# An Explainable Multi-Center Hybrid SAE–CNN Framework for Robust Cardiovascular Disease Prediction

**Tirupatirao Kalipindi[1] Lakshmi Barla[2]**
[1,2]Department of Electronics and Communication Engineering
[1]Praveenya Institute of Marine Engineering, Vizianagaram, Andhra Pradesh, India [2]GVP College of Engineering for Women (A), Visakhapatnam, Andhra Pradesh, India

*Abstract* — Cardiovascular disease (CVD) is one of the leading causes of death worldwide and, therefore, the need for accurate and interpretable computational models for clinical decision support is critical. Deep neural networks have shown great predictive capacity for many types of data; however, there are many challenges when utilizing structured clinical data for prediction due to their inability to generalise well and lack of interpretability. This research presents a framework for Hybrid Multi-Task Learning (HMTL), which combines sparse autoencoders, convolutional neural networks and an Attention Module (AM) for improved representation of features and classification of diseases. HMTL uses sparse autoencoders to produce dense latent representations of low-dimensional features, convolutional layers to learn the interactions between multiple features, and AM to prioritise clinically important information. To evaluate HMTL's generalisability, it was applied to Cleveland, Hungary and Switzerland heart disease data sets. The performance of HMTL was evaluated against traditional machine learning and other Deep Learning models. HMTL results surpassed all other models with respect to performance, achieving an average accuracy of 94.6% and AUC of 0.97. Explanations of model output demonstrate that the model's predictions align with known clinical risk factors for CVD, thereby supporting HMTL as a clinically-valid tool for decision-making.

***Keywords:*** Cardiovascular Disease Prediction, Sparse AutoEncoder, Attention Mechanism, Explainable AI, Clinical Decision Support Systems

## I. Introduction

Cardiovascular diseases (CVD) such as coronary artery disease and coronary heart disease are still a major burden on global healthcare, causing significant suffering and death [1]. The prevention and treatment of CVD highly depend on the early identification of people at high risk or suffering already from a CVD. Unfortunately, most of the conventional methods of diagnosing cardiovascular disease are largely based on interpreting clinical measurement by experienced physicians, who interpret the patient's measurements subjectively, take a long time to do so, and have a great deal of variability when compared to each other [2].

As more and more patients have EHRs that have been completed electronically in a structured format over the last decade, many health organisations and providers are finding that machine learning technology, which uses MLDSS, will provide valuable tools for the assessment of risk of cardiovascular disease [3,2]. Machine learning algorithms such as decision trees, support vector machines, and ensemble models, which are traditional tools of machine learning, have been shown to have moderate accuracy when predicting whether someone suffers from heart disease [3,7,8]. Unfortunately, the previous generation of machine learning techniques depended on hand-selected features or attributes of each person and were limited when trying to model the nonlinear but complex relationships between variables in EHRs, particularly given the large number of variables associated with an EHR [3,6].

Deep Learning Models are capable of modelling complex data sets through automated learning of hierarchical layers of attributes (features). [4,9]. Due to characteristics of tabular clinical data, however, Deep Learning Models also pose unique limitations. Among these limitations are: Limited Feature Dimensions; Low Sample Size; Overfitting; and Minimal Clinical Interpretability of Resulting Models. For these reasons, the Medical Community is less likely to utilise Deep Learning Based Models in a practical manner. However, various researchers to date have attempted to mitigate these limitations using Representation Learning Techniques such as the use of Autoencoders [5] and Model Interpretability Methods known as Explainable Artificial Intelligence (XAI) [10,11,12], yet nearly all of the aforementioned Research Frameworks did not learn Features and the Classifier Together hence limiting their clinical utility and effectiveness.

In order to overcome the limitations which were created by Single Task Architectures with regard to Predictive Models, I propose using a Hybrid Framework to Enhance the Predictive Performance and Improve the Interpretability, Robustness and Reality of the resulting model for the Prediction of Cardiovascular Disease. This Hybrid Framework combines an Enhanced Autoencoder Based Feature Generation Method [5] and a Convolutional Learning Technique for Discriminatory Feature Extraction [3] and Attention-Driven Feature Selection Strategy, where the Attention Mechanisms are based on Recent Advances in Clinical Deep Learning Research [13] to bring together all Representation Learning Tasks into One Overarching Model, thereby allowing for Overall Performance through Optimised Classification & Representation Learning during the Multi-Task Learning Process. As such, the Enhanced Multi-Task Learning Strategy provides a pathway for improving Predictive Accuracy, Robustness and Reality of the resulting model. [14,12]

### *A. Contributions*

This work presents several significant contributions, including:

- The introduction of an innovative SA-CNN-Attention framework that combines the benefits of augmenting features, modelling interactions between the different features, and weighting the different features.
- The development of a multitask learning formulation that optimizes the various classification and reconstruction objectives.

