Performance Analysis of Spectrogram-Based Versus Raw Waveform-Based Deep Learning Models for Smoker Detection from Cough Audio

Widi Nugroho[1], Alhadi Bustamam[2]*, Rinaldi Anwar Buyung[3]
Department of Mathematics, Universitas Indonesia, Depok, Indonesia[1, 2]
Data Science Center, Universitas Indonesia, Depok, Indonesia[2]
Department of Data Science, Seleris Meditekno Internasional, Jakarta, Indonesia[3]

***Abstract*—The classification of cough sounds for smoker detection represents a challenging task in audio processing that compares different data representation methods. This study presents a performance analysis of two prominent deep learning approaches: a spectrogram-based model, the Audio Spectrogram Transformer (AST), and a raw waveform-based model, Wav2Vec2. We used 7,561 smoker and 7,561 non-smoker samples from the CODA TB DREAM Challenge dataset. Both models were trained with five-fold cross-validation and data augmentation (SpecAugment for AST; noise, pitch, and time shifts for Wav2Vec2). The raw waveform-based Wav2Vec2 model achieved the best performance, with an average accuracy of 86.5%, an F1-score of 0.862, and an Area Under the Curve (AUC) of 0.945, completing training in approximately 49 minutes per fold. In contrast, the spectrogram-based AST model reached around 76-77% accuracy and an AUC of 0.85 in approximately 78 minutes per fold. These findings indicate that the raw waveform-based approach is significantly more effective and computationally efficient than the spectrogram-based approach for this task, offering a robust method for non-invasive smoker classification through the analysis of vocal biomarkers.**

***Keywords*—Smoker detection; cough audio classification; deep learning; Audio Spectrogram Transformer; Wav2Vec2; vocal biomarker**

I. Introduction

Tobacco consumption continues to be a major contributor to preventable illness and death on a global scale. As reported by the World Health Organization (WHO), tobacco-related illnesses affect more than 8 million individuals each year, including approximately 1.3 million non-smokers who die as a result of exposure to secondhand smoke [1]. The long-term health effects of smoking, such as cardiovascular diseases, chronic obstructive pulmonary disease (COPD), and multiple types of cancer, pose a substantial burden on both individuals and healthcare systems [2].

Tobacco smoke comprises a complex blend of more than 7,000 chemical substances, including at least 70 that are carcinogenic, all of which significantly contribute to its wide-ranging adverse health effects. These substances can damage nearly every organ in the body, leading to a wide range of diseases [3]. For example, smoking is a leading cause of several types of cancer, including those affecting the lungs, oral cavity, larynx, bladder, and pancreas. It also markedly elevates the risk of cardiovascular diseases by contributing to atherosclerosis development and impairing endothelial function. Moreover, smoking impairs respiratory health, leading to conditions like COPD and worsening of asthma symptoms [4]. The harmful effects of smoking are not limited to active smokers; secondhand smoke exposure also poses serious health risks, including increased incidence of lung cancer and heart disease in non-smokers.

Despite the well-documented health consequences, the accurate and efficient verification of smoking status remains a significant challenge. Conventional methods primarily rely on self-reported declarations, which are prone to misrepresentation and bias. Studies indicate that a substantial number of smokers misrepresent their status on various applications. To overcome this, biochemical tests, such as cotinine detection in saliva or urine, are often used for objective assessment. While reliable, these biochemical methods have notable drawbacks: they are invasive, costly, and often impractical for large-scale or remote screening scenarios [5]. These limitations highlight the need for an alternative detection method that is non-invasive, objective, and scalable.

Recent advancements in digital health have introduced vocal biomarkers as a promising, non-invasive solution. Smoking induces physiological changes in the respiratory tract and vocal folds, leading to measurable alterations in voice and cough characteristics such as fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio [6]. Smokers tend to exhibit rougher voice textures, prolonged coughing bouts, and altered pitch or spectral features, which have been empirically shown to distinguish them from non-smokers with significant accuracy [7]. These acoustic features can be captured and analyzed using data science to differentiate between smokers and non-smokers, offering a cost-effective and accessible approach to verification.

The application of data science and deep learning is expected to improve the efficiency of the underwriting process. As an interdisciplinary field, data science uses scientific methodologies, algorithms, and systems to extract insights from both structured and unstructured data. It draws upon mathematics, statistics, information science, and computer science principles and is closely associated with techniques such as data mining, machine learning, and big data analytics. These tools support the creation of predictive models, uncover

*Corresponding authors

patterns and trends, and facilitate data-driven decision-making. Deep learning, as outlined in [8], is a subset of machine learning that enables a system to learn from previous data and understand complex concepts without explicit programming. By learning from experience, computers can autonomously perform tasks, reducing the need for manual input. This capability holds promise for enabling real-time underwriting processes. In previous studies [9] and [10], deep learning, particularly Convolutional Neural Networks (CNNs), was employed to detect diabetic retinopathy using retinal fundus images.

