香 港 大 學
THE UNIVERSITY OF HONG KONG

LAIDLAW FOUNDATION

Utilizing Metamorphic Testing to assess generalizability of medical sound systems: a case study in heart sound classification.

Yang Liuqing

Year 2, BEng (EngSc)

Laidlaw Undergraduate Research and Leadership

Scholarship Program Report

Supervised by Dr Joshua HO

School of Biomedical Sciences, LKS Faculty of Medicine

The University of Hong Kong

Table of Contents

Abstract

Due to complex input spaces and difficulty determining correct outputs, traditional testing strategies can hardly evaluate biomedical machine-learning (ML) systems effectively. Metamorphic testing (MT), a software testing technique that generates follow-up test cases from the source test case, is widely employed to solve the oracle problem. However, the research of MT in the medical domain mostly focuses on image analysis systems, and its application in sound processing systems has yet to be carefully investigated.

In view of this situation, we propose using metamorphic testing to assess the generalizability of ML-based sound-processing models by designing MRs related to object mutations. It will potentially also contribute to explainable Artificial Intelligence by enhancing the interpretability of such models. We hypothesize that we can evaluate model generalizability, which is the ability to classify or predict new data different form the training samples, by comparing system outputs of the original and mutated audio files.

Focusing on heart sound classification models as our experimental subject, we strategically mutate the most activated regions in inputs with class activation mapping and gradient masking. Subsequently, we assess the consistency of prediction results of the test case pair between the original and mutated data. A model with excellent generalizability should output inconsistent predictions due to the dampened impact of the most activated regions in inputs, while a model with poorer generalizability may remain insensitive to mutations and still output the same classification result.

To verify our proposal, we tested two DenseNet models across three datasets, revealing average consistency rates of 10.12% and 25.79% respectively. The divergent outcomes reveal different generalizability of the models. Our approach also enhances the interpretability of deep learning models by visualizing model reactions within the input spectrograms. Thus, the proposed testing technique is not only effective in assessing generalizability of heart sound classification model, but also beneficial for the realization of explainable AI in biomedical systems.

Keywords: *metamorphic testing; heart sound classification; software testing; explainable AI*

1. Introduction

Metamorphic testing (MT) is a software testing technique that generates follow-up test cases from the source test case according to the pre-designed metamorphic relations (MRs). MRs describes the anticipated output properties of a correct program, and the system fails when any pair of source and follow-up test cases violates their corresponding MR [1]. MT effectively alleviates the oracle problem by analyzing output relations between multiple program executions, rather than evaluating each individual test case in isolation [2]. It also offers a solution to the reliable test set problem as it enhances the failure detection capability of original test cases by generating potent follow-up ones [3].

MT has wide applications in various domains such as web services, autonomous machinery, and machine learning. Its advantages of efficient test-case generation and omission of testing oracles especially highlights in the biomedical field, where scarcity of testing data and complexity of system input spaces pose rough challenges. However, existing MT research predominantly concentrated on image-related programs with limited investigation into sound-processing models, even though sound auscultation is vital [4]. Drawbacks in medical sound auto-analysis systems can lead to serious misdiagnoses and erroneous clinical decisions, threatening patients' safety and precipitating unnecessary anxiety [5]. Thus, developing effective testing techniques for sound processing systems is imperative.

Among the sound auscultation systems, heart sound classification models are of particular attention. Heart sounds mainly comprise the first heart sound (S1) and the second heart sound (S2) in sequential order. Other sounds include third and fourth heart sounds, heart murmurs, and adventitious sounds may also occur during the cardiac cycle. Normal heart sound components also have specific ranges of intensity,

frequency, and durations [6]. Furthermore, background interfering noises such as breathing sounds and medical instrument noises are commonplace in heart sound recordings. An optimal model should focus on prominent properties and features of heart sound events while demonstrating resilience against extraneous noise.

Traditional metrics such as accuracy and precision assess the program's performance on specific test cases but fall short in evaluating the adaptability of deep learning systems to novel data. However, handling new data independent from the training set is particularly crucial for biomedical models, which must carry out clinical tasks for diverse patients. Generalizability, therefore, serves as a metric to determine whether a model can efficiently make predictions or diagnoses on data beyond its training dataset [7]. A model with poor generalizability may excel on training samples but performs poorly on dissimilar inputs. Only a model with good generalizability can effectively handle diverse data and be confidently deployed in clinical settings [8].

In view of these aspects, we propose using metamorphic testing to assess the generalizability of heart sound classification models. We mutate the most activated regions in the inputs via gradient masking and detect consistency rate of model predictions towards paired test cases. A low consistency rate indicates that the model has good generalizability as it is sensitive to newly introduced mutations.

