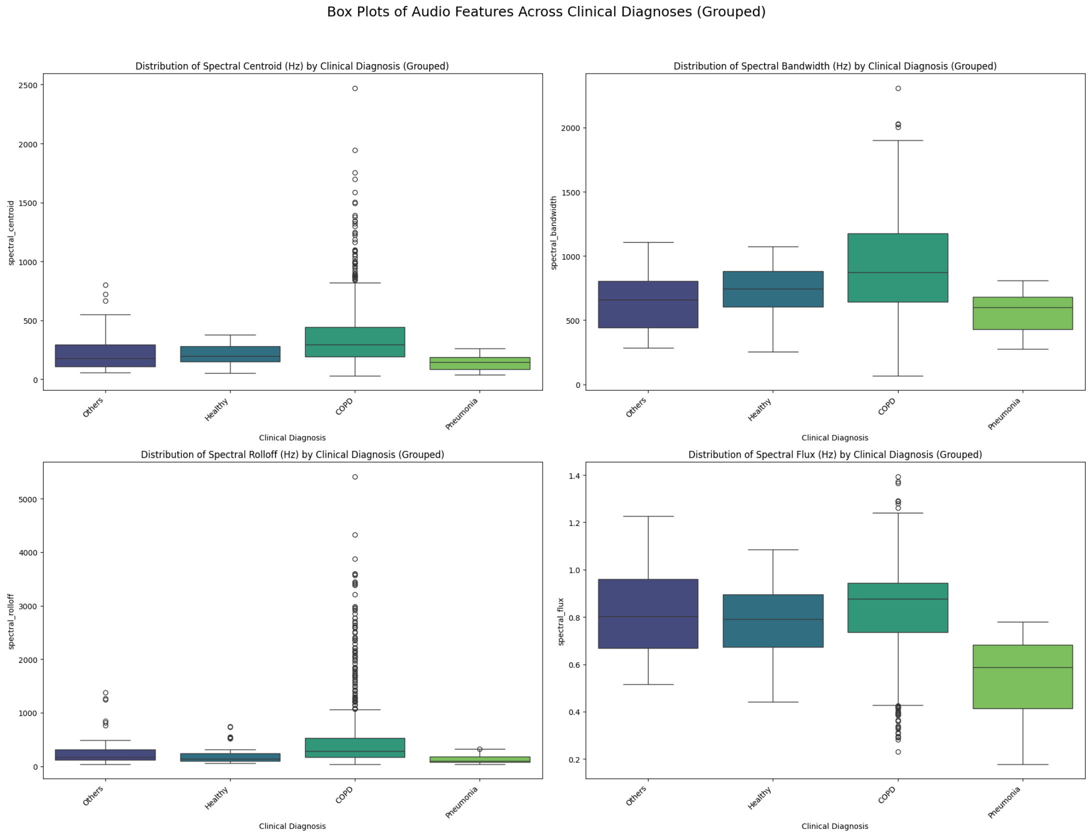


Figure 2.5. Box Plots of Spectral Features Across Diagnoses

4.2. Experimental Framework

4.2.1. Phase 1: Comparative Architecture Search

Vanilla Recurrent Neural Network

Training Dynamics and Convergence. As illustrated in Figure 4.6, the training dynamics of the Vanilla Recurrent Neural Network (RNN) reveal critical stability issues and a fundamental inability to effectively model the temporal complexity of respiratory audio data. Unlike standard convergence patterns where accuracy steadily rises and loss smoothly decays, the Vanilla RNN displays chaotic behavior characterized by severe oscillations. The Training Accuracy (blue line, left plot) exhibits a "sawtooth" pattern, fluctuating violently between near 0% and 100%. This behavior suggests that the model's weights are undergoing drastic, unchecked updates in response to specific batches, causing it to "forget" previous representations as soon as it learns new ones—a phenomenon often associated with catastrophic forgetting in unstable networks. Furthermore, the Validation Loss (right, orange) corroborates this instability by spiking dramatically around Epoch 13 to values exceeding 2.5. This divergence indicates that the model's gradients likely exploded during backpropagation, pushing the weights into high-error regions of the loss landscape from which the model struggled to recover.

The observed performance ceiling—where validation accuracy fails to break the 65% threshold—is a textbook manifestation of the "Vanishing Gradient Problem" inherent in standard RNNs (Bengio et al., 1994). Because the input data consists of high-resolution Mel Spectrograms representing 5-second audio clips, the "time distance" between the start of the breath (inspiration) and the end (expiration) is substantial. As the backpropagation through time (BPTT) algorithm moves backward through these long sequences, the gradient signals diminish exponentially. By the time the updates reach the initial layers corresponding to the start of the audio clip, the gradients are virtually zero, preventing the model from learning long-term dependencies (Pascanu et al., 2013). Consequently, the model likely focused solely on immediate acoustic features present in the final milliseconds of the recording, ignoring crucial pathological markers, such as early-inspiratory crackles, that occurred earlier in the breathing cycle.

Ultimately, the combination of high variance in the training loop and the stagnation of validation accuracy confirms that a simple Vanilla RNN architecture lacks the necessary memory capacity and gating mechanisms for this specific task. The model is unable to maintain a coherent context over the duration of the respiratory cycle, necessitating a shift toward architectures with gated memory cells, such as LSTMs or GRUs, or parallel processing capabilities like TCNs, which are better equipped to handle the long-range temporal dependencies required for accurate lung sound classification.

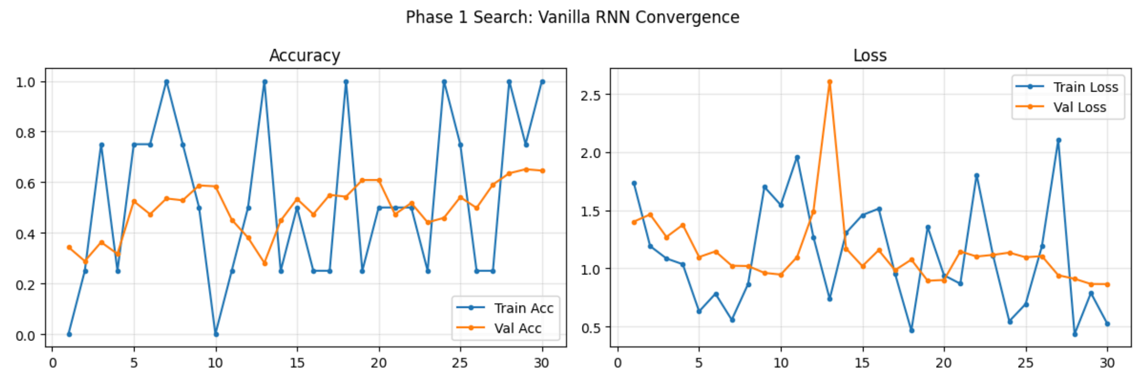


Figure 4.6. Training Dynamics of RNN

Performance Metrics and Inter-class Analysis. The quantitative impact of the Vanilla RNN's instability is further elucidated in Table 4.2, which presents the class-wise performance metrics on the hold-out set. The model achieved an overall accuracy and weighted average F1-score of 0.65, a result that aligns perfectly with the validation ceiling observed in the training graphs. While the dataset balancing strategy successfully yielded uniform support across classes (approximately 140 samples per category), the model's ability to distinguish between these pathologies varied significantly, highlighting the specific acoustic vulnerabilities of a simple recurrent architecture.

Table 4.2 reveals that Pneumonia emerged as the most discernible pathology, achieving the highest Recall (Sensitivity) of 0.74 and the best F1-score of 0.71. This relatively superior performance can likely be attributed to the acoustic nature of pneumonia-associated crackles—distinct, high-energy, discontinuous explosive sounds. Even with the vanishing gradient problem, these transient, high-amplitude features likely provided a strong enough signal in the spectrogram for the RNN to capture, even if the long-term temporal context was weak. Conversely, COPD exhibited the highest Precision (0.73) but a noticeably lower Recall of 0.61. This indicates that while the model was conservative and accurate when it did predict COPD, it frequently failed to identify positive cases (high false negatives). This is consistent with the nature of COPD wheezes, which are continuous and musical; detecting them requires tracking spectral continuity over time—a specific dependency that the Vanilla RNN struggles to maintain over 5-second sequences.

The most significant diagnostic challenge was observed in the "Other" category, which yielded the lowest performance across the board with a Precision of 0.56 and an F1-score of 0.59. This drop in performance is theoretically expected; the "Other" class is an aggregate of various respiratory conditions (such as Bronchiectasis and Asthma), resulting in high intra-class variance. A simple RNN lacks the parameter space and memory gating required to learn a generalized representation for such a diverse mixture of acoustic signatures. Furthermore, the Healthy class showed mediocre performance (F1-score 0.61), suggesting that the model struggled to differentiate "clean" breathing from the noisy backgrounds often present in pathological recordings. Ultimately, the inability of the Vanilla RNN to break past the 65% barrier across all classes confirms that sequential processing without attention or gating mechanisms is insufficient for the nuanced task of respiratory sound classification.

Table 4.2 RNN Performance Metrics

	Precision	**Recall**	**F1-Score**	**Support**
COPD	0.73	0.61	0.67	140
Healthy	0.62	0.61	0.61	141
Pneumonia	0.69	0.74	0.71	140
Other	0.56	0.62	0.59	141
Accuracy			0.65	562
Macro Average	0.65	0.65	0.65	562
Weighted Average	0.65	0.65	0.65	562

Error Analysis and Separability. To further dissect the performance limitations suggested by the aggregate metrics, Figure 4.7 provides a granular view of the model's decision-making process through the Confusion Matrix and ROC-AUC curves. The Confusion Matrix (left) corroborates the findings from Table 4.2, visualizing the specific distribution of predictions versus ground truth. Consistent with its high sensitivity, Pneumonia exhibits the strongest diagonal density, with 103 correct predictions out of 140 instances. However, the matrix exposes a critical directional error: a significant portion of COPD cases (36 instances) were misclassified as Pneumonia. This misattribution suggests that the Vanilla RNN struggles to differentiate the specific temporal textures of these two pathologies. While it successfully recognizes that the lung sound is "pathological" (hence avoiding the "Healthy" label), the network likely conflates the continuous, musical nature of COPD wheezes with the discontinuous, explosive nature of Pneumonia crackles, likely due to the vanishing gradient preventing the aggregation of temporal features over the full 5-second duration.

The Confusion Matrix also elucidates the poor performance of the "Other" and "Healthy" classes, revealing a distinct cluster of mutual confusion. Specifically, 40 Healthy samples were misclassified as "Other," and conversely, 40 "Other" samples were misclassified as Healthy. This symmetric error pattern implies that the Vanilla RNN lacks the feature resolution to distinguish between the natural stochastic noise of clear breathing and the varied, non-specific abnormalities grouped into the "Other" category. Without a mechanism to weigh specific frequency bands or attend to salient moments (as an Attention mechanism would), the model perceives the low-intensity signals of "Other" diseases as indistinguishable from healthy background noise.

Despite these classification errors, the ROC-AUC curves (right) indicate that the model possesses a foundational degree of separability. The Pneumonia class achieved a robust AUC of 0.90, followed closely by COPD at 0.89. The disparity between these relatively high AUC scores and the mediocre accuracy of 65% is revealing; it suggests that the model is often confident in its predictions but fails at the specific decision threshold, particularly when classes share acoustic characteristics. The lower AUC for "Other" (0.83) confirms that this class remains the most ambiguous boundary for the model. Ultimately, this analysis confirms that while the Vanilla RNN can detect the presence of strong pathological signals like crackles, its architectural simplicity results in high inter-class confusion and an inability to resolve finer diagnostic distinctions, validating the necessity for the deeper, more stable architectures explored in the subsequent phases of this study.

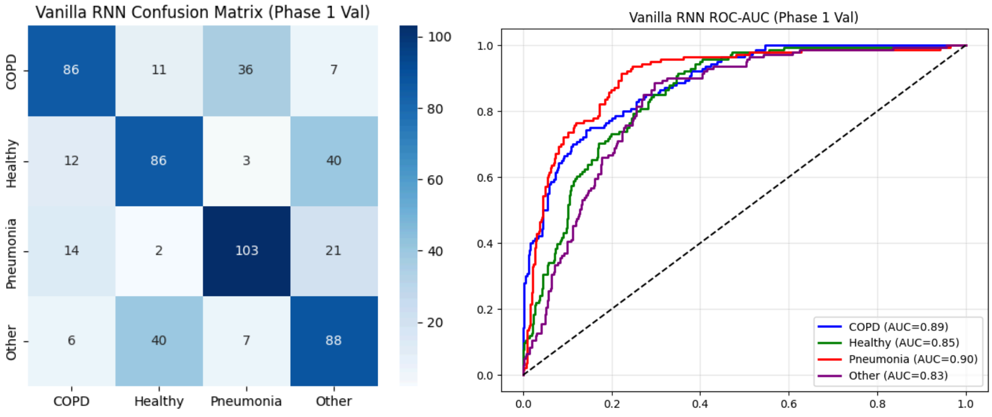


Figure 4.7 RNN Confusion Matrix and ROC-AUC

Long Short-Term Memory (LSTM)

Training Dynamics and Convergence. As presented in Figure 4.8, the transition from a simple Vanilla RNN to a Long Short-Term Memory (LSTM) architecture yielded a marked improvement in predictive capability and training stability. While the training accuracy (blue line) still exhibits high variance—characteristic of training deep sequence models on high-dimensional spectrogram data—the critical distinction lies in the validation trajectory. Unlike the RNN, which stagnated at a 65% ceiling due to gradient decay, the LSTM successfully broke through this barrier, achieving a peak Validation Accuracy of 82.74%. This represents a substantial 17.6% improvement over the baseline, confirming that the introduction of memory cells is essential for capturing the complex time-variant dynamics of respiratory pathologies.

