Hypernetwork-Driven Multimodal Integration for Cardiac Diagnosis Using Phonocardiograms and Demographic Data

January 2025

Abstract

Deep learning has been successfully used in medicine for a number of years [4] [2] for diagnosis and identification of conditions [8] [10]. While there exists an abundance of patient records, the usage of such information is plagued by a variety of issues, resulting in a lack of data for both train and inference time. In this paper, we seek to improve training and inference with partial data through multimodality and a hypernetwork to generate classification weights specific to the patient's demographic characteristics. We test our model on the CirCor DigiScope Phonocardiogram dataset [9] for normal or abnormal cardiac function, while demonstrating predictions competitive with state-of-the-art (SOTA) results from the PhysioNet 2022 challenge [11].

1 Introduction

Deep learning has been used in medical image recognition for a number of years [8] [12] and has demonstrated exceptional accuracy in patient diagnosis across multiple domains [2] [4] [7]. However, medical models have not been able to share the same benefits of scaling that we have seen with recent large language models (LLMs) and vision models. This is due to a range of challenges specific to the medical field. One of the primary issues is the decentralized nature of healthcare data, which makes it difficult to collect large datasets necessary for effective model training. Moreover, privacy concerns related to patient information impose additional barriers, and there are multiple regulatory hurdles placed by governments and institutions that complicate the process of obtaining this data. While there are a few centralized platforms that aim to facilitate medical research, they remain limited in scope and access.

Further complicating matters is the lack of standardization across various regions, leading to inconsistencies in how data is documented and recorded. These inconsistencies are exacerbated by privatization and data access control, which hinder the flow of information across institutions. Additionally, the quality of medical data is often compromised by incomplete or erroneous records, and the commercial interests of organizations create further access barriers and data silos. Technical challenges, such as interoperability between different systems and the handling of unstructured data, also present significant obstacles. Resource inequality among healthcare institutions leads to disparities in data access, and the available datasets often exhibit population bias, limiting the generalizability of AI models. Lastly, vendor lock-in with proprietary electronic medical record (EMR) systems further restricts the exchange and utilization of valuable healthcare data.

Building on common challenges encountered in the field of multimodal artificial intelligence (AI), specifically those related to handling missing modalities, noisy data, and data fusion, we extended the HyperFusion framework by Duenias et al. [1], originally designed for brain MRI tasks such as Alzheimer's disease classification. Our work adapts the framework to cardiology, harnessing auditory data, namely phonocardiograms (PCGs) while integrating richer metadata to overcome real-world limitations, such as missing echocardiograms or clinical data. By employing a computationally efficient and simple hypernetwork architecture, we tackle the problem of learning robust cross-modal representations with minimal computational overhead. This approach not only processes PCGs effectively for diagnostic purposes, but also demonstrates resilience in addressing the absence of advanced diagnostic tools, ensuring adaptability and scalability in multimodal healthcare scenarios, and facilitating access to quality healthcare in underserved regions.

2 Background

Advancements in AI have been heavily driven by scaling up models, datasets, and computational resources. Research has shown that increasing these factors leads to predictable improvements in model performance across a variety of tasks. For example, studies by OpenAI on scaling laws [6] demonstrated how larger models, trained on vast datasets using powerful compute infrastructure, achieve better results. This aligns with Rich Sutton's "Bitter Lesson [15]," which emphasizes that the most significant AI advancements often come from leveraging computation rather than relying on handcrafted features or algorithms.

In the field of medical AI, deep learning models have primarily focused on tasks such as image recognition, particularly in medical imaging for disease detection. Architectures like U-Net [12] have been groundbreaking for medical image segmentation, enabling high precision in identifying anatomical structures or abnormalities. Despite this success, there remains a gap in utilizing demographic and clinical data, such as patient age, gender, and medical history, to inform predictions. While such relationships are well-studied in traditional medicine, they are not as thoroughly explored in current AI models. Data sources like PhysioNet [3] provide structured physiological and clinical datasets, but their specificity often poses challenges for general-purpose modeling. Bridging the gap between medical knowledge and AI capabilities could unlock significant advancements in personalized healthcare.