2. Related information

2.1 Datasets

Our research employs heart sound sets sourced from the PhysioNet/CinC Challenge 2016 dataset. Each of these training sets comprises multiple waveform audio files alongside a corresponding CSV file recording sound quality assessments and classification labels (normal/abnormal). To establish a robust experimental framework, we created one training and three validation sets for each model under

investigation.

2.2 Mel spectrogram

For clearer visualization and more streamlined processing, we transform waveform audios into mel spectrograms for model inputs. The transformation applies frequency-domain filter banks to windowed audio signals and generates spectrograms with frequencies mapped to the mel scale. This conversion offers notable advantages over employing raw sound files as model inputs. Mel spectrograms have significantly lower dimensions than audio data, resulting in expedited and more efficient model training. Furthermore, the interpretability of spectrogram analysis surpasses that of sound files due to the availability of various visualization techniques that can display the model's responsiveness to input images. Foe instance, saliency maps or heatmaps can effectively accentuate the activated and focused patterns within the inputs. Remarkably, the utilization of established libraries and tools, such as librosa, further streamlines the preprocessing of spectrograms.

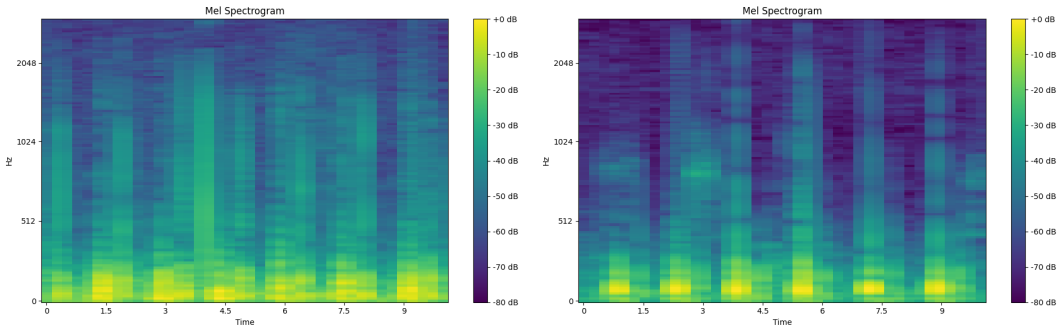

Figure 1. Examples of input mel spectrograms

2.3 Dense Neural Network

Among the neural network architectures for heart sound classification, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are most frequently employed. Our study selected DenseNet121 and DenseNet169 from the Keras application library, having 242 and 578 layers respectively. DenseNet, short for "Densely Connected Convolutional Networks," adopts a

densely interconnected structure by fostering dense connections among the network wherein each layer receives inputs from all preceding ones. This approach balances model depth and computational efficiency and has gained acclaim for its exceptional performance across various computer vision tasks.

2.4 Activation map

A neural network is composed of multiple layers of neurons, with the activation degree of each neuron dependent on connections to preceding inputs. An activation map denotes the spatial distribution of neuron activations within a given layer and highlights regions where neural responses are most pronounced. It is a coherent representation of the feature responses evoked by distinct input elements and visualizes the network's underlying feature detection mechanism in an explicit way, enabling insights into the hierarchy of learned representations.

Techniques to map activation patterns back into input images include gradient activation mapping and saliency map. These methods reinterpret activation profiles of network layers within the input space, thereby emphasizing regions that exert maximal influence on the network's responses. In addition to visualizing input features that command the network's foremost attention, these techniques allow researchers to decipher the model's decisions, scrutinizing whether the model captures the anticipated input features accurately.

3. Methodology

3.1 Objective

The primary research question of our study is: Can mutational MR based on activation profiles assess the generalizability of heart sound classification models effectively?

3.2 Metamorphic Relation

i. Heart Sound Properties

The heart sound classification models are expected to focus on prominent features of essential cardiac cycle events. For instance, a representation learning model should analyze properties such as relative amplitudes, frequencies, and durations of S1, S2, and murmur. In contrast, a shortcut learning model may be perturbed by irrelevant and interfering noises such as background sounds.

ii. Mutation techniques

Considering the difficulty of directly mutating the waveform properties of heart sound components, we designed a mutation strategy in reference to the activation profile of neural network layer. We utilized Grad-CAM (Gradient Class Activation Mapping) as the explainer to access the activation map of the output dense layer in DenseNet, and then applied gradient masking to the original spectrogram to generate follow-up test cases. This way, the impact of the highly activated regions in the original input was reduced in the mutated image.

iii. Definition of MR

MR	Transformation	Explanation
Reduce the most activated region, the model's prediction should be inconsistent.	Apply gradient masking to the original spectrogram regarding Grad-CAM.	Mutate the intensity of input feature elements, the activation profile of last dense layer should change, and the model's output prediction should be different.