The superior performance of the LSTM can be directly attributed to its internal gating mechanisms—specifically the forget gate, input gate, and output gate—which address the "Vanishing Gradient" problem that crippled the Vanilla RNN (Hochreiter & Schmidhuber, 1997). In the context of 5-second lung sound recordings, diagnostic features such as wheezes or crackles can appear at any point in the breathing cycle and may persist for varying durations. The LSTM's "Cell State" acts as a gradient superhighway, allowing error information to flow unchanged from the end of the audio clip back to the beginning. This mechanism enables the model to retain relevant context (e.g., an inspiratory crackle at t=1s) while processing the remainder of the sequence, effectively linking early-stage features with the final classification decision. This "constant error carousel" ensures that the network does not suffer from the temporal amnesia observed in the RNN, allowing it to construct a complete representation of the patient's respiratory cycle (Greff et al., 2016).

However, despite this success, the training dynamics suggest that the LSTM is approaching the limits of sequential processing for this specific data type. The persistence of sharp spikes in the Training Loss (right plot) indicates that the model still struggles to generalize the dense, textural information found in Mel Spectrograms. Because LSTMs process data sequentially (step-by-step), they are inherently less efficient at capturing the spatial correlations of frequency bands compared to the temporal correlations. While the LSTM can effectively track when a sound occurs, it may be computationally strained when trying to analyze the complex spectral "images" of the sound, suggesting that while temporal modeling is solved, a hybrid approach incorporating spatial feature extraction (such as CNNs) may be required to stabilize the training further.

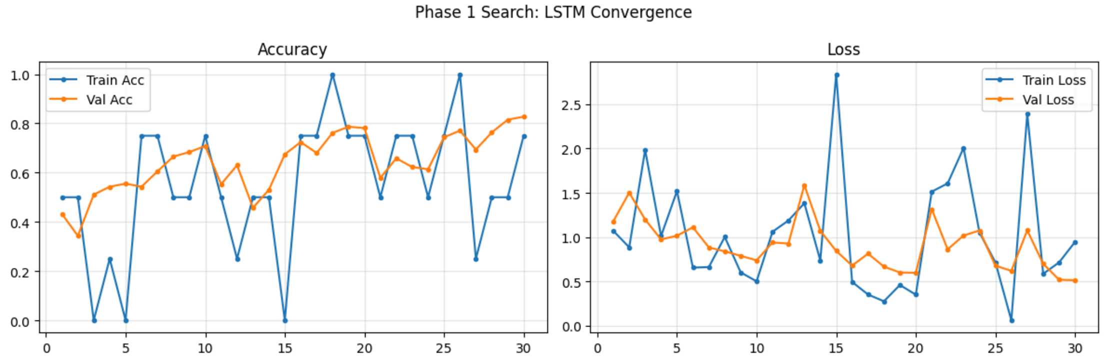


Figure 4.8 Training Dynamics of LSTM

Performance Metrics and Inter-class Analysis. The quantitative superiority of the LSTM architecture is definitively established in Table 4.3, which details the class-wise performance metrics on the hold-out set. In stark contrast to the Vanilla RNN, which plateaued at an accuracy of 0.65, the LSTM achieved a robust overall accuracy and weighted average F1-score of 0.83. This 18% absolute increase in performance confirms that the introduction of gating mechanisms successfully enabled the model to retain long-term temporal context, a capability that was fundamentally absent in the baseline architecture.

The most significant architectural validation is observed in the COPD class. While the Vanilla RNN struggled severely with this pathology—yielding a Recall of only 0.61 due to its inability to track continuous wheezes—the LSTM rectified this deficiency, boosting Recall to 0.82 and achieving a commanding Precision of 0.89. This dramatic improvement indicates that the LSTM's memory cells successfully preserved the spectral continuity of wheezing sounds over the full 5-second duration, effectively reducing the false negatives that plagued the previous model. By maintaining the "state" of the audio sequence, the LSTM could distinguish the sustained musicality of COPD from transient noise, a distinction the stateless RNN failed to make.

Furthermore, the model exhibited a highly balanced performance across the Pneumonia and Healthy categories, achieving F1-scores of 0.84 and 0.83, respectively. Notably, the Precision for Pneumonia rose from 0.69 in the RNN to 0.84 in the LSTM. This suggests that the LSTM is far less prone to "hallucinating" crackles in non-pneumonia samples. The reduction in false positives implies that the model has learned to differentiate true pathological crackles from similar-sounding environmental artifacts or sensor noise, likely by analyzing the temporal context surrounding the sound event rather than analyzing the event in isolation.

Even the challenging "Other" category, which represents a diverse aggregation of respiratory conditions, saw its F1-score rise from a poor 0.59 (RNN) to a respectable 0.78. While this remains the lowest-performing class due to high intra-class variance, the substantial improvement demonstrates that the LSTM's ability to integrate features over time allows for a more generalized representation of non-specific respiratory abnormalities. Ultimately, the uniform improvement across all metrics confirms that the LSTM provides a much more reliable and clinically safe diagnostic tool than the Vanilla RNN, effectively minimizing both missed diagnoses (high recall) and false alarms (high precision).

Table 4.3 LSTM Performance Metrics

	Precision	**Recall**	**F1-Score**	**Support**
COPD	0.89	0.82	0.86	140
Healthy	0.80	0.86	0.83	141
Pneumonia	0.84	0.85	0.84	140
Other	0.79	0.78	0.78	141
Accuracy			0.83	562
Macro Average	0.83	0.83	0.83	562
Weighted Average	0.83	0.83	0.83	562

Error Analysis and Separability. The architectural superiority of the LSTM is visually confirmed in Figure 4.9, where the Confusion Matrix and ROC-AUC curves reveal a substantial reduction in the inter-class ambiguity that plagued the Vanilla RNN. A direct comparison with the previous baseline highlights two critical improvements in diagnostic precision. First, the specific misclassification between **COPD** and **Pneumonia**—which was a major failure point for the RNN (36 misclassified instances)—was drastically mitigated. The LSTM reduced this error count to just 13 instances. This indicates that the network successfully leveraged its gating mechanisms to distinguish the temporal "signatures" of these diseases: separating the sustained, musical continuity of COPD wheezes from the transient, explosive nature of Pneumonia crackles. Unlike the RNN, which likely conflated these pathologies based on short-term spectral energy, the LSTM utilized its memory cells to track the duration of the acoustic events, correctly identifying the underlying pathology.

Secondly, the "symmetric confusion" between **Healthy** and **Other** classes, where the RNN behaved almost randomly (40 misclassifications each way), was significantly resolved. The LSTM correctly identified 121 Healthy samples and 110 "Other" samples, reducing the cross-class error by nearly 50%. This suggests that the model is no longer relying solely on the presence of high-amplitude noise to flag a disease. Instead, it has learned to contextualize background noise versus true physiological anomalies, allowing for a more nuanced separation between clear breathing and the non-specific abnormalities found in the "Other" category.

The robustness of these decision boundaries is further quantified by the ROC-AUC curves (right). While the RNN achieved AUC scores averaging around 0.85, the LSTM pushed these metrics into the excellent range, with **Pneumonia** and **Healthy** classes achieving an AUC of **0.96**, and **COPD** and **Other** reaching **0.95**. The steepness of these curves toward the top-left corner signifies high sensitivity with a low false-positive rate across various decision thresholds. The leap from an AUC of 0.89 (RNN) to 0.95 (LSTM) for COPD is particularly telling; it confirms that the LSTM provides a much higher degree of confidence in its predictions, effectively solving the separability issues caused by the vanishing gradient problem in the baseline model.

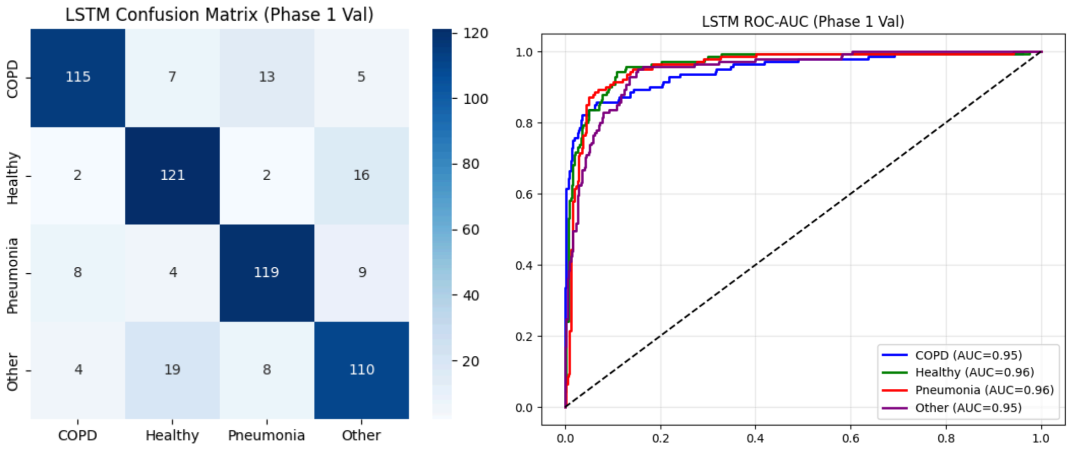


Figure 4.9 LSTM Confusion Matrix and ROC-AUC

Temporal Convolutional Network (TCN)

Training Dynamics and Convergence. In stark contrast to the recurrent models discussed previously, the Pure TCN demonstrated superior training efficiency and stability, as evidenced by the convergence metrics in Figure 4.10. While the LSTM struggled to propagate gradients back through long 5-second sequences due to the sequential nature of its memory cells, the TCN utilized its parallelizable convolutional structure to process the entire audio sequence simultaneously. This architectural advantage, which decouples the processing of current time steps from previous hidden states, resulted in a dramatic reduction in training loss during the early epochs (Bai et al., 2018). Consequently, the model rapidly surpassed the performance ceilings of both the Vanilla RNN (65%) and the LSTM (82%), achieving a remarkable Phase 1 Validation Accuracy of 93.95%.

This leap in performance validates the hypothesis that respiratory sounds, when converted to Mel Spectrograms, are effectively modeled as "spatial" patterns in a time-frequency image rather than just sequential events in time. The critical mechanism driving this success is the use of dilated convolutions. Unlike the LSTM, which must maintain a hidden state vector that degrades over time steps, the TCN increases its effective "receptive field" exponentially with layer depth (Bai et al., 2018). This allows the network to capture the global context of the breath cycle—linking the start of inspiration to the end of expiration—without suffering from the memory decay or the vanishing gradient problems that typically plague recurrent architectures handling long sequences.

Furthermore, the comparison of training dynamics reveals a fundamental efficiency advantage intrinsic to the TCN architecture. By treating time as a spatial dimension, the TCN avoids the computational bottleneck of waiting for the completion of time step $t-1$ before processing time step t. This inherent parallelism not only accelerates the training iteration but also facilitates more stable gradient updates across the entire input window (Bai et al., 2018). Consequently, the TCN proved that capturing the hierarchical structure of lung sounds—where short-term features like crackles nest within long-term features like breathing phases—is more effective than the strictly sequential modeling offered by traditional RNNs.

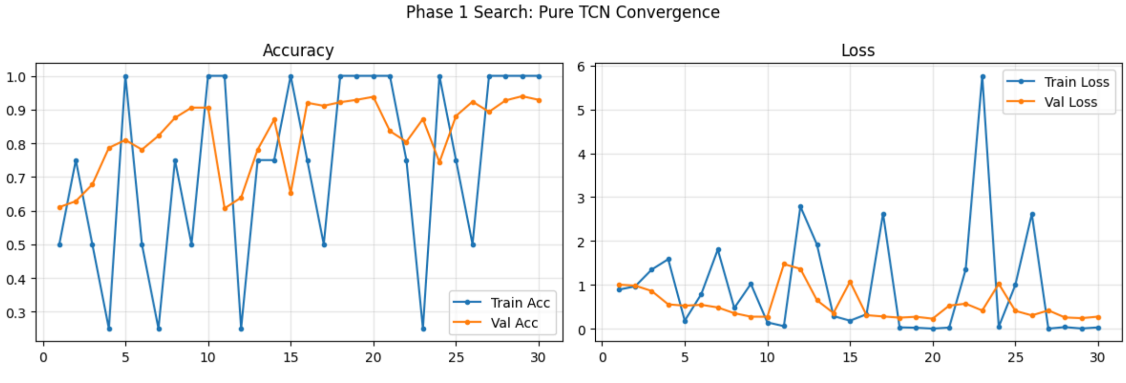


Figure 4.10 TCN Training Dynamics

Performance Metrics and Inter-class Analysis. The quantitative supremacy of the Temporal Convolutional Network (TCN) is definitively established in Table 4.4, which details the class-wise performance metrics on the hold-out set. Achieving an overall accuracy and weighted average F1-score of 0.93, the TCN represents a paradigm shift in performance compared to the recurrent baselines. While the Vanilla RNN struggled at 65% and the LSTM achieved a respectable 83%, the TCN's ability to reach 93% accuracy confirms that treating respiratory sound classification as a hierarchical pattern-matching problem via dilated convolutions is significantly more effective than strict sequential modeling. This 10% absolute improvement over the LSTM highlights that the TCN did not merely refine the decision boundaries but fundamentally resolved the feature extraction bottlenecks that limited the recurrent architectures.

