Multimodal Deep Learning for Tuberculosis Detection Using Cough Audio and Clinical Data with Health Acoustic Representations (HeAR)

Rinaldi Anwar Buyung*, Widi Nugroho
Department of Data Science, Seleris Meditekno Internasional, Jakarta, Indonesia

***Abstract*—Tuberculosis (TB) remains a significant global health challenge, necessitating rapid and accessible screening methods. This study proposes a multimodal deep learning model for non-invasive TB detection by fusing acoustic features from cough sounds with clinical metadata. We utilize the pre-trained Health Acoustic Representations (HeAR) model as a powerful backbone to extract features from mel-spectrograms of cough audio. These acoustic features are combined with clinical data, including sex, age, and key symptoms through a late-fusion architecture. The model was trained and evaluated on a balanced dataset of 16,000 samples derived from the CODA TB DREAM Challenge dataset. Our proposed multimodal approach achieved a high overall accuracy of 90% on the unseen test set, with balanced precision, recall, specificity, and F1-scores of 0.90 for both TB-positive and non-TB classes. These results demonstrate the effectiveness of using cough sound as a non-invasive vocal biomarker, amplified by combining advanced acoustic representations with clinical context. This highlights the potential of our method as a robust, low-cost, and scalable tool for early TB screening.**

***Keywords*—Tuberculosis; cough detection; Health Acoustic Representation; multimodal; vocal biomarker**

I. Introduction

Tuberculosis (TB) continues to be a devastating global health crisis. It is caused by the bacterium Mycobacterium tuberculosis, the disease primarily affects the lungs (pulmonary TB) and spreads through the air when an infected person coughs, sneezes, or speaks [1]. In 2023 alone, an estimated 10.8 million people fell ill with the disease, and it claimed the lives of 1.25 million people, reaffirming its status as a leading cause of death from a single infectious agent worldwide. The epidemic's immense scale underscores the critical need for global efforts to combat TB. These efforts align with the United Nations Sustainable Development Goal of ending the epidemic by 2030 [2]. The burden of this disease is unevenly distributed, with the majority of cases occurring in developing countries. Indonesia ranks as the country with the second-highest TB burden in the world after India, which underscores the urgency to develop more effective, rapid, and accessible screening and diagnostic methods within the country [3].

Conventional diagnostic methods for TB, such as sputum smear microscopy, bacterial culture, and rapid molecular tests (e.g., Xpert MTB/RIF), often have limitations. Sputum smear microscopy has variable sensitivity and may fail to detect cases with a low bacterial load. Bacterial culture, while being the gold standard, requires several weeks to yield results, which can delay treatment and increase the risk of transmission [4]. Meanwhile, molecular tests and radiological examinations like chest X-rays, although more accurate, require expensive laboratory infrastructure, specialized equipment, and expert personnel that are not always available in primary healthcare facilities or remote areas. These limitations hinder efforts for early detection and massive-scale containment of TB spread [5].

The pathological progression of pulmonary TB significantly alters the respiratory system's mechanics and acoustics. The formation of granulomas, inflammation of the airways, and accumulation of fluid or sputum change the physical properties of the lungs and trachea. These changes directly impact the sound produced during a cough, which is a forceful expulsion of air [6]. Consequently, coughs from TB-infected individuals can exhibit distinct acoustic patterns, such as variations in spectral energy, frequency components, and temporal characteristics compared to those from healthy individuals [7]. This acoustic signature provides a physiological basis for using cough sound as a non-invasive biomarker for TB screening [8].

Along with advancements in digital technology, sound analysis using artificial intelligence (AI) has emerged as a promising alternative solution. Deep learning, a subset of machine learning, enables systems to learn from data and understand complex concepts without explicit programming [9]. By learning from experience, computers can autonomously perform tasks, making them highly suitable for analyzing complex patterns in medical data. In previous studies, for instance, deep learning particularly Convolutional Neural Networks (CNNs) was successfully employed to detect diabetic retinopathy using retinal fundus images [10], [11]. Similarly, cough sounds, as a primary symptom of TB, contain rich acoustic information and can serve as a non-invasive vocal biomarker [12]. This capability offers the potential for a low-cost, rapid, and remotely deployable screening tool using common devices such as smartphones [13].

To maximize the potential of audio analysis, this study proposes the use of Health Acoustic Representations (HeAR), a foundation model developed by Google AI specifically for health-related sounds. HeAR is built on a transformer-based architecture and employs a self-supervised learning strategy, having been pre-trained on over 300 million two-second audio clips encompassing sounds like coughs and breaths. This

*Corresponding author.

extensive pre-training enables HeAR to generate powerful and generalized acoustic embeddings. In benchmark evaluations, simple linear classifiers trained on HeAR embeddings achieved state-of-the-art or competitive performance across various health tasks, including the detection of conditions like COVID-19 and the inference of smoking status from cough recordings [14].