This study used audio data to implement a deep learning approach to identify individuals' smoking status. The proposed approach is expected to enable accurate, real-time identification of smoking status, thereby enhancing the underwriting process. Moreover, by automating a task that is traditionally performed in the laboratory, artificial intelligence offers a cost-saving opportunity for insurance companies through increased operational efficiency.

Based on this context, this study aims to answer the following primary research questions: 1) How effectively can deep learning models differentiate between smokers and non-smokers using only the sound of their cough? 2) Which audio data representation: a spectrogram-based model (Audio Spectrogram Transformer) or a raw waveform-based model (Wav2Vec2) offers superior classification performance and computational efficiency for this task?

Two frameworks for audio classification were used in this study. The first kind operates using the mel-spectrogram image as input. The second kind uses the raw waveform as the input. Representative models for the first approach include CNN, VGGish, YAMNet, and AST. For the second approach, CNN1D, WaveNet, SampleRNN, and Wav2Vec2 are commonly used models [11]. While these architectures have been benchmarked for detecting overt respiratory diseases, their comparative efficacy for identifying the subtler physiological markers of smoking status in cough audio remains an open and important research question.

AST (Audio Spectrogram Transformer) is a deep learning model that adapts the transformer architecture, originally developed for natural language processing, to audio classification tasks by processing mel-spectrogram representations of audio signals. Unlike conventional convolutional models, AST leverages self-attention mechanisms to capture long-range temporal and frequency dependencies in audio data. AST has demonstrated strong performance in the context of detecting anomalies or categorizing specific audio events, such as identifying coughs. For instance, research in [12] used AST to classify various types of cough sounds and achieved an F1-score of 0.804.

Wav2Vec2 is a self-supervised learning model developed for speech representation learning that has shown remarkable performance in various audio classification tasks. It operates directly on raw audio waveforms and learns contextualized representations using a combination of convolutional feature encoders and transformer-based architectures. In contrast to traditional methods that depend extensively on handcrafted features or spectrogram representations, Wav2Vec2 learns meaningful patterns from raw, unlabeled audio data autonomously, thereby substantially minimizing the reliance on large annotated datasets [13]. In the domain of audio-based classification, including speaker identification, emotion recognition, and health-related applications such as cough sound classification, Wav2Vec2 has demonstrated high levels of accuracy. For instance, a study [14] employed Wav2Vec2 with minimal preprocessing to detect COVID-19 coughs, achieving competitive results with an Area Under the Curve (AUC) of 0.7810.

In this study, the AST and Wav2Vec2 algorithms were employed to classify coughs of smokers. The AST model was selected because of its strong performance in audio classification tasks, particularly those using mel-spectrogram inputs. Then, the Wav2Vec2 algorithm, which is one of the best models for audio representation that uses raw waveform as the input, is used. The performances of AST and Wav2Vec2 will be compared with those of prior approaches for classifying smokers from audio recordings.

The remainder of this paper is organized as follows. Section II provides an overview of related work in the field. Section III details the dataset, data preprocessing techniques, and the architecture of the deep learning models used. Section IV presents the experimental results and a comparative analysis of the models' performance. Finally, Section V concludes the paper by summarizing the key findings and suggesting directions for future research.

II. Related Work

Research into identifying smoking status from vocal cues has evolved from traditional acoustic feature analysis to the application of deep learning. Early studies focused on identifying perturbations in voice characteristics, such as fundamental frequency, jitter, shimmer, and harmonics-to-noise ratio, to distinguish smokers from non-smokers. These approaches, while foundational, often relied on handcrafted features and classical machine learning models.

Subsequent research moved towards machine learning techniques applied to more complex vocal representations as summarized in Table I. Poorjam et al. employed i-vector embeddings representing whole vocal tract characteristics combined with Logistic Regression classification, achieving an AUC of 0.74. This approach demonstrated the value of compact speaker-level representations for smoking status detection. Ma et al. applied convolutional neural networks (ResNet18) to Mel-frequency cepstral coefficients (MFCCs) and other acoustic features, reporting an Accuracy of 82% on a dataset comprising 1,194 samples. This deep learning approach leveraged spatial patterns in spectrogram representations to improve classification performance. Furthermore, Ayadi et al. explored a hybrid model leveraging pretrained Wav2Vec as a feature extractor directly from raw waveforms, followed by a Support Vector Machine (SVM) classifier, which achieved an accuracy of 72%. The use of Wav2Vec allowed for the extraction of rich temporal and spectral features without reliance on handcrafted features, highlighting the promise of end-to-end learned representations for this task.