Table 1. Description of Metamorphic Relation

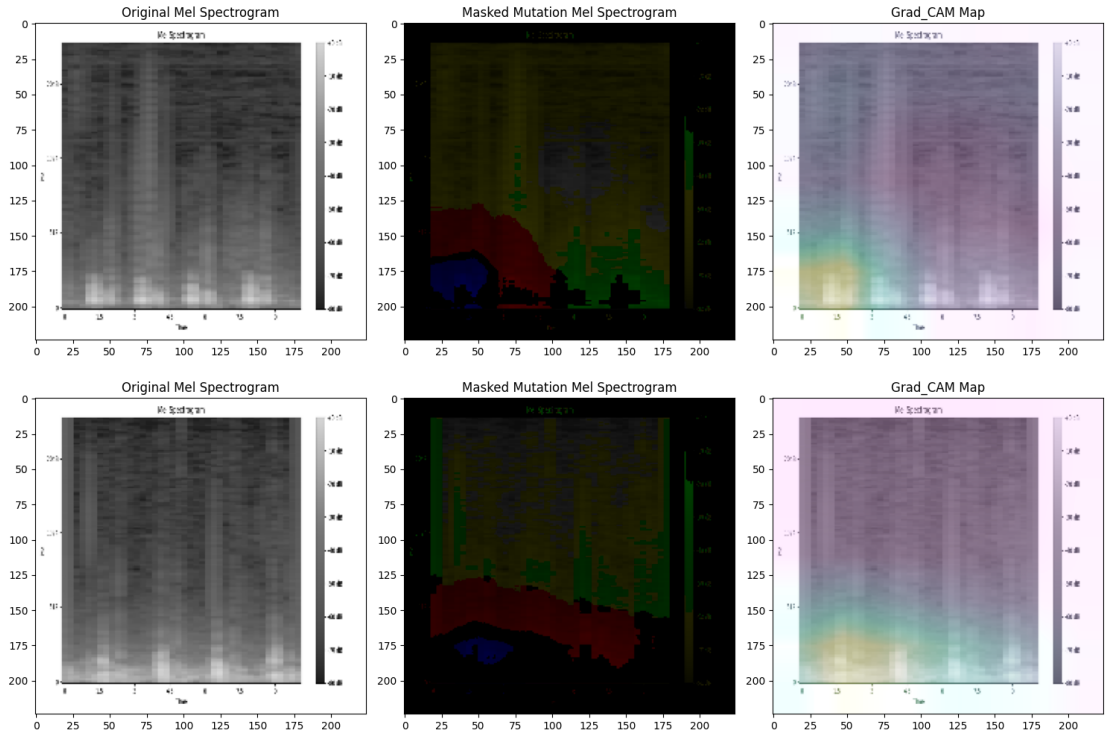

Figure 2. Waveforms of original and mutated spectrograms together with the activation map

iv. Definition of metric

The model prediction assigns confidence scores to the two classes, 'normal' and 'abnormal', and classifies the heart sound into the higher-score category. To gauge prediction consistency across paired test cases, we define a metric as follows: for a pair of original and mutated mel spectrograms, if the absolute value of the difference in confidence scores for the 'normal' class between the original and mutated samples falls within a threshold of 0.2, we classify the model's results as consistent. Given three testing datasets, we first count consistent test case pairs in each set and subsequently compute a weighted consistency rate according to the number of test cases in the three sets.

3.3 Procedure

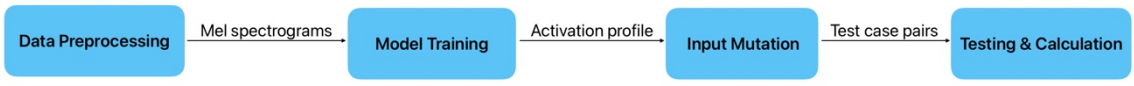

Figure 3. The research framework / workflow

i. Data Preprocessing

We constructed a raw data frame from the training set and filtered out low-quality audio files. Given that the recordings are in different lengths and too long to analyze, we standardized the duration by padding audio files to 5 seconds. Subsequently, we transformed the padded audio files into mel spectrograms, forming the dataset comprising images and labels. Further, we divided this dataset into training and validation subsets. We constructed other testing sets with the same procedure.