By leveraging these rich representations as a backbone, this study aims to build a robust classifier for TB [15]. Furthermore, this study adopts a multimodal approach by not only relying on cough sounds but also integrating essential clinical data such as sex, age, and key symptoms (hemoptysis, night sweats, weight loss, and fever). The combination of advanced acoustic features from HeAR and contextual clinical information is expected to significantly improve detection accuracy and reliability, leading to a more holistic and accurate screening model [16].

This study makes several key contributions. First, we introduce a multimodal deep learning architecture that effectively fuses powerful pre-trained acoustic features from the HeAR model with essential clinical metadata. Second, we validate our approach on a large, balanced dataset, demonstrating its robustness and high accuracy. The remainder of this paper is organized as follows: Section II reviews related work in acoustic-based TB detection. Section III details the dataset, preprocessing steps, and our proposed model architecture. Section IV presents the experimental results and analysis. Finally, Section V discusses the clinical implications of our findings. Section VI concludes the study, and outlines directions for future research.

II. Related Work

The use of cough sound analysis as a low-cost, non-invasive tool for tuberculosis (TB) screening has been an active area of research. Early studies established the foundational potential of this approach by applying traditional machine learning models to handcrafted acoustic features. For instance, Botha et al. employed statistical classifiers like Logistic Regression on short-term spectral features combined with clinical data, achieving a sensitivity of 96% at a specificity of 72% on a dataset of 518 coughs. While pioneering, this work relied on manually engineered features (e.g., MFCC), which may not fully capture the complex, subtle acoustic patterns indicative of TB [17].

Subsequent research shifted towards deep learning to automate feature extraction and improve performance. Xu et al. developed a sophisticated fusion model using a Bidirectional Long Short-Term Memory (Bi-LSTM) network combined with a Convolutional Neural Network (CNN). Their approach, which fused traditional features like MFCC and Zero-Crossing Rate (ZCR) with features learned from spectrograms, reported a very high accuracy of 96.33%. However, this impressive result was obtained from a relatively small and homogenous dataset of 456 coughs from a limited number of participants. This raises important questions about the model's scalability and ability to generalize to more diverse, real-world populations [18].

In contrast, recent studies have demonstrated the power of simpler architecture when applied to large-scale datasets. Yadav et al. utilized a massive dataset of over 500,000 cough recordings and found that a simple 1D CNN trained on Mel-Frequency Cepstral Coefficients (MFCCs) achieved a high accuracy of 91%, exceeding the WHO's requirements for a screening test. This work underscores that with sufficient data, less complex models can be highly effective. Similarly, Kafentzis et al. also worked with a large, real-world dataset, achieving a respectable Area Under the Curve (AUC) of approximately 0.80 by combining audio features with clinical metadata, further highlighting the importance of a multimodal approach [19].

Addressing the need for larger and more diverse datasets, Kafentzis et al. utilized a substantial dataset of over 9,000 coughs collected "in the wild" via a mobile application. Using a combination of spectral features and clinical metadata, their statistical models achieved a respectable Area Under the Curve (AUC) of approximately 0.80 [20]. Their work highlighted the feasibility of large-scale data collection but also demonstrated that performance on noisy, real-world data remains a significant challenge, suggesting that more powerful feature representation is needed. A summary of these key studies is presented in Table I.

TABLE I. State-of-the-Art

Study	Methodology	Audio Features	Dataset	Evaluation Metrics
[17]	Logistic Regression	MFCC, Log Spectral Energies, Clinical Data	518 Coughs	Sensitivity = 96% Specificity = 72%
[18]	BiLSTM + Conv2D	MFCC, ZCR, RMS, Chroma Cens, Short Time Energy, Spectrogram	456 Coughs	Accuracy = 96.33% Specificity = 94.99%
[19]	1D CNN	MFCC	502,252 Coughs	Accuracy = 91% AUC = 0.8
[20]	CNN	LLDs, Clinical Data	9772 Coughs	AUC = 0.8

While these studies have established the potential of cough sound analysis for TB screening, clear research gaps remain. Previous works have either relied on handcrafted features or custom-trained CNNs. Furthermore, these investigations were often conducted on considerably smaller datasets, which may not capture the full complexity of acoustic biomarkers and could limit the generalizability of their findings.