Despite these promising results, current research is limited by several factors. Many studies utilize relatively small and imbalanced datasets, which may restrict the generalizability of models. Moreover, most frameworks employ multi-stage pipelines, separate feature extraction followed by classification or adapt architectures originally designed for image data (e.g., ResNet) on spectrogram inputs. There remains a lack of direct, rigorous comparative evaluations between state-of-the-art end-to-end audio-specific architectures. Specifically, models that process spectrograms such as the Audio Spectrogram Transformer (AST) and those that work directly on raw audio waveforms like Wav2Vec2 have not yet been comprehensively contrasted for smoker detection.

This study aims to address these gaps by conducting a systematic performance comparison of these two leading approaches on a large, balanced dataset. The results will provide critical insights into the optimal audio representation and model design for accurate and robust smoking status classification from voice signals.

TABLE I. STATE-OF-THE-ART

Study	Methodology	Audio Features	Dataset	Evaluation Metrics
[15]	Wav2Vec + SVM	Raw-waveform	4917 data	Accuracy = 72% AUC = 0.76
[16]	ResNet18	MFCC, FBank, F0, jitter, and shimmer	-	Accuracy = 82% F1-score = 0.823
[17]	Logistic Regression	i-vector	1194 data	AUC = 0.74

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.

A. Dataset

The dataset used in this research is the CODA TB DREAM Challenge dataset [18] comprises 29,768 cough recordings from seven countries (India, Philippines, South Africa, Uganda, Vietnam, Tanzania, and Madagascar). Each recording is associated with demographic and clinical metadata. For the purposes of our study, we removed all tuberculosis-positive cases and performed random undersampling of the majority class, resulting in 7,561 smoker and 7,561 non-smoker cough samples. The dataset is illustrated in Fig. 2.

B. Data Preprocessing and Augmentations

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 [19]. 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 0.5 seconds.

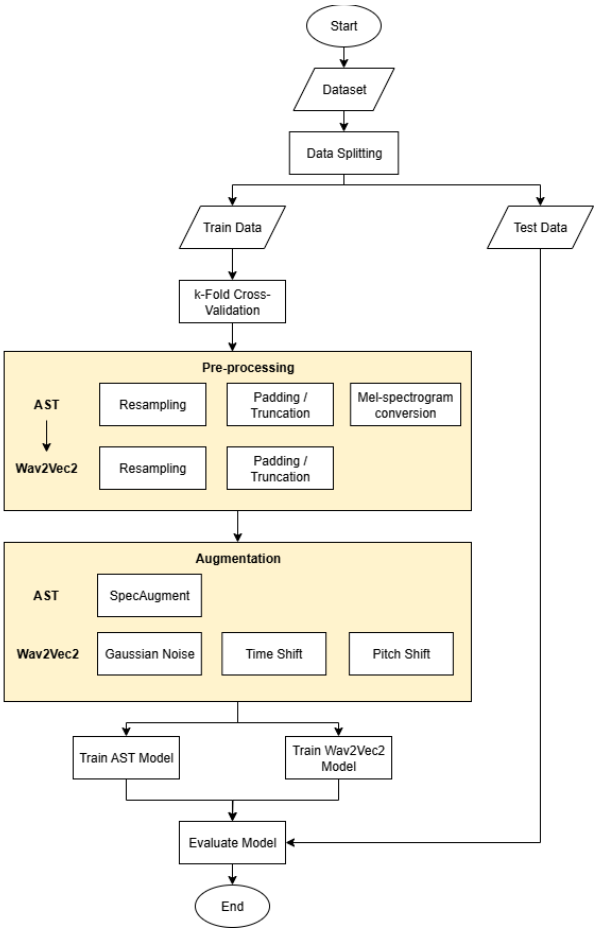

Fig. 1. Research workflow.

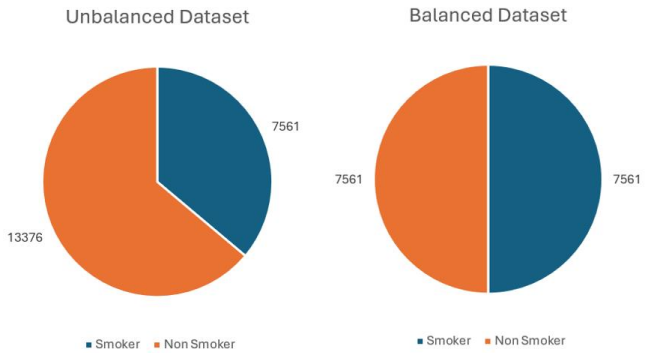

Fig. 2. Dataset distribution.

