



Article

Advancing Cough Classification: Swin Transformer vs. 2D CNN with STFT and Augmentation Techniques

Malak Ghourabi , Farah Mourad-Chehade  and Aly Chkeir

Computer Science and Digital Society (LIST3N), University of Technology of Troyes, 10000 Troyes, France; farah.chehade@utt.fr (F.M.-C.); aly.chkeir@utt.fr (A.C.)

* Correspondence: malek.ghourabi@utt.fr

Abstract: Coughing, a common symptom associated with various respiratory problems, is a crucial indicator for diagnosing and tracking respiratory diseases. Accurate identification and categorization of cough sounds, specially distinguishing between wet and dry coughs, are essential for understanding underlying health conditions. This research focuses on applying the Swin Transformer for classifying wet and dry coughs using short-time Fourier transform (STFT) representations. We conduct a comprehensive evaluation, including a performance comparison with a 2D convolutional neural network (2D CNN) model, and exploration of two distinct image augmentation methods: time mask augmentation and classical image augmentation techniques. Extensive hyperparameter tuning is performed to optimize the Swin Transformer's performance, considering input size, patch size, embedding size, number of epochs, optimizer type, and regularization technique. Our results demonstrate the Swin Transformer's superior accuracy, particularly when trained on classically augmented STFT images with optimized settings (320×320 input size, RMS optimizer, 8×8 patch size, and an embedding size of 128). The approach achieves remarkable testing accuracy (88.37%) and ROC AUC values (94.88%) on the challenging crowdsourced COUGHVID dataset, marking improvements of approximately 2.5% and 11% increases in testing accuracy and ROC AUC values, respectively, compared to previous studies. These findings underscore the efficacy of Swin Transformer architectures in disease detection and healthcare classification problems.



Citation: Ghourabi, M.; Mourad-Chehade, F.; Chkeir, A. Advancing Cough Classification: Swin Transformer vs. 2D CNN with STFT and Augmentation Techniques. *Electronics* **2024**, *13*, 1177. <https://doi.org/10.3390/electronics13071177>

Academic Editors: Fan Yang, Zongwei Wu, Virginie Fresse, Chao Li and Slaviša Jovanović

Received: 8 February 2024

Revised: 18 March 2024

Accepted: 21 March 2024

Published: 22 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: 2D CNN; cough classification; COUGHVID dataset; disease detection; data augmentation; hyperparameter tuning; short-time Fourier transform (STFT); Swin Transformer

1. Introduction

Coughing is a prevalent symptom linked to various respiratory conditions and serves as a crucial indicator for diagnosing and monitoring respiratory diseases. Accurate detection and classification of cough sounds offer valuable insights into individuals' health, particularly in distinguishing between wet and dry coughs. This distinction is essential as it assists in determining the nature and severity of respiratory infections.

Wet coughs, characterized by the presence of mucus or phlegm, are often associated with conditions like bronchitis or pneumonia. In contrast, dry coughs, which do not produce mucus, typically result from irritation or inflammation in the airways, as seen in allergies, asthma, or viral infections.

This differentiation is not just crucial for diagnosis but also for the management of respiratory conditions, providing valuable insights into the underlying causes and guiding treatment approaches. Coughs can be characterized by different phases and frequencies, and these characteristics vary between wet and dry coughs. In dry coughs, all three phases of a cough signal are usually observed, with the initial burst of energy followed by a phase of relatively less energy, especially at higher frequencies. This may indicate a less productive and drier cough. In wet coughs, associated with mucus, the second phase often

shows increased energy and activity, especially at higher frequencies, indicative of a more productive cough associated with clearing the airways [1].

Traditional diagnostic approaches for coughs rely heavily on subjective assessments by healthcare professionals, posing challenges such as time consumption, high costs, and potential human errors. However, recent advancements in audio analysis and machine learning present opportunities for automated cough detection and classification. Utilizing digital signal processing and pattern recognition, audio-based cough analysis offers objective and efficient tools for healthcare practitioners to assess cough characteristics [2].

Several methods have been explored for classifying wet and dry coughs through audio analysis. One prevalent technique involves extracting mel-frequency cepstral coefficients (MFCCs) from cough sounds. MFCCs capture audio signal spectral characteristics and have proven successful in various audio processing applications. Applying machine learning algorithms, such as support vector machines (SVM) or convolutional neural networks (CNN), to these MFCC features enables accurate wet and dry cough classification [3]. Also, approaches that employ continuous wavelet transform (CWT) and short-time Fourier transform (STFT), providing time-frequency representations of coughs signals, reflect discriminative features for wet and dry cough classification [4].

Moreover, statistical features computed from cough signals, such as mean, standard deviation, energy, and zero-crossing rate, have been utilized in classification tasks. These features capture different aspects of the cough waveform and serve as inputs to different machine learning algorithms [5].

Moving to deep learning models, particularly recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, have shown promise in audio classification tasks. Processing raw audio samples or spectrogram representations of cough sounds, these models demonstrate high accuracy in wet and dry cough classification [6].

In recent years, the application of state-of-the-art deep learning architectures, such as Swin Transformers, has gained prominence in various domains, including medical audio analysis. The unique ability of these transformer-based models to capture long-range dependencies and hierarchical features makes them particularly promising for tasks like cough classification. Swin Transformers, with their shifted window self-attention mechanism, offer an innovative approach to processing image patches, allowing for efficient feature extraction across diverse spatial scales. Leveraging the pre-trained representations from these models or fine-tuning them on cough-specific datasets can potentially enhance the robustness and accuracy of cough classification systems. The exploration of transformer architectures in the realm of medical audio analysis signifies a promising direction for advancing automated diagnostic tools and improving the understanding of respiratory conditions based on cough sound patterns [7,8].

The application of sophisticated machine and deep learning techniques in the classification of wet and dry coughs based on audio signals has seen significant advancements. Various studies have explored diverse methodologies, data types, and metrics, particularly in the context of respiratory illnesses and infectious diseases. This section reviews the key contributions and methodologies related to the classification of wet and dry coughs using artificial intelligence techniques. In our literature, we focused on the approaches that have used the COUGHVID crowdsourced data, as in our case. The COUGHVID data represent the widest, most diverse, and challenging public coughs data.

The COUGHVID public dataset comprises more than 25,000 crowdsourced cough audio samples from diverse age groups, genders, and countries. This dataset encompasses coughs from both COVID-19 patients and healthy individuals. Collected through an online platform, participants recorded their coughs and provided information about their age, gender, and health status. The dataset includes a meta file containing details such as file name, subject status, age, gender, health condition, cough type, and a cough detection score [9]. This score, determined by an automatic detection algorithm, represents the probability that the audio contains cough sounds. The cough type label indicates the cough as wet or dry. This makes the dataset a good fit for wet and dry cough classification tasks. Different

approaches for wet and dry coughs classification using the crowdsourced COUGHVID dataset have been presented in the literature. In Table 1 below, we summarize the previous methods and results for wet and dry cough classification using COUGHVID dataset.

Table 1. Previous trials for wet and dry cough classification using COUGHVID dataset.