Hypernetworks, introduced by David Ha and colleagues, present an innovative approach to addressing such domain-specific challenges [5]. These models learn relationships between input data and network weights, enabling the dynamic generation of parameters for other networks. This adaptability makes hypernetworks particularly well-suited for applications requiring flexibility, such as handling diverse datasets or transferring knowledge across domains. Building on earlier work by researchers like Ken Stanley [14], hypernetworks have demonstrated potential in improving the adaptability and efficiency of neural networks, especially in cases where data is sparse or domain-specific.

Traditional deep learning approaches for sequences and images often rely on Convolutional Neural Networks (CNNs) for spatial data and Bidirectional Long Short-Term Memory (Bi-LSTM) networks for temporal sequences [13]. These methods are widely used and effective but can be limited by their inability to generalize across highly specific datasets. Hypernetworks offer a complementary solution by enabling models to adapt dynamically to new domains or data distributions, addressing some of these limitations.

The interplay between scaling, medical deep learning, and hypernetworks offers exciting opportunities. Scaling enables the training of powerful general-purpose models, while hypernetworks provide the adaptability needed for domain-specific tasks like personalized medicine. However, leveraging these advancements in medical AI also brings challenges, including managing the specificity of medical data, ensuring interpretability, and addressing ethical considerations in healthcare applications.

3 Dataset

This study utilises the CirCor DigiScope Phonocardiogram Dataset [9], version 1.0.3, published on PhysioNet[3]. This dataset is the largest publicly available collection of paediatric heart sound recordings, designed to support research into automated auscultation-based health recommendation systems. Its extensive annotations, demographic coverage, and real-world noise conditions make it particularly valuable for the development of machine learning models for heart murmur detection and classification.

3.1 Dataset overview

This dataset contains a total of 5272 heart sound recordings, collected from the 4 main auscultation locations of 1568 pediatric patients in rural Brazil. The dataset was collected during two screening campaigns in Paraíba, Brazil from July 2014 to August 2014 and from June 2015 to July 2015. The resultings (PCG) were assessed for signal quality and semi-automatically segmented, and corrected when necessary by a cardiac physiologist. Murmur events within the PCG were manually classified and characterized (location, timing, shape, pitch, quality, and grade) by a single cardiac physiologist independently of the clinical notes and PCG segmentation. The murmur annotations did not indicate whether a murmur was pathological or innocent.

During the data collection session, participants also answered a socio-demographic questionnaire, and subsequently underwent a clinical examination (anamnesis and physical examination), a nursing assessment

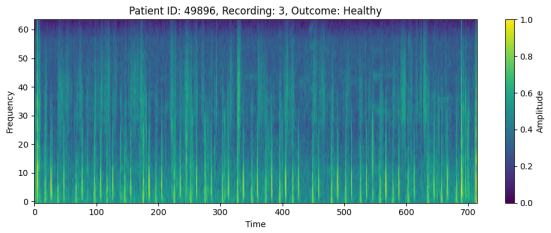

Figure 1: Spectrogram of a healthy patient's phonocardiogram (PCG), displaying time-frequency patterns of heart sounds with no abnormal murmurs detected.

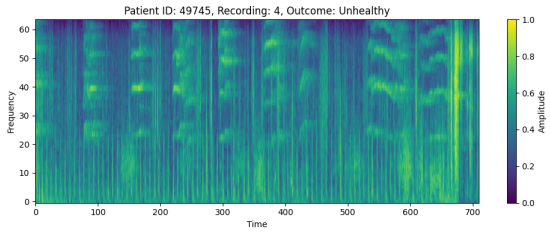

Figure 2: Spectrogram of an unhealthy patient's phonocardiogram (PCG), illustrating time-frequency patterns with abnormalities indicative of cardiac murmurs

(physiological measurements), and cardiac investigations (chest radiography, electrocardiogram, and echocardiogram). In addition to the phonocardiograms, demographic data of the patient was also recorded: weight, height, age group and pregnancy status.

The clinical outcome annotations were determined by a cardiac physiologist using all available notes, using a binary label indicating normal or abnormal cardiac function.