- Mel-Spectrogram Conversion: For the AST model, the input must be in the mel-spectrogram format. This process involves applying the Short-Time Fourier

Transform (STFT) to decompose the audio into its frequency components over time. Subsequently, the spectrogram was mapped onto the Mel scale, which more closely aligns with human auditory perception by emphasizing lower frequencies. Then, the mel-spectrogram is converted to a decibel scale [20]. In this study, we used a mel-spectrogram with 128 mels bands, 128 FFT components, and a hop length of 128.

- Training, Validation, and Test Data Splitting: Dividing the dataset into training, validation, and test sets is a crucial step in developing deep learning models. The training set, which comprises most of the data, is used to teach the model to recognize patterns and extract relevant features. During training, validation is employed to monitor the model's performance and mitigate issues such as overfitting or underfitting. The test set is used to evaluate the model's generalization ability on previously unseen data. In this study, the dataset was partitioned with 25% used as the test set, and the remainder split using 5-fold cross-validation.
- Subsequently, augmentation techniques were applied to the training set to enhance data diversity and improve the model's generalization performance. The augmentation methods include SpecAugment for the AST algorithm, and Gaussian noise, pitch shift, and time shift for the Wav2Vec2 algorithm [21]. The following provides a detailed explanation of each augmentation technique:
- SpecAugment: SpecAugment is a data augmentation technique that modifies audio data spectrogram representations. It applies three primary transformations: time warping, frequency masking, and time masking. Time warping involves stretching or compressing the spectrogram along the temporal axis. Frequency masking randomly obscures certain frequency bands. Time masking similarly hides segments along the time axis. These augmentations compel models to focus on more general patterns within the data, rather than overfitting to specific features, thereby improving generalization to diverse and unseen audio inputs. In this study, we used eight bands for frequency masking and eight frames for time masking.
- Gaussian Noise: Gaussian noise augmentation involves the addition of random noise drawn from a Gaussian distribution to the original audio signal. This technique simulates real-world acoustic conditions such as background chatter, machinery hum, or environmental disturbances, thereby increasing the robustness of models against noisy inputs. By slightly perturbing the waveform, the Gaussian noise encourages the model to generalize better by learning essential features rather than memorizing clean training data. The noise mean and standard deviation can be adjusted to control the noise intensity. In this study, the minimum and maximum amplitudes for the Gaussian noise are 0.001 and 0.015, respectively.
- Pitch Shift: Pitch shifting is an augmentation technique that modifies the pitch of an audio signal while preserving its temporal duration. This is achieved by raising or lowering the frequency components of the audio waveform, effectively simulating vocal tone, gender, or emotional state variations. By introducing pitch variability, this technique enhances the model's ability to generalize across different speakers and speaking styles. In this study, we used pitch shift with a minimum of -4 semitones and a maximum of 4 semitones.
- Time Shift: Time shift augmentation involves shifting the audio waveform by a small amount along the time axis. This technique simulates natural variations in speech or sound onset, such as differences in timing or speaker response delays, which often occur in real-world recordings. By randomly altering the start point of the audio signal, time shifting encourages the model to become less sensitive to the exact temporal position of features and focus more on the content itself. In this study, we used time shift augmentation with a minimum shift of -10% and a maximum shift of 10%.

C. Building Models

This study adopts two distinct deep learning approaches for classifying smoking status based on cough audio. The first approach employs the AST, which converts raw audio signals into mel-spectrograms—a visual time-frequency representation—before feeding them into a vision transformer-based model. This method treats audio as an image and leverages computer vision techniques to extract spatial patterns within the spectrogram. In contrast, the second approach uses Wav2Vec2, a transformer-based architecture designed to directly process raw audio waveforms without the need for spectrogram conversion. This end-to-end method allows the model to learn relevant features directly from the waveform through self-supervised pretraining followed by fine-tuning, integrating feature extraction and classification within a single unified stage.

For the AST model, the input mel-spectrogram is first divided into non-overlapping 16 x 16 patches. Each patch is flattened into a one-dimensional vector and then passed through a linear projection layer to obtain a fixed-dimensional embedding. The architecture of AST can be seen in Fig. 3. Mathematically, for each patch P_i , the embedding vector E_i is computed using a linear projection, as defined in (1):

$$E_i = W \cdot vec(P_i) + b \quad (1)$$

where, W denotes the projection weight matrix, b represents the bias vector, and $vec(P_i)$ corresponds to the flattened image patch. A learnable classification token, denoted as [CLS] is prepended to the patch embedding sequence to enable global representation learning for the input. Then, positional embeddings are added to each token to encode the relative position of each patch within the mel-spectrogram.