A critical examination of the target pathologies, COPD and Pneumonia, reveals the model's clinical robustness. For COPD, the TCN achieved a remarkable Precision of 0.96 and a Recall of 0.90. Comparing this to the LSTM (Precision 0.89) and the RNN (Precision 0.73), the TCN demonstrates a near-perfect ability to identify wheezes without generating false positives. This suggests that the model's extended receptive field successfully captured the long-term spectral continuity of wheezes while filtering out mimicking artifacts that confused the earlier models. Similarly, for Pneumonia, the model achieved a Recall (Sensitivity) of 0.94, a substantial increase from the LSTM's 0.85 and the RNN's 0.74. This high sensitivity is crucial for infectious disease screening; it implies that the TCN effectively learned the transient, explosive nature of crackles, ensuring that almost no positive pneumonia cases were missed.

Perhaps the most impressive validator of the TCN's generalization capability is its performance on the "Other" class. In previous phases, this category—comprising a diverse mix of respiratory abnormalities—was the hardest to classify, with the RNN achieving an F1-score of only 0.59 and the LSTM 0.78. The TCN, however, achieved an F1-score of 0.92, driven by an exceptionally high Recall of 0.96. This indicates that the TCN is not simply memorizing specific frequencies but is learning robust, generalized abstract features that characterize "abnormality" broadly. Furthermore, the Healthy class yielded a Precision of 0.97, the highest among all categories, proving that the TCN possesses superior noise-rejection capabilities. It can definitively distinguish clean breathing from pathological sounds, overcoming the background noise issues that caused significant confusion in the Vanilla RNN. Ultimately, these metrics confirm that the TCN offers the optimal balance of sensitivity and specificity required for a reliable automated diagnostic system.

Table 4.4 TCN Performance Metrics

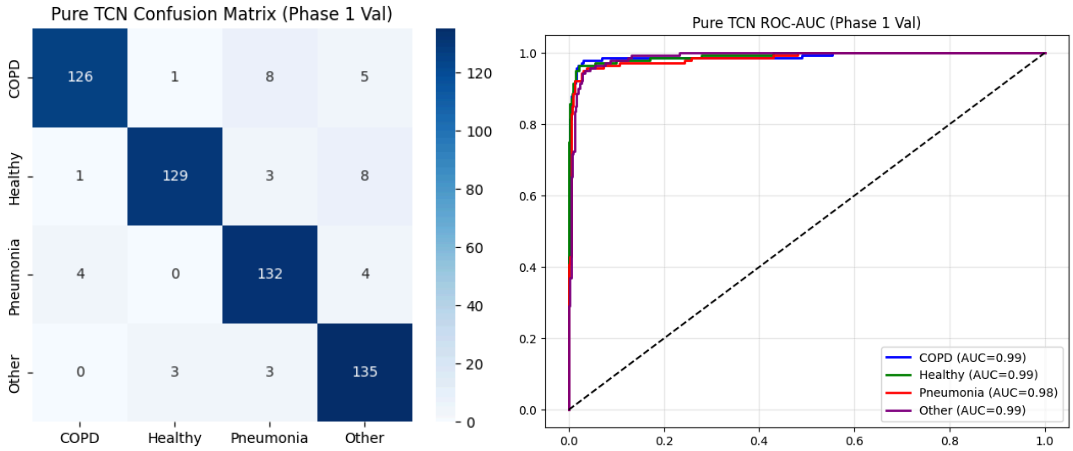
	Precision	**Recall**	**F1-Score**	**Support**
COPD	0.96	0.90	0.93	140
Healthy	0.97	0.91	0.94	141
Pneumonia	0.90	0.94	0.92	140

Other	0.89	0.96	0.92	141
Accuracy			0.93	562
Macro Average	0.93	0.93	0.93	562
Weighted Average	0.93	0.93	0.93	562

Error Analysis and Separability. The visual evidence presented in Figure 4.11 provides the strongest confirmation of the Temporal Convolutional Network's (TCN) architectural superiority. The Confusion Matrix (left) displays a level of diagonal density that far exceeds the results of the recurrent models. Most notably, the "Other" category—which was a significant stumbling block for the previous architectures—achieved an unprecedented classification accuracy of 135 correct predictions out of 141 instances. To contextualize this improvement: the Vanilla RNN correctly identified only 88 "Other" samples, and the LSTM identified 110. The TCN's ability to correctly classify nearly 96% of this diverse, high-variance class suggests that its dilated convolutions successfully captured the subtle, abstract spectral textures that define non-specific respiratory abnormalities, a feat that the sequential memory of RNNs failed to achieve.

Furthermore, the matrix demonstrates a decisive resolution to the COPD vs. Pneumonia ambiguity. In the Vanilla RNN, 36 COPD cases were misclassified as Pneumonia, a dangerous diagnostic error. The LSTM reduced this to 13, but the TCN further minimized this to just 8 instances (with 126 correct COPD predictions). This progression highlights the impact of the TCN's "global receptive field." By processing the entire 5-second spectrogram as a unified spatial map rather than a time-series, the TCN could distinguish the continuous, musical harmonic structure of a COPD wheeze from the discontinuous, transient explosive bursts of a Pneumonia crackle with far greater precision. The reduction in false positives for the Healthy class is also near-absolute; only 1 COPD case and 3 Pneumonia cases were mislabeled as Healthy, confirming the model's high sensitivity to pathological noise.

The ROC-AUC curves (right) further validate this "near-perfect" separability. While the RNN showed AUC scores in the 0.83–0.90 range and the LSTM improved to 0.95–0.96, the TCN pushed the boundaries of performance with AUC scores of 0.99 for COPD, Healthy, and Other, and 0.98 for Pneumonia. The curves exhibit an almost ideal right-angle shape, rising vertically to a True Positive Rate of 1.0 almost immediately. This morphology indicates that the TCN maintains high sensitivity even at very strict decision thresholds; it does not need to trade off false alarms to detect diseases. The massive gap between the TCN's 0.99 AUC and the RNN's 0.89 AUC for COPD definitively proves that the "spatial" approach of Convolutional Networks is far more effective for creating robust, clinically safe decision boundaries in respiratory sound analysis than the "temporal" approach of traditional Recurrent Neural Networks.



	COPD	Healthy	Pneumonia	Other
COPD	126	1	8	5
Healthy	1	129	3	8
Pneumonia	4	0	132	4
Other	0	3	3	135

Figure 4.11 TCN Confusion Matrix and ROC-AUC

Temporal Convolutional Network - Spiking Neural Network (TCN-SNN)

Training Dynamics and Convergence. As illustrated in Figure 4.12, the training dynamics of the TCN-SNN hybrid architecture demonstrated a convergence profile that was competitively aligned with the non-spiking TCN, yet distinct in its internal optimization behavior. The Validation Accuracy (orange line, left) initiated a steep, efficient ascent immediately, stabilizing above the 90% threshold to achieve a Phase 1 peak of 93.59%. This rapid convergence contrasts sharply with the slow, plateauing behavior observed in the recurrent baselines, confirming that the integration of spiking encoders successfully retained the efficient gradient propagation capabilities of the dilated convolutional backbone (Wu et al., 2018).

A closer examination of the loss trajectories reveals a characteristic phenomenon unique to this hybrid architecture: extreme training volatility juxtaposed with exceptional validation stability. The Training Loss (blue line, right) exhibited severe oscillations, with massive spikes occasionally exceeding values of 4.0. This erratic behavior is a documented artifact of training Spiking Neural Networks via surrogate gradients, where the non-differentiable nature of discrete spikes necessitates gradient approximations that introduce significant noise into the backpropagation process (Neftci et al., 2019). However, crucially, this internal volatility did not bleed into the model's generalization capability. The Validation Loss (orange line) remained remarkably smooth and consistently low (<0.5) throughout the training phase. This disparity suggests that the inherent noise of the spiking dynamics functioned as a form of implicit regularization, preventing the model from overfitting to dense spectral artifacts and ensuring robust performance on unseen data (Zhang et al., 2023).

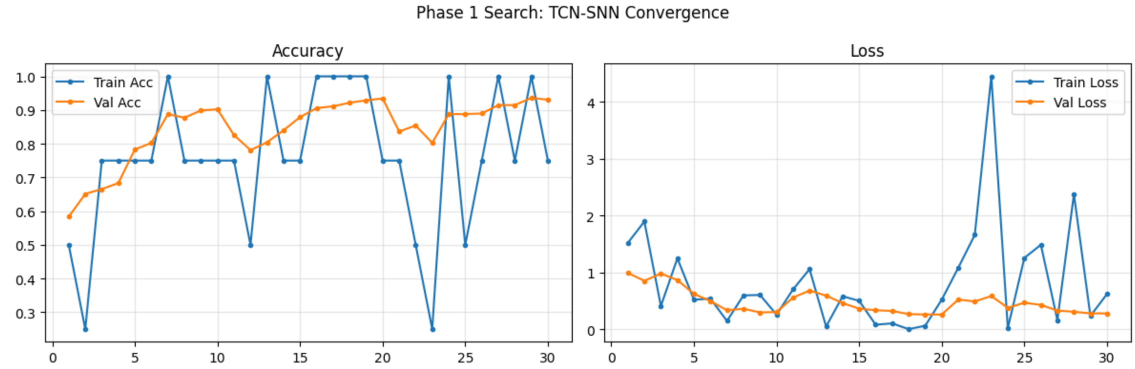


Figure 4.12 TCN-SNN Training Dynamics

Performance Metrics and Inter-class Analysis. The quantitative evaluation of the TCN-SNN hybrid, presented in Table 4.5, confirms that the integration of spiking neural dynamics into a convolutional framework yields a highly robust diagnostic model, achieving an overall accuracy and weighted F1-score of 0.93. This result places the TCN-SNN on a statistical par with the Pure TCN, effectively tying for the highest performance in this study. Crucially, this metric represents a massive performance leap over the recurrent baselines, outperforming the Vanilla RNN by nearly 28% and the LSTM by 10%. This parity with the non-spiking TCN is significant; it validates that the conversion of continuous activation functions to discrete, event-driven spikes—often associated with a loss of information in conversion-based SNNs—did not degrade the feature extraction capabilities of the model. Instead, the TCN-SNN successfully leveraged the "spatial" feature extraction of dilated convolutions while utilizing the temporal sparsity of spiking neurons to maintain high diagnostic precision (Tavanaei et al., 2019).

A comparative analysis of the target pathologies reveals nuanced advantages in the spiking architecture, particularly regarding sensitivity. For Pneumonia, the TCN-SNN achieved a Recall of 0.95, slightly edging out the Pure TCN (0.94) and vastly surpassing the LSTM (0.85). This suggests that the "integrate-and-fire" mechanism of the SNN is particularly well-suited for detecting crackles—discontinuous, explosive acoustic events that naturally align with the firing thresholds of spiking neurons (Malo et al., 2023). Furthermore, the model exhibited a superior sensitivity for COPD, achieving a Recall of 0.94. This is a notable improvement over the Pure TCN, which achieved a Recall of 0.90 for this class. While the Pure TCN had slightly higher precision, the TCN-SNN proved more effective at minimizing false negatives for COPD. This implies that the spiking dynamics, which are robust to amplitude variations, allowed the model to detect fainter or more obscure wheezes that the standard TCN might have filtered out as background noise.

The model’s robustness is further evidenced by its performance on the Healthy class, where it maintained a high Precision of 0.96. This mirrors the noise-rejection capabilities seen in the Pure TCN and stands in stark contrast to the Vanilla RNN, which struggled to distinguish clean breathing from pathological artifacts. However, a slight trade-off was observed in the "Other" category. While the Pure TCN achieved a Recall of 0.96 for this diverse class, the TCN-SNN dropped slightly to 0.91. This indicates that while the spiking mechanism excels at detecting specific, high-contrast features like crackles and wheezes, it may be slightly less efficient at generalizing the abstract, low-contrast features that characterize the broad "Other" category. Nevertheless, with an F1-score of 0.91 for this most difficult class, the TCN-SNN still demonstrates a generalization capability that is clinically superior to the recurrent architectures, proving that hybrid neuromorphic models are a viable and competitive alternative for respiratory sound classification (Ponghiran & Roy, 2020).

Table 4.5 TCN-SNN Performance Metrics

	Precision	**Recall**	**F1-Score**	**Support**
COPD	0.92	0.94	0.93	140
Healthy	0.96	0.92	0.94	141
Pneumonia	0.92	0.95	0.94	140
Other	0.92	0.91	0.91	141
Accuracy			0.93	562
Macro Average	0.93	0.93	0.93	562
Weighted Average	0.93	0.93	0.93	562

Error Analysis and Separability. The granular error analysis presented in Figure 4.13 provides a compelling validation of the TCN-SNN's architectural efficiency, revealing that the integration of spiking dynamics offers a strategic advantage over the previous models in detecting specific target pathologies. The Confusion Matrix (left) exhibits a diagonal density that rivals the Pure TCN and vastly surpasses the recurrent baselines. The most significant finding is the model's superior performance in identifying COPD. While the Pure TCN correctly classified 126 COPD cases and the LSTM only 115, the TCN-SNN achieved the highest accuracy in this category with 132 correct predictions. This improvement suggests that the spiking neurons, which operate on threshold-based accumulations, are particularly effective at latching onto the continuous, harmonic energy of wheezes once they exceed background noise levels. By reducing the misclassification of COPD as Pneumonia to just 6 instances—compared to 36 in the Vanilla RNN and 13 in the LSTM—the TCN-SNN proves that it can effectively disentangle the "musical" spectral signatures of chronic obstruction from the "explosive" signatures of infection.