Reference	Method	Year	Dataset Information	Results
[10]	Audio features and XGBoost using Bayesian optimization	2021	1659 coughs	Testing accuracy: 66%
[11]	Adaptive synthetic (ADASYN) oversampling cepstral-based statistical features and MLP	2022	1145 dry cough and 409 wet cough samples with 80–20 training–testing split	Testing accuracy: 85.84%
[12]	Oversampling via the SMOTE algorithm and ANN, or kNN/spectral and cepstral features	2021	1659 with and without segmentation	KNN classifier: AUC score of 0.61/ANN model an AUC score of 0.63
[13]	Cochleagram image classification using CNN with data augmentation during training using SMOTE	2022	396: 34 wet vs. 362 dry with eight-fold cv cough samples from 77 recordings (7 wet cough recordings and 70 dry cough recordings) (all four physicians agreement)	AUC ROC of 80.71 for automatic segmentation and 83.76 in manual segmentation
[14]	Using spectral, cepstral, fractal, and nonlinear time-series analyses and neural network pattern recognition	2023	70 wet vs. 30 dry cough samples 70 signals are used for training; 15 signals are used for testing and the last 15 signals are used for validation	Prediction accuracy = 99%
[15]	Energy envelope peaks, crest factors, zero-crossings, and formant frequencies 1–4 introduced to SVM and LR	2023	870 cough samples: 347 wet cough and 523 dry cough samples	SVM achieves an average testing accuracy of 71.26% and an F1-score of 67.94%; LRM classifier achieves an accuracy of 71.26% and an F1-score of 68.45%

In this paper, we introduce an outperforming approach in classifying dry and wet coughs. Our method involves utilizing short-time Fourier transform (STFT) representations of the crowdsourced COUGHVID cough data as input to a Swin Transformer model. This model is compared with a traditional CNN 2D besides incorporating different image augmentation techniques, i.e., time-masking and classical image augmentation. Also, fine-tuning is implemented to enhance the Swin Transformer architecture and performance.

2. Materials and Methods

The methodology of this work involves comparing two distinct approaches for classification task of wet and dry coughs using STFT representations: a traditional 2D CNN model and a Swin Transformer model. Both models undergo hyperparameter tuning and are subjected to different image augmentation techniques at the input level.

2.1. Dataset

The dataset utilized in this study is the COUGHVID public dataset, consisting of more than 25,000 cough audio recordings contributed by individuals of various age groups, genders, and geographical locations. This dataset comprises coughs from both COVID-19 patients and individuals in good health. Data collection occurred through an online platform where participants recorded their coughs and provided information about their age, gender, and health status [9].

Each audio recording is associated with metadata which include details such as the file name, the participant's health status, age, gender, a cough detection score, and other cough characteristics (i.e., wet or dry cough). The cough detection score represents the likelihood that the audio contains cough sounds, as determined by an automated detection algorithm. To ensure that all the audios used in our analysis genuinely include coughs, we specifically included those with a score of 80% or higher. Regarding the wet/dry labeling, it is conducted by four health practitioners. A voting mechanism is introduced to indicate the cough as wet or dry, where the cough is included in the study only if three out of four physicians agreed on the type of the cough. Therefore, after applying the cough detection limit and the labels' voting mechanism, the work included a total of 1432 dry coughs and 427 wet coughs. Two approaches are implemented and tested in this study based on the dataset version. The first approach performs random undersampling to balance the classes of the dataset. However, another approach tackles the unbalanced dataset.

2.2. Audio Data Preprocessing

To prepare the audio signals for analysis, we begin by removing any initial and final periods of silence to reduce the computational power. Afterward, we make sure that all signals share a common sampling frequency of 48,000 Hz for uniformity. Next, we apply a pre-emphasis technique to highlight essential features within the audio data [16]. Following this, we standardize the audio data by normalizing them, ensuring that the values fall within a range of -1 to 1 . This normalization simplifies the data for neural networks, making it more conducive for learning [17].

Also, one of the preprocessing steps is unifying the lengths of the cough audios. This is achieved by applying spline interpolation, where the lengths of all audios are unified to the average length [18]. Next, a pre-emphasis filter is applied to the signals, which increases the amplitude of high-frequency bands and decreases the amplitudes of the lower bands of a signal (Equation (1)) [17].

$$y(t) = x(t) - \alpha x(t - 1), \quad (1)$$

where, $y(t)$ is the output signal, $x(t)$ is the input of the pre-emphasis filter, and α is the pre-emphasis coefficient = 0.97

2.3. Extraction of Spectrogram Representations

The STFT representations are generated to capture the local frequency content over time, essential for audio classification. We use 1024 frequency bins (n_{fft}) for high-frequency resolution and set the hop length to 512 for a 50% overlap between frames, balancing spectral precision with computational efficiency. Hann windowing is applied to minimize artifacts in the Fourier transform [17]. Post-generation, STFT images undergo normalization based on the mean and standard deviation of the three input channels individually and we apply the formula presented below (Equation (2)). In our case, having image input, we normalize them to make the model converge faster. For the STFT images, we compute the mean and standard deviation independently for each of the three channels and then perform normalization accordingly [19]. The normalization formula is as follows:

$$x_{\text{norm}} = \frac{x - \text{mean}(x)}{\text{std}(x)}, \quad (2)$$

where x denotes the values in the three different channels, x_{norm} signifies the normalized values, and $\text{std}(x)$ is the standard deviation of x .

To address class imbalance, random downsampling is employed, resulting in 427 wet cough and 432 dry cough STFT images for model input.

2.4. Implementing the 2D-CNN-Based Model

The 2D convolutional neural network (CNN) is a deep learning approach used for analyzing and classifying images which consists of various layers such as convolutional

layers, pooling layers, and fully connected layers. These layers work together to learn local patterns, downsample the data, and classify the images. By applying filters and pooling operations, the network extracts features and hierarchies of information, enabling it to comprehend complex visual patterns and make accurate predictions [20]. Interestingly, research has shown that shallow networks can be more effective than deep networks when working with small datasets [16].

Based on this insight, we constructed a model consisting of two 2D CNN layers. A fine-tuning of the model is applied using Keras tuner library in python; the tuner searches for the optimal combination of the layers of the model, as well as the kernel, pooling, and filter dimensions. Accordingly, a 3×3 kernel size is set for the two 2D CNN layers, each having 256 filters and 128 filters, respectively. In addition, a batch normalization layer is inserted between the two 2D CNN layers. The model also incorporates two max pooling layers with 3×3 pooling size. Additionally, two ReLU dense layers of size 64 are included with a dropout layer after each of them. Finally, a fully connected layer utilizing sigmoid activation function is inserted for the final classification. The architecture of the model is depicted in the flowchart below (Figure 1). The STFT images serve as an input to the model, where an 80–20–10 training–validation–testing split is inherited. The validation and testing accuracies of the model serve as evaluation metrics to assess its performance.

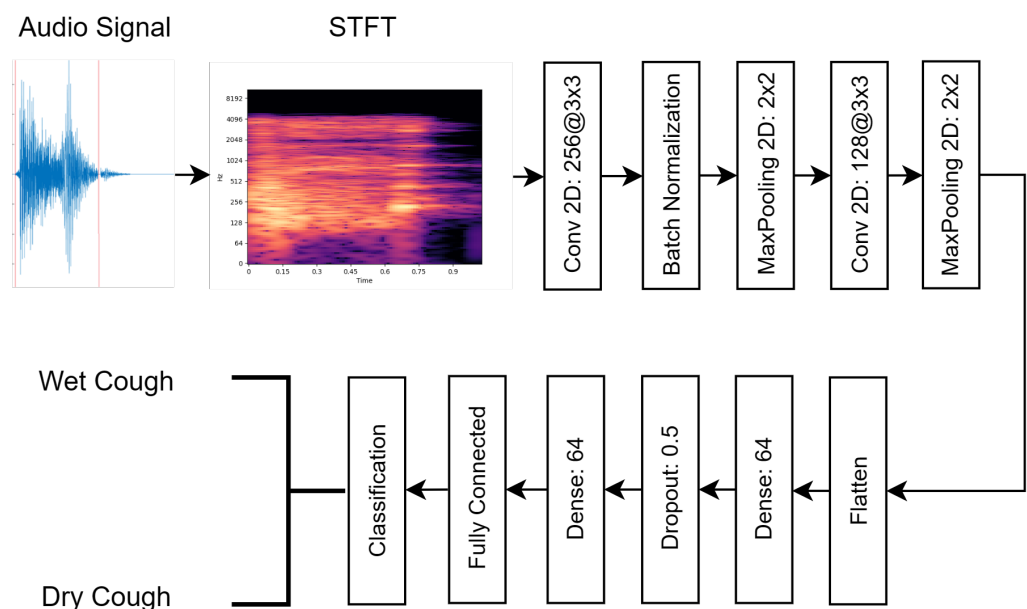


Figure 1. Wet–dry classification workflow and the CNN 2D model architecture.