To our knowledge, the application of a specialized, pre-trained health acoustic foundation model like HeAR within a multimodal framework on a large-scale dataset for TB detection has not yet been thoroughly investigated. This study aims to fill these gaps by evaluating the performance of HeAR as a powerful feature extractor, combined with clinical data on a large and diverse dataset of 16,000 samples, to potentially set a new benchmark for accuracy and robustness in non-invasive TB screening.

III. Data and Methodology

This section details the data sources and methodological approaches employed to address the core objectives of this research. Fig. 1 illustrates the procedural framework guiding this study.

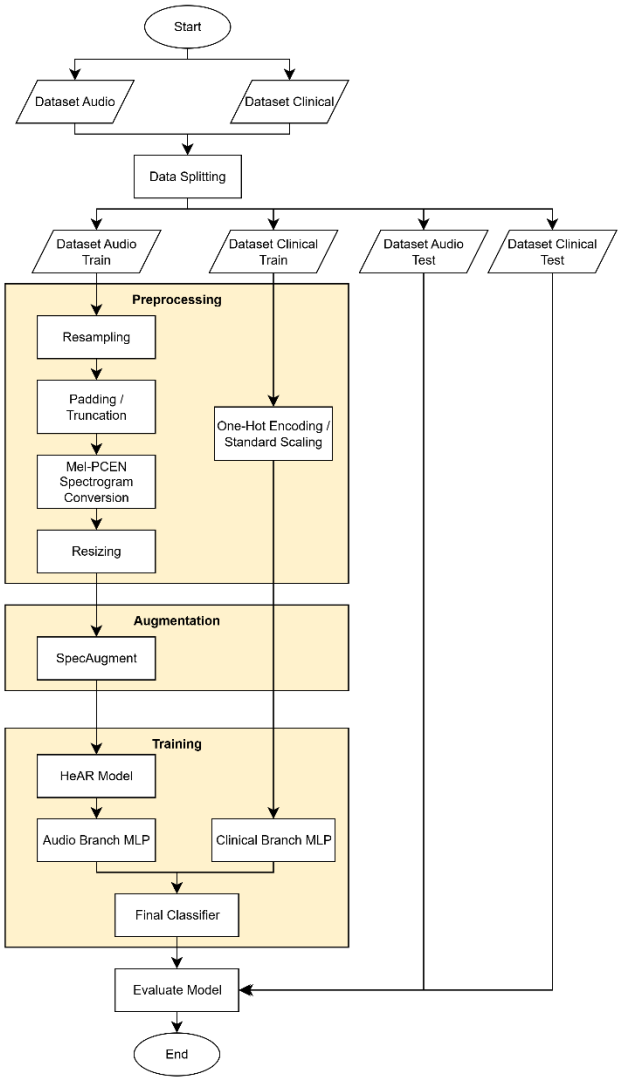

Fig. 1. Research workflow.

A. Dataset

The dataset employed in this research is the CODA TB DREAM Challenge dataset [21], a comprehensive collection of solicited cough recordings designed to advance TB triage testing. The dataset comprises 29,768 cough recordings from individuals across seven countries: India, the Philippines, South Africa, Uganda, Vietnam, Tanzania, and Madagascar.

Each recording is accompanied by detailed metadata, including the participant's confirmed TB status (positive or negative) and a range of demographic and clinical information.

To address the class imbalance often present in medical datasets, we performed random undersampling on the majority class. This process resulted in a balanced final dataset for our study, consisting of 8,000 samples for the TB-positive class and 8,000 samples for the TB-negative class, thereby preventing potential model bias towards the more frequent class. The dataset is illustrated in Fig. 2. While this strategy is effective for balancing classes, we acknowledge that it may discard potentially useful data from the majority class. Alternative techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE) or using focal loss during training, were considered but random undersampling was chosen for its simplicity and computational efficiency in this initial large-scale study. The richness and scale of this balanced 16,000-sample dataset, sourced from seven diverse countries, provide a robust foundation for training a generalizable AI-based screening tool.

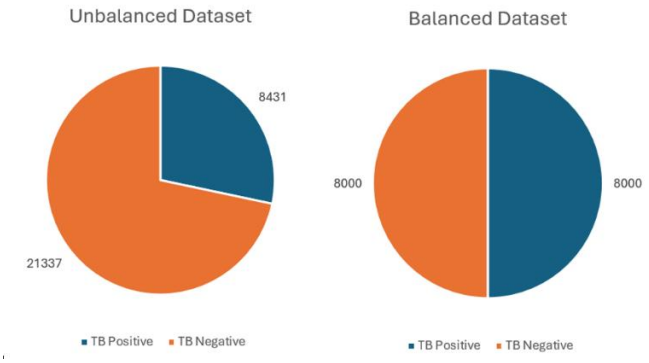

Fig. 2. Dataset distribution.