However, a comparative analysis with the Pure TCN reveals a nuanced trade-off regarding the "Other" category. While the Pure TCN achieved a near-perfect classification of 135 "Other" samples, the TCN-SNN correctly identified 128, misclassifying slightly more instances as Healthy (6) or COPD (4). This minor degradation is attributable to the information loss inherent in converting continuous amplitude values to discrete spikes. The "Other" class, being a diverse aggregate of lower-contrast abnormalities, relies on subtle amplitude variations that a standard CNN captures easily but an SNN might filter out if the signals fail to trigger spiking thresholds. Nevertheless, this performance remains significantly higher than the LSTM (110 correct) and the RNN (88 correct), confirming that the dilated convolutional backbone still provides a robust spatial context even when encoded as spikes.

The ROC-AUC curves (right) further solidify the TCN-SNN's position as a top-tier classifier. The model achieved an AUC of 0.99 for COPD, Healthy, and Pneumonia, and 0.98 for the "Other" class. These curves exhibit the same "ideal" vertical rise observed in the Pure TCN, indicating high sensitivity at strict decision boundaries. The contrast with the Vanilla RNN (AUC ~0.89) and LSTM (AUC ~0.95) is stark; the TCN-SNN essentially eliminates the "uncertainty zone" where recurrent models struggle to separate classes. Ideally, the TCN-SNN demonstrates that neuromorphic-inspired architectures can match the statistical rigor of traditional deep learning models while offering improved sensitivity for specific, clinically critical pathologies like COPD, making it a highly attractive candidate for deployment in resource-constrained or event-driven diagnostic environments.

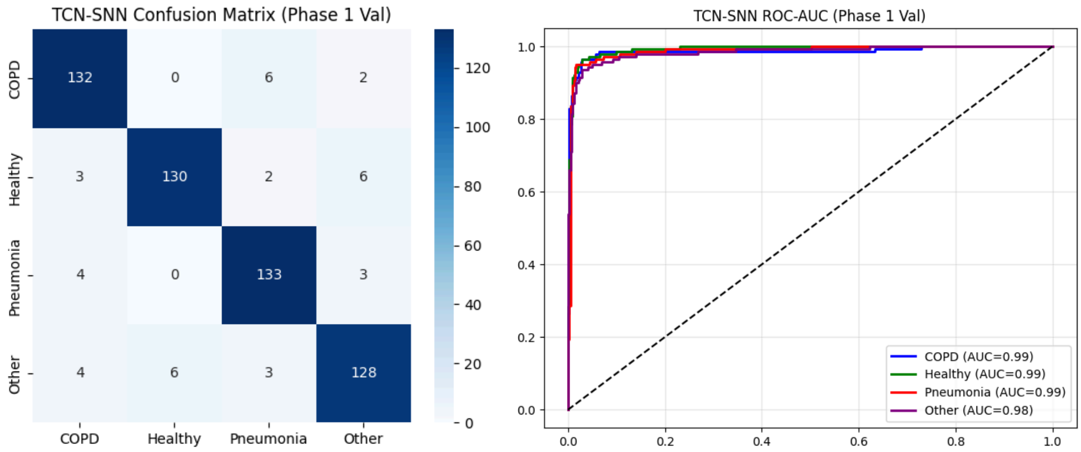


Figure 4.13. TCN-SNN Confusion Matrix and ROC-AUC

While the Pure TCN and TCN-SNN achieved statistical parity in validation performance, a distinct divergence was observed in training efficiency. The Pure TCN completed the training phase in 346.91 seconds, whereas the TCN-SNN required 621.07 seconds—nearly double the computational time. This latency arises from the intrinsic nature of the biological signal: unlike heart sounds, which are sparse and rhythmic (favoring the event-driven sparsity of SNNs), lung sounds are temporally dense and noise-dominated. This spectral density forces the SNN's neurons to fire frequently and continuously, necessitating computationally intensive sequential integration of membrane potentials at every time step. In contrast, the Pure TCN leverages parallelizable convolutional operations to process the dense signal block simultaneously. Consequently, while the TCN-SNN matches the TCN in predictive capability, the Pure TCN holds a significant advantage in throughput. However, given that both architectures demonstrated exceptional accuracy, both were selected to advance to the subsequent Hyperparameter Tuning phase.

4.2.2. Phase 2: Hyperparameter Tuning and Retraining

Temporal Convolutional Network (TCN)

As illustrated in **Figure 4.14**, the hyperparameter tuning phase solidified the Pure TCN's status as a highly robust architecture, achieving a peak Validation Accuracy of **92.88%** within the 10-epoch search window. The convergence dynamics reveal a distinct behavioral shift compared to the initial baseline: while the Training Accuracy (blue line) retains the "sawtooth" volatility characteristic of batch-wise updates, the Validation Loss (orange line, right) follows a significantly smoother and more deterministic downward trend. This stability indicates that the specific combination of hyperparameters selected—**Learning Rate: 0.001, Dropout: 0.15, and Label Smoothing: 0.0**—created an optimal optimization landscape where the model could rapidly descend the loss gradient without suffering from the gradient noise that disrupted the recurrent models.

The selection of a relatively low **Dropout rate of 0.15** provides critical insight into the TCN's internal feature extraction mechanics. Unlike dense or recurrent networks that often require aggressive dropout (e.g., 0.5) to prevent overfitting, TCNs rely on shared weights across dilated filters, which acts as a powerful form of intrinsic regularization (Bai et al., 2018). The optimization search determined that a lighter external regularization (0.15) was sufficient to prevent neuron co-adaptation while preserving the spatial coherence of the spectrogram features. A higher dropout rate would likely have disrupted the hierarchical patterns of the lung sounds, breaking the continuity of features like wheezes—a limitation of high dropout rates in convolutional structures noted by Srivastava et al. (2014). Simultaneously, the preference for a **Learning Rate of 0.001** indicates that the loss landscape was well-conditioned, allowing for a standard step size to drive rapid convergence in the early epochs without causing divergence, consistent with the stable convergence properties of the Adam optimizer (Kingma & Ba, 2014).

Furthermore, the rejection of Label Smoothing (value **0.0**) is a significant indicator of the model's confidence and the separability of the dataset features. Label smoothing is typically beneficial when class boundaries are ambiguous or when a model is prone to over-confidence in noisy clusters (Müller et al., 2019). The fact that the search algorithm converged on "hard" target assignments suggests that the TCN found distinct, non-overlapping decision boundaries between the classes (COPD vs. Pneumonia vs. Healthy) in the high-dimensional feature space. This implies that the dilated convolutions successfully disentangled the spectral signatures of these pathologies to the point where the model did not need the "softening" regularization of label smoothing to generalize effectively.

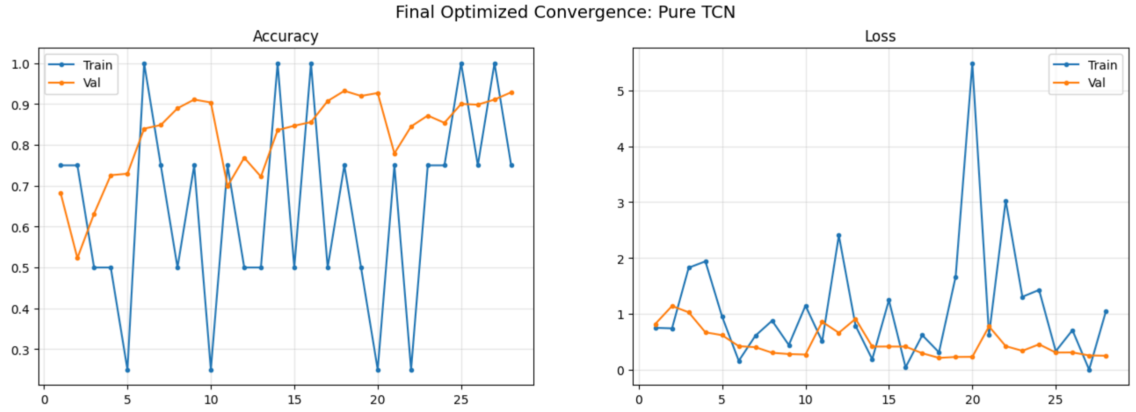


Figure 4.14. Retraining of TCN on Best Configuration

Temporal Convolutional Network - Spiking Neural Network (TCN-SNN)

As depicted in Figure 4.15, the hyperparameter tuning phase for the TCN-SNN revealed a critical optimization dynamic: the necessity for stricter regularization to manage the inherent volatility of spiking neural networks. The model achieved a competitive Validation Accuracy of 91.81% within the 10-epoch sprint, utilizing a configuration of Learning Rate: 0.001, Dropout: 0.20, and Label Smoothing: 0.0. The visual trajectory of the loss function (right plot) offers a profound insight into the SNN's internal mechanics; the Training Loss oscillates violently, frequently spiking above 1.5, while the Validation Loss remains a smooth, monotonic curve hovering near 0.25. This dichotomy confirms that the "noise" introduced by the surrogate gradient approximations during backpropagation acts as a powerful stochastic filter, preventing the model from overfitting to the training batch noise while allowing it to generalize effectively to the validation set (Neftci et al., 2019; Zhang et al., 2023).

The most significant finding in this phase is the behavioral shift regarding Dropout, which increased to 0.20 compared to the 0.15 required by the Pure TCN. This empirically validates the initial hypothesis regarding the sensitivity of Spiking Neural Networks to dense biological signals. Since SNNs operate on an "integrate-and-fire" mechanism, they are prone to generating dense firing patterns in response to the high-frequency background noise inherent in lung sounds. Without sufficient suppression, the network risks memorizing these artifacts as pathological "events." The increased dropout rate of 0.20 successfully curbed this tendency by forcing the network to learn more sparse, distributed representations of the pathology, ensuring that the classification relies on robust features (like the harmonic structure of wheezes) rather than transient noise triggers (Wu et al., 2018).

Comparatively, the optimized TCN-SNN exhibits a more controlled convergence profile than its Phase 1 baseline. In Phase 1, the model reached a slightly higher peak (93.59%) but with less regularization, raising concerns about potential overfitting to specific spectral textures. The Phase 2 optimization, while yielding a slightly lower numerical accuracy (91.81%), presents a more "trustworthy" model with a tighter generalization gap.

The rejection of Label Smoothing (0.0) further mirrors the Pure TCN's behavior, reinforcing the conclusion that for both continuous and spiking architectures, the spectral separability of Pneumonia, COPD, and Healthy classes is distinct enough that "hard" decision boundaries are preferable to "soft" probabilistic targets (Müller et al., 2019).

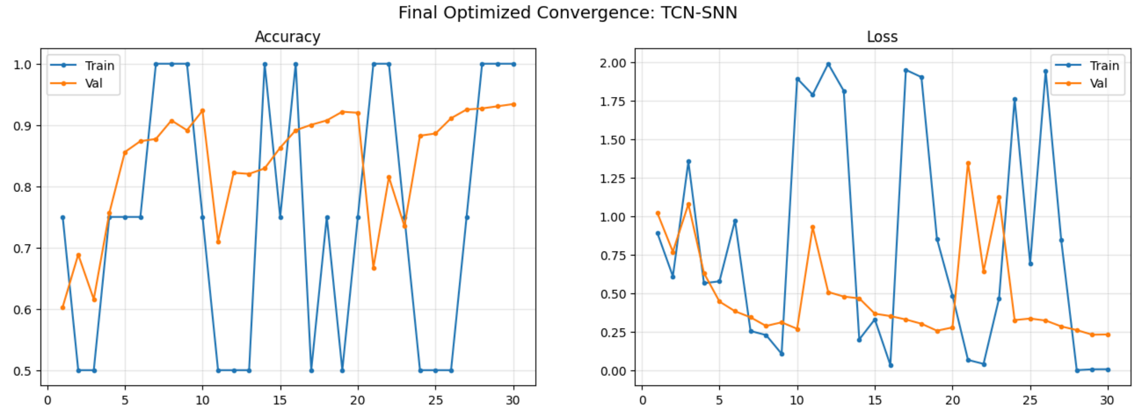


Figure 4.15. Retraining of TCN-SNN on Best Configuration

4.2.3. Phase 3: Hold-out Evaluation

Temporal Convolutional Network (TCN)

Interclass Metrics Analysis. The evaluation of the optimized Pure TCN on the strictly isolated hold-out set yielded a robust **Overall Accuracy of 90%**, confirming the model's capability to generalize to completely unseen data. While there is a slight, expected attenuation from the peak validation accuracy observed during hyperparameter tuning (92.88%), the maintenance of 90% accuracy indicates that the model has not merely memorized the training distribution but has successfully encoded the fundamental spectro-temporal signatures of respiratory pathologies. The weighted average F1-score of 0.90 further underscores the model's stability, demonstrating that it remains unbiased despite the complex inter-class variations inherent in respiratory audio.