ii. Model Training

We imported well-established DenseNet models from the Keras library and added our customized output layer. The models were compiled using binary cross-entropy loss and trained with the training and validation datasets. We ran 15 epochs and plotted the accuracy rate curve.

iii. Mutation Based on Activation Profile

With Grad-CAM, we accessed the activation profile of the last dense layer, proceeding to apply gradient masking to the original spectrogram with a mutation factor of 0.5. By reducing the influence of input features most focused by the model, the activation map for the mutated input is expected to change, thereby leading to different classification outcomes.

iv. Calculation of the consistency rate

We iterated through the validation datasets I, II, and III, recording counts of consistent test case pairs. The ultimate consistency rate for each model was computed as the weighted percentage of consistent pair counts across the complete validation sets.

4. Results and Discussion

4.1 Results

		DenseNet121	DenseNet169	Total testcases
Number of consistent testcase pairs	Set I	209	374	1747
	Set II	17	149	750
	Set III	64	216	368
Weighted consistency rate		10.12%	25.79%	/

Table 2. Testing results of consistency for each model

As shown in Table 2, DenseNet169 has a much higher consistency rate of 25.79%, indicating poorer model generalizability, while DenseNet121 demonstrates a comparatively lower consistency rate of 10.12%.

One possible explanation for the worse performance of DenseNet169 under the MR may be attributed to its complexity and depth. The intricate structure of DenseNet169 may be too complicated for the input data, risking in overfitting and excessive learning. To verify this assumption, we visualized the saliency maps of multiple samples under both models and compared their activation patterns. We noted that DenseNet121 focuses concentratedly on pronounced events in a single cardiac cycle, whereas the activation regions of DenseNet169 covers almost the entire recoding, indicating overfitting. This suggests that DenseNet169 performs well on training samples by simply memorizing complete patterns, so it falls short when dealing with new data after activation mutations were introduced.

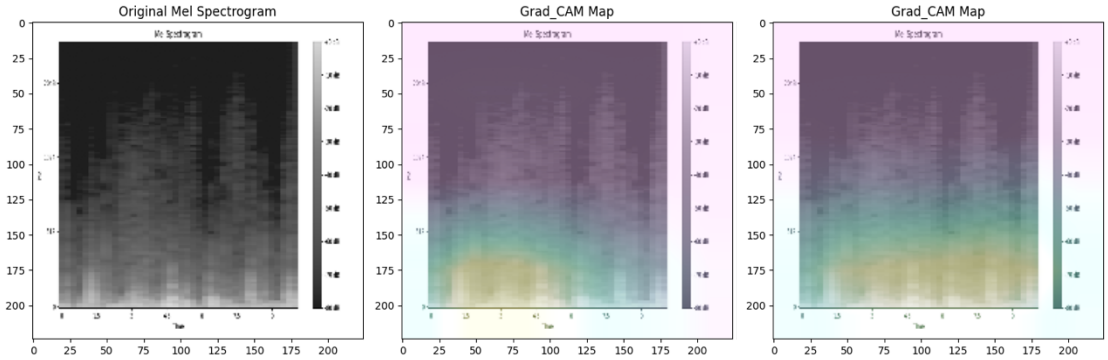

Figure 4. Comparison of the two models' class activation maps

4.2 Findings

It's noteworthy that despite the minimal disparity in validation accuracies between the models—71.60% for DenseNet121 and 72.38% for DenseNet169—the Metamorphic Testing-derived consistency rates reveal a distinct difference between the two models.

This divergence reflects potential shortcomings of traditional metrics and highlights the effectiveness of our proposed Metamorphic Relation in unearthing deeper flaws in program architecture. It is effective in identifying invisible learning process deficiencies of neural networks beneath the ostensible surface of "high-accuracy models."

The mutation method adopted in the research has also improved the interpretability of deep learning models. Saliency maps of original and mutated mel spectrograms explicitly visualized the model's activation responses towards distinctive input features, which offer a comprehensive understanding of the underlying mechanics of model classification process to the researcher.

4.3 Contributions

Our proposed Metamorphic Relation has demonstrated remarkable efficacy in assessing the generalizability of the DenseNet models and displayed potential for explainable AI applications. This successful implementation of MT in biomedical deep-learning sound models is meaningful. We anticipate that the MR can be extrapolated for broader use, encompassing diverse neural network architectures, and potentially extending to other biomedical programs such as lung sound classification models.