For this study, the TB status will serve as the primary target label for classification. To develop our multimodal model, we will utilize both the cough audio recordings, and the following supplementary clinical data provided in the metadata: sex, age, and the presence of key TB-related symptoms, namely hemoptysis (coughing up blood), night sweats, weight loss, and fever. The richness and diversity of this dataset make it an ideal resource for training and validating a robust, generalizable AI-based screening tool for tuberculosis.

B. Data Preprocessing and Augmentation

This study applied several data preprocessing techniques, including the following:

- Resampling: This process involves adjusting the sampling rate to match the model's requirements, ensuring input consistency, reducing memory usage, and improving computational efficiency [22]. In this study, we used a sampling rate of 16,000 Hz.
- Padding and Truncation: Padding and truncation are processes used to standardize the duration of audio samples, ensuring that they can be uniformly processed by the model. Padding involves adding zero values to audio samples that are shorter than the target length, whereas truncation trims samples that exceed the specified duration. In this study, the audio length was set to 2 seconds according to the model specification.
- Mel-PCEN Spectrogram Conversion: To create a suitable input representation for the model, the standardized waveforms are converted into a mel-spectrogram format. This process involves applying the Short-Time Fourier Transform (STFT) to decompose

the audio into its frequency components over time, as defined in Eq. (1).

$$X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \quad (1)$$

where, $x[n]$ is the input signal, $w[n]$ is the window function, and the output $X(m,w)$ is the complex value for the frequency bin w at time frame m. Subsequently, the spectrogram is mapped onto the Mel scale, which more closely aligns with human auditory perception by emphasizing lower frequencies. Following this, Per-Channel Energy Normalization (PCEN) is applied as an advanced dynamic range compression method to enhance signal robustness against noise and volume variations [23]. In this study, the conversion utilized a frame length of 400 samples, a frame step of 160 samples, and generated 128 mel bands.

- Input Resizing: To ensure a consistent input size for the deep learning model, the final mel-PCEN spectrogram was resized to a fixed dimension of 192x128, representing time and frequency axes, respectively.
- SpecAugment: To enhance data diversity and improve the model's generalization capabilities, we applied SpecAugment to the training set. This data augmentation technique operates directly on the spectrogram by randomly applying time masking (obscuring a range of consecutive time steps) and frequency masking (hiding a block of consecutive mel frequency channels). This process encourages the model to learn more robust features and reduces the risk of overfitting [24].

C. Building Model

This study adopts a multimodal deep learning approach that leverages both acoustic features from cough sounds and structured clinical data for TB classification. The architecture consists of two parallel processing streams: one for audio and one for clinical metadata which are then combined through a fusion mechanism before a final classification is made.

The audio processing stream utilizes the pre-trained HeAR model as its feature extraction backbone. The choice of HeAR is deliberate; as a foundation model pre-trained on over 300 million health-related audio clips, it provides powerful and generalized acoustic embeddings, allowing us to achieve high performance without training a deep architecture from scratch. Architecturally, HeAR is a transformer-based model, like a Vision Transformer (ViT), designed to process audio spectrograms as images [25]. The Vision Transformer architecture described in Fig. 3.

The input 192x128 mel-PCEN spectrogram is first divided into a sequence of smaller, non-overlapping patches P_i.

Mathematically, for each patch, the embedding vector E_i is computed using a linear projection, as defined in Eq. (2).

$$E_i = W \cdot vec(P_i) + b \quad (2)$$

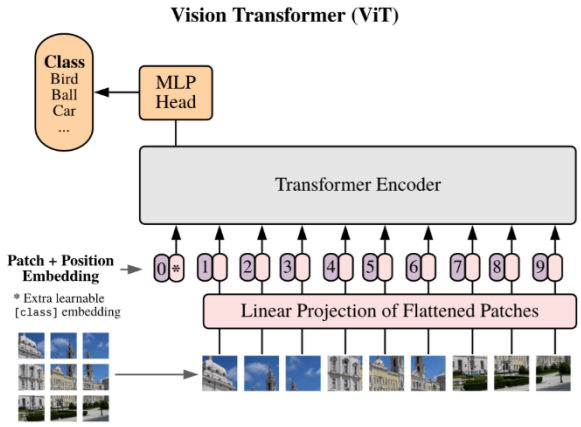

Fig. 3. Vision transformer architecture.

where, W denotes the projection weight matrix, b represents the bias vector, and P_i corresponds to the flattened image patch. To retain spatial information, a learnable positional encoding is then added to each embedding vector. These encodings provide the model with information about the position of each patch in the sequence, using sine and cosine functions of different frequencies, as defined in Eq. (3).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{100000^{\frac{2i}{d}}}\right)$$
$$PE_{(pos,2i+1)} = \sin\left(\frac{pos}{100000^{\frac{2i}{d}}}\right) \quad (3)$$

where, pos denotes the position index, i represents the dimension index, and d is the embedding dimension.

