Metamorphic Testing in Heart Sound Model

Utilize Metamorphic Testing to evaluate the generalizability of medical sound models: a case study in heart sound classification.

Student: Liuqing Yang, Year 2, Faculty of Engineering
Supervisor: Dr. Joshua W.K. HO, LKS Faculty of Medicine

Host Department: School of Biomedical Sciences, LKS Faculty of Medicine, The University of Hong Kong

A. Introduction

Due to complex input spaces and difficulty determining correct outputs, traditional testing strategies can hardly evaluate biomedical systems effectively. Metamorphic Testing (MT) alleviates the problem by analyzing output relations between multiple program executions instead of checking individual output. It generates follow-up test cases from source test cases according to pre-designed Metamorphic Relation (MR), and the model fails when test case pairs violate their corresponding MR.

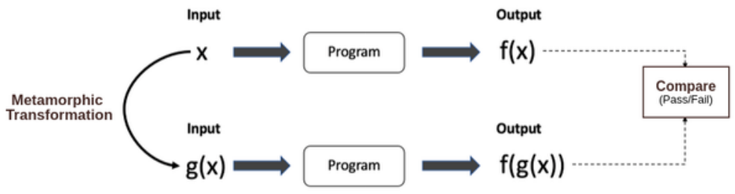

Figure 1. Demonstration of Metamorphic Testing [1]

B. Objective

We propose using MT to assess the generalizability of heart sound classification models, which is the ability to perform prediction tasks on clinical data different from training.

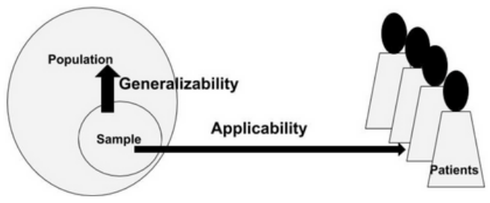

Figure 2. Demonstration of model generalizability [2]

The method involves mitigating highly activated regions in input spectrograms and calculating model prediction consistency for test case pairs.

C. Procedure

MR Definition: Reduce the intensity of most activated input regions; the classification results of original and mutated spectrograms should be inconsistent.

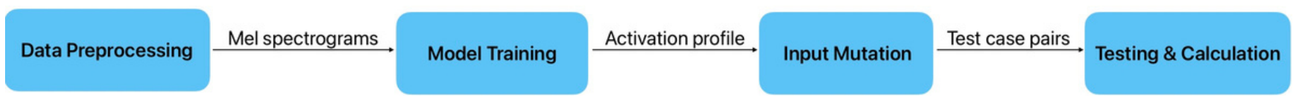

Figure 3. Research workflow

Step 1: Data Preprocessing

Filter out low-quality files, standardize the duration to 5 seconds, and transform padded audios into mel spectrograms. (Source: PhysioNet/CinC Challenge 2016)

Step 2: Model Training

Import Keras models DenseNet121 and DenseNet169 and add a customized output layer. Train each model with 15 epochs. The validation accuracies for DenseNet121 and DenseNet169 are 71.60% and 72.38% respectively.

Step3: Mutation Based on Activation Map

Access the activation profile of the last dense layer with Grad-CAM (Gradient Class Activation Mapping), and apply gradient masking to the original spectrogram with a mutation factor of 0.5 to reduce the influence of most focused input features.

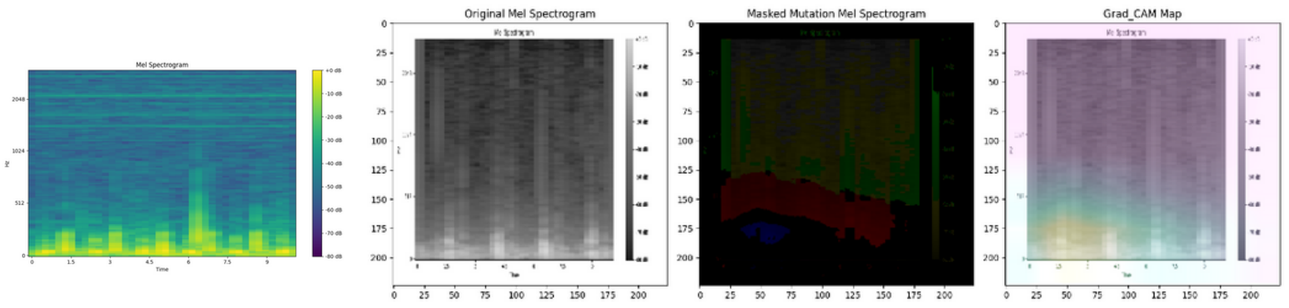

Figure 4. Spectrograms of original and mutated inputs together with the activation map

Step 4: Consistency Rate Calculation

Iterate each model through three testing datasets and record consistent test case pairs. The consistency rate is the weighted percentage of consistent pair count across the entire testing dataset.

D. Result & Analysis

		Model 1 - 121	Model 2 - 169	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 1. Results of experiment

DenseNet121 has a lower consistency rate of 10.12%, suggesting better model generalizability. On the contrary, DenseNet169 has a higher consistency rate when handling the testing data, which indicates poorer model transferability. To verify the assumption, we visualized saliency maps of samples under both models and compared their activation patterns.

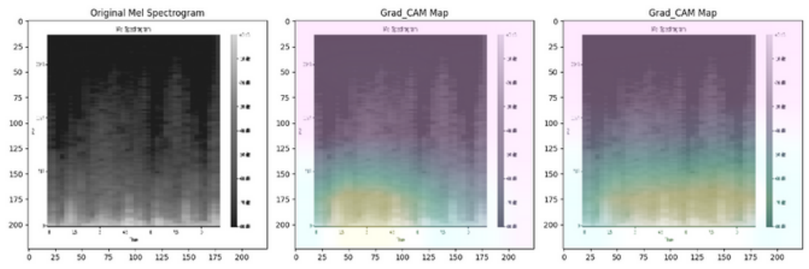

Figure 5. Comparison of two models' saliency maps (The map on the left is from DenseNet121, and the one on the right is from DenseNet169.)

We observed that DenseNet121 focuses on pronounced events in a single cardiac cycle, whereas the activation region of DenseNet169 covers almost the entire recording. This suggests potential overfitting in DenseNet169 and may explain the divergence in consistency rates, which supports our results.

E. Conclusion

Our proposed MR of mutating highly activated input regions effectively evaluated the generalizability of heart sound classification models. It facilitates the utilization of MT in sound model assessment and promotes the integration of explainable AI in the biomedical domain. Hopefully, the MR can be tested on more data and extended to other sound-processing systems for wider applications.

F. Future Work

Based on the feasibility of generating testing data via activation mutation on inputs, we propose developing an innovative software testing technique other than hold-out and cross-validation. Hopefully, it will use all data for training and generate all testing cases from the original data, thus alleviating the scarcity of medical data and enabling more resource-efficient model testing.

References:

[1] Nair, A., Meinke, K. and Eldh, S. (2019) 'Leveraging mutants for automatic prediction of Metamorphic Relations using machine learning', Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation [Preprint]. doi:10.1145/3340482.3342741.
[2] "Generalizability in Qualitative Research," Helping Research writing for student & professional researchers, Apr. 24, 2019. http://researcharticles.com/index.php/generalizability-qualitative-research/
[3] 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.] T.
[4] 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.
[5] 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.