Vusani Radzilani
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
2431400@students.wits.ac.za

Introduction

Heart disease remains the preeminent cause of mortality worldwide, as substantiated by various global health reports, including those from the Centers for Disease Control and Prevention (CDC). The diagnostic process for heart disease is fraught with complexities due to the multifaceted nature of its symptoms and the myriad ways it interacts not just with the heart, but with other organ systems as well. As a result, there are often significant delays in achieving accurate diagnoses, which can adversely impact patient outcomes.

Problem Statement:

Contemporary medical decision support systems, particularly those integrated within Hospital Information Systems (HIS), demonstrate notable deficiencies. They are adept at performing straightforward queries, such as identifying specific demographic groups affected by heart issues, but they struggle with more complex predictive tasks. These tasks—essential for timely and accurate diagnosis—require systems to predict the likelihood of heart disease from comprehensive and varied patient data. Moreover, the reliance on physician intuition and heuristic judgment in making clinical decisions can introduce bias, increase error rates, and inflate the costs associated with diagnosis and treatment, ultimately compromising the quality of care.

Proposal

In light of these challenges, this paper proposes the development of an innovative diagnostic tool utilizing a Bayesian Network approach. The choice of a Bayesian Network is strategic; despite its infrequent application in medical diagnostic systems, its probabilistic framework is particularly well-suited for handling the uncertainties and variabilities associated with the symptoms and progression of heart disease. Bayesian Networks offer a powerful means of encapsulating complex relationships between various health indicators and heart disease outcomes. They facilitate not only a robust assessment of risk factors but also enable the integration of continuous learning processes as new data becomes available, thereby refining their predictive accuracy. Furthermore, Bayesian Networks allow for the incorporation of prior knowledge or expert insights, which can be particularly valuable in the medical domain where domain expertise plays a crucial role.

This approach aims to revolutionize how diagnoses are formulated, moving away from intuition-based methods towards a more deterministic, data-driven model. By leveraging the capabilities of Bayesian Networks, the proposed system seeks to reduce diagnostic times, minimize medical errors, and enhance the overall safety and satisfaction of patients, thereby addressing the critical gaps identified in existing diagnostic frameworks.

Heart disease prediction is a crucial aspect of this endeavor, as it can significantly improve patient outcomes and reduce healthcare costs. In this study, we propose a probabilistic graphical model (PGM) approach for heart disease prediction, leveraging a well-known dataset containing relevant features for heart disease classification.

Dataset Information

The Heart Disease dataset consists of 76 attributes originally collected from 1025 patients from Cleveland, Hungary, Switzerland, and the VA Long Beach databases. Among these, the Cleveland dataset stands out as a notable benchmark utilized for heart disease diagnosis systems. This dataset includes 303 instances and features 13 medical attributes sourced from the Cleveland Heart Disease dataset [Janosi and Detrano, 1988]. ML researchers have predominantly focused on a subset of 14 attributes within this extensive dataset, and so will we. The ”goal” field signifies the presence of heart disease, categorized with integer values ranging from 0 (absence) to 4. Notably, studies with the Cleveland dataset have primarily centered on distinguishing between the presence (values 1, 2, 3, 4) and absence (value 0) of heart disease. The original dataset contained patient names and social security numbers, which were anonymized and replaced with dummy values to prioritize data privacy and ethical considerations.

The dataset's availability for research and analysis, donated by Andras Janosi, William Steinbrunn, Matthias Pfisterer and Robert Detrano, on 6/30/1988, underscores its significance in advancing heart disease diagnostics. Furthermore, the removal of identifying information aligns with ethical standards and safeguards patient confidentiality, making it suitable for research use.

In summary, the heart disease dataset, particularly the subset derived from the Cleveland database, has been pivotal for machine learning research in this domain, offering a rich source of data for developing diagnostic tools and advancing healthcare analytics.