The resulting sequence of vectors is then processed by 24 transformer encoder layers. Each layer comprises a multi-head self-attention mechanism with 16 attention heads, followed by a position-wise feed-forward network (FFN) with a hidden size of 4096. The self-attention mechanism calculates attention scores based on the query Q, key K, and value V to weigh the importance of different patches relative to each other, allowing the model to capture complex dependencies across the spectrogram as defined in Eq. (4).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where, $Q, K,$ and V are the query, key, and value matrices, and d_k is the dimension of the key.

Following the attention sub-layer, the output is passed to an FFN. This network consists of two linear transformations with a GELU activation in between as shown in Eq. (5).

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \quad (5)$$

where, x denotes the input token representation, W_1 and W_2 are learnable weight matrices, b_1 and b_2 are the corresponding bias vectors, and GELU is the Gaussian Error Linear Unit activation as defined in Eq. (6).

$$GELU(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \quad (6)$$

where, $\Phi(x)$ is the standard Gaussian cumulative distribution function. The FFN allows for nonlinear transformations and enhances the model's capacity to learn complex representations. Residual connections are applied around each of the two sub-layers, followed by layer normalization, as defined in Eq. (7). This step is crucial for stabilizing the training of deep networks.

$$Layernorm(x + Sublayer(x)) \quad (7)$$

For this study, the output embedding from the final transformer layer is used as the acoustic feature representations, and the weights of the HeAR backbone are kept frozen. The output which is a 512-dimensional embedding vector is then passed through a dedicated audio branch, which consists of a linear layer that projects the features from 512 to 128 dimensions, followed by Batch Normalization, a ReLU activation, a dropout layer with a rate of 0.4, and a final linear layer that outputs a 64-dimensional audio feature vector.

Concurrently, the clinical data stream handles the tabular metadata. After preprocessing (one-hot encoding for categorical features and standard scaling for age), the vector is fed into a clinical branch. This branch is composed of a linear layer mapping the input to 64 dimensions, followed by a ReLU activation, a dropout layer with a rate of 0.3, and another linear layer that produces a 32-dimensional clinical feature vector.

In the fusion stage, the 64-dimensional audio vector and the 32-dimensional clinical vector are concatenated, creating a combined 96-dimensional feature vector. This vector is then passed to the final classifier which consists of a Batch Normalization layer, a ReLU activation, a high-rate dropout layer (0.5) for regularization, and a final linear layer that outputs the logits for the two classes (TB-positive and TB-negative). The final prediction probabilities are obtained using a softmax function, as shown in Eq. (8).

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (8)$$

where, z is the input vector of logits and K is the number of classes. The entire model is trained end-to-end using a cross-entropy loss function to quantify the difference between the predicted probability and the actual class labels, as defined in Eq. (9).

$$\mathcal{L} = -\sum_{i=0}^{1} y_i \log(\hat{y}_i) \quad (9)$$

D. Training Setup

The AdamW optimizer was used to train the models, incorporating a weight decay of 0.01 and an initial learning rate of 1×10^{-4}. Training was conducted for 50 epochs with a batch size of 32. The objective function was cross-entropy loss. Early stopping was applied based on validation loss to prevent overfitting, with a patience of 7 epochs and minimum delta of 0.001. Model performance was comprehensively evaluated using several metrics, including accuracy, precision, recall, and F1-score, with 80/10/10 train, validation and test dataset. All experiments were conducted on a NVIDIA RTX A4000 GPU with 16 GB VRAM.

IV. Experimental Results

A. Evaluation Parameters

This study employed several evaluation metrics, including accuracy, F1-score, precision, and recall. The confusion matrix is a fundamental tool for assessing classification model performance, offering a comprehensive comparison between the model's predictions and the actual labels. The confusion matrix comprises four components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive, thereby providing insight into the model's susceptibility to false positive errors. Recall, on the other hand, evaluates the proportion of actual positive instances that the model correctly identified, thereby reflecting its effectiveness in minimizing false negative errors. Specificity measures the proportion of actual negatives that are correctly identified as such by the model. The F1-score, computed as the harmonic means of precision and recall, provides a balanced assessment of the model's performance by simultaneously accounting for both FP and FN. These evaluation metrics are computed using the formulas presented in Eq. (10) through Eq. (14).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$F1 - Score = \frac{2 \times (precision \times recall)}{precision + recall} \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

B. Experimental Analysis

In this study, simulations and model development for detecting TB based on cough sounds-, as well as the creation of a digital audio-based detection system-, were conducted using hardware and software with specific configurations. Table II presents the detailed specifications of the equipment and tools used.