4.4 Limitations

i. MR design

While our MR remains effective for most models, there exists an extreme case that our MR may fail to detect. It is when the model concentrate totally on irrelevant features or background. Even though the model does not capture expected input features, it is still likely to output inconsistent results for testcase pairs if it has high sensitivity. The MR of mutating the most activated regions is still applicable in this worst circumstance, making it challenging to distinguish between a model that is completely correct and one that is entirely erroneous. Although such extreme cases are unlikely to exist, we still acknowledge the limitation for the sake of comprehensiveness.

ii. Threat to validity

The study's reliance on only two models each tested across three datasets might be insufficient for generating generalized conclusions. Additionally, the sole reliance on single-sourced heart sound data from the PhysioNet/CinC Challenge 2016 might limit the scope of findings. The utilization of models from the same DenseNet Keras library with similar structures also introduces potential bias to the results.

iii. Model limitation

The constraint of the DenseNet models to three input dimensions necessitated the conversion of mel spectrograms (with four dimensions) to grayscale to align with the model requirements. This conversion might compromise the quality of input data. Moreover, the models in our research achieved validation accuracies only around 70%, suggesting they were not optimally trained and exhibited modest performance on heart sound data.

iv. Further improvement

Based on the limitations, possible improvements involve diversifying neural

network types, expanding dataset sources, and training more-fitted models for heart sound classification.

5. Conclusion

Our proposed Metamorphic Relation of mutating highly activated input regions has effectively evaluated the generalizability of heart sound classification models. It not only facilitates the utilization of Metamorphic Testing in generalizability assessment of sound-related programs but also promotes the integration of explainable AI in the medical domain. Despite the limitations in model selection and methodological design, this study is a valuable exploration of Metamorphic Testing in biomedical audio programs. Further studies could validate and refine this MR across a wider range of models and datasets and investigate its transferability to diverse sound-processing systems for extended applications.

6. Future work

Building upon the foundation laid by this study, we propose the development of an innovative software testing technique deviating from traditional hold-out and cross-validation. We will use all data for training instead of splitting them into training and testing dataset. This idea relies on the feasibility of generating testing data from the training data by performing mutations related to activation profiles. In fact, the testing set I in our research were exactly created from the training dataset. Eliminating the need for separate and additional testing datasets, this methodology will address the scarcity of medical data and enable swifter and more resource-efficient software testing, maximizing the utility of available data. While this proposed approach requires much further assumptions and experimentation, we believe it is possible to develop a Metamorphic Testing-based testing technique that surpasses existing methods through innovation and research.

Acknowledgment

This research has been made possible through the generous support of the Laidlaw Research Scholarship and was conducted in the Biomedical Science Lab at the University of Hong Kong.

References

[1] S. Segura, D. Towey, Z. Q. Zhou, and T. Y. Chen, "Metamorphic Testing: Testing the Untestable," *IEEE Softw.*, vol. 37, no. 3, pp. 46–53, May 2020, doi: 10.1109/MS.2018.2875968.

[2] T. Y. Chen *et al.*, "Metamorphic Testing: A Review of Challenges and Opportunities," *ACM Comput. Surv.*, vol. 51, no. 1, p. 4:1-4:27, 2018, doi: 10.1145/3143561.

[3] T. Y. Chen and T. H. Tse, "New visions on metamorphic testing after a quarter of a century of inception," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Athens Greece: ACM, Aug. 2021, pp. 1487–1490. doi: 10.1145/3468264.3473136.

[4] Ma, Y., Pan, Y. and Fan, Y. (2022) *Metamorphic testing for the medical image classification model*, *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. Available at: https://ieeexplore.ieee.org/document/10077041/.

[5] S. H. N. Santos, B. N. C. Da Silveira, S. A. Andrade, M. Delamaro, and S. R. S. Souza, "An Experimental Study on Applying Metamorphic Testing in Machine Learning Applications," in *Proceedings of the 5th Brazilian Symposium on Systematic and Automated Software Testing*, Natal Brazil: ACM, Oct. 2020, pp. 98–106. doi: 10.1145/3425174.3425226.

[6] J. M. Felner, "The First Heart Sound," in *Clinical Methods: The History, Physical, and Laboratory Examinations*, H. K. Walker, W. D. Hall, and J. W. Hurst, Eds., 3rd

ed.Boston: Butterworths, 1990. Accessed: Aug. 02, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK333/

[7] J. Yang, A. A. S. Soltan, and D. A. Clifton, “Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening,” *Npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41746-022-00614-9.

[8] F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, and R. Forghani, “Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls,” *Radiol. Artif. Intell.*, vol. 5, no. 1, p. e220028, Jan. 2023, doi: 10.1148/ryai.220028.