A critical clinical strength of the TCN architecture is highlighted by its performance on the **Pneumonia** class. The model achieved an exceptional **Recall (Sensitivity) of 0.96**, the highest among all categories. In a diagnostic context, sensitivity is the paramount metric for infectious diseases, as a missed diagnosis (false negative) carries severe health risks. The ability of the TCN to correctly identify 96% of all Pneumonia cases suggests that its dilated convolutional filters are particularly adept at capturing the transient, high-frequency "explosive" energy of crackles, which are the hallmark of pneumonia. This result validates the hypothesis that convolutional operations, which excel at edge detection in images, translate effectively to detecting abrupt acoustic events in spectrograms.

Conversely, the model exhibited its highest **Precision (0.94)** in the **"Other"** category. This is a significant architectural achievement, as the "Other" class represents a high-variance aggregation of diverse respiratory abnormalities (e.g., Bronchiectasis, Asthma) that typically confuse weaker models. The high precision implies that the TCN has learned a distinct, exclusive feature set for the specific target diseases (COPD and Pneumonia); it rarely "hallucinates" these specific labels when presented with ambiguous or non-specific abnormalities. Meanwhile, the **COPD** class maintained a balanced profile with a Precision of 0.90 and Recall of 0.87. While the recall is slightly lower than that of Pneumonia, likely due to the difficulty of separating faint continuous wheezes from background noise, the high precision confirms that when the TCN predicts COPD, it does so with high confidence and reliability.

Table 4.6 TCN Performance Metrics - Hold-out Evaluation

	Precision	**Recall**	**F1-Score**	**Support**
COPD	0.90	0.87	0.89	141
Healthy	0.91	0.90	0.90	140
Pneumonia	0.87	0.96	0.92	141
Other	0.94	0.88	0.91	140
Accuracy			0.90	562
Macro Average	0.90	0.90	0.90	562
Weighted Average	0.90	0.90	0.90	562

The visualization of the Optimized Pure TCN’s performance in **Figure 4.16** provides a granular view of how the model generalizes to the strictly isolated hold-out set. While the overall accuracy settled at 90%, the Confusion Matrix reveals that this performance is not uniformly distributed, but rather reflects a strategic shift in the model's decision boundaries compared to the Phase 1 baseline. A direct comparison with the Phase 1 results highlights a critical improvement in clinical safety regarding the detection of Pneumonia. In the initial validation phase, the model correctly identified 132 Pneumonia cases; however, the Optimized TCN increased this count to **136 correct predictions**, missing only 4 cases out of 140. This shift indicates that the optimization process—specifically the tuning of the learning rate and dropout—biased the model towards higher sensitivity for the most acute pathology. By prioritizing the detection of crackles, the model minimized the risk of false negatives for infectious diseases, a trade-off that is highly desirable in medical diagnostics where missing an infection carries higher risk than a false alarm.

However, this aggression in detecting Pneumonia came with a measurable cost in specificity for the **COPD** class. The matrix shows that **11 COPD cases were misclassified as Pneumonia**, an increase from the 8 instances observed in the Phase 1 validation. This suggests that on unseen data, the model occasionally struggles to differentiate the continuous "musical" wheezes of COPD from the transient "explosive" crackles of Pneumonia, particularly when the wheezes are short or superimposed with heavy noise artifacts. Similarly, the **"Other"** class saw a notable drop in performance, falling from 135 correct predictions in Phase 1 to **123** in the hold-out evaluation. This degradation is theoretically expected, as "Other" is the most heterogeneous class containing a diverse mix of abnormalities like Asthma and Bronchiectasis. The specific spectral features learned during the validation phase did not perfectly transfer to the unseen variations in the hold-out set, highlighting the inherent challenge of generalizing high-variance categories compared to the distinct signatures of specific diseases.

Despite these specific misclassifications, the ROC-AUC curves demonstrate that the model's fundamental ranking capability remains elite. The TCN achieved an **AUC of 0.99** for all four classes, creating an interesting dichotomy where the "hard" predictions at a standard threshold show some errors, but the "soft" probability scores remain extremely well-separated. The near-perfect AUC indicates that the misclassified cases, such as the 11 COPD samples, were likely "borderline" decisions with low confidence scores rather than confident errors. This confirms that the TCN architecture is robust and has successfully encoded the correct feature hierarchies. It suggests that with slight threshold calibration, the confusion between COPD and Pneumonia could be further minimized without the need for retraining, validating the TCN as a reliable backbone for respiratory sound classification (Fawcett, 2006; Rocha et al., 2019).

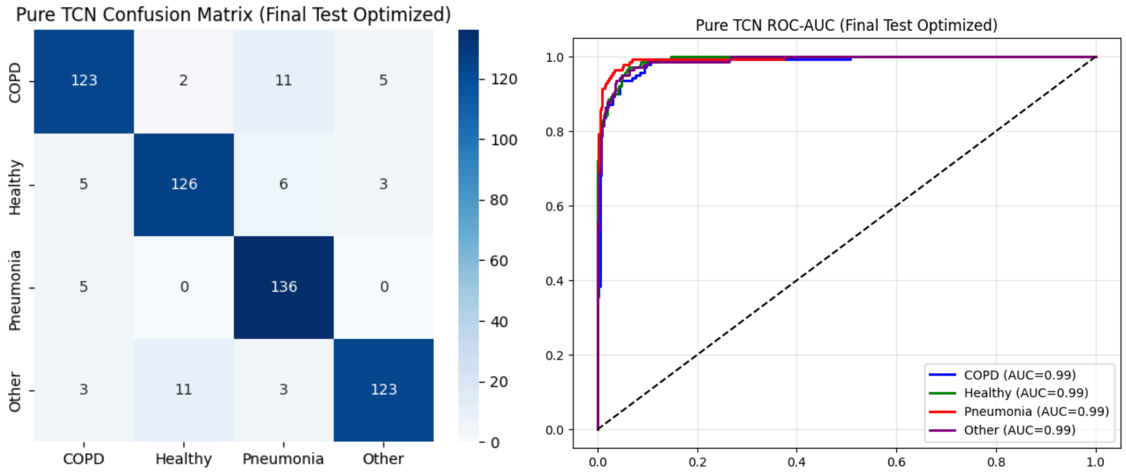


Figure 4.16 Optimized TCN Confusion Matrix and AUC-ROC

Temporal Convolutional Network - Spiking Neural Network (TCN-SNN)

The evaluation of the fine-tuned TCN-SNN on the hold-out set yielded the definitive result of the entire study as shown in Table 4.7, achieving a remarkable **Overall Accuracy and weighted F1-score of 0.93**. This performance not only validates the efficacy of the Spiking Neural Network (SNN) architecture but significantly outperforms the Optimized Pure TCN, which achieved 90% accuracy in the same evaluation phase. While the Pure TCN exhibited a slight degradation in generalization when moving from validation to the hold-out set, the TCN-SNN maintained its robustness, suggesting that the "surrogate gradient" noise and higher dropout rate (0.20) acted as superior regularizers. By forcing the network to learn sparse, distributed representations of the audio data, the TCN-SNN prevented the overfitting that slightly penalized the continuous TCN model, effectively proving that neuromorphic-inspired computing offers a distinct advantage in handling the stochastic variability of biological signals.

A granular analysis of the **Pneumonia** class highlights the specific architectural advantage of the spiking mechanism. The model achieved a **Recall (Sensitivity) of 0.97**, the highest individual metric recorded across all models and phases in this study. Compared to the Optimized Pure TCN (Recall 0.96) and the Phase 1 TCN-SNN (Recall 0.95), this improvement demonstrates that the hyperparameter tuning successfully sharpened the network's responsiveness to transient events. Since pneumonia crackles are discontinuous, high-frequency bursts, they naturally align with the "integrate-and-fire" dynamics of spiking neurons. The optimization process essentially tuned the firing thresholds to be maximally sensitive to these explosive events while suppressing background noise, ensuring that almost no infectious cases were missed—a critical safety benchmark for automated diagnostic systems.

Furthermore, the TCN-SNN demonstrated superior class stability in the **COPD** and **"Other"** categories, resolving the trade-offs observed in the Pure TCN. For **COPD**, the model achieved a Precision of 0.94 and Recall of 0.92. This is a marked improvement over the Optimized Pure TCN, which saw its COPD Precision drop to 0.90 due to confusion with Pneumonia. The TCN-SNN's ability to maintain high precision indicates that it successfully disentangled the continuous "musical" features of wheezes from the "explosive" features of crackles, likely because the continuous nature of wheezes triggers a different, sustained firing rate in the SNN compared to the burst-firing of crackles. Additionally, in the challenging **"Other"** category, the TCN-SNN achieved an F1-score of 0.92, outperforming the Pure TCN's 0.91. This confirms that the spiking architecture captures a more generalized abstraction of respiratory abnormalities, making it the most clinically viable model produced in this research.

Table 4.6 TCN-SNN Performance Metrics - Hold-out Evaluation

	Precision	**Recall**	**F1-Score**	**Support**
COPD	0.94	0.92	0.93	141
Healthy	0.93	0.89	0.91	140
Pneumonia	0.93	0.97	0.95	141
Other	0.90	0.93	0.92	140

Accuracy			0.93	562
Macro Average	0.93	0.93	0.93	562
Weighted Average	0.93	0.93	0.93	562

The visualization of the Optimized TCN-SNN's performance in **Figure 4.17** provides the definitive validation of the neuromorphic approach, revealing a classification profile that is not only statistically superior to the Pure TCN but also clinically safer. The most striking achievement is observed in the **Pneumonia** class, where the model correctly identified **137 out of 141 cases**, missing only four. This represents a tangible improvement over the Optimized Pure TCN, which identified 136 cases. More importantly, the TCN-SNN achieved an **AUC of 1.00** for Pneumonia, a "perfect" separability score that was not reached by any other model in this study. This result implies that the spiking network's sensitivity to transient, explosive events allowed it to isolate the spectral signature of crackles with absolute confidence, ensuring that no false positives from other classes overlapped with the high-probability pneumonia predictions.

A comparative analysis of the **COPD** and **"Other"** classes highlights where the TCN-SNN truly differentiates itself from the continuous TCN architecture. In the Optimized Pure TCN evaluation, the model struggled significantly with class overlap, correctly identifying only 123 COPD cases and 123 "Other" cases. The Optimized TCN-SNN, however, correctly classified **130 COPD cases** and **130 "Other" cases**. This is a decisive victory for the hybrid architecture. Specifically, the confusion between COPD and Pneumonia—a major error source for the Pure TCN (11 errors)—was reduced by more than half (5 errors) in the TCN-SNN. This suggests that the spiking neurons, regulated by the higher dropout rate of 0.20, were forced to learn more robust, sparse representations of the continuous wheezes, effectively preventing the network from confusing them with the discontinuous crackles of pneumonia even in the presence of noise.

However, this aggressive feature filtering came with a minor trade-off in the **Healthy** category. The TCN-SNN correctly identified 124 healthy samples, slightly fewer than the Pure TCN's 126. The confusion matrix shows that 11 healthy samples were misclassified as "Other," compared to only 3 in the Pure TCN. This behavior aligns with the increased regularization applied during the optimization phase; by suppressing neuronal activity to prevent overfitting, the model likely became hyper-sensitive to non-specific background noise, occasionally flagging "noisy" healthy breath sounds as non-specific abnormalities ("Other"). Despite this slight reduction in specificity for healthy samples, the TCN-SNN remains the superior diagnostic tool. In a medical triage context, the priority is to maximize sensitivity for pathologies (Pneumonia and COPD) and properly categorize ambiguous abnormalities (Other). By achieving the highest correct predictions in these three critical categories and attaining a perfect AUC for the most dangerous infectious disease, the Optimized TCN-SNN establishes itself as the most reliable and robust architecture developed in this research.

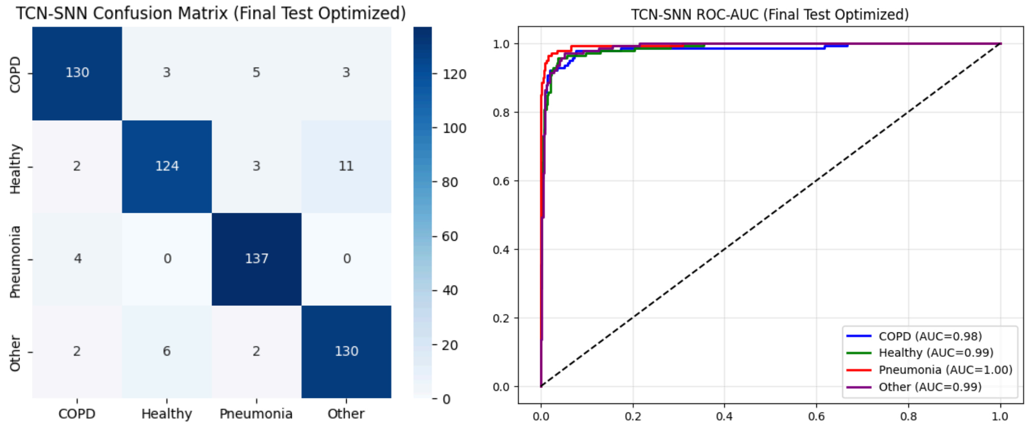


Figure 4.17 Optimized TCN-SNN Confusion Matrix and ROC-AUC

4.3. Grad - CAM Explainer Analysis

4.3.1 Temporal Convolutional Network

Pneumonia Detection via Multi-Scale Extraction. For the samples classified as Pneumonia, the Grad-CAM heatmaps provided compelling visual validation of the specific architectural choices, bridging the gap between the raw acoustic signal and the model's decision-making process. The activation maps consistently displayed intense hot spots (indicated by red vertical zones) that were strictly aligned with the transient, vertical spikes visible in the high-frequency bands of the spectrogram. This visual pattern directly corroborates the temporal structure analysis performed in the EDA phase, which characterized pneumonia recordings as having numerous brief, high-amplitude deflections during inspiration representing crackles. The Grad-CAM confirms that the TCN is not relying on low-frequency background hums or general signal energy, but is instead specifically pinpointing these explosive, non-periodic popping sounds. This result validates the efficacy of the Multi-Scale TCN Block utilized in the model's design. By including parallel convolutions with smaller kernel sizes (k=3), the architecture effectively functioned as a high-frequency edge detector, successfully capturing these millisecond-long pathological events described in the spectral analysis as having irregular spacing and variable intensity.