The authors configured the training process with 50 epochs, a learning rate of 0.0001, and a batch size of 32, using the AdamW optimizer.

As shown in Table III, the proposed multimodal deep learning model demonstrated strong performance in classifying tuberculosis from cough audio and clinical data. The model achieved an overall accuracy of 90% on the test set. The results

show a well-balanced performance across both classes, with precision, recall, and F1-scores of 0.90 for both the "Non-TB" and "TB" classes. This symmetry indicates that the model is equally effective at identifying both positive and negative cases and is not biased towards any single class, which is a crucial attribute for a reliable medical screening tool.

TABLE II. DEVICE SPECIFICATION

Specifications	
GPU	NVIDIA RTX A4000
GPU Memory	16 GB
RAM	16 GB
Disk	1 TB
Programming Language	Python 3.10

TABLE III. MODEL PERFORMANCE

Class	Accuracy	Precision	Recall	F1-Score	Specificity
Non-TB	0.90	0.90	0.90	0.90	0.9
TB	0.90	0.90	0.90	0.90	0.9
Average	**0.90**	**0.90**	**0.90**	**0.90**	**0.9**

The confusion matrix, shown in Fig. 4, provides a more granular view of the model's predictions. Out of 800 non-TB samples, the model correctly identified 723 as Non-TB (True Negatives) and misclassified 77 as TB (False Positives). Similarly, for the 800 TB samples, the model correctly identified 723 as TB (True Positives) and misclassified 77 as Non-TB (False Negatives). The low number of misclassifications for both classes further validates the model's high accuracy and balanced predictive power.

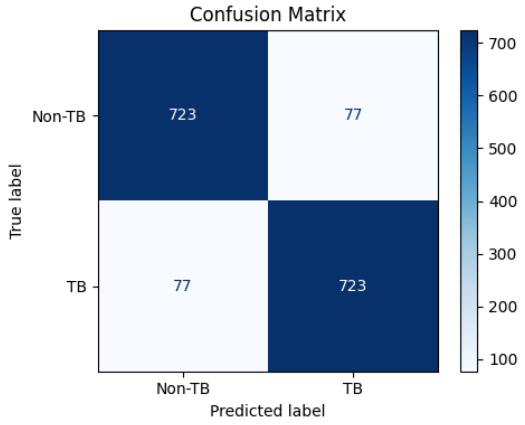

Fig. 4. Confusion matrix of the proposed model on the test set.

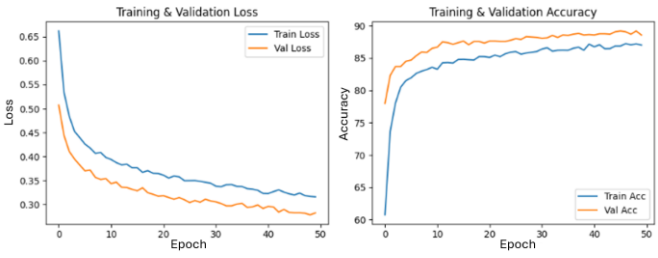

Fig. 5. Training and validation loss and accuracy curves over 50 epochs.

Fig. 5 illustrates the model's learning process over 50 epochs. The training and validation loss curves show a consistent downward trend, indicating that the model was effectively learning from the data. Importantly, the validation loss closely follows the training loss without significant divergence, which suggests that the model generalized well to unseen data and was not overfitting. The accuracy curves mirror this positive trend, with both training and validation accuracy steadily increasing and converging to approximately 90%. This stable training behavior highlights the effectiveness of the chosen architecture and regularization techniques (e.g., dropout) in building a robust classifier.

V. DISCUSSION

Our proposed model achieved a strong overall accuracy of 90%, with balanced performance across all key metrics. This section discusses the interpretation of these findings, the study's limitations, and challenges for real-world deployment.

A. Interpretation of Results and Clinical Implications

Our model's 90% overall accuracy, with balanced precision, recall, and F1-scores of 0.90 across both TB and non-TB classes, demonstrates its reliability and lack of bias. For clinical applications, this performance is highly promising. The 90% sensitivity is crucial for identifying most positive cases early, aiding in timely treatment and reducing transmission. Simultaneously, the 90% specificity effectively rules out the disease in most healthy individuals, minimizing unnecessary follow-up tests, patient anxiety, and healthcare costs. While the 10% false-negative rate requires further work to mitigate public health risks, these findings strongly validate that fusing advanced acoustic features with clinical data offers a holistic and accurate approach for a scalable TB screening tool.