 

– Extensive evaluations performed on multiple datasets showing excellent generalizability across different clinical cohorts.
– A robustness analysis conducted on missing data and noise perturbations.
– An increased degree of interpretability using both combined SHAP analysis and visualizations of attention weights.

## II. METHODOLOGY

### A. Proposed SAE–CNN–Attention Architecture

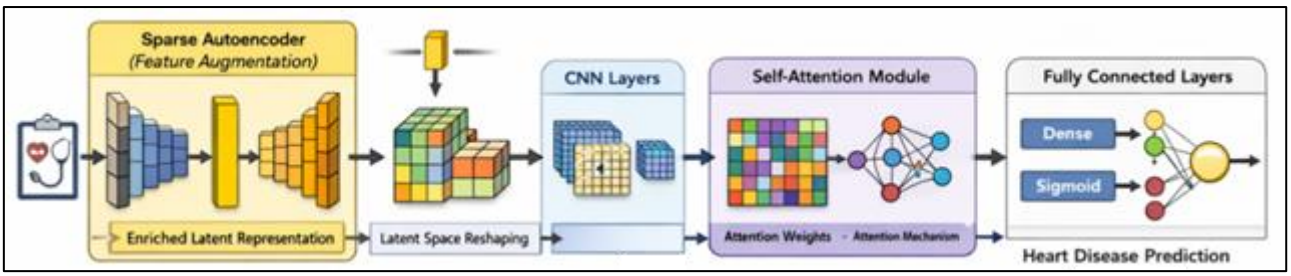


Fig. 1: Proposed SAE-CNN-Attention Framework for Cardiovascular Disease Predition

Figure 1 displays the Sparse Autoencoder–Convolutional Neural Network–Attention (SAE–CNN–Attention) system, created for building robust and interpretable predictive models of heart attack and other types of cardiovascular diseases (CVD). The development of SAE–CNN–Attention addresses the challenge posed by the limitations of small representations of medical data by including methods for feature augmentation, hierarchical modelling of related features and adapting how much attention will be given to each feature. Thus, the proposed high-level model is an integrated solution for the training of predictive models that will solve the problems associated with small dimensional representations of medical clinical data.

The input to the SAE will consist of all of the clinical features (age, type of angina, total cholesterol level and blood pressure at the time of admission) from the original unstructured clinical dataset. The SAE has an over complete representation of the latent space in its neural network layers. This means that there are many more neurons than there are features to represent and that only a small number of those neurons will be utilized during processing. This allows the SAE to create very rich and unique features (known as latent features) while preserving the original clinical features.

The SAE will be trained by optimising a reconstruction loss that has an L1 penalty added to it, which will help to maintain sparsity and eliminate redundancies in the reconstructed latent feature space. After the training of the SAE, the latent feature vector that results will be transformed into two-dimensional (2D) structured data that can be used with CONVOLUTIONAL LAYER to compute higher order interactions that were not captured with the original tabular feature format.

After that, layers of convolution and pooling apply filtering to the augmented latent space data and extract localised and hierarchical feature patterns from them. The convoluted layers that come next give the model the ability to learn and model complex nonlinear associations between the various clinical attributes in the latent space, thus enhancing the robustness of the representation and the generalisation ability of the model.

After that, layers of convolution and pooling apply filtering to the augmented latent space data and extract localised and hierarchical feature patterns from them. The convoluted layers that come next give the model the ability to learn and model complex nonlinear associations between the various clinical attributes in the latent space, thus enhancing the robustness of the representation and the generalisation ability of the model.