3.2 Data Preprocessing

The raw phonocardiogram (PCG) audio recordings undergo a series of preprocessing steps to normalise their format, reduce noise, and align them with the input requirements of the model. First, the audio recordings, originally sampled at 4,000 Hz, are downsampled to 1,000 Hz. Most heart sound frequencies lie below 1,000 Hz, ensuring no loss of critical signal information while reducing computational complexity. Subsequently, spectrograms are generated from the audio to create time-frequency representations, which serve as a robust format for feature extraction and model processing. Converting audio data to spectrograms reduces noise by isolating relevant frequency patterns of heart sounds while attenuating irrelevant high-frequency artefacts and background disturbances. A windowing technique, akin to mean pooling, is applied to the spectrograms to reduce the temporal dimension and minimise noise, enhancing the model's ability to capture relevant signal features.

Following these transformations, the spectrograms are zero-padded to match the length of the longest audio recording, ensuring uniform dimensions and preserving the original information across all input data. This step also allows the concatenation of recordings for each patient into a single tensor. For patients with missing auscultation recordings, additional zero-padding is applied to maintain consistency across samples.

4 Experiments and Discussion

The experiments carried out aimed to assess the performance of three model configurations for classifying heart sound recordings, focusing on how auxiliary components—a hypernetwork or an multilayer perceptron (MLP)—enhanced the primary network's ability to tackle challenges such as incomplete modalities, noise, and data fusion. Additionally, the architecture of the primary network evolved to optimize its performance, incorporating changes such as kernel adjustments in the convolutional neural network (CNN) to better capture local and global features of heart sounds. Furthermore, spatial attention mechanisms were introduced to highlight critical regions within the spectrogram representations, while temporal attention mechanisms were added to capture the dynamic patterns in heart sound recordings over time. These enhancements collectively improved the model's ability to make accurate predictions by refining feature extraction and focusing on the most relevant aspects of the data.

We employed 5-fold cross-validation to assess the performance of the model. The models were trained using Adam optimizer in conjunction with a learning rate scheduler, which reduces the learning rate after every 20 epochs. A custom loss function was devised that incorporates binary cross-entropy loss, with an increased weight assigned to the positive class. This adjustment results in a greater penalty for false negatives, which is preferred in medical applications where the cost of missing a positive instance is higher.

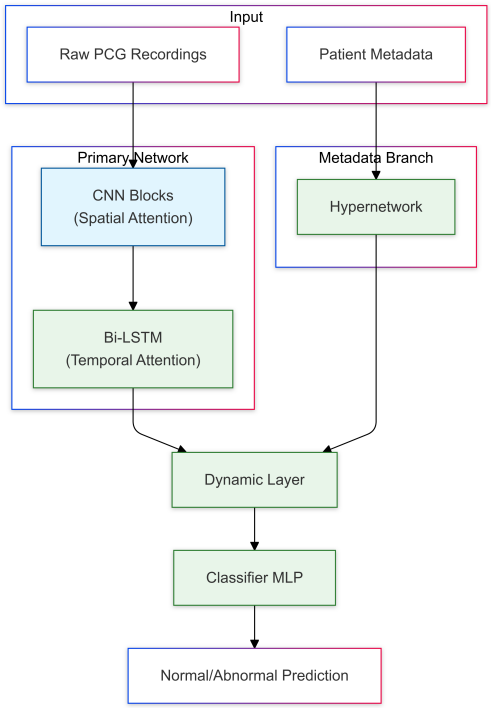

Figure 3: Model Architecture

This weighting strategy is also consistent with the formula for the clinical outcome identification task cost proposed by the authors of the PhysioNet challenge. All training was conducted locally on a system equipped with an NVIDIA RTX 4070 GPU.

Loss function

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(-\text{target}_i \cdot \log(\text{output}_i) \cdot w_{\text{pos}} - (1 - \text{target}_i) \cdot \log(1 - \text{output}_i) \right) \tag{1}$$