B. Real-World Implementation Scenarios and Challenges

For deployment in real-world TB screening, several challenges must be addressed. Data privacy is paramount, requiring secure data handling. The model's robustness to background noise in uncontrolled environments needs rigorous testing. Furthermore, hardware variability, particularly the quality of microphones in different smartphones, could impact performance and requires investigation to ensure consistent results.

C. Limitations

A primary limitation is our use of random undersampling, which, as noted, may discard potentially useful data. Future studies could address this by implementing alternative balancing techniques like SMOTE or focal loss.

VI. CONCLUSION AND FUTURE WORK

This study successfully demonstrated the effectiveness of a multimodal deep learning approach for tuberculosis detection by combining cough audio signals with clinical metadata. By leveraging the powerful Health Acoustic Representations (HeAR) model as a feature extractor and integrating it with key patient symptoms, our proposed model achieved an impressive accuracy of 90% on a balanced test set. The balanced precision, recall, and F1-scores further underscore the model's capability to reliably identify both Non-TB and TB cases,

highlighting its potential as a robust and unbiased screening tool. These findings confirm that the fusion of advanced acoustic features with contextual clinical data provides a more holistic and accurate representation for diagnosis than either modality alone.

From a clinical perspective, the balance between sensitivity (recall) and specificity is critical for a screening tool. The model's 90% sensitivity indicates that it successfully identifies 9 out of 10 individuals with TB, which is vital for early treatment initiation. However, the 10% of cases missed (false negatives) represent a significant public health concern, as these individuals may unknowingly continue to transmit the disease. Conversely, the 90% specificity means that 9 out of 10 healthy individuals are correctly identified, but the 10% of false positives would be subjected to unnecessary follow-up tests, causing patient anxiety and burdening healthcare resources. For a preliminary screening tool, this performance represents a promising balance, but future work should prioritize further reducing the false-negative rate to maximize the tool's public health impact.

Several avenues for future research can be explored to build upon these promising results. First, extensive hyperparameter tuning of the fusion model, including the architecture of the audio and clinical branches as well as the dropout rates, could further optimize performance. Second, exploring different fusion strategies, such as attention-based mechanisms, may allow the model to dynamically weigh the importance of audio versus clinical features for each sample. Additionally, evaluating the model's generalization on external, unseen datasets from different demographic or geographic populations is crucial to ensure its real-world applicability. Finally, incorporating explainable AI (XAI) techniques is crucial for clinical adoption. By providing insights into which acoustic or clinical features most influence the model's predictions, XAI can build trust among healthcare professionals, aid in validating the model's decision-making process, and facilitate its integration into clinical workflows. The development of such a tool holds significant promise for creating an accessible, low-cost, and scalable solution to support global efforts in early TB detection and containment.

ACKNOWLEDGMENT

The authors like to acknowledge the support of PT Seleris Meditekno Internasional.

REFERENCES

[1] E. C. Jones-L Opez, O. Namugga, F. Mumbowa, M. Ssebidandi, and O. Mbabazi, “Coughing Is Not Required to Transmit Mycobacterium tuberculosis: Another Nail in the Coffin,” https://doi.org/10.1164/rccm.202204-0645ED, vol. 206, no. 2, pp. 141–143, Jul. 2022, doi: 10.1164/RCCM.202204-0645ED.

[2] World Health Organization, “Tuberculosis,” Sep. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[3] World Health Organization, “1.1 TB incidence.” Accessed: Sep. 18, 2025. [Online]. Available: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024/tb-disease-burden/1-1-tb-incidence

[4] M. MacGregor-Fairlie, S. Wilkinson, G. S. Besra, and P. G. Oppenheimer, “Tuberculosis diagnostics: overcoming ancient challenges with modern solutions,” Emerg Top Life Sci, vol. 4, no. 4, p. 435, Dec. 2020, doi: 10.1042/ETLS20200335.

[5] K. K. Chopra and S. Singh, “Tuberculosis: Newer diagnostic tests: Applications and limitations,” Indian Journal of Tuberculosis, vol. 67, no. 4, pp. S86–S90, Dec. 2020, doi: 10.1016/j.ijtb.2020.09.025.