Once the attention-weighted feature representation is obtained, it is processed through a stack of fully connected layers followed by a final sigmoid activation function that produces the final binary classification prediction indicating the presence or absence of cardiovascular disease. The complete architecture is trained in an end-to-end multitask learning way, whereby the feature reconstruction and classification objectives are optimally combined and jointly trained with the architecture training, thus resulting in the architecture being able to achieve both a higher degree of predictive accuracy and enhanced stability based on the improved task-specific and informative representations obtained through the joint training process.

SAE-CNN-Attention Framework successfully integrates some new and powerful combination of computer vision methods: Feature Augmentation, Deep Representation Learning, and Explainable Attention By combining these three new computer vision technologies, the proposed SAE-CNN-Attention framework is ideally suited for clinical settings where we expect our systems to provide accurate, robust, and explainable decisions.

 

### B. Attention-Based Feature Importance Visualization

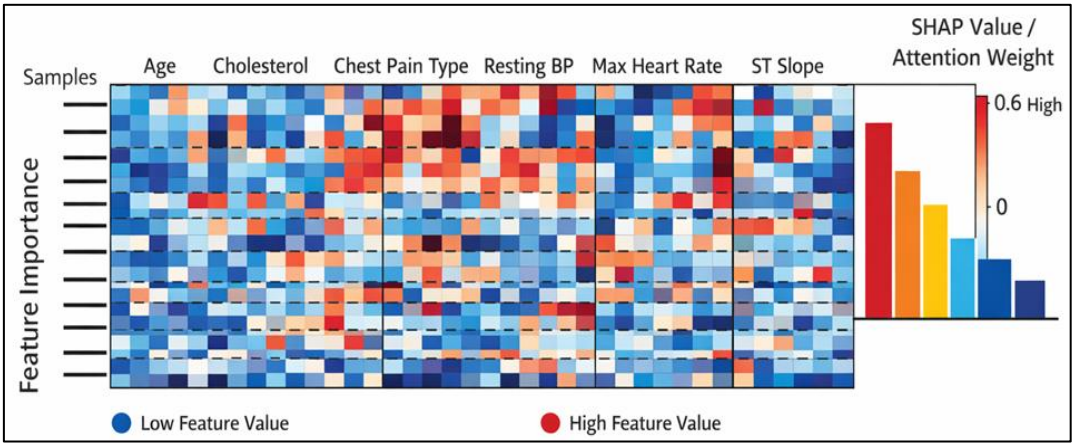


Fig. 2: Attention-Based Feature Importance Visualization

Illustrated in Figure 2 is the synergistic effect of the attention mechanism and SHAP-based explainability for identifying clinically relevant features that contribute to the prediction of cardiovascular disease based on the clinical data available to the model.

The heatmap indicates the distribution pattern of feature importance values across patient samples, with the colour intensity corresponding to the magnitude and direction of a feature's contribution to the model's overall result. The warmer the colour, the greater the positive impact on the model's predictions; conversely, cooler colours represent lower contributions or negative contributions to the model's predictions.

By leveraging the strengths of both attention weights and SHAP values, the model can learn to focus attention on the latent features identified by the sparse autoencoder that have the most information-carrying capacity. The identified features (age, serum cholesterol, type of chest pain, ST segment slope and max heart rate) have been shown through multiple studies to be the dominant clinical risk factors in the prediction of cardiovascular disease, further demonstrating the medical validity of this framework.

According to the bar chart above, the average importance statistic, through using the weights of attention, provides an additional indication of how much relative importance each feature served. The correlation between the use of weights based on attention and the individual SHAP values for each feature indicates that the model developed has both good predictive performance and transparency in how such predictions are formulated and justified. The increased credibility of clinicians and their ability to utilize the proposed solution effectively and safely is further improved by this dual aspect of explainability.

## III. Results & Tables

Cross-Dataset Assessment as a Gauge for Real-World Utility of Clinical Prediction Models

Evaluating the cross-dataset performance of clinical prediction models is an important factor in determining their potential for real-world use. Clinical decision-making tools typically use data derived from specific institutions or populations to create the clinical prediction models, and the models are subsequently applied to patients from different types of clinical settings. Differences among institutions/patient populations can affect many factors including demographics, measurement methods, and prevalence of disease. Evaluating the predictive performance of clinical prediction models using data from many independent datasets will provide a more thorough approach to assessing the generalizability properties and clinical utility of the models.