Attribute Information

The dataset includes the following attributes:

- **Age:** Age of the patient (in years)
- **Sex:** Gender of the patient (0 = female, 1 = male)
- **Chest Pain Type (cp):** Type of chest pain experienced (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)

- **Resting Blood Pressure (trestbps):** Resting blood pressure (in mm Hg)
- **Serum Cholesterol (chol):** Serum cholesterol level (in mg/dl)
- **Fasting Blood Sugar (fbs):** Fasting blood sugar level (0 = normal, 1 = elevated)
- **Resting ECG (restecg):** Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy)
- **Maximum Heart Rate (thalach):** Maximum heart rate achieved
- **Exercise-Induced Angina (exang):** Exercise-induced angina (0 = no, 1 = yes)
- **ST Depression (oldpeak):** ST depression induced by exercise relative to rest
- **Slope of ST Segment (slope):** Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)
- **Number of Vessels (ca):** Number of major vessels colored by fluoroscopy (0-3)
- **Thal:** Defect type (3 = normal, 6 = fixed defect, 7 = reversible defect)
- **Class (num):** Diagnosis of heart disease (0 = absence, 1-4 = presence)

Dataset Characteristics

- **Data Type:** Predictive and Descriptive: Nominal, ordinal, and continuous
- **Task:** Classification
- **Attribute Type:** Categorical and continuous
- **Area:** Medical
- **Missing values?** Yes
- **Number of Instances:** 303
- **Number of Attributes:** 14 (including the class variable)

METHODOLOGY

Data Preprocessing

Data preprocessing is a crucial step in machine learning and predictive modeling, as it ensures that the data is in a suitable format for the subsequent modeling steps. In this study, we performed several preprocessing techniques to prepare the heart disease dataset for analysis.

Missing Value Handling: The initial step in data preprocessing involved handling missing values in the dataset. We used the following approach:

- The `dropna()` function from pandas was applied to remove any instances (rows) containing missing values. While this approach results in a loss of data, it ensures that the remaining instances have complete information, which is essential for accurate model training and inference.

Feature Discretization: Several features in the dataset, such as age, resting blood pressure, serum cholesterol level, maximum heart rate, and ST depression, are continuous numerical variables. To incorporate these features into the Bayesian Network model effectively, we discretized them into ordinal categories using the KBinsDiscretizer from scikit-learn.

The KBinsDiscretizer is a robust discretization technique that divides the continuous features into k bins based on quantile information. We employed the following settings:

- **Number of bins (n_bins):** We set $n\_bins = 12$, which discretized each continuous feature into 12 ordinal categories or bins. This choice of a higher number of bins compared to the default setting of 4 bins was made to capture the granularity and nuances of the continuous features more effectively.
- **Encoding ($encode$):** We set $encode =' ordinal'$ to transform the discretized data into ordinal attributes, preserving the inherent ordering of the values.
- **Strategy ($strategy$):** We used the 'quantile' strategy, which mimics the behavior of pandas' `qcut` function, ensuring an approximately equal number of instances in each bin.

The discretization process involved the following steps:

1) Fitting the KBinsDiscretizer on the continuous features using the `fit_transform()` method, which learns the bin boundaries from the data and transforms the continuous features into discrete ordinal categories.
2) Converting the discretized array back into a pandas DataFrame, with column names indicating the binned features (e.g., `age_bin`, `trestbps_bin`, etc.).
3) Concatenating the binned features with the original dataset, replacing the continuous features with their discretized counterparts.

Discretization offers several benefits, including:

- Improved handling of non-linear relationships between continuous features and the target variable.
- Reduced computational complexity and enhanced interpretability of the learned Bayesian Network structure.
- Potential mitigation of the impact of outliers or skewed distributions on the model's performance.

The choice of a higher number of bins (15) was made to capture the granularity and nuances of the continuous features more effectively, while still maintaining a reasonable level of discretization. This approach strikes a balance between preserving the information content of the continuous variables and the benefits of discretization for the Bayesian Network model.

After preprocessing, the dataset was ready for further analysis, including structure learning, parameter estimation, and model evaluation.