To further understand the benefits of the hypernetwork, we conducted an ablation study by comparing the performance of the primary network with and without the hypernetwork, and with an alternative architecture for metadata processing-an MLP. The inclusion of the MLP introduces a simple but effective auxiliary model to learn relationships between features, but it is clear from the results that the hypernetwork outperforms the MLP in both cost and AUROC. This finding supports our hypothesis that a hypernetwork, which dynamically generates the weights for the dynamic layer tailored to patient demographics, results in more precise and personalized predictions. Unlike traditional approaches, such as using MLPs to process metadata followed by feature-level concatenation, the hypernetwork approach can capture complex relationships between demographic attributes and model parameters. This enables a more nuanced and adaptable representation of patient-specific information, enhancing prediction accuracy in comparison to static processing methods. The hypernetwork itself is built upon an MLP architecture, consisting of two fully connected layers with a hidden representation activated by a ReLU function, followed by dropout regularization to prevent overfitting. The first layer transforms the input metadata into a compact latent space, effectively capturing essential features and their relationships. The second layer processes this representation to produce a high-dimensional output, which is split into weights and biases. Unlike the approach of directly embedding metadata, concatenating it with the Bi-LSTM output, and passing it to the classifier MLP, this design allows the hypernetwork to dynamically generate parameters for a downstream layer. This adaptability provides a more nuanced integration of metadata into the model, enabling improved performance across diverse popu-

lations.

Model	Average Cost	AUROC
Primary + Hypernetwork	10717.6048	0.6419
Primary + MLP	11597.96	0.587
Primary Only	13922.5460	0.5584

Table 1: Model Performance Comparison. Lower cost is better and higher AUROC is better.

The key difference in performance arises from how the metadata is utilized in each approach. In the first method, where the metadata is concatenated directly with the Bi-LSTM output before being passed to the classifier, the model treats the metadata as a separate feature. While this provides some added information, it doesn't allow for a dynamic interaction between the metadata and the Bi-LSTM features, limiting the model's flexibility. On the other hand, when the hypernetwork generates weights for a downstream layer that processes the Bi-LSTM output, the metadata is integrated more deeply into the model's predictions. This approach allows the metadata to directly influence how the Bi-LSTM's sequence-based features are processed, enabling the model to capture more complex and dynamic relationships between the metadata and sequence data. This results in a more personalized and adaptable model, ultimately improving performance across diverse populations.

5 Results

Cost Computation for Screening and Treatment
1. **Algorithmic Prescreening Cost**:

$$C_{\text{algorithm}}(m) = 10 \cdot m$$

2. **Expert Screening Cost** (for m patients out of n total patients):

$$C_{\text{expert}}(m, n) = \left(25 + 397 \cdot \frac{m}{n} - 1718 \cdot \left(\frac{m}{n}\right)^2 + 11296 \cdot \left(\frac{m}{n}\right)^4\right) \cdot n$$

3. **Treatment Cost**:

$$C_{\text{treatment}}(m) = 10000 \cdot m$$

4. **Missed or Late Treatment Cost**:

$$C_{\text{error}}(m) = 50000 \cdot m$$

The **Challenge Cost Metric** is computed as follows:
- Total patients:

$$N = TP + FP + FN + TN$$

- Compute the **Total Cost**:

$$C_{\text{total}} = C_{\text{algorithm}}(N) + C_{\text{expert}}(TP + FP, N) + C_{\text{treatment}}(TP) + C_{\text{error}}(FN)$$

- Compute the **Mean Cost Per Patient**:

$$\text{Mean Cost} = \frac{C_{\text{total}}}{N}$$

From the results presented in Table 1, it is clear that incorporating the hypernetwork architecture significantly improves the performance of the model compared to the primary network alone. The average cost for the model with the hypernetwork is 10717, which is the lowest among all configurations, indicating that it is able to make more accurate predictions with a lower penalty for false negatives. Moreover, the AUROC score

of 0.6419 is substantially better than the primary network alone (0.5584) and the primary network combined with an MLP (0.587). These results suggest that the hypernetwork's ability to generate patient-specific weights based on demographic data contributes significantly to improving the model's diagnostic accuracy, particularly in situations where the training data may be insufficient or noisy.

Our model ranks among the top-performing submissions in the challenge. Although we did not have access to the hidden test dataset to directly compare our model's ranking with that of other submissions, a comparison based on the validation results—specifically the cost and AUROC metrics from 5-fold cross-validation—demonstrates that our model exhibits performance consistent with the best-performing entries. Previous best-performing studies have reported AUROC values in the range of 0.58 to 0.63 for this task. Our model achieves an AUROC of 0.6419, aligning well with these results and highlighting the effectiveness of our hypernetwork approach in addressing the unique challenges of the dataset. Additionally, the range of cost outcomes for the best-performing models spans from 9178 to 12515. Although there is some variability, our model's outcome cost of 10717 remains well within this range, further demonstrating its competitive performance.