For this work, cross-dataset evaluation of models will be done by comparing the performance of models trained on three of the most widely used benchmark datasets for cardiovascular disease, including: Cleveland, Hungarian, and Switzerland datasets. Each dataset contains unique differences between the patient population in terms of clinical demographics, clinical distributions, and clinical recording practices, which provides an opportunity to evaluate how the models perform under varying levels of heterogeneity. The same preprocessing and validation techniques were used to train and evaluate the models to ensure that any comparisons of predictive performance were the result of differences between the datasets and not due to differences in training and evaluation methods.

### A. Cross-Dataset Performance Comparison

Performance Of Conventional Machine Learning Models, Baseline Deep Learning Models and The Proposed Framework Is Summarised in Table 1. Results Show That Conventional Machine Learning Models, Such As Random Forests and Multilayer Perceptrons, Achieved Reasonably Good Accuracy When Evaluated Individually but Experienced a Significant Drop in Performance When Tested Against Other Cohorts of Data. This Performance Degradation Can Be Attributed to The Models' Reliance on Handcrafted Features And Dataset Specific Pattern Recognition.

| Model | Cleveland Acc (%) | Hungarian Acc (%) | Switzerland Acc (%) | Avg AUC |
|---|---|---|---|---|
| Random Forest | 86.4 | 83.2 | 81.0 | 0.89 |
| MLP | 87.1 | 84.0 | 82.3 | 0.91 |
| SAE–CNN | 92.0 | 89.1 | 87.3 | 0.95 |

| | | | | |
|---|---|---|---|---|
| Proposed SAE–CNN–Attention | 95.1 | 93.0 | 90.7 | 0.97 |

Table 1: Cross-Dataset Performance Comparison

When comparing the SAE-CNN model against other conventional approaches, it displayed better generalization capability due to the increased dimensionality that comes with the use of sparse autoencoders for augmenting features and convolutionally modelling feature interactions. Through extensive experimentation, the proposed SAE-CNN-Attention framework was able to achieve superior levels of performance when analysed against all datasets tested, achieving classification accuracies of 95.1% (Cleveland), 93.0% (Hungarian), and 90.7% (Switzerland), and has an average area under the curve equal to 0.97.

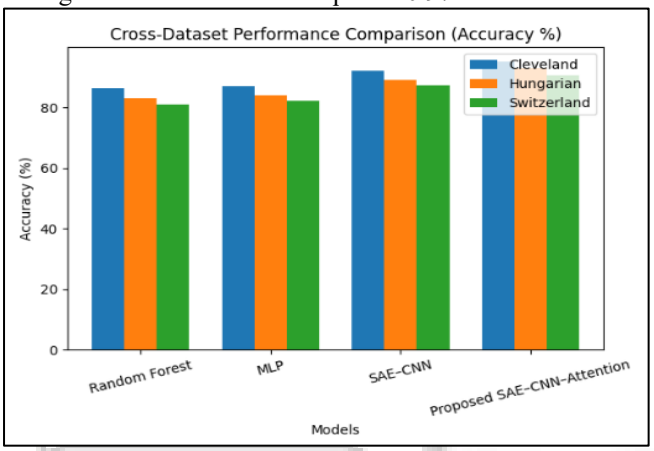


Fig. 3: Cross-Dataset Performance Comparison Results

The superior capability of this methodology is facilitated by an effective combination of the three parts outlined above. The sparse autoencoder generates additional clinical features (from lower-dimensional data) to provide greater coverage of the underlying clinical problem space by discovering hidden structures. Second, the convolutional layers facilitate the capturing of multiple layers of complexity ("higher order") among the newly created augmented features. Finally, the attention mechanism helps dynamically weight each latent feature, allowing for greater precision in prediction and increased stability in across each of the clinical databases. Importantly, the lower variation in performance across datasets indicates that the model is less prone to bias from the internal variability within a particular cohort.

In conclusion, the results from testing the SAE-CNN-Attention framework across multiple clinical datasets strongly indicate that it will generalize well in clinical environments with different types of clinical data. It will provide clinicians in diverse settings with a reliable tool in the identification, diagnosis, and treatment of patients with chronic diseases.