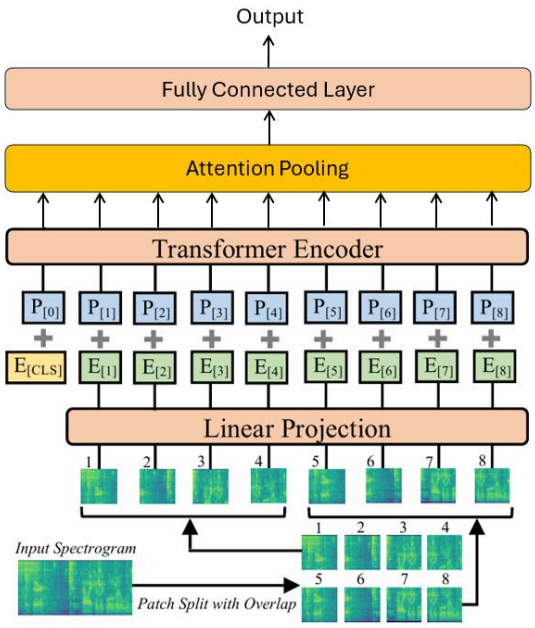

Fig. 3. AST architecture.

The positional encoding, which uses sine and cosine functions at varying frequencies to capture the order of sequence elements, is mathematically defined in (2):

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

where, pos denotes the position index, i represents the dimension index, and d is the embedding dimension. The positional encodings are added element-wise to incorporate the positional information into the model, enabling it to capture the sequential order of the input data.

Subsequently, the resulting sequence is input into the transformer encoder, which comprises 24 stacked layers. Each layer includes a multi-head self-attention mechanism with 16 attention heads, followed by a position-wise feed-forward network with a hidden size of 1024. The self-attention mechanism calculates attention scores based on the query Q, key K, and value V, and is mathematically expressed in (3):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where, d_k denotes the dimensionality of the key vectors. The multi-head attention mechanism computes multiple parallel attention outputs, each corresponding to a distinct attention head. These outputs are then concatenated, and a linear transformation is performed to produce the final attention representation.

Following the self-attention mechanism, each token is processed through a position-wise feed-forward network (FFN), which comprises two linear transformations separated by as activation function, and non-linear activation function. This operation can be mathematically described by (4):

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \quad (4)$$

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. The GELU activation, used in the FFN, is defined in (5):

$$GELU(x) = x \cdot \Phi(x) =$$
$$x \cdot \frac{1}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715\,x^3)\right)\right) \quad (5)$$

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 followed by layer normalization are applied after the self-attention and feed-forward sublayers. This architectural design helps stabilize the training process and facilitates efficient gradient flow throughout the network. After processing through 24 transformer encoder layers, the model does not rely on the conventional classification token [CLS] to represent global information. In this study, we employ attention pooling, which dynamically aggregates the sequence of patch embeddings into a single vector based on learned attention weights. The attention pooling mechanism assigns a weight a_i to each token embedding h_i, and computes the aggregated representation h_{attn} as a weighted sum of these embeddings, shown in (6):

$$h_{attn} = \sum_{i=1}^{N} a_1 h_i \quad (6)$$

This representation is then passed through a final fully connected layer, followed by a softmax activation to perform a binary classification between smokers and non-smokers. The output logits $z = [z_0, z_1]$ are converted into the class probability using the softmax function, as shown in (7):

$$\sigma(z)_i = \frac{e^{z_i}}{e^{z_0} + e^{z_1}}, i \in \{0,1\} \quad (7)$$

During model training, the cross-entropy loss function was employed to quantify the difference between the predicted probability distribution and the actual class labels. The loss $\mathcal{L}$, which the model aims to minimize during training, is mathematically expressed in (8):

$$\mathcal{L} = -\sum_{i=0}^{1} y_i \log(\hat{y}_i) \quad (8)$$

The input of the Wav2Vec2 model is a raw audio waveform with a fixed length of 8,000 samples and a normalized amplitude within the range [-1, 1]. The Wav2Vec2 architecture comprises two primary components: a convolutional feature encoder and a transformer-based context network. The architecture of Wav2Vec2 is shown in Fig. 4.