Structure Learning

Structure learning for the Bayesian Network is performed using the HillClimbSearch algorithm with the Bayesian Information Criterion (BIC) score. The BIC score balances model fit and complexity, helping to identify the most informative features and their dependencies while avoiding overfitting. Bayesian Networks are well-suited for modeling complex relationships and capturing conditional dependencies between variables, making them an appropriate choice for the heart disease prediction task.

The HillClimbSearch algorithm with random restarts is employed for structure learning, which helps explore the search space more thoroughly and mitigate the risk of getting trapped in local optima. This algorithm is a greedy search method that iteratively makes local improvements to the structure, and the random restarts help to improve the chances of finding a better solution. The choice of Bayesian Networks and the HillClimbSearch algorithm for structure learning is strategic, as it allows for capturing the intricate relationships and uncertainties associated with heart disease diagnosis, leveraging the probabilistic framework of Bayesian Networks.

The choice of the maximum in-degree (indegree) and the tabu list length (length) parameters in the HillClimbSearch algorithm plays a crucial role in controlling the complexity of the learned structure. These parameters were selected on a trail and error method to strike a balance between model complexity and performance.

A higher value of indegree allows for more complex structures with a larger number of parent nodes for each variable, potentially capturing intricate dependencies but increasing the risk of overfitting. Conversely, a lower value of indegree simplifies the structure but may fail to capture important relationships. Similarly, the length parameter determines the length of the tabu list, which helps the search algorithm escape local optima by preventing the revisiting of previously explored structures.

In this study, we set indegree to 5 and length to 7. These values were chosen to balance the model's complexity and generalization ability, allowing for the representation of important dependencies while avoiding overfitting on the training data.

By carefully tuning these parameters and leveraging the capabilities of the HillClimbSearch algorithm and Bayesian Networks, we aimed to learn a structure that accurately captures the relationships between various health indicators and heart disease outcomes, ultimately enabling more reliable and interpretable predictions.

Parameter Learning

Following the structure identification of the Bayesian Network, parameter estimation was performed using the Bayesian Estimation method. Unlike Maximum Likelihood Estimation (MLE), Bayesian Estimation incorporates prior information through specified prior distributions, which is particularly beneficial in scenarios where domain knowledge is available or the dataset size is limited.

The Bayesian Estimator was utilized with the $BDeu$ (Bayesian Dirichlet equivalent uniform) prior, which facilitates a balanced learning from the data by treating all potential structures uniformly initially before learning from the data. An equivalent sample size (ESS) of 1 was chosen to ensure a weak influence of the prior, allowing the data itself to play a significant role in shaping the final probabilistic model while maintaining a degree of regularization to prevent overfitting.

The general approach to parameter learning with Bayesian Estimation is not only principled but also robust, providing a comprehensive framework that merges data-driven insights with pre-existing knowledge. This methodology is particularly advantageous in our context, where complex dependencies exist, and the underlying distributions of the data might not be sufficiently captured by the simplicity of MLE. Bayesian methods provide a flexible yet theoretically rigorous means to quantify uncertainty, which aligns well with the probabilistic nature of Bayesian networks and enhances model interpretability and decision-making reliability in practical applications.

Model Evaluation

The learned model is used for inference and prediction on the test data. The predictions are compared against the true labels, and various performance metrics are calculated, such as accuracy, precision, recall, F1-score, and the confusion matrix.

The choice of performance metrics provides a comprehensive assessment of the model's performance and allows for comparisons with other models or baselines. Accuracy measures the overall correctness of the predictions, while precision, recall, and F1-score provide insights into the model's performance for specific classes (e.g., heart disease presence or absence). The confusion matrix further breaks down the predictions into true positives, false positives, true negatives, and false negatives, enabling a detailed analysis of the model's strengths and weaknesses.

Implementation Details

The implementation of the probabilistic graphical model for heart disease prediction is carried out using Python and several libraries, primarily the pgmpy library for Probabilistic Graphical Models. The pgmpy library offers comprehensive functionality for structure learning, parameter estimation, inference, and visualization of Bayesian Networks, making it an appropriate choice for this project.