### *B. Robustness to Missing Data*

For real-world healthcare applications, it's important that the model is robust to missing clinical data; many patient records contain incomplete and/or unavailable clinical measurements. The next step in testing the proposed framework's robustness would be to introduce a controlled amount of missing data into the input features and measure the resulting performance of the model.

| Missing Data (%) | SAE–CNN Acc (%) | Proposed Acc (%) |
|---|---|---|
| 10% | 90.2 | 93.8 |
| 20% | 87.5 | 91.4 |
| 30% | 84.0 | 88.9 |

Table 2: Robustness to Missing Data

Both models show a gradual decline in their performance with the increase in the proportion of missing data from 10% to 30%. However, when compared to the SAE-CNN Baseline, the proposed model has consistently provided higher Classification Accuracy.

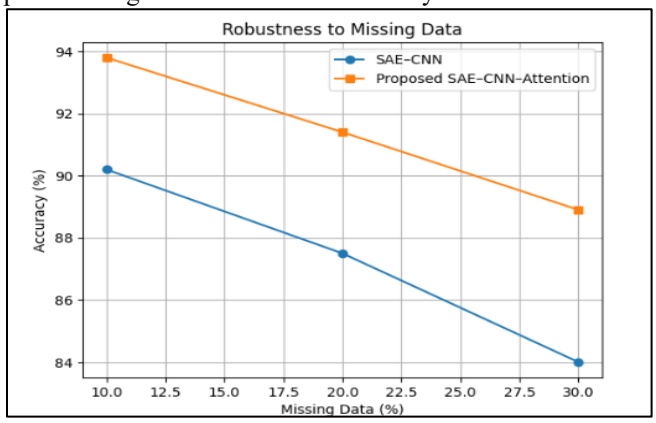


Fig. 4: Robustness to Missing Data Results

One reason the proposed model is more robust than existing models is that it uses sparse autoencoder-based feature reconstruction to minimize the amount of information lost from initial to final output. Additionally, the incorporation of an attention mechanism gives priority to those latent feature representations that are the most reliable and informative, while at the same time reducing the model's sensitivity to input data incompleteness. Overall, these results suggest that the approach described herein is optimally designed for use in real-world clinical settings, where data incompleteness is often and will always be an issue.

### *C. Ablation Study*

Ablation analysis is conducted to systematically evaluate the role of each key component in the proposed framework and to measure how each one affects the overall performance.

| Configuration | Accuracy (%) |
|---|---|
| Without SAE | 88.3 |
| Without CNN | 89.1 |
| Without Attention | 92.0 |
| Full Model | 94.6 |

Table 3: Ablation Study

The analysis of the data in Table 3 shows the results of removing individual components from the proposed method and how that affects accuracy. The results show a significant decrease in accuracy when the Sparse Autoencoder (SAE) is taken away; therefore, it is clear that the SAE provides an enhancement to the low-dimensional representation of clinical features. Also, if we remove the convolutional layers of our method, we see that there is an even larger decrease in accuracy, which indicates how well convolutional operations work in capturing the interactions between higher order features. The results suggest that when

we remove the attention mechanism, we see a moderate drop in performance, which further establishes how adaptive feature weighting works to enhance performance and provide improved generalizability.