2.5. Implementing the Swin-Transformer-Based Model

We propose the classification of wet and dry STFT representations using a Swin-Transformer-based model. The Swin Transformer is a novel architecture based on the Transformer for visual recognition tasks. It aims to overcome the limitations of traditional convolutional neural networks (CNNs) by introducing a hierarchical structure and utilizing the self-attention mechanism.

The Swin transformer is structured into multiple stages, each featuring hierarchical windows that segment the input image into non-overlapping patches, enabling the capture of local and global context. Within each stage, a self-attention mechanism with multi-head attention assesses dependencies among patches, facilitating the understanding of complex image relationships by considering long-range interactions. A distinctive shifting operation between stages allows patches to change positions within windows, promoting the integration of information across different parts of the image. Additionally, the Swin Transformer incorporates a hierarchical approach to feature maps, progressively reducing

their spatial resolution to decrease computational demands while preserving detailed image representation [21].

Our Swin Transformer model's architecture experimentation involves testing various input sizes to identify the optimal configuration for image processing. Specifically, the input layer accommodates images of sizes 224×224 , 256×256 , 320×320 , or 384×384 , each represented as a $M \times N \times 3$ tensor (Table 2). The primary goal is to determine the most effective image size for our model's architecture. Image augmentation techniques, including random cropping and flipping, are employed to enhance the model's ability to learn from different image variations. This process is essential for the initial layers of our architecture (Table 2, Layers 2–3).

Table 2. The Swin Transformer model's architecture.

Swin Transformer Architecture	
1.	Input Layer
2.	Random Crop Layer
3.	Random Flip Layer
4.	Patch Extract Layer
5.	Patch Embedding Layer
6.	Swin Transformer Layer 1
7.	Swin Transformer Layer 2
8.	Patch Merging
9.	Global Average Pooling
10.	Fully Connected Dense Layer

Following augmentation, the architecture processes the images through patch extraction and embedding stages. Each image patch is converted into a one-dimensional vector and then linearly projected using a learnable weight matrix in the patch embedding layer (Table 2, Layers 4–5). The core of the model consists of two Swin Transformer layers, which are pivotal for the model's ability to capture complex features and relationships within the data (Table 2, Layers 6–7). These layers employ shifted windows-based multi-head self-attention (MSA) modules, MLPs with GELU activation, layer normalization, and residual connections to enhance learning efficiency and model performance.

The concluding phases of our model involve patch merging and global average pooling, leading up to the final classification layer. The patch merging step aggregates patch-level information, which is then compacted through global average pooling. This compact representation is crucial for the model to make accurate predictions based on the entire image. The final classification is executed with a dense layer utilizing a sigmoid activation function, tailored for binary classification tasks like distinguishing between wet and dry STFT images (Table 2, Layers 8–10).

Fine-tuning the model involves adjusting several parameters, including input size, patch size, embedding dimension, the number of epochs, and the optimizer choice (Adam or RMS) [22]. These adjustments are crucial for optimizing the model's performance, ensuring accurate predictions while avoiding overfitting or underfitting. The results, discussed later, highlight the effectiveness of our chosen parameters in achieving high accuracy and efficient training times.

2.6. Application of Augmentation Techniques on STFT Representations

Different versions of the STFT representations are introduced to the 2D-CNN model and the Swin Transformer Model. The first two versions contain the original batches of balanced and unbalanced wet/dry STFT images. Regarding the other two versions, two augmentation techniques are performed in order to increase the training set size and eventually enhance the generalization ability of the model [23].

2.6.1. Time Mask Augmentation

Time mask augmentation is a powerful technique used in audio processing tasks to enhance the performance of deep learning models [24]. In our case, this is achieved by randomly masking consecutive time segments within the audio signal. These masks are applied to random time intervals of the training STFT representations of coughs audio signals. This augmentation helps the model become more robust and generalized by allowing the training of the model on various sets of data. Figure 2 shows an example of a time-masked STFT image. The time mask augmentation is performed on the training set using the “torchaudio.transforms” module’s time masking class in python, where the time mask size is about 0.03 s, applied two times for each image with a 0.27 s interval between each [25]. The training set is augmented to double its size (i.e., a total of 1708 STFT representations).

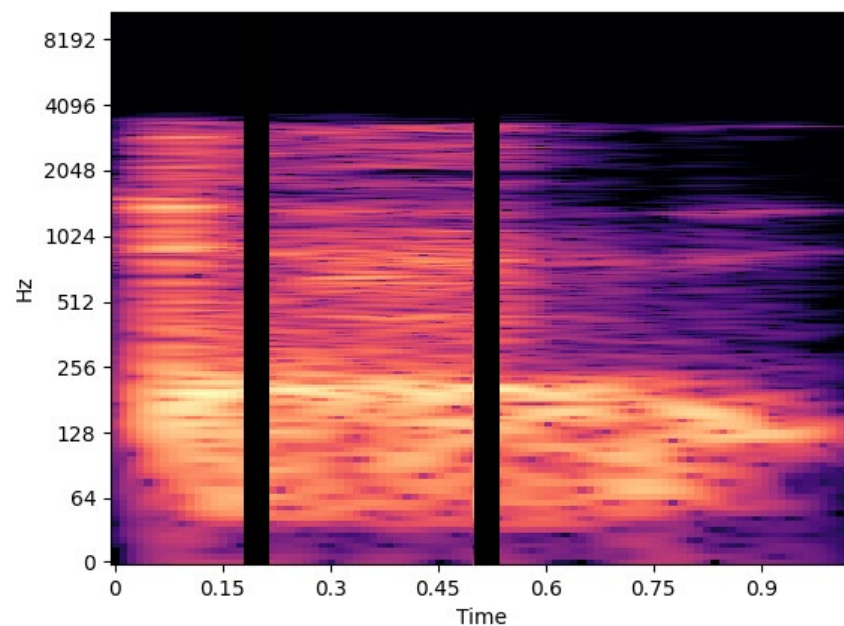


Figure 2. Time-masked STFT image.

2.6.2. General Image Augmentation Techniques

We move to the second approach to STFT image augmentation, where edits and effects like crop, rotation, shear, saturation, brightness, and noise addition are performed using the Roboflow online datasets processing and management tool [26].

The augmentation settings and parameters performed on the wet and dry cough STFT representations are listed below:

- Crop: 0% minimum zoom; 20% maximum zoom;
- Rotation: between -15° and $+15^\circ$;
- Shear: $\pm 15^\circ$ horizontal; $\pm 15^\circ$ vertical;
- Saturation: between -25% and $+25\%$;
- Brightness: between -25% and $+25\%$;
- Noise: up to 5% of pixels.

Accordingly, the number of STFT representations is augmented by a factor of three (i.e., $3 \times$ original representations count). We originally had 429 dry and 422 wet cough STFT images, which are increased in size to 1287 dry and 1266 wet cough STFT representations.

3. Evaluation Metrics

In our analysis, we implemented a robust set of metrics to rigorously evaluate the performance of our classification models. The metrics used are accuracy, ROC AUC, and F1-score [27–29]. Accuracy measures the number of correctly classified samples; however,

the ROC curve is a plot of the true positive rate (TPR) versus the false positive rate (FPR). Below are represented the equations of accuracy, TPR, FPR, and F1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{TPR} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (5)$$

$$\text{F1 - Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

Accuracy provides a straightforward measure of overall correctness but may be inadequate for unbalanced datasets. ROC AUC assesses the model's ability to distinguish between wet and dry coughs across various thresholds, making it robust for unbalanced data by considering true positive and false positive rates. Meanwhile, the F1-score balances precision and recall, offering a single metric that addresses the trade-off between false positives and false negatives, thus ensuring a comprehensive evaluation of the classifier's performance across different aspects essential for wet/dry cough classification tasks. These metrics provide a comprehensive understanding of the model's capabilities across different extents and offer a comprehensive evaluation of the different models' performance across different datasets, architectures, and parameters for classification.