The implementation follows these key steps:

1) **Data Fetching and Preprocessing:** The heart disease dataset is fetched from the UCI Machine Learning Repository using the `fetch_ucirepo` function. The dataset is then preprocessed by handling missing values and discretizing continuous features using the `KBinsDiscretizer` from scikit-learn. The discretization process transforms continuous features like age, resting blood pressure, serum cholesterol, maximum heart rate, and ST depression into ordinal categories using quantile-based binning with 12 bins for each feature.
2) **Data Splitting:** The preprocessed dataset is split into training and test sets using the `train_test_split` function from scikit-learn, with a 20% test set size and a fixed random state for reproducibility.
3) **Structure Learning:** The structure of the Bayesian Network is learned from the training data using the Hill Climb Search algorithm (`HillClimbSearch`) with the Bayesian Information Criterion (BIC) score as the scoring method. The search algorithm explores the space

of possible network structures, making local improvements to the structure while considering the trade-off between model fit and complexity. The maximum in-degree and tabu list length parameters are set to 5 and 7, respectively, to control the complexity of the learned structure. To improve the chances of finding a better solution, multiple random restarts were perfomred,we experimented with 20, 200,20000 and 20 and the run 20 was used, and the structure with the highest BIC score is selected.

4) **Parameter Learning:** After learning the structure of the Bayesian Network, the conditional probability distributions of the variables are estimated using the Bayesian Estimator (`BayesianEstimator`) with the BDeu prior type and an equivalent sample size of 5. This step learns the parameters of the model based on the training data.
5) **Visualization:** The learned Bayesian Network structure is visualized using the `networkx` and `matplotlib` libraries, creating a directed graph representation of the network. The visualization aids in interpreting the learned dependencies and relationships between the features and is saved as a PDF file (`learned_bn_structure.pdf`).
6) **Inference and Evaluation:** The learned Bayesian Network model is used for inference and prediction on the test data using the Variable Elimination algorithm (`VariableElimination`). For each instance in the test set, the model computes the most probable value of the target variable (heart disease diagnosis) given the evidence from the other features. The predictions are then evaluated against the true labels using various performance metrics, such as accuracy, precision, recall, F1-score, and confusion matrices, both for multi-class and binary classification tasks.

The implementation leverages the flexibility and capabilities of the pgmpy library, allowing for efficient structure learning, parameter estimation, inference, and evaluation of the Bayesian Network model for heart disease prediction.

Evaluation and Results

The PGM-based model achieved the following performance metrics on the test data:

Binary Classification

For the binary classification task, where the target is simplified to the presence or absence of heart disease, the model achieved the following performance:

- Accuracy: 0.7500
- Precision: 0.9444
- Recall: 0.7234
- F1 Score: 0.8193

The confusion matrix for the binary-class classification task is as follows:

TABLE I
Confusion Matrix

	Predicted	
Actual	11	2
	13	34

The learned structure of the Bayesian Network for the binary classification task is shown in Figure 1. The structure highlights the conditional dependencies and relationships between the various features and the target variable (heart disease presence or absence).

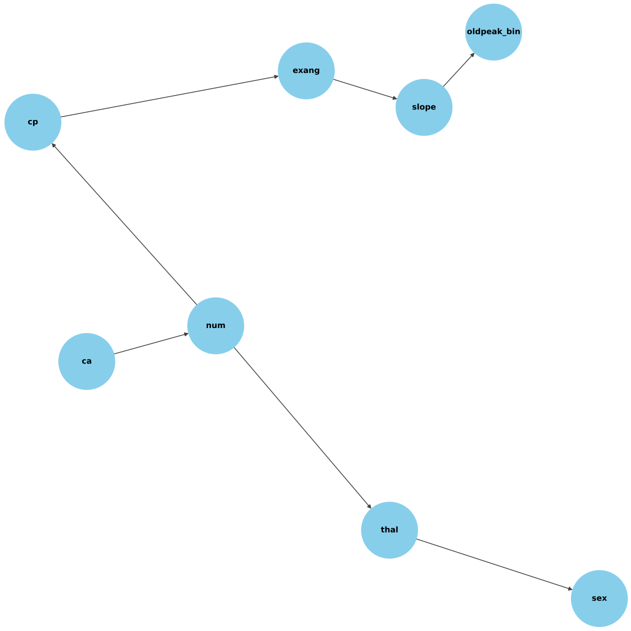

Fig. 1. Learned Bayesian Network structure for heart disease prediction.