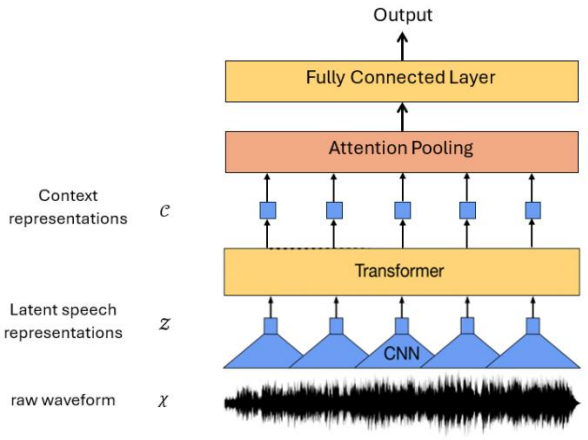

Fig. 4. Wav2Vec2 architecture.

The feature encoder consists of seven one-dimensional convolutional layers that transform the raw audio signal into a sequence of latent acoustic representations. Each convolutional layer applies a kernel across the temporal dimension and reduces the sequence length through stride-based downsampling, which is computed as shown in (9):

$$Y(i) = (x * w)(i) = \sum_{k=0}^{K-1} x(i+k) \cdot w(k) \quad (9)$$

where, x is the input signal, w is the convolution kernel, and K is the kernel size. The output is a lower-resolution but higher-dimensional representation capturing the local temporal features of the waveform. Each convolutional layer is immediately followed by the GELU activation function and layer normalization, which is computed as shown in (10):

$$LayerNorm(x) = \gamma \cdot \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}} + \beta \quad (10)$$

where, μ and σ^2 represent the mean and variance of the input x, respectively, ϵ is a small constant added to ensure numerical stability, and γ and β are learnable scaling and shifting coefficients that allow the model to adapt to the normalized outputs during training.

The output of the final convolutional layer is a sequence of latent vectors $z = (z1, z2, \dots z_T)$, which is passed to the context network— a stack of 24 transformer encoder layers with 16 attention heads and a position-wise FFN with a hidden size of 1024. Positional encodings are first added to the sequence, as in the AST model.

Each transformer block contains two sublayers: a multi-head self-attention mechanism and an FFN. The attention mechanism enables each time step to attend to all others, thereby capturing global contextual dependencies. This self-attention mechanism is mathematically expressed in (11):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

The FFN in each transformer layer comprises two linear layers separated by a GELU activation, as described in (12):

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \quad (12)$$

After all transformer layers are processed, the output is a sequence of contextualized embeddings $c = (c_1, c_2, \dots, c_T)$, where each vector $c_t \in \mathbb{R}^d$ encodes the acoustic information at time step t. Instead of using standard mean pooling, this study implements attention pooling to allow the model to learn which time steps are most informative for classification.

This attention mechanism enables the model to focus on the most relevant portions of the cough signal, thereby improving its ability to differentiate between smokers and non-smokers. The attention-weighted vector c_{attn} is passed through a linear classification layer followed by a softmax activation to produce the final prediction, as defined in (13):

$$\sigma(z)_i = \frac{e^{z_i}}{e^{z_0}+e^{z_1}}, i \in \{0,1\} \quad (13)$$

During training, the cross-entropy loss is used to measure the discrepancy between the predicted probabilities and true labels. The loss function $\mathcal{L}$ for the Wav2Vec2 model is defined in (14):

$$\mathcal{L} = -\sum_{i=0}^{1} y_i \log(\hat{y}_i) \quad (14)$$

D. Training Setup

The AdamW optimizer was used to train the models, incorporating a weight decay of 0.001 and an initial learning rate of 1×10^{-4}. Training was conducted for 20 epochs with a batch size of 16. The objective function was cross-entropy loss. Early stopping was applied based on validation loss to prevent overfitting, with a patience of four epochs. Model performance was comprehensively evaluated using several metrics, including accuracy, precision, recall, F1-score, and AUC, based on a 5-fold cross-validation with 75/25 train and validation/test split [22]. All experiments were conducted on an NVIDIA RTX A4000 GPU with 16 GB VRAM.

IV. Experimental Results

A. Evaluation Parameters

This study employed several evaluation metrics, including the accuracy, F1-score, precision, recall, and AUC. 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. The F1-score, computed as the harmonic mean of precision and recall, provides a balanced assessment of the model's performance by simultaneously accounting for both FP and FN. Additionally, the AUC, derived from the Receiver Operating Characteristic (ROC) curve, and measures the model's ability to discriminate between classes across various threshold settings. A higher AUC value indicates better overall classification performance, especially in imbalanced dataset scenarios. These evaluation metrics are computed using the formulas presented in (15) through (21).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (15)$$

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

$$F1-Score = \frac{2\times(precision\times recall)}{precision+recall} \quad (18)$$

$$TPR = \frac{TP}{TP+FN} \quad (19)$$

$$FPR = \frac{FP}{FP+TN} \quad (20)$$

$$AUC = \int_0^1 TPR(FPR)dFPR \quad (21)$$

B. Experimental Analysis

In this study, simulations and model development for detecting smokers 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.