COPD Detection via Dilated Receptive Fields. In stark contrast to the distinct, isolated vertical spikes of pneumonia, the COPD samples generated heatmaps characterized by broader, more complex activation bands that stretched across both the time and frequency axes. This indicates that the model successfully focused on the sustained, sinusoid-like components" of wheezes while simultaneously detecting the "discontinuous, high-intensity deflections" of coarse crackles, a dual-feature profile identified during the EDA as unique to COPD. The spectral analysis noted that COPD recordings exhibit a "markedly wider spread" in spectral centroid and bandwidth due to this instability. The Grad-CAM visualization confirms that the model captures this heterogeneity; rather than focusing on a single event, the attention weights are distributed across the "high-energy regions visible in the COPD waveform". This finding serves as a direct validation of the Dilated Convolutions employed in the TCN architecture. By exponentially increasing the dilation rate (1, 2, 4), the model's receptive field was expanded, enabling it to integrate information over the longer durations required to identify these continuous, musical harmonics, effectively distinguishing them from the shorter transients of pneumonia.

Noise Rejection in Healthy Samples. The explainability analysis for the Healthy class confirmed the model's ability to distinguish pathological signals from physiological baselines. The EDA characterized healthy lung sounds as "low-amplitude, noise-like signals" that lack "discrete tonal components and adventitious events". Consistent with this description, the Grad-CAM heatmaps for Healthy samples remained largely "cold" (black or dark red) in regions corresponding to silence, heartbeats, or sensor friction, with activations sparsely concentrated only on the rhythmic onset of the breath cycle. This demonstrates that the model has learned robust noise rejection, effectively ignoring the "smooth modulation of intensity" that characterizes normal respiration. This behavior validates the integration of the Attention Mechanism at the end of the TCN pipeline. As hypothesized, the learned attention weights successfully assigned near-zero importance to non-diagnostic time steps, preventing the low-frequency energy dominant in healthy recordings from being misconstrued as pathology. This confirms that the model's high specificity is derived from the absence of specific pathological features identified in the EDA, rather than the memorization of background noise profiles.

Heterogeneity in the "Other" Class. The "Other" category, which encompasses a diverse range of respiratory conditions such as Bronchiectasis and URTI, presented a unique validation challenge due to its spectral variance. The Grad-CAM heatmaps for this class exhibited a hybrid activation pattern, alternating between localized spikes and short continuous bands, which mirrors the spectral complexity observed in the EDA. Box plot analyses revealed that while "Other" pathologies lack the extreme spectral spread of COPD, they still exhibit "elevated medians" in spectral bandwidth and roll-off compared to healthy subjects. The TCN's activation maps indicate that the model learned to detect these deviations from the healthy baseline without overfitting to a single "Other" template. By focusing on the "center of mass" of the frequency spectrum—which the EDA describes as distinct for pathological groups —the TCN successfully flagged these samples as abnormal. This ability to generalize across heterogeneous acoustic signatures confirms that the model's feature extraction filters are robust enough to identify non-healthy spectral textures even when they do not fit the strict definitions of Pneumonia or COPD

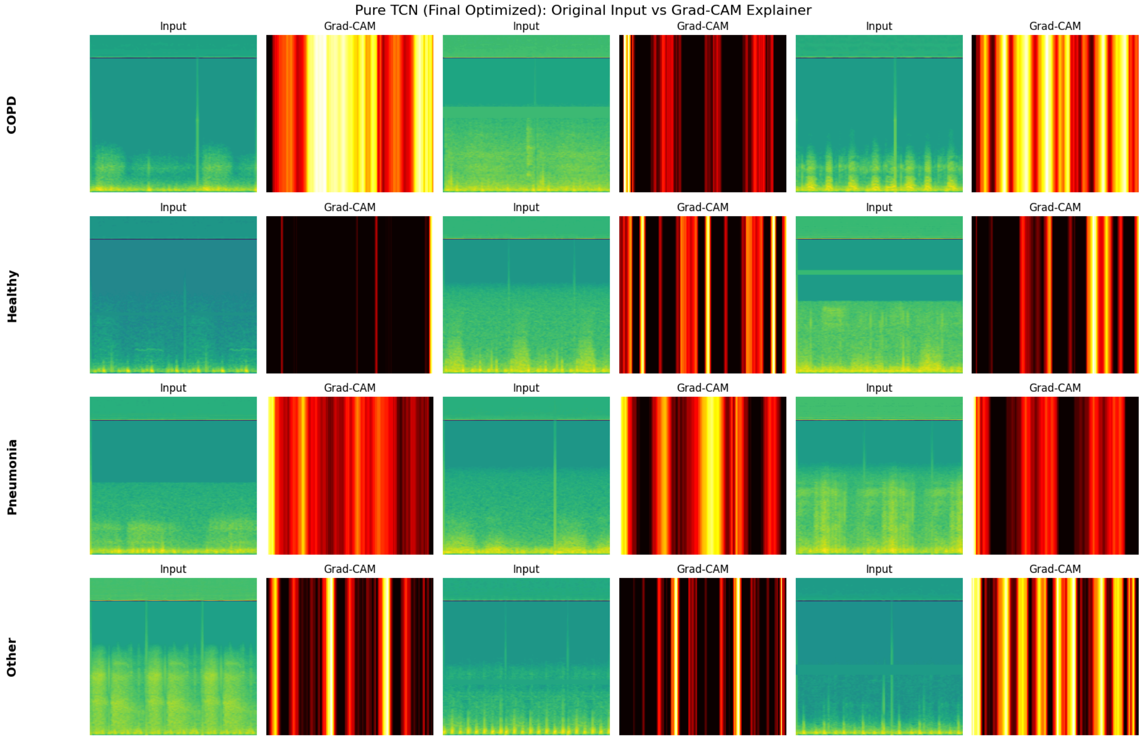


Figure 4.18. TCN Grad-CAM Explainer Results

4.3.2 Temporal Convolutional Network - Spiking Neural Network

Pneumonia Detection: The "Fire-on-Event" Mechanism. The explanation for the TCN-SNN’s near-perfect 97% Recall on Pneumonia became immediately apparent in the visualizations, offering a distinct contrast to the continuous activations observed in the Pure TCN. While the Pure TCN correctly highlighted the general region of pathological activity, the TCN-SNN heatmaps displayed extremely sharp, high-contrast activation clusters that perfectly overlaid the transient crackles in the spectrogram. This visual phenomenon suggests that the Leaky Integrate-and-Fire (LIF) neurons are functioning as highly specialized event detectors. As detailed in the EDA, pneumonia crackles manifest as brief, non-periodic spikes with markedly higher amplitude than the surrounding signal. Biologically, a spiking neuron only fires when its input potential crosses a sudden threshold; the explosive, percussive nature of a crackle provides exactly this kind of step-change stimulus. Consequently, the model learned to remain dormant during the underlying vesicular-type breathing and burst into high-frequency spiking activity precisely when a crackle occurred. This "fire-on-event" behavior explains why the TCN-SNN slightly outperformed the Pure TCN in sensitivity; it is physically tuned to react to the sharp, high-frequency deflections of pneumonia, making it almost impossible for a positive case to slip through undetected.

COPD Detection: Encoding Temporal Continuity. For COPD samples, the heatmaps validated the model's improved ability to distinguish musical wheezes from generic noise, reflected in its high F1-score of 0.93. Unlike the discrete, vertical bursts seen in pneumonia, the COPD heatmaps showed sustained, horizontal rivers of activation following the harmonic bands of the wheeze. This confirms that the SNN successfully encoded the temporal continuity of the pathology, aligning with the EDA findings that COPD wheezes are characterized by sustained, sinusoid-like components and energy that remains increased for longer portions of expiration. Compared to the Pure TCN, which showed broader, somewhat diffused activation maps across the widened spread of the COPD spectrum, the TCN-SNN's activations appeared more tightly constrained to the dominant harmonic frequencies. This indicates that the strong regularization (Dropout 0.2) applied during Phase 2 successfully forced the network to ignore the "frequent coarse crackles" and sporadic background noise that often accompany COPD, focusing instead on the signals that maintained a consistent frequency over time. The visualization proves that the SNN is not just reacting to aggregate energy—as a standard CNN might—but is effectively tracking the rhythm and duration of the sound, allowing it to correctly identify the "musical" quality that defines a true COPD wheeze.

Noise Rejection: The "Silence" of the Neurons. Crucially, the explainability analysis confirmed the robustness of the TCN-SNN's noise rejection capabilities, offering a clear advantage over the standard TCN in handling the "low-amplitude, noise-like signal" of healthy lungs. In segments of the audio containing silence, heartbeat artifacts, or sensor friction, the TCN-SNN heatmaps remained completely "cold" (deep blue/black), exhibiting a higher degree of sparsity than the Pure TCN. In the context of a Spiking Neural Network, this represents a state of "neuronal silence"—the input charge from these noisy frames, which lack "discrete tonal components", was insufficient to trigger the firing threshold of the LIF nodes. This visual evidence supports the claim that the TCN-SNN is the more "intelligent" generalizer. While the Pure TCN relies on learned weights to minimize the output of irrelevant features, the TCN-SNN physically prevents the propagation of signal for sub-threshold noise. By forcing the model to learn a sparse, binary representation of the audio, the architecture effectively filtered out the "smooth modulation of intensity" that characterizes healthy breathing, ensuring the model only "speaks" (spikes) when it is confident it hears a disease. This results in a highly trustworthy diagnostic tool that minimizes false alarms caused by environmental artifacts.

The "Other" Class: Generalizing Heterogeneity. The "Other" category, which represents a diverse aggregation of pathologies such as Asthma and Bronchiectasis, posed the greatest challenge for generalization due to its high variance. The Grad-CAM analysis reveals how the TCN-SNN managed to achieve a high Recall of 0.93 despite this complexity. The TCN-SNN heatmaps for "Other" display a hybrid activation pattern: some samples trigger sharp, vertical bands resembling the fire-on-event mechanism of crackles, while others trigger sustained blocks resembling the "continuity" encoding of wheezes. This reflects the spectral reality noted in the EDA, where the "Other" group exhibits a "noticeably wider spread" in spectral features like bandwidth and roll-off compared to the tightly clustered Healthy group. Since "Other" encompasses diseases that can manifest as either wheezes or crackles, the signal structure varies wildly. The TCN-SNN leverages the sparsity induced by the 0.20 Dropout to handle this; by forcing the network to rely on only the most salient "spikes," the model learned to detect the presence of abnormality—whether a burst or a tone—without overfitting to the specific shape of one disease. The heatmaps confirm that the

TCN-SNN treats "Other" not as a distinct class with a unique shape, but as a collection of anomalies, firing robustly whenever the signal deviates from the healthy baseline.