Additional Experiment: K-Fold Cross-Validation

To further evaluate the performance and generalization capabilities of the PGM-based model, we conducted an additional experiment using K-fold cross-validation. This technique involves partitioning the dataset into k equal-sized subsets or folds, and iteratively training and evaluating the model on different combinations of these folds. In this experiment, we set $k = 5$, resulting in a 5-fold cross-validation process. The dataset was split into five equal folds, and the model was trained and evaluated five times, with each fold serving as the test set once, and the remaining four folds constituting the training set. The motivation behind using K-fold cross-validation is twofold: (1) it provides a more robust and reliable estimate of the model's performance by evaluating it on multiple train-test splits, and (2) it helps mitigate potential issues related to a single train-test split, such as biases or inadequate representation of the data in the test set. The results of the 5-fold cross-validation experiment are summarized in the provided output. For each fold, the following metrics were calculated and reported:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

Additionally, the average values of these metrics across all folds were computed and reported, providing an overall assessment of the model's performance. The results demonstrate the variability in performance across different folds, with accuracy ranging from 0.65 to 0.8644, precision ranging from 0.8438 to 0.9706, recall ranging from 0.5814 to 0.825, and F1 score ranging from 0.7042 to 0.8919. This variability underscores the importance of evaluating the model on multiple train-test splits to obtain a more comprehensive understanding of its capabilities. Overall, the average metrics across all folds suggest that the PGM-based model achieves reasonable performance, with an average accuracy of 0.7549, average precision of 0.9039, average recall of 0.7149, and average F1 score of 0.7962. The inclusion of the K-fold cross-validation experiment and its results in the report provides a more robust and comprehensive evaluation of the PGM-based model's performance, strengthening the overall analysis and conclusions drawn from this study. The learnt network during this experiment can be seen in figure 2.

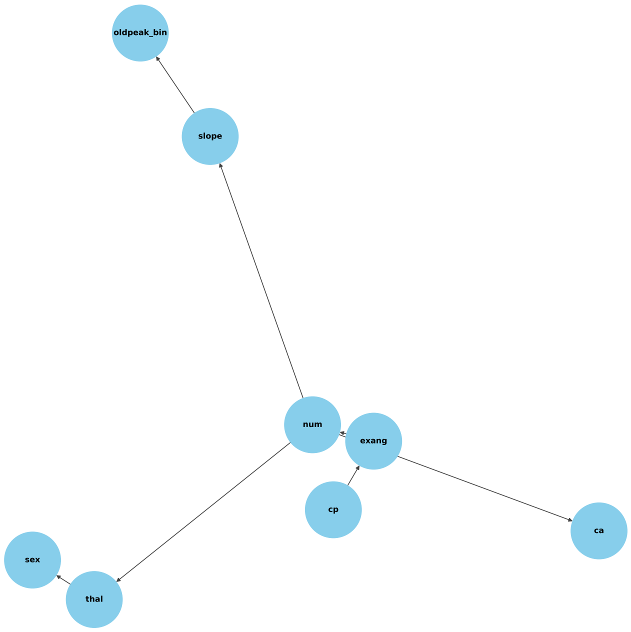

Fig. 2. Learnt Bayesian Network

Discussion

The results obtained from the evaluation of the probabilistic graphical model (PGM) for heart disease prediction provide valuable insights into the model's performance and its potential implications for clinical practice.

Firstly, the binary classification task, where the model distinguishes between the presence and absence of heart disease, achieved a respectable accuracy of 75%. This indicates that the model is capable of making correct predictions on a significant proportion of cases, which is encouraging for its potential application in real-world scenarios. The high precision (94.44%) further underscores the model's ability to correctly identify instances of heart disease when it is present, minimizing the number of false positives. This is particularly crucial in a medical context, where false positives can lead to unnecessary interventions and patient anxiety.