TABLE II. DEVICE SPECIFICATION

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

The authors configured the training process with 20 epochs, a learning rate of 0.0001, and a batch size of 16, using the AdamW optimizer. The raw waveform audio was resampled to 16 kHz for input into the Wav2Vec2 model and subsequently transformed into mel-spectrograms with 128 frequency bins for use with the AST model.

As shown in Table III, the Wav2Vec2 model without augmentation achieved the highest overall performance, attaining average accuracy, recall, precision, F1-score, and AUC values of 86.48%, 0.866, 0.868, 0.862, and 0.945, respectively. This model also required the shortest training time, completing each fold in an average of 49 minutes. Although the Wav2Vec2 with augmentation showed comparable AUC values (consistently 0.94 across all folds), its overall performance metrics were slightly lower, and its average training time increased to 65 minutes per fold. On the other hand, the AST model, both with and without SpecAugment, demonstrated moderate performance, with average F1-scores around 0.76-0.77 and longer training times of approximately 78 minutes per fold. Notably, the addition of SpecAugment to the AST did not yield significant improvements and, in some folds, even resulted in minor declines in accuracy. Overall, the Wav2Vec2-based approach outperformed the AST-based method in terms of both predictive performance and computational efficiency, indicating its suitability for fast and accurate detection of smokers based on cough audio.

TABLE III. MODEL'S PERFORMANCE

Model	Fold	Evaluation Parameters					Training Time
		Acc	Rec	Pre	F1-Score	AUC	
AST (w/o Augmentation)	1	76.65 %	0.77	0.77	0.77	0.84	78 min
	2	76.62 %	0.76	0.76	0.76	0.84	78 min
	3	76.49 %	0.76	0.77	0.76	0.85	78 min
	4	78.07 %	0.78	0.78	0.78	0.86	78 min
	5	76.33 %	0.76	0.78	0.76	0.85	78 min
AST (w/ Augmentation)	1	75.75 %	0.76	0.76	0.76	0.84	78 min
	2	77.70 %	0.78	0.78	0.78	0.86	77 min
	3	77.31 %	0.77	0.79	0.77	0.86	77 min
	4	77.52 %	0.78	0.78	0.78	0.85	78 min
	5	76.75 %	0.77	0.77	0.77	0.84	78 min
Wav2Vec2 (w/o Augmentation)	1	87.46 %	0.87	0.87	0.87	0.94	49 min
	2	85.69 %	0.86	0.86	0.86	0.93	49 min
	3	86.54 %	0.87	0.87	0.86	0.94	49 min
	4	86.51 %	0.87	0.87	0.86	0.94	49 min
	5	86.19 %	0.86	0.87	0.86	0.94	49 min
Wav2Vec2 (w/ Augmentation)	1	87.17 %	0.87	0.87	0.87	0.94	65 min
	2	84.53 %	0.85	0.85	0.84	0.94	65 min
	3	86.49 %	0.86	0.87	0.86	0.94	66 min
	4	87.23 %	0.87	0.87	0.87	0.94	65 min
	5	85.35 %	0.85	0.86	0.85	0.94	65 min

V. CONCLUSION AND FUTURE WORK

This study successfully demonstrated that cough sounds can serve as a noninvasive signal for detecting smoking status using deep learning approaches. Among the models evaluated, the Wav2Vec2 architecture without augmentation yielded the best performance, achieving an average accuracy of 86.48%, F1-score of 0.862, and AUC of 0.945, while also requiring the shortest training time of only 49 minutes per fold. This model significantly outperformed the AST-based models, both with and without SpecAugment, which achieved only around 76%–77% accuracy with a longer training duration of approximately 78 minutes per fold. These findings highlight the effectiveness

of raw waveform-based representations like Wav2Vec2 in capturing audio characteristics related to smoking, surpassing the performance of spectrogram-based methods such as AST. Therefore, the raw waveform-based approach using Wav2Vec2 presents a viable and efficient direction for developing automated, non-invasive health screening tools. In practice, this technology could be integrated into mobile applications for the insurance industry, offering an objective method for underwriting and risk assessment that moves beyond self-reported data. In clinical settings, it could serve as a rapid, preliminary screening tool in primary care. However, significant barriers to real-world adoption must be addressed. These include ensuring model robustness against varying audio quality from different consumer-grade microphones, handling background noise, and validating performance across diverse demographic populations.