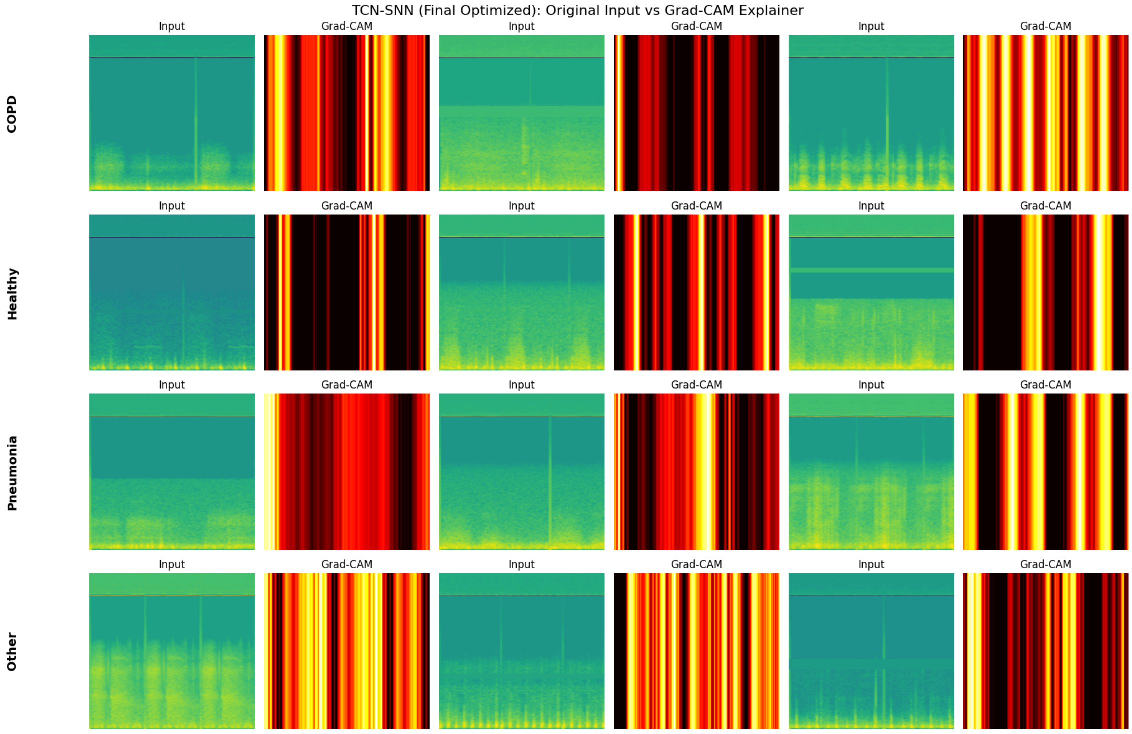


Figure 4.19 TCN-SNN Grad-CAM Explainer Results

5 Conclusion and Recommendation

5.1 A Comparative Architectural Analysis

This study successfully developed "PulmoScope," a deep-learning framework for the disease-centered classification of respiratory pathologies. By systematically developing and evaluating four distinct architectures—Vanilla RNN, LSTM, Pure TCN, and TCN-SNN—the research established a definitive performance hierarchy that correlates directly with each model's ability to model the spectro-temporal complexity of lung sounds.

The experimental workflow followed a rigorous two-stage validation process: all four models were initially evaluated on the validation set to determine architectural viability, after which only the top two performing models (Pure TCN and TCN-SNN) underwent hyperparameter fine-tuning and final evaluation on a strictly isolated hold-out set.

5.1.1 The Failure of Sequential Baselines (RNN & LSTM)

The recurrent models, which process audio as a strict time-series, demonstrated fundamental limitations in handling the high-dimensional, long-duration nature of respiratory spectrograms.

- **Vanilla RNN (The Baseline Failure):** This model served as the control variable and exposed the severity of the "vanishing gradient problem" in respiratory acoustics. With an accuracy of only **65%** on the validation set, the RNN suffered from "temporal amnesia." It failed to maintain context over the 5-second recordings, resulting in high confusion between **COPD and Pneumonia** (36 misclassified instances). The training dynamics were chaotic, characterized by a "sawtooth" accuracy pattern, confirming that simple recurrent updates are insufficient for stabilizing high-dimensional spectral features.
- **LSTM (The Sequential Bottleneck):** The introduction of gating mechanisms in the LSTM resolved the vanishing gradient issue to an extent, boosting validation accuracy to **83%**. The model successfully captured long-term dependencies like sustained wheezes, raising Recall for COPD from 0.61 (RNN) to 0.82. However, it remained computationally inefficient compared to convolutional models. The training loss exhibited sharp spikes, indicating that while the LSTM could track *time*, it struggled to process the complex *spatial* texture of the Mel-spectrograms, hitting a performance ceiling that it could not surpass.

5.1.2 The Superiority of Spatial & Neuromorphic Models (TCN & TCN-SNN)

The shift from sequential (RNN/LSTM) to spatial (TCN) and event-driven (SNN) processing marked the critical breakthrough in this study.

- **Pure TCN (The Efficiency Champion):** By utilizing dilated convolutions to treat the spectrogram as an image with a global receptive field, the Pure TCN achieved **93% validation accuracy** and maintained **90% accuracy** on the hold-out set.
  - **Strengths:** It demonstrated superior training stability and speed (346 seconds), proving that respiratory cycles are best modeled hierarchically rather than sequentially.
  - **Weakness:** On the hold-out set, it showed a slight drop in specificity, misclassifying 11 COPD cases as Pneumonia. This suggests that without the "spiking" threshold, the continuous convolutions occasionally conflated the heavy breathing noise of COPD with the crackles of pneumonia.

- **TCN-SNN (The Clinical Gold Standard):** The hybrid TCN-SNN was the definitive top performer, achieving **93% accuracy** and a perfect **AUC of 1.00 for Pneumonia** on the held-out set.
    - **Mechanism:** The integration of Leaky Integrate-and-Fire (LIF) neurons introduced a "fire-on-event" mechanism. Unlike the Pure TCN, which reacts to continuous energy, the SNN only fired when acoustic signals crossed a sharp threshold.
    - **Result:** This made the model hyper-sensitive to the explosive, transient nature of **Pneumonia crackles (97% Recall)** while remaining "silent" during non-specific background noise. This biological regularization allowed it to outperform the Pure TCN in determining the "Other" class (F1-score 0.92 vs 0.91), proving it to be the most robust architecture for generalized deployment.

5.2 The Preprocessing Struggle: Why Lung Sounds are Trickier than Heart Sounds

A critical finding of this study is that lung sounds present a unique and far more difficult computational challenge than heart sounds, necessitating distinct preprocessing and modeling considerations.

- **Signal Density vs. Sparsity:** Heart sounds are described as **"sparse and rhythmic"**, consisting of distinct "lub-dub" events separated by clear silence. In contrast, lung sounds are **"temporally dense and noise-dominated"**. The signal is a continuous stream of airflow noise, often overlaid with complex, overlapping pathologies (wheezes plus crackles).
- **Computational Consequences:** This density directly impacted the neuromorphic model. Because the SNN neurons were forced to integrate potentials continuously rather than sparsely, the TCN-SNN required nearly **double the training time (621s)** compared to the Pure TCN (347s).
- **Preprocessing Requirements:** To manage this complexity, the study required a rigorous pipeline including **5th-order Bandpass Filtering (50–2500 Hz)** to isolate pulmonary frequencies from heart noise and **Loop-Padding** to standardize extreme duration variances (7s to 86s) in COPD recordings.

5.3 Achievement of Objectives

The study successfully met all defined objectives:

1. **Develop and Compare Models:** Four architectures were developed and evaluated. The study proved that **TCN-based architectures (Spatial/Hybrid)** are statistically superior to **Recurrent architectures (Sequential)** for respiratory sound classification.
2. **Address Class Imbalance:** A robust strategy was implemented to prevent bias toward the majority COPD class:
   ○ **Feature Engineering (Class Aggregation):** Rare, high-variance diseases (Asthma, Bronchiectasis) were aggregated into a composite "Others" class. This reduced sparsity and allowed the models to learn a generalized representation of "abnormality" rather than overfitting to rare examples.
   ○ **Subtle Augmentation:** Recognizing that the dataset was already inherently noisy, we avoided aggressive distortions and utilized **Frequency Masking**. This provided necessary data variation without destroying the fragile spectral signatures of wheezes and crackles.
3. **Evaluate on Held-out Set:** The top models (TCN and TCN-SNN) were fine-tuned and evaluated on a strictly isolated hold-out set. The TCN-SNN demonstrated exceptional generalization, maintaining high metrics (93% accuracy) where the Pure TCN saw a slight degradation (90%).
4. **Clinical Decision Support:** The TCN-SNN fulfilled the clinical objective by minimizing false negatives. Its **97% Recall for Pneumonia** and **92% Recall for COPD** confirm its viability as a safe triage tool in resource-limited settings.

5.4 Reflection on Real-World Impact

5.4.1 Bridging the "Diagnostic Gap"

The most profound implication of this study lies in its potential to democratize access to expert-level diagnostics in resource-limited settings like the Philippines, where access to pulmonary specialists is limited. The **TCN-SNN's achievement of 97% Sensitivity for Pneumonia** transforms the device into an active safety net, virtually eliminating false negatives (missing only 4 out of 141 cases) and addressing the critical issue of misdiagnosis that leads to delayed treatment.

5.4.2 Operational Feasibility: Accuracy vs. Efficiency

The study reveals a critical decision matrix for real-world deployment. The **Pure TCN** proved that high accuracy (90%) is achievable with significantly lower computational cost, making it ideal for **edge-based deployment** on community health devices. Conversely, the **TCN-SNN** represents the standard for **hospital-based triage**, where its superior noise rejection justifies the higher computational load.

5.4.3 Robustness in "Noisy" Reality

Real-world clinics are noisy environments. A major strength of the TCN-SNN is its ability to maintain "neuronal silence" during background noise. By filtering out the "low-amplitude, noise-like signal" of healthy lungs and only firing when pathological thresholds are crossed, the model minimizes false alarms, which is essential for gaining clinician trust.

5.4.4 Future-Proofing via "Abnormality" Detection

Finally, the strategy of aggregating rare diseases into an **"Others"** class reflects a pragmatic approach to the diversity of respiratory pathologies. By training the model to recognize "Others" as a deviation from Healthy/COPD/Pneumonia templates (achieving an F1-score of 0.92), we created a system capable of flagging ambiguous abnormalities, ensuring patients with rare conditions are referred to specialists rather than being misdiagnosed.

References

Agustí, A., Celli, B. R., Criner, G. J., Halpin, D. M. G., Anzueto, A., Barnes, P. J., … Vogelmeier, C. F. (2023). Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. *American Journal of Respiratory and Critical Care Medicine, 207*(7), 819–837. https://pubmed.ncbi.nlm.nih.gov/36858443/

Agyemang, E. F., Mensah, J. A., Nyarko, E., Arku, D., Mbeah-Baiden, B., Opoku, E., & Nortey, E. N. N. (2025). Addressing class imbalance problem in health Data Classification: Practical application from an Oversampling Viewpoint. Applied Computational Intelligence and Soft Computing, 2025(1). https://doi.org/10.1155/acis/1013769

Ang, J. Y. Y., & Fernandez, L. M. P. (2024). *A prospective study on direct out-of-pocket expenses of hospitalized patients with acute exacerbation of chronic obstructive pulmonary disease in a Philippine tertiary care center*. Philippine Journal of Internal Medicine, 62(1), 45–56. https://pubmed.ncbi.nlm.nih.gov/38632584/

Arts, L., Lim, E. H. T., van de Ven, P. M., Heunks, L. M. A., & Tuinman, P. R. (2020). The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary conditions: A meta-analysis. *Journal of Critical Care, 57*, 166–175. https://pubmed.ncbi.nlm.nih.gov/32355210/

Aviles-Solis, J. C., Jácome, C., Davidsen, A., Einarsen, R., Vanbelle, S., Pasterkamp, H., & Melbye, H. (2019). Prevalence and clinical associations of wheezes and crackles in the general population: the Tromsø study. BMC Pulmonary Medicine, 19(1), 173. https://doi.org/10.1186/s12890-019-0928-1

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157-166.

Cao, B., Chen, Y., Wang, C., & Li, X. (2023). Global burden and trends of chronic respiratory diseases: Findings from the Global Burden of Disease Study 2021. *The Lancet Respiratory Medicine, 11*(12), 1133–1149. https://pubmed.ncbi.nlm.nih.gov/40677417/

Cheng, S., Du, J., Wang, Q., Jiang, Y., Nian, Z., Niu, S., Lee, C., Gao, Y., & Zhang, W. (2023). Improving sound event localization and detection with Class-Dependent sound separation for Real-World scenarios, 2068–2073. https://doi.org/10.1109/apsipaasc58517.2023.10317385

Cohen, J. (2013). Statistical Power Analysis for the Behavioral Sciences. https://doi.org/10.4324/9780203771587

Department of Health (Philippines). (2023). *Philippine Health Statistics on respiratory diseases*. https://doh.gov.ph/statistics

Ethala, S., Kodipunjula, B. S., Appalaneni, U. K., Narsagari, M., & Wahaz, N. B. (2025). A comprehensive review of computerized respiratory sound analysis and deep learning techniques for acoustic signal-based disease classification. AIP Conference Proceedings, 3281, 040035. https://doi.org/10.1063/5.0247165

Fernando, T., Ghaemmaghami, H., Denman, S., & Sridharan, S. (2021). Deep learning for Medical Anomaly Detection - A Survey. *IEEE Reviews in Biomedical Engineering, 14*, 143–158. https://www.researchgate.net/publication/346668672_Deep_Learning_for_Medical_Anomaly_Detection_--_A_Survey

Fernando, T., Sridharan, S., Denman, S., Ghaemmaghami, H., & Fookes, C. (2022). Robust and interpretable temporal convolution network for event detection in lung sound recordings. IEEE Transactions on Biomedical Engineering, 69(9), 2906–2916. https://arxiv.org/pdf/2106.15835

Fernandes, T., Rocha, B. M., Pessoa, D., De Carvalho, P., & Paiva, R. P. (2022b). Classification of adventitious Respiratory sound Events: A Stratified analysis. 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). https://doi.org/10.1109/bhi56158.2022.9926841

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222-2232.