4. Results

In this study, we assessed the performance of both the 2D CNN and the Swin Transformer models in classifying wet and dry coughs using STFT representations. Each model underwent various training trials with different configurations. For the 2D CNN model, trials included using unbalanced, original, and time-mask-augmented STFT representations. The Swin Transformer model was tested with different input sizes and augmentation, including time mask and general image augmentations.

Figures 3 and 4 present the evaluation metrics of the three trials. The models are trained using 100 epochs and Adam optimizer with an input size of 224×224 . The change in the optimizer's choice and input shape did not significantly affect the results; however, upon increasing the number of epochs, an overfitting was noticed; this is the reason behind fixing the number of epochs to 100.

Moving to Swin-Transformer-based architecture, different models are trained with different versions of the dataset, different input image size, as well as a different optimizer. Also, a fine-tuning of the patch size and embedding dimension is performed, leading to a choice of 8×8 for the patch size and 128 as the embedding dimension. Ten different combinations of the Swin Transformer model are trained and established. In the first four combinations, we test for the optimal input size. No significant change is noticed upon changing the optimizer, and a number of epochs equal to 100 is set to prevent overfitting. Table 3 depicts the different model results obtained upon varying the input shape using the balanced version of the dataset. The increase in the input image size is constrained by the maximum allowable dimensions that ensure training proceeds without crashing due to hardware limitations.

Due to the memory requirements, computational cost, and model size constraints, a compromise between these constraints and the accuracy of the model is established for later stages, where a size of 320×320 is employed for the models' trained using the further trials.

Now, after training using the balanced dataset, a trial using the unbalanced dataset is performed with a 320×320 input size, and the results are also depicted in Table 3.

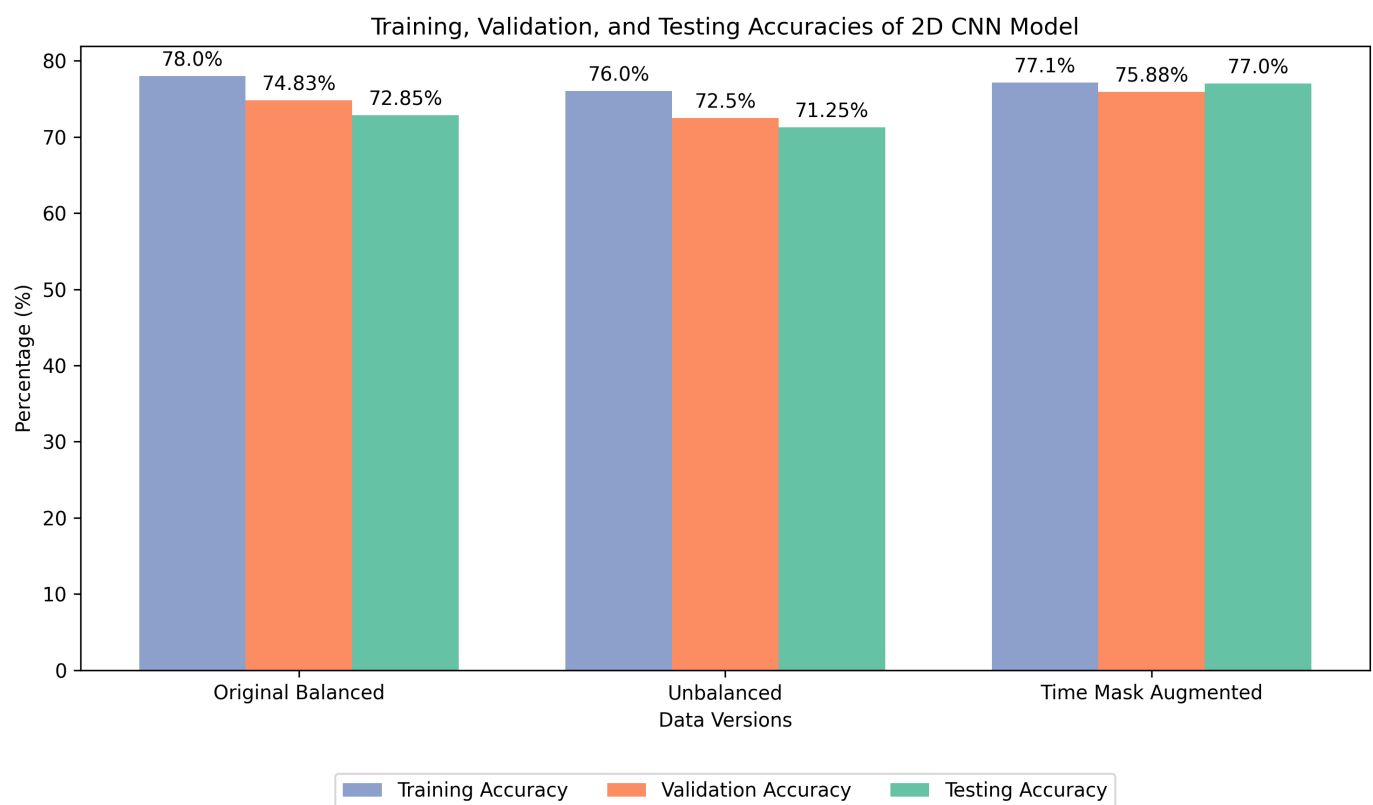


Figure 3. Training, validation, and testing accuracies of 2D CNN model using the different data versions: balanced, unbalanced, and time-mask-augmented.