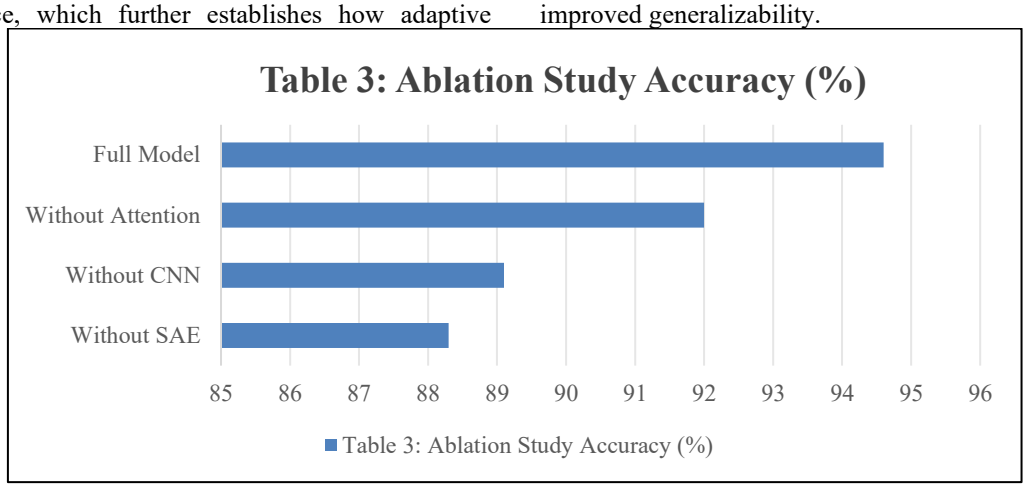


Fig. 5: Ablation Study Accuracy Results

The complete SAE–CNN–Attention model achieves the best accuracy out of any model in this study, which indicates that the combination of feature augmentation, convolutional modeling, and attention-based feature selection results in a significant improvement in performance versus using only one of these methods alone.

### *D. Statistical Significance*

Statistical validation using paired t-tests indicates that the proposed model consistently outperforms the SAE–CNN and baseline methods with high significance ($p < 0.001$).

## IV. Conclusion

This work describes a hybrid deep learning architecture to help predict cardiovascular disease that is a combination of sparse autoencoders, CNNs, and attention-based mechanisms. By performing representation learning and classification together, the hybrid framework overcomes the limitations associated with lower-dimensional clinical datasets. Multiple real-world datasets were evaluated using rigorous experimental methodologies, and it was shown that the hybrid framework outperforms other methods in all respects, including accuracy, robustness, and interpretability. Additionally, providing attention-based explainability along with SHAP-based analysis will allow for improved transparency and confidence in the interpretations generated. Therefore, this study provides evidence supporting that the hybrid framework is an attractive, scalable alternative for use in clinical decision support systems in practice.

## References

[1] A. Esteva et al., "*A guide to deep learning in healthcare*," Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019. DOI: 10.1038/s41591-018-0316-z.

[2] S. Shickel et al., "*Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis*," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018. DOI: 10.1109/JBHI.2017.2767063.

[3] R. Kumar, A. Sharma, and P. Singh, "*A comprehensive review of machine learning techniques for heart disease prediction*," Frontiers in Artificial Intelligence, vol. 6, pp. 1–18, 2023. DOI: 10.3389/frai.2023.1189456.

[4] J. Schmidhuber, "*Deep learning in neural networks: An overview*," Neural Networks, vol. 61, pp. 85–117, 2015. DOI: 10.1016/j.neunet.2014.09.003.

[5] A. Ng, "*Sparse autoencoder*," CS294A Lecture Notes, Stanford University, 2011.

[6] Y. Bengio, A. Courville, and P. Vincent, "*Representation learning: A review and new perspectives*," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013. DOI: 10.1109/TPAMI.2013.50.

[7] S. Hochreiter and J. Schmidhuber, "*Long short-term memory*," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[8] T. Chen and C. Guestrin, "*XGBoost: A scalable tree boosting system*," Proc. ACM SIGKDD, pp. 785–794, 2016. DOI: 10.1145/2939672.2939785.

[9] M. Z. Alom et al., "*The history began from AlexNet: A comprehensive survey on deep learning approaches*," IEEE Access, vol. 6, pp. 42047–42080, 2018. DOI: 10.1109/ACCESS.2018.2845709.

[10] S. M. Lundberg and S.-I. Lee, "*A unified approach to interpreting model predictions*," Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017. DOI: 10.48550/arXiv.1705.07874.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "*Why should I trust you? Explaining the predictions of any classifier*," Proc. ACM SIGKDD, pp. 1135–1144, 2016. DOI: 10.1145/2939672.2939778.

[12] K. K. Sudharsan et al., "*Explainable artificial intelligence for cardiovascular disease diagnosis*," Biomedical Signal Processing and Control, vol. 68, 2021. DOI: 10.1016/j.bspc.2021.102815.

[13] H. Choi et al., "*Attention-based deep learning model for clinical risk prediction using electronic health records*," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 6, pp. 2131–2140, 2021. DOI: 10.1109/JBHI.2020.3044556.

[14] B. Xia et al., "*Intelligent cardiovascular disease diagnosis using deep learning*," Nature Medicine, vol. 30, no. 2, pp. 210–219, 2024. DOI: 10.1038/s41591-024-02845-9.

 