Future work should focus on these challenges. First, hyperparameter tuning is necessary to further optimize model performance. Second, model ensembling techniques could enhance predictive stability. Finally, and most critically, evaluating cross-dataset generalization is crucial to ensure the model performs reliably on cough samples from different sources and acoustic environments before it can be considered for deployment in real-world scenarios.

ACKNOWLEDGMENT

This research was funded by the 2025 DIKTI Research Grant, under contract number PKS-582/UN2.RST/HKP05.00/2025. The authors also like to acknowledge the support of the Data Science Center FMIPA UI, BrainAI Lab (Bioinformatics Research, Data Intelligence, and AI Innovation Laboratory), PT Seleris Meditekno Internasional, and PT Global Risk Management.

REFERENCES

[1] World Health Organization, “Tobacco.” Accessed: Nov. 12, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco

[2] World Health Organization, “Tobacco.” Accessed: Nov. 12, 2024. [Online]. Available: https://www.who.int/europe/health-topics/tobacco

[3] C. B. Sherman, “Health effects of cigarette smoking,” Clin Chest Med, vol. 12, pp. 643–658, May 1991, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1747984/

[4] S. P Saha, D. K Bhalla, T. F Whayne, and C. Gairola, “Cigarette smoke and adverse health effects: An overview of research trends and future needs,” International Journal of Angiology, vol. 16, no. 03, pp. 77–83, Sep. 2007, doi: 10.1055/s-0031-1278254.

[5] N. L. Benowitz, K. E. Schultz, C. A. Haller, A. H. B. Wu, K. M. Dains, and P. Jacob, “Prevalence of Smoking Assessed Biochemically in an Urban Public Hospital: A Rationale for Routine Cotinine Screening,” Am J Epidemiol, vol. 170, no. 7, pp. 885–891, Oct. 2009, doi: 10.1093/aje/kwp215.

[6] Z. Ma, C. Bullen, J. T. W. Chu, R. Wang, Y. Wang, and S. Singh, “Towards the Objective Speech Assessment of Smoking Status based on Voice Features: A Review of the Literature,” Journal of Voice, vol. 37, no. 2, pp. 300.e11-300.e20, Mar. 2023, doi: 10.1016/j.jvoice.2020.12.014.

[7] L. Chai, A. J. Sprecher, Y. Zhang, Y. Liang, H. Chen, and J. J. Jiang, “Perturbation and Nonlinear Dynamic Analysis of Adult Male Smokers,” Journal of Voice, vol. 25, no. 3, pp. 342–347, May 2011, doi: 10.1016/j.jvoice.2010.01.006.

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

[9] 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.

[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. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, “A review of deep learning techniques for speech processing,” Information Fusion, vol. 99, p. 101869, Nov. 2023, doi: 10.1016/j.inffus.2023.101869.

[12] K. Habashy et al., “Cough Classification Using Audio Spectrogram Transformer,” in 2022 IEEE Sensors Applications Symposium (SAS), IEEE, Aug. 2022, pp. 1–6. doi: 10.1109/SAS54819.2022.9881344.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[14] W.-J. Chung, M. Kim, and H.-G. Kang, “C2C: Cough to COVID-19 Detection in BHI 2023 Data Challenge,” 2023. [Online]. Available: https://arxiv.org/abs/2311.00364

[15] H. Ayadi, A. Elbéji, V. Despotovic, and G. Fagherazzi, “Digital Vocal Biomarker of Smoking Status Using Ecological Audio Recordings: Results from the Colive Voice Study,” Digit Biomark, vol. 8, no. 1, pp. 159–170, Aug. 2024, doi: 10.1159/000540327.

[16] Z. Ma et al., “Automatic Speech-Based Smoking Status Identification,” Lecture notes in networks and systems, pp. 193–203, Nov. 2022, doi: 10.1007/978-3-031-10467-1_11.

[17] A. H. Poorjam, S. Hesaraki, S. Safavi, H. van Hamme, and M. H. Bahari, “Automatic Smoker Detection from Telephone Speech Signals,” 2017, pp. 200–210. doi: 10.1007/978-3-319-66429-3_19.

[18] 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.

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

[20] D. Desblancs, “Self-Supervised Beat Tracking in Musical Signals with Polyphonic Contrastive Learning,” May 2022. doi: 10.48550/arXiv.2201.01771.

[21] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, “Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review,” Electronics (Basel), vol. 11, no. 22, p. 3795, Nov. 2022, doi: 10.3390/electronics11223795.

[22] M. Grandini, E. Bagli, and G. Visani, “Metrics for Multi-Class Classification: an Overview,” ArXiv, vol. abs/2008.05756, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:221112671