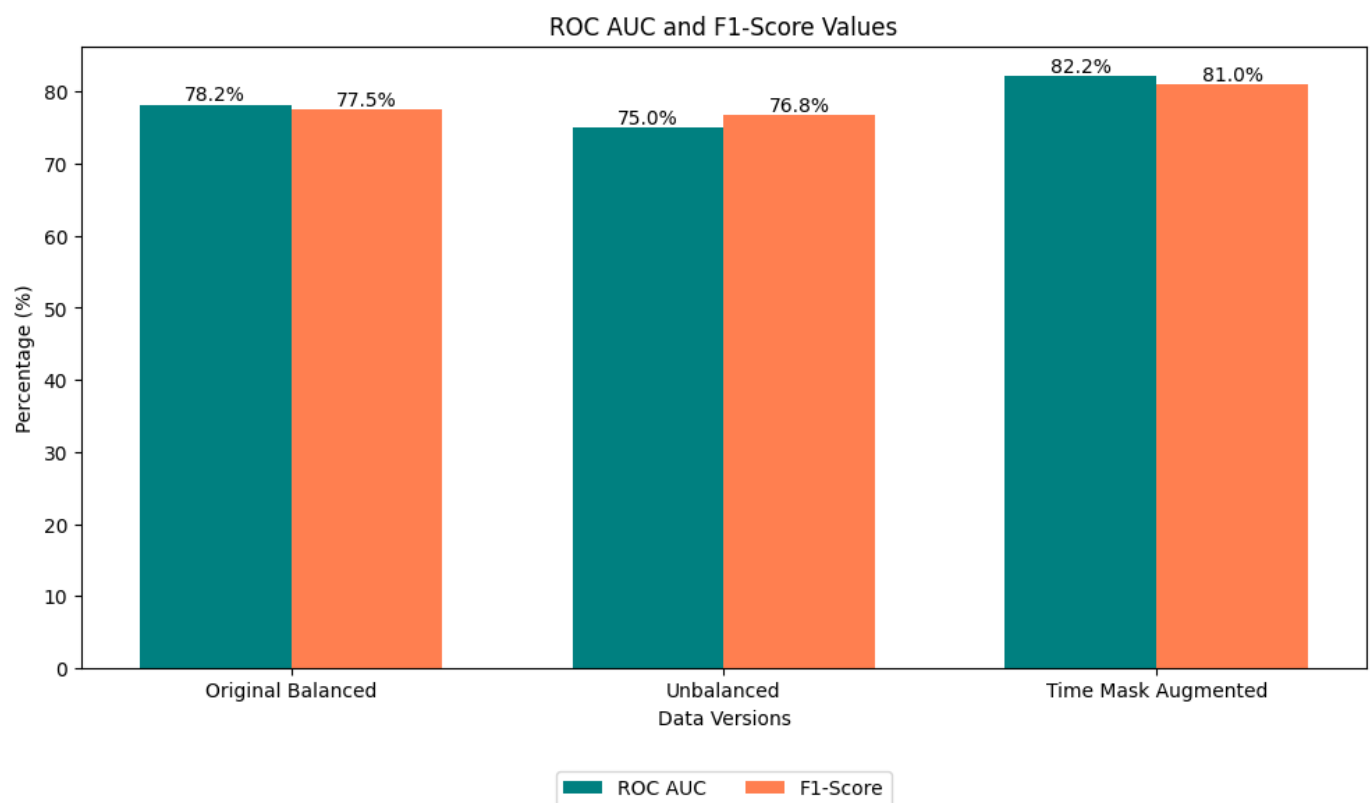


Figure 4. ROC AUC and F1-score values of 2D CNN model using the different data versions: balanced, unbalanced, and time-mask-augmented.

Table 3. Swin Transformer models' accuracies upon different input sizes using the balanced and unbalanced datasets.

Swin Transformer Models Trained Using Balanced STFT Images on Different Input Shapes	Training	Validation	Testing
224 × 224	78.45	74.50%	73.92%
256 × 256	78.96	73.49	74.73%
320 × 320	82.07%	75.84%	74.19%
384 × 384	82.49%	75.84%	76.08%
Swin Transformer Models Trained using Unbalanced STFT Images on Different Input Shapes			
320 × 320	76.3%	72.6%	73.07%

Similarly, another four combinations aim to test the effect of time mask augmentation on the models. Table 4 shows the results of the Swin Transformer models upon training using the time-mask-augmented dataset.

Table 4. Results of training the Swin Transformer models using the time-mask-augmented dataset and different input shapes.

Swin Transformer Models Trained Using Time-Mask Augmented STFT Images on Different Input Shapes	Training Accuracy	Validation Accuracy	Testing Accuracy
224 × 224 and Time Mask Augmentation	77.34	75.88	76.47
256 × 256 and Time Mask Augmentation	77.24	75.88	76.47
320 × 320 and Time Mask Augmentation	78.04	76.27	79.57
384 × 384 and Time Mask Augmentation	78.24	76.12	79.57

In the two last models' combinations, training is performed with the augmented dataset of the general image augmentation techniques (crop, rotation, shear, saturation, brightness, and noise addition). The performance of the models is monitored upon changing the number of epochs and the optimizers. Figure 5 below shows the evolution of the model accuracy and loss as more epochs are performed using Adam as an optimizer. Similarly, the same monitoring is presented in Figure 6; however, the RMS optimizer is implemented instead of Adam. In both models, the learning rate and weight decay are set to 10^{-3} . A learning rate and weight decay of 10^{-3} is a commonly chosen value that balances fast convergence with stability and falls within the typical range for these hyperparameters, which often vary from 10^{-1} to 10^{-6} in machine learning applications. Relatively low learning rate and weight decay help with optimization dynamics; also, they contribute to controlling the complexity of the model and improving generalization performance [30,31]. Figure 7 presents the training, validation, and testing accuracies of both models as well as the testing ROC AUC and F1-score values.

It is important to mention the difference in the computational time between the different scenarios tested on 300 training epochs, where the average training computational time for the 2D-CNN-based models is 13 min for the different datasets. This training time increases significantly when implementing Swin-Transformer-based models to reach an average of 30 min for 320 × 320 input size. However, the time required for model evaluation is approximately equal for both scenarios, with an average of 0.7 s for 2D CNN models and 1 s for Swin Transformer models.