6 Conclusion

In this study, we presented a novel approach for enhancing deep learning models in medical applications, specifically focusing on classifying heart sound recordings in the PhysioNet dataset. Our proposed framework integrates a hypernetwork to generate patient-specific classification weights based on demographic data, addressing challenges such as incomplete modalities, noisy data, and demographic variability.

Through extensive experiments and comparative analysis, we demonstrated that the hypernetwork-based model significantly outperforms traditional approaches, including models with standard MLPs for metadata processing and those that use only one modality. The hypernetwork's ability to dynamically adjust weights tailored to individual patient characteristics led to improvements in both prediction accuracy and cost efficiency. Our model achieved an AUROC of 0.6419 and a mean cost of 10717, positioning it as a competitive solution within the range of SOTA results reported in the PhysioNet challenge.

Future Work

While our approach shows promising results, areas for future research and improvement remain:

- **Cross-Dataset Validation**: Testing the model on additional datasets from different clinical environments will help assess its robustness and generalizability across diverse populations, as the current dataset is focused on children and young adults, ensuring its broad applicability.
- **Expanding to Other Modalities**: Extending the model to handle other forms of medical data, such as ECG signals, imaging, or clinical notes, could offer more comprehensive insights for diagnosis and treatment prediction.

Through these future directions, we aim to enhance the model's robustness, and applicability in personalized medicine, contributing to more accurate and efficient healthcare solutions in diverse and resource-constrained settings.

References

[1] Tal Arbelb Tammy Riklin Raviva ADNI Daniel Dueniasa, Brennan Nichyporukb. Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. *arXiv:2403.13319v1*, 2024. URL: `https://arxiv.org/pdf/2403.13319`.

[2] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. `doi:10.1038/nature21056`.

[3] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, et al. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. URL: https://physionet.org/, doi:10.1161/01.CIR.101.23.e215.

[4] Varun Gulshan, Lily Peng, Marc Coram, Michael C Stumpe, Derek Wu, Arunachalam Narayanaswamy, and Dale R Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016. doi:10.1001/jama.2016.17216.

[5] David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. URL: https://arxiv.org/abs/1609.09106.

[6] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL: https://arxiv.org/abs/2001.08361.

[7] June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: General overview. *Korean J. Radiol.*, 18(4):570, 2017.

[8] Geert Litjens, Thijs Kooi, Babak E Bejnordi, Arnaud AA Setio, Francesco Ciompi, Mohsen Ghafoorian, and Jeroen AW van der Laak. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi:10.1016/j.media.2017.07.005.

[9] Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, Andoni Elola, Ali Bahrami Rad, Reza Sameni, Gari D. Clifford, and Miguel T. Coimbra. The circor digiscope dataset: From murmur detection to murmur classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2524–2535, 2022. doi:10.1109/JBHI.2021.3137048.

[10] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017. URL: http://arxiv.org/abs/1711.05225, arXiv:1711.05225.

[11] Matthew A. Reyna, Yashar Kiarashi, Andoni Elola, Jorge Oliveira, Francesco Renna, Annie Gu, Erick A. Perez Alday, Nadi Sadr, Ashish Sharma, Jacques Kpodonu, Sandra Mattos, Miguel T. Coimbra, Reza Sameni, Ali Bahrami Rad, and Gari D. Clifford. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. *medRxiv*, 2023. URL: https://www.medrxiv.org/content/early/2023/04/05/2022.08.11.22278688, doi:10.1101/2022.08.11.22278688.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 9351:234–241, 2015. doi:10.1007/978-3-319-24574-4_28.

[13] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. 28:802–810, 2015. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

[14] Kenneth O. Stanley, Jeff Clune, Joel Lehman, and Risto Miikkulainen. *Designing Neural Networks Through Neuroevolution*, volume 1. 2019. doi:10.1038/s42256-018-0006-z.

[15] Richard Sutton. The bitter lesson. *Incomplete Ideas (Blog)*, 2019. URL: http://www.incompleteideas.net/IncIdeas/BitterLesson.html.