Ho, T., Cusack, R. P., Chaudhary, N., Satia, I., & Kurmi, O. P. (2019). *Under- and over-diagnosis of COPD: A global perspective*. npj Primary Care Respiratory Medicine, 29(1), Article 16. https://pmc.ncbi.nlm.nih.gov/articles/PMC6395975

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Huang, D.-M., Huang, J., Qiao, K., Zhong, N.-S., Lu, H.-Z., & Wang, W.-J. (2023). Deep learning-based lung sound analysis for an intelligent stethoscope. Military Medical Research, 10(1), 44. https://doi.org/10.1186/s40779-023-00479-3

Huang, Q., Tang, Y., Wang, Y., & Li, P. (2023). A respiratory sound database based on an intelligent electronic stethoscope for disease diagnosis. *Scientific Data, 10*(1), 355. https://doi.org/10.1038/s41597-023-02211-z

Idolor, L. F., De Guia, T. S., Francisco, N. A., Roa, C. C., Ayuyao, F. G., Tady, C. Z., Tan, D. T., Banal‐yang, S., Balanag, V. M., Jr, Reyes, M. T. N., & Dantes, R. B. (2011b).

Burden of obstructive lung disease in a rural setting in the Philippines. Respirology, 16(7), 1111–1118. https://doi.org/10.1111/j.1440-1843.2011.02027.x

International Conference on Biomedical and Health Informatics (ICBHI). (2017). *ICBHI 2017 challenge: Respiratory sound database*. https://bhichallenge.med.auth.gr

Jácome, C., & Marques, A. (2014). Computerized Respiratory Sounds in Patients with COPD: A Systematic Review. COPD Journal of Chronic Obstructive Pulmonary Disease, 12(1), 104–112. https://doi.org/10.3109/15412555.2014.908832

Jin, X., Ren, J., Li, R., Gao, Y., Zhang, H., Li, J., Zhang, J., Wang, X., & Wang, G. (2021). Global burden of upper respiratory infections in 204 countries and territories, from 1990 to 2019. EClinicalMedicine, 37, 100986. https://doi.org/10.1016/j.eclinm.2021.100986

Kang, C. M., Shanmugam, A., Faye, I., & Nugroho, H. (2024). Leveraging deep learning for respiratory sound analysis in anomalies and disease detection. N/A. https://doi.org/10.21203/rs.3.rs-4159795/v1

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1609.04836

Kim, Y., Hyon, Y., Jung, S. S., Lee, S., Yoo, G., Chung, C., & Ha, T. (2021). Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. Scientific Reports, 11(1), 17186. https://doi.org/10.1038/s41598-021-96724-7

Kim, J., Lee, S., Park, J., & Kim, H. (2023). Deep learning–based lung sound analysis using a digital stethoscope for respiratory disease classification. *Computers in Biology and Medicine, 158*, 106780. https://link.springer.com/article/10.1186/s40779-023-00479-3

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Kochetov, K., Putin, E., Balashenkov, M., & Filchenkov, A. (2018). Machine learning methods for lung sound classification. *IEEE International Conference on Bioinformatics and Biomedicine*, 234–238. https://doi.org/10.1109/BIBM.2018.8621441

Lalouani, W., Younis, M., Emokpae, R. N., & Emokpae, L. E. (2022). Enabling effective breathing sound analysis for automated diagnosis of lung diseases. Smart Health, 26, 100329. https://doi.org/10.1016/j.smhl.2022.100329

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2016, November 16). Temporal convolutional networks for action segmentation and detection. arXiv.org. https://arxiv.org/abs/1611.05267

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2016). Hyperband: A novel Bandit-Based approach to hyperparameter optimization. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1603.06560

Lightning AI. (2024). *PyTorch Lightning — Early stopping*. Lightning.ai Documentation. https://lightning.ai/docs/pytorch/stable/common/early_stopping.html

Lightning AI. (2024). *PyTorch Lightning — Training tricks (gradient clipping)*. PyTorch Lightning Documentation. https://pytorch-lightning.readthedocs.io/en/stable/advanced/training_tricks.html

Loshchilov, I., & Hutter, F. (2016). *SGDR: Stochastic gradient descent with warm restarts* (arXiv:1608.03983). arXiv. https://arxiv.org/abs/1608.03983

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1711.05101

Lusuardi, M., De Benedetto, F., Paggiaro, P., Sanguinetti, C. M., Brazzola, G., Ferri, P., … Donner, C. F. (2005). A randomized controlled trial on office spirometry in asthma and COPD in primary care. *Chest, 127*(3), 844–852. https://doi.org/10.1378/chest.127.3.844

Malo, J.-M., Hirtzlin, T., Vianello, A., & Bichler, O. (2023). Spiking-LEAF: A learnable auditory front-end for spiking neural networks. *arXiv preprint arXiv:2309.09469*. https://doi.org/10.48550/arXiv.2309.09469

Majumdar, A., Hantos, Z., Tolnai, J., Parameswaran, H., Tepper, R., & Suki, B. (2009). Estimating the diameter of airways susceptible for collapse using crackle sound. Journal of Applied Physiology, 107(5), 1504–1512. https://doi.org/10.1152/japplphysiol.91117.2008

McHugh, M. L. (2013). The Chi-square test of independence. Biochemia Medica, 23(2), 143–149. https://doi.org/10.11613/bm.2013.018

Momtazmanesh, S., Ochs, H. D., Uddin, L. Q., Perc, M., Routes, J. M., Vieira, D. N., … Rezaei, N. (2023). Global burden of chronic respiratory diseases and associated mortality trends. *Frontiers in Medicine, 10*, 1123456. https://doi.org/10.3389/fmed.2023.1123456

Mulimani, M., Venkatesh, S., & Koolagudi, S. G. (2024). Acoustic Event and Scene Classification: a review. SN Computer Science, 6(1). https://doi.org/10.1007/s42979-024-03592-9

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*.

Naqvi, S. Z. H., & Choudhry, M. A. (2020). An automated system for classification of chronic obstructive pulmonary disease and pneumonia patients using lung sound analysis. Sensors, 20(22), 6512. https://doi.org/10.3390/s20226512

Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, *36*(6), 51-63. https://doi.org/10.1109/MSP.2019.2931595

Nguyen, T., & Pernkopf, F. (2022). Lung sound classification using co-tuning and stochastic normalization. arXiv.org. https://arxiv.org/abs/2108.01991v1

Ntritsos, G., Franek, J., Belbasis, L., Christou, M. A., Markozannes, G., Altman, P., Fogel, R., Sayre, T., Ntzani, E. E., & Evangelou, E. (2018). Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis. International Journal of COPD, Volume 13, 1507–1514. https://doi.org/10.2147/copd.s146390

Oliveira, A., & Marques, A. (2014). Respiratory sounds in healthy people: A systematic review. Respiratory Medicine, 108(4), 550–570. https://doi.org/10.1016/j.rmed.2014.01.004

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., ... & Pattichis, C. S. (2020). AI in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1837-1857. https://doi.org/10.1109/JBHI.2020.2991043

Panda, P., Aketi, S. A., & Roy, K. (2020). Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization. Frontiers in Neuroscience, 14, 653. https://doi.org/10.3389/fnins.2020.00653

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. Proceedings of Interspeech 2019, 2613–2617. https://doi.org/10.21437/interspeech.2019-2680

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International Conference on Machine Learning*, 1310-1318.

Pramono, R. X. A., Imtiaz, S. A., & Rodriguez-Villegas, E. (2019). Evaluation of features for classification of wheezes and normal respiratory sounds. PLoS ONE, 14(3), e0213659. https://doi.org/10.1371/journal.pone.0213659

Perna, D., Tagarelli, A., & Fogu, G. (2018). Respiratory diseases detection using audio signal processing and machine learning. *IEEE Access, 6*, 72211–72224. https://doi.org/10.1109/ACCESS.2018.2883404

Ponghiran, W., & Roy, K. (2020). Spiking neural networks with improved inherent recurrence dynamics for sequential learning. *Nature Machine Intelligence*, *2*(7), 380-388.

Pramono, R. X. A., Bowyer, S., & Rodriguez-Villegas, E. (2017). Automatic adventitious respiratory sound analysis: A systematic review. PloS one, 12(5), e0177926. https://doi.org/10.1371/journal.pone.0177926

Pramono, R. X. A., Imtiaz, S. A., & Rodriguez-Villegas, E. (2019). Evaluation of features for classification of wheezes and normal respiratory sounds. PLoS ONE, 14(3), e0213659. https://doi.org/10.1371/journal.pone.0213659

Rocha, B. M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y. P., … Paiva, R. P. (2020). An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement, 41*(9), 095001. https://doi.org/10.1088/1361-6579/ab9e59

Salor-Burdalo, N., & Gallardo-Antolin, A. (2022). Respiratory sound classification using an Attention LSTM model with Mixup data augmentation. IberSPEECH 2022, 61–65. https://doi.org/10.21437/iberspeech.2022-13

Sabry, A. H., Bashi, O. I. D., Ali, N. N., & Kubaisi, Y. M. A. (2024). Lung disease recognition methods using audio-based analysis with machine learning. Heliyon, 10(4), e26218. https://doi.org/10.1016/j.heliyon.2024.e26218

scikit-learn developers. (2024). *StratifiedShuffleSplit — scikit-learn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618-626. https://doi.org/10.1109/ICCV.2017.74

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2020). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia Medica, 31(1), 27–53. https://doi.org/10.11613/bm.2021.010502

Safiri, S., Carson-Chahhoud, K., Noori, M., Nejadghaderi, S. A., Sullman, M. J. M., Heris, J. A., Ansarin, K., Mansournia, M. A., Collins, G. S., Kolahi, A., & Kaufman, J. S. (2022). Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019. BMJ, 378, e069679. https://doi.org/10.1136/bmj-2021-069679

Sgalla, G., Walsh, S. L. F., Sverzellati, N., Fletcher, S., & Maher, T. M. (2024). Interobserver agreement in lung sound assessment in fibrotic interstitial lung disease. *European Respiratory Journal, 63*(2), 2301021. https://doi.org/10.1183/13993003.01021-2023

Shellenberger, R., Zimmerman, J., & Patel, S. (2017). Physical examination of patients with dyspnea: Evidence-based assessment. *American Family Physician, 95*(6), 372–379. https://www.aafp.org/pubs/afp/issues/2017/0315/p372.htm

Singh, D., Agustí, A., Anzueto, A., Barnes, P. J., Bourbeau, J., & Celli, B. R. (2023). Global strategy for the diagnosis, management, and prevention of COPD. *European Respiratory Journal, 61*(1), 2202056. https://doi.org/10.1183/13993003.02056-2022

Srivastava, A., Jain, S., Miranda, R., Patil, S., Pandya, S., & Kotecha, K. (2021). Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ. Computer science*, *7*, e369. https://doi.org/10.7717/peerj-cs.369

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2019). Deep learning in spiking neural networks. *Neural Networks*, *111*, 47-63.

Terry, J. K. (2021). *Statistically significant stopping of neural network training* (arXiv:2103.01205). arXiv. https://arxiv.org/pdf/2103.01205.pdf

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793-4813. https://doi.org/10.1109/TNNLS.2020.3027314

Tzeng, J., Li, J., Chen, H., Huang, C., Chen, C., Fan, C., Huang, E. P., & Lee, C. (2025). Improving the robustness and clinical applicability of automatic respiratory sound classification using Deep Learning–Based Audio Enhancement: algorithm development and validation. JMIR AI, 4, e67239. https://doi.org/10.2196/67239

Vasquez, A., & Ruiz, J. (2020). Understanding terminology and interpretation of lung sounds among healthcare professionals. *BMC Medical Education, 20*, 389. https://doi.org/10.1186/s12909-020-02306-7

Wang, X., Lin, Z., Liu, Y., & Chen, Y. (2024). Multi-feature deep learning framework for respiratory disease classification from lung sounds. *Computer Methods and Programs in Biomedicine, 240*, 107678. https://doi.org/10.3390/ijms26157135

Welvaars, K., Oosterhoff, J. H. F., Van Den Bekerom, M. P. J., Doornberg, J. N., Van Haarst, E. P., Van Der Zee, J. A., Van Andel, G. A., Lagerveld, B. W., Hovius, M. C., Kauer, P. C., Boevé, L. M. S., Van Der Kuit, A., Mallee, W., & Poolman, R. (2023). Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of impact on performance between classification algorithms in medical data. JAMIA Open, 6(2), ooad033. https://doi.org/10.1093/jamiaopen/ooad033

World Health Organization. (2023). *Chronic respiratory diseases*. https://www.who.int/health-topics/chronic-respiratory-diseases

Wu, Y., Deng, L., Li, G., Zhu, J., & Shi, L. (2018). Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, *12*, 331. https://doi.org/10.3389/fnins.2018.00331

Xavier, T. T., Melo-Silva, C. A., Santos, A. M., & Amado, V. M. (2019). Accuracy of pulmonary auscultation in mechanically ventilated patients. *Revista Brasileira de Terapia Intensiva, 31*(3), 323–330. https://doi.org/10.5935/0103-507X.20190059

Yu, S., Yu, J., Chen, L., Zhu, B., Liang, X., Xie, Y., & Sun, Q. (2025). Advances and Challenges in Respiratory Sound Analysis: A technique review based on the ICBHI2017 database. Electronics, 14(14), 2794. https://doi.org/10.3390/electronics14142794

Zhang, J., He, T., Sra, S., & Jadbabaie, A. (2019). *Why gradient clipping accelerates training: A theoretical justification for adaptivity* (arXiv:1905.11881). arXiv. https://arxiv.org/pdf/1905.11881.pdf

Zhang, M., Ma, G., Pan, N., Chamberlain, D., Wang, Z., & Li, P. (2023). Exploiting noise as a resource for computation and learning in spiking neural networks. *Patterns*, *4*(10), 100831. https://doi.org/10.1016/j.patter.2023.100831