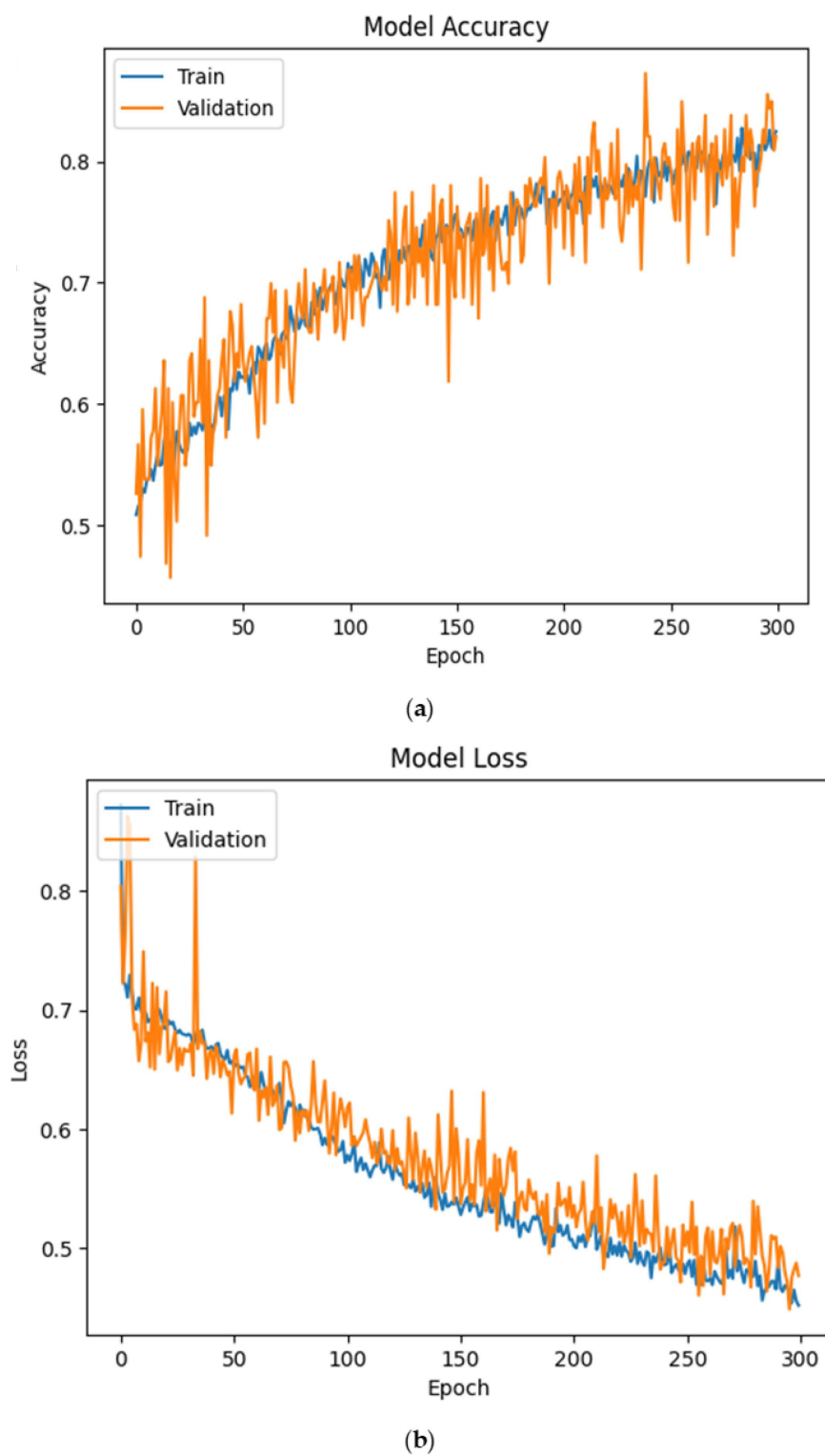
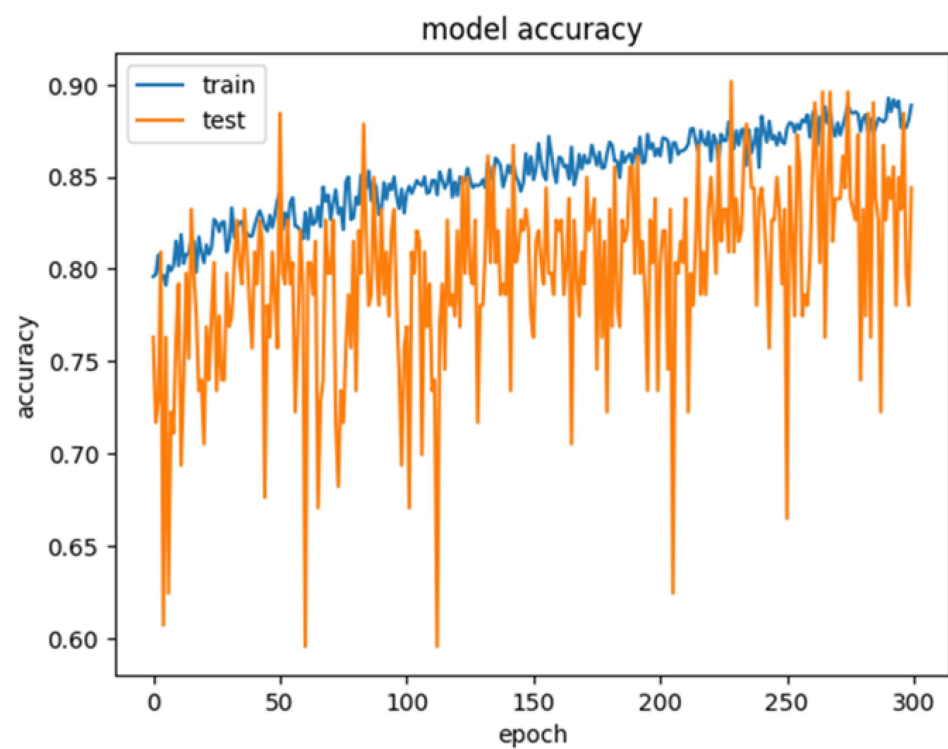
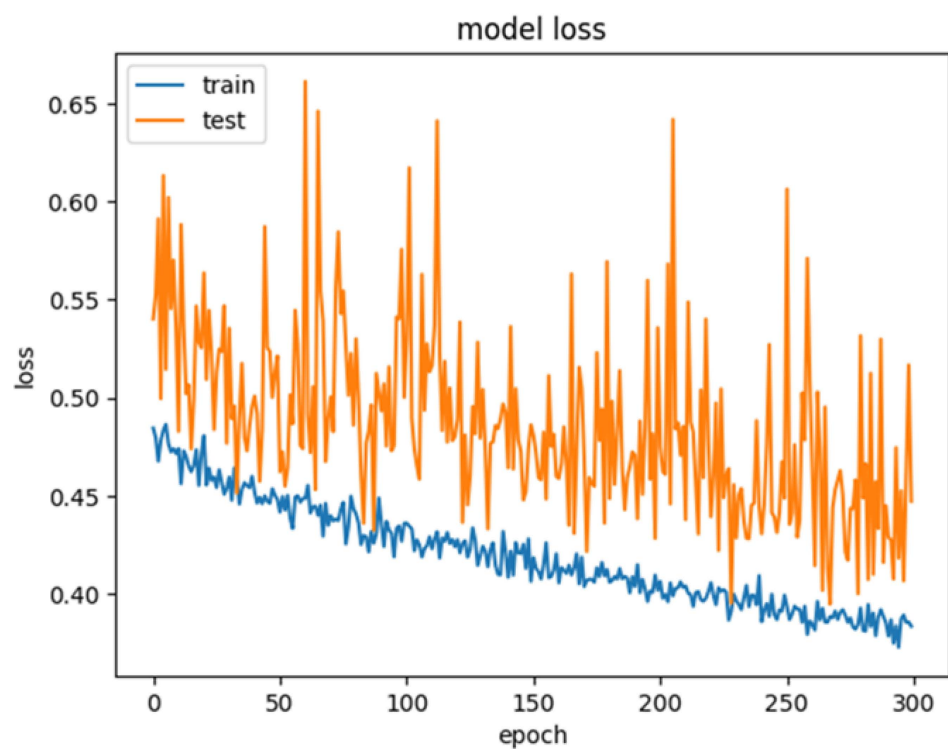


Figure 5. Model's training and validation accuracy (a) and loss (b) evolution when using the Adam optimizer for 300 epochs of training.



(a)



(b)

Figure 6. Model's training and validation accuracy (a) and loss (b) evolution when using the RMSprop optimizer for 300 epochs of training.

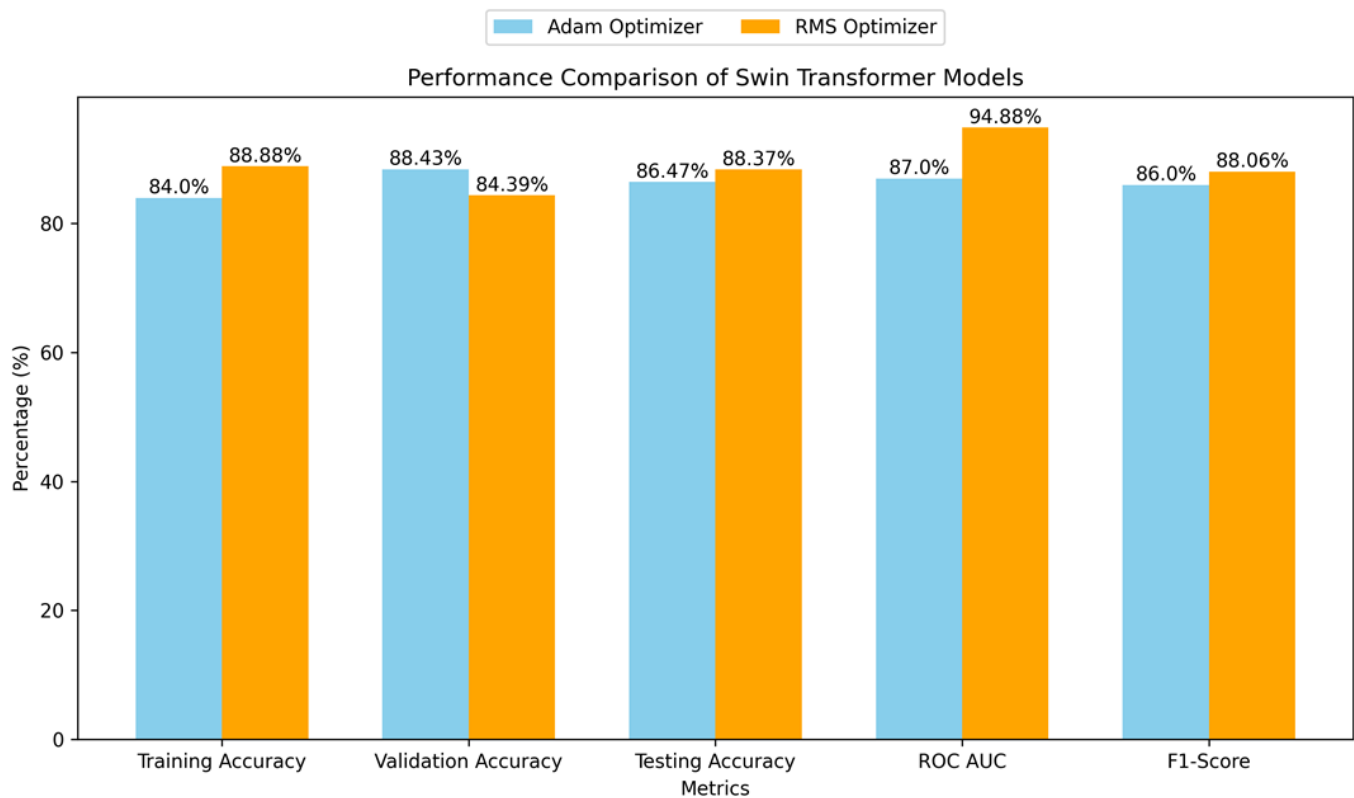


Figure 7. Swin Transformer models results for a 320×320 input size and the proposed image augmentation techniques using either the Adam or RMSprop optimizer at 300 epochs.