[6] M. G. Moule and J. D. Cirillo, “Mycobacterium tuberculosis Dissemination Plays a Critical Role in Pathogenesis,” Front Cell Infect Microbiol, vol. 10, Feb. 2020, doi: 10.3389/fcimb.2020.00065.

[7] K. C. Rahlwes, B. R. S. Dias, P. C. Campos, S. Alvarez-Arguedas, and M. U. Shiloh, “Pathogenicity and virulence of Mycobacterium tuberculosis,” Virulence, vol. 14, no. 1, Dec. 2023, doi: 10.1080/21505594.2022.2150449.

[8] B. Mathema et al., “Drivers of Tuberculosis Transmission,” J Infect Dis, vol. 216, no. suppl_6, pp. S644–S653, Nov. 2017, doi: 10.1093/infdis/jix354.

[9] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2016.

[10] A. Salma, A. Bustamam, A. Yudantha, A. Victor, and W. Mangunwardoyo, “Artificial Intelligence Approach in Multiclass Diabetic Retinopathy Detection Using Convolutional Neural Network and Attention Mechanism,” International Journal of Advances in Soft Computing and its Applications, vol. 13, no. 3, pp. 101–114, Dec. 2021, doi: 10.15849/IJASCA.211128.08.

[11] A. Bustamam, D. Sarwinda, R. H. Paradisa, A. A. Victor, A. R. Yudantha, and T. Siswantining, “Evaluation of convolutional neural network variants for diagnosis of diabetic retinopathy,” Communications in Mathematical Biology and Neuroscience, 2021, doi: 10.28919/cmbn/5660.

[12] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, and T. Niesler, “Automatic Cough Classification for Tuberculosis Screening in a Real-World Environment,” Physiol Meas, vol. 42, no. 10, pp. 10.1088/1361-6579/ac2fb8, Oct. 2021, doi: 10.1088/1361-6579/AC2FB8.

[13] R. K. Sahoo, A. Sinha, M. Mishra, D. Bhattacharya, S. Pati, and K. C. Sahoo, “A Systematic Review and Meta-Analysis of the Diagnostic Accuracy of Artificial Intelligence in Detecting Tuberculosis Using Cough Sounds,” 2025, doi: 10.2139/SSRN.5242653.

[14] S. Baur et al., “HeAR-Health Acoustic Representations,” 2024.

[15] A. Ehtesham, A. Singh, S. Kumar, and T. T. Khoei, “Early Detection of Pediatric Pneumonia Using Google’s HeAR Model: A Respiratory Sound Embedding Approach,” 2025 IEEE World AI IoT Congress (AIIoT), pp. 0185–0191, May 2025, doi: 10.1109/AIIOT65859.2025.11105317.

[16] G. P. Kafentzis et al., “Predicting Tuberculosis from Real-World Cough Audio Recordings and Metadata,” Jul. 2023, Accessed: Sep. 19, 2025. [Online]. Available: https://arxiv.org/pdf/2307.04842

[17] S. Huddart et al., “Solicited Cough Sound Analysis for Tuberculosis Triage Testing: The CODA TB DREAM Challenge Dataset,” Mar. 28, 2024. doi: 10.1101/2024.03.27.24304980.

[18] A. Dadu et al., “PASS to End TB in Europe: Accelerated efforts on prevention and systematic screening to end tuberculosis in the WHO European Region by 2030,” International Journal of Infectious Diseases, vol. 141, Apr. 2024, doi: 10.1016/J.IJID.2024.02.023.

[19] G. H. R. Botha et al., “Detection of tuberculosis by automatic cough sound analysis,” Physiol Meas, vol. 39, no. 4, p. 045005, Apr. 2018, doi: 10.1088/1361-6579/AAB6D0.

[20] W. Xu et al., “Feature fusion method for pulmonary tuberculosis patient detection based on cough sound,” PLoS One, vol. 19, no. 5, p. e0302651, May 2024, doi: 10.1371/journal.pone.0302651.

[21] G. P. Kafentzis et al., “Predicting Tuberculosis From Real-World Cough Audio Recordings And Metadata”.

[22] Y. C. Eldar, “Resampling,” Cambridge University Press eBooks, pp. 323–367, Dec. 2014, doi: 10.1017/cbo9780511762321.010.

[23] V. Lostanlen et al., “Per-Channel Energy Normalization: Why and How,” IEEE Signal Process Lett, vol. 26, no. 1, pp. 39–43, Jan. 2019, doi: 10.1109/LSP.2018.2878620.

[24] D. S. Park et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," 2019, doi: 10.21437/Interspeech.2019-2680.

[25] A. Dosovitskiy et al., "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE", Accessed: Sep. 25, 2025. [Online]. Available: https://github.com/