5. Discussion

In our study, we employed a Swin-Transformer-based model to classify wet and dry coughs using the COUGHVID dataset. Our approach, distinguished in the literature of 2023, utilizes augmented short-time Fourier transform (STFT) images derived from 429 dry and 422 wet cough recordings. These were expanded to 1287 dry and 1266 wet cough STFT representations through augmentation. The model was trained over 300 epochs using an RMS optimizer, enhanced with weight decay regularization. This methodology yielded a training accuracy of 88.88%, a validation accuracy of 84.39%, and a testing accuracy of 88.37%. Notably, our model achieved an ROC AUC of 94.88% and an F1-score of 88.06%.

The results of our study yield valuable insights:

1. Effectiveness of time mask augmentation: time mask augmentation led to a notable improvement of 4% in testing accuracy (Figures 3 and 4 and Tables 3 and 4).
2. Swin Transformer's performance: The performance of Swin Transformer models was notably influenced by two parameters: input shape and optimizer. Increasing the input shape led to a substantial increase in accuracy, as observed in the testing accuracies of both original and STFT time-mask-augmented images, with improvements of 2.16% and 3.1%, respectively (Tables 3 and 4). Comparison between architectures: The Swin-Transformer-based architecture outperformed the 2D-CNN-based architecture in classifying dry/wet STFT images. When implementing the original dataset, there was an increase in testing accuracy from 72.85% to 76.08% (Figure 3 and Table 3). This difference was even more pronounced (2.57% improvement) when applying time mask augmentation, demonstrating the superiority of the Swin Transformer architecture in wet/dry representation classification.
3. Balanced and unbalanced datasets: Although a balanced dataset has yielded better performance, performing the same trials on an unbalanced dataset reflects a more realistic approach. While balanced datasets ensure equal representation of each

class, mimicking ideal conditions for training, unbalanced datasets better mirror real-world scenarios, where certain classes may be more prevalent than others. Therefore, evaluating model performance on unbalanced data provides insights into how well the classifier generalizes to real-world conditions, where class distributions are often skewed. It helps to identify potential biases and weaknesses in the model's ability to handle unbalanced data, thus guiding improvements in robustness and reliability. Moreover, solutions developed on unbalanced datasets are more likely to be applicable in practical settings, where unbalanced data are common, enhancing the model's practical utility and effectiveness.

4. Robustness enhancement through image augmentation: Image augmentation techniques applied to the training dataset significantly improved the model's robustness and accuracy. There were substantial improvements of 6.4%, 8.56%, and 12% in training, validation, and testing accuracies, respectively (Figure 3 and Table 4).
5. Optimizer selection: While the Adam optimizer provided a smoother training process, the RMS optimizer outperformed it in terms of accuracy. A noticeable increase of approximately 5% in training accuracy and 2% in testing accuracy was observed when using the RMS optimizer. High ROC AUC and F1-score values of 94.88% and 88.06%, respectively, were achieved (Figures 5–7).
6. Computational complexity: Generally, 2D CNNs are considered more computationally efficient than Swin Transformers for image processing tasks. The 2D CNNs utilize efficient operations like convolutions that exploit local dependencies in images and are well-optimized for hardware acceleration on GPUs. However, Swin Transformers rely on self-attention mechanisms, which involve heavier computations compared to convolutions, especially for high-resolution images; the Swin Transformer has achieved better accuracy but at the cost of higher computational demands [32].
7. Comparative analysis with literature: Unlike many literature approaches that often start with unbalanced datasets and apply data augmentation techniques like ADASYN and SMOTE, our approach began with a balanced dataset. Data augmentation was applied symmetrically to both classes to preserve uniformity in dealing with STFT representations of wet and dry coughs, reducing the risk of model bias. Additionally, we ensured that augmented data were only integrated into the training split, maintaining the integrity of validation and testing datasets for unbiased performance evaluation. These findings highlight the efficacy of various strategies in improving cough sound classification using deep learning models, with a focus on Swin Transformer architecture, input shape, image augmentation, and optimizer selection. Comparing the accuracies in [10,11,15], our model outperformed them by an increase of 22%, 2.5%, and 16.74% in the testing accuracy, respectively. Now, moving to the comparison of AUC ROC in [13] with our approach, we surpassed the results by an improvement of 11.12%. The testing accuracy in [14] is found to be 99%; however, a total of only 100 cough audios was used for this approach, with only 15 audios dedicated for testing.

6. Conclusions and Perspectives

Our investigation into the application of Swin Transformers for cough classification marks a significant advancement in employing sophisticated machine learning techniques within healthcare diagnostics. The study titled “Advancing Cough Classification: Swin Transformer vs. 2D CNN with STFT and Augmentation Techniques” reveals the superior capability of Swin Transformers over conventional 2D CNN models, especially in distinguishing between wet and dry coughs through STFT representations. The Swin Transformer model, particularly when trained on augmented STFT images, exhibited exceptional proficiency in classifying cough types. This augmented approach significantly boosted testing accuracy and ROC AUC values, underscoring the model's enhanced predictive performance. Additionally, the employment of innovative methodologies like time mask and

general image augmentation techniques has substantially increased the model's robustness and its ability to generalize across various cough sound representations.

A detailed comparative analysis between the Swin Transformer and traditional 2D CNN models has been provided, laying down a comprehensive benchmark for future endeavors in the domain of cough sound analysis. Looking forward, several promising directions are envisioned for extending the scope of our research. Applying our model to a broader spectrum of cough datasets could offer a more in-depth evaluation of its generalization capability and its relevance to diverse demographic and geographic populations. Delving into hyperparameter optimization could unlock further performance enhancements, with the exploration of alternative optimizers, learning rate schedules, and regularization techniques potentially providing deeper insights into improving model accuracy and efficiency. Moreover, integrating time-masked with general image augmentation strategies might establish a more robust framework for cough sound classification, leveraging the strengths of both approaches to set new performance benchmarks.

However, our study is not without its limitations. The computational demands of Swin Transformers, notably higher than those of 2D CNNs, present a significant challenge. While Swin Transformers excel in capturing complex patterns within STFT representations, their reliance on self-attention mechanisms necessitates considerable computational resources, particularly for processing high-resolution images. Furthermore, the extensive range of models and configurations tested posed a formidable challenge in re-evaluating computational times for previously examined scenarios, rendering a complete reassessment unfeasible within our study's scope. Additionally, while our model demonstrates impressive accuracy on the COUGHVID dataset, its performance across other cough sound datasets has yet to be thoroughly investigated, highlighting a need for further research to ascertain the model's predictive capabilities across varying datasets.

In conclusion, our study not only underscores the effectiveness of Swin Transformers in cough sound classification but also presents new opportunities for advancing disease detection and healthcare classification systems. Despite facing challenges related to computational efficiency, our work significantly contributes to the ongoing efforts in harnessing advanced deep learning architectures for medical diagnostics. We are optimistic that future progress in computational technology and algorithmic efficiency will amplify the impact of Swin Transformers in healthcare applications, paving the way for the development of improved diagnostic tools and enhancing patient outcomes.

Author Contributions: The idea and methodology of this research was established by M.G., as well as the analysis and validation and the writing the original draft preparation. All this research work was validated, supervised, and administered by A.C. and F.M.-C., A.C. and F.M.-C. worked on the review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: We would like to thank the region Grand Est for funding this work/CD10.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study by the time of data collectors in [9].

Data Availability Statement: The COUGHVID dataset implemented in this study is adapted from [9].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, M.; Sykes, D.L.; Brindle, K.; Sadofsky, L.R.; Morice, A.H. Chronic cough—The limitation and advances in assessment techniques. *J. Thorac. Dis.* **2022**, *14*, 5097–5119. [CrossRef] [PubMed]
2. Huang, Y.-P.; Mushi, R. Classification of Cough Sounds Using Spectrogram Methods and a Parallel-Stream One-Dimensional Deep Convolutional Neural Network. *IEEE Access* **2022**, *10*, 97089–97100. [CrossRef]
3. Amrulloh, Y.A.; Wati, D.A.R.; Pratiwi, F.; Triasih, R. A novel method for wet/dry cough classification in pediatric population. In Proceedings of the 2016 IEEE Region 10 Symposium (TENSymp), Bali, Indonesia, 9–11 May 2016; pp. 125–129.
4. Erdoğan, Y.E.; Narin, A. COVID-19 detection with traditional and deep features on cough acoustic signals. *Comput. Biol. Med.* **2021**, *136*, 104765. [CrossRef] [PubMed]

5. Lim, W.L.; Chan, W.Y.; Sani, L.L.; Chew, J.T.H.; Chua, K.C.; Ng, B.K.; Tan, S.S.L. Automatic Cough Detection in COVID-19 Patients: A Machine Learning Approach. *Front. Med.* **2021**, *8*, 693809.
6. Valdes, J.; Habashy, K.; Xi, P.; Cohen-McFarlane, M.; Wallace, B.; Goubran, R.; Knoefel, F. Cough Classification with Deep Derived Features using Audio Spectrogram Transformer. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 1729–1739. [\[CrossRef\]](#)
7. Mauricio, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [\[CrossRef\]](#)
8. Garg, M.; Gajjar, P.; Shah, P.; Shukla, M.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A. Comparative Analysis of Deep Learning Architectures and Vision Transformers for Musical Key Estimation. *Information* **2023**, *14*, 527. [\[CrossRef\]](#)
9. Orlandic, L.; Teijeiro, T.; Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Sci. Data* **2021**, *8*, 156. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Leirgulen, J.; Nuris-Souquet, M.; Lévy-Fidel, C.; Orlandic, L. Dry vs. Wet Cough Automatic Classification Using the COUGHVID Dataset. 2021. Available online: <https://www.semanticscholar.org/paper/Dry-vs-Wet-Cough-Automatic-Classification-using-the-Leirgulen-Nuris-Souquet/8ca8cf2ab92cb77b016de875522ad3ac2f21840b> (accessed on 7 February 2024).
11. Pande, S.; Patil, A.; Petkar, S. Dry and Wet Cough Detection using Fusion of Cepstral base Statistical Features. In Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 23–25 March 2022; pp. 874–878. [\[CrossRef\]](#)
12. Celik, D.; Mainusch, N.; Oliva, X.; Jurgens, I. Cough Classifier CS-433 Machine Learning: Project 2. 2021. Available online: <https://www.epfl.ch/labs/mlp/wp-content/uploads/2021/05/crpmlcourse-paper834.pdf> (accessed on 7 February 2024).
13. Sharan, R.V. Productive and Non-Productive Cough Classification Using Biologically Inspired Techniques. *IEEE Access* **2022**, *10*, 133958–133968. [\[CrossRef\]](#)
14. Renjini, A.; Swapna, M.N.S.; Kumar, K.N.S.; Sankararaman, S.I. Time series and mel frequency analyses of wet and dry cough signals: A neural net classification. *Phys. A Stat. Mech. Its Appl.* **2023**, *626*, 129039. [\[CrossRef\]](#)
15. Andrei, P.C.S.; Madamba, C.A.J.; Guico, M.L.C.; Galicia, J.K.A. Wet and Dry Cough Classification System Using Support Vector Machine and Logistic Regression. In Proceedings of the 2023 9th International Conference on Computer and Communication Engineering (ICCE), Kuala Lumpur, Malaysia, 15–16 August 2023; pp. 252–257. [\[CrossRef\]](#)
16. Prabakaran, D.; Sriuppili, S. Speech Processing: MFCC Based Feature Extraction Techniques—An Investigation. *J. Phys. Conf. Ser.* **2021**, *1717*, 012009. [\[CrossRef\]](#)
17. Nema, B.M.; Abdul-Kareem, A.A. Preprocessing signal for Speech Emotion Recognition. *Al-Mustansiriyah J. Sci.* **2018**, *28*, 157–165. [\[CrossRef\]](#)
18. Keesling. Cubic Splines. University of Florida. Available online: <https://people.clas.ufl.edu/kees/files/CubicSplines.pdf> (accessed on 9 June 2023).
19. Albert, S.; Wichtmann, B.D.; Zhao, W.; Maurer, A.; Hesser, J.; Attenberger, U.I.; Schad, L.R.; Zöllner, F.G. Comparison of Image Normalization Methods for Multi-Site Deep Learning. *Appl. Sci.* **2023**, *13*, 8923. [\[CrossRef\]](#)
20. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. 2021. Available online: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper (accessed on 7 February 2023).
22. Kandel, I.; Castelli, M.; Popović, A. Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images. *J. Imaging* **2020**, *6*, 92. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [\[CrossRef\]](#)
24. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779. [\[CrossRef\]](#)
25. TimeMasking—Torchaudio 2.2.0.dev20231121 Documentation. Available online: <https://pytorch.org/audio/main/generated/torchaudio.transforms.TimeMasking.html> (accessed on 21 November 2023).
26. Roboflow. Available online: <https://app.roboflow.com> (accessed on 7 February 2023).
27. Soliński, M.; Łepeć, M.; Kołtowski, Ł. Automatic cough detection based on airflow signals for portable spirometry system. *Inform. Med. Unlocked* **2020**, *18*, 100313. [\[CrossRef\]](#)
28. Wikipedia Contributors. Receiver Operating Characteristic. Wikipedia. 20 March 2019. Available online: https://en.wikipedia.org/wiki/Receiver_operating_characteristic (accessed on 7 February 2023).
29. Pérez-Sala, L.; Curado, M.; Tortosa, L.; Vicent, J.F. Deep learning model of convolutional neural networks powered by a genetic algorithm for prevention of traffic accidents severity. *Chaos Solitons Fractals* **2023**, *169*, 113245. [\[CrossRef\]](#)
30. Zhang, G.; Wang, C.; Xu, B.; Grosse, R. Three Mechanisms of Weight Decay Regularization. *arXiv* **2018**, arXiv:1810.12281.

31. Wilson, D.R.; Martinez, T.R. The need for small learning rates on large problems. In Proceedings of the IJCNN'01, International Joint Conference on Neural Networks (Cat. No.01CH37222), Washington, DC, USA, 15–19 July 2001; Volume 1, pp. 115–119. [[CrossRef](#)]
32. Wu, P.; Pan, Z.; Tang, H.; Hu, Y. Cloudformer: A Cloud-Removal Network Combining Self-Attention Mechanism and Convolution. *Remote Sens.* **2022**, *14*, 6132. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.