The results from the K-fold cross-validation experiment provide further insights into the model's generalization capabilities. The variability in performance metrics across different folds, with accuracy ranging from 0.65 to 0.8644, precision ranging from 0.8438 to 0.9706, recall ranging from 0.5814 to 0.825, and F1 score ranging from 0.7042 to 0.8919, highlights the importance of evaluating the model on multiple train-test splits. The average metrics across all folds, with an accuracy of 0.7549, precision of 0.9039, recall of 0.7149, and F1 score of 0.7962, suggest reasonable overall performance. However, these results should be interpreted in light of the limitations discussed, such as the limited feature set and dataset size.

However, the recall rate of 72.34% suggests that the model has room for improvement in capturing all instances of heart disease, as it misses approximately 27.66% of cases. Enhancing the recall rate would be beneficial for ensuring that no cases of heart disease are overlooked, thereby maximizing the model's utility as a diagnostic tool.

The confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. From the confusion matrix, it is evident that the model performs well in correctly classifying instances where heart disease is absent (true negatives), with a count of 34. However, it struggles more with instances where heart disease is present, particularly in terms of false negatives (13 cases), indicating instances where the model fails to identify actual cases of heart disease.

The learned Bayesian Network structure provides valuable insights into the conditional dependencies and relationships between the various features and the target variable (heart disease presence or absence). The visualization of the learned structure allows clinicians and researchers to interpret the model's decision-making process and understand which features play a significant role in predicting heart disease.

Overall, while the PGM-based model shows promise for heart disease prediction, further refinement and validation are necessary to enhance its performance and ensure its suitability for clinical deployment. Future work could involve incorporating additional features, exploring alternative modeling techniques, and conducting extensive validation studies using diverse datasets.

Expected Benefits of Bayesian Network Modeling

The choice of a Bayesian Network (BN) for heart disease prediction was grounded in several anticipated benefits:

1) **Handling Uncertainty:** Bayesian Networks excel at representing and reasoning under uncertainty, which is inherent in medical diagnosis. By explicitly modeling uncertain relationships between symptoms and disease outcomes, BNs can provide more robust predictions.
2) **Incorporating Prior Knowledge:** BNs allow for the incorporation of prior knowledge or domain expertise into the model, providing a mechanism to encode known relationships between variables and improve prediction accuracy.
3) **Interpretability:** The graphical nature of BNs facilitates the interpretation of learned dependencies between features and the target variable. Clinicians can gain insights into the underlying mechanisms driving predictions, enhancing trust and acceptance of the model in clinical practice.

Potential Reasons for Suboptimal Performance

Despite the anticipated benefits, several factors may contribute to the model's suboptimal performance:

1) **Limited Feature Set**: The heart disease dataset used for modeling may lack certain features that are crucial for accurately determining whether a patient has heart disease. Originally comprising 76 attributes, the dataset was reduced to 13 features during preprocessing. Furthermore, after structure learning occurred, the resulting Bayesian Network included only 8 attributes, including the target variable (cp, exang, slope, oldpeak_bin, ca, num, thal, sex). This reduction in the number of features may have impacted the model's performance, as important predictors might have been omitted.
2) **Limited Dataset Size:** Another factor contributing to the suboptimal performance could be the relatively small size of the dataset. With only 303 instances initially, and further reduction to 297 instances after removing rows with missing data, the dataset size might not have been sufficient to effectively capture the complexity of the relationships between the features and the target variable. A larger dataset could potentially provide more diverse examples for the model to learn from, leading to improved performance.

Conclusion

In conclusion, this study presents a novel approach to heart disease prediction using a probabilistic graphical model (PGM) based on Bayesian Networks. Leveraging a well-known dataset containing relevant features for heart disease classification, the PGM-based model demonstrates promising performance in distinguishing between the presence and absence of heart disease.

Through meticulous data preprocessing, including handling missing values and discretizing continuous features, the dataset was prepared for analysis. Structure learning and parameter estimation techniques were then employed to learn the dependencies between features and heart disease outcomes, resulting in a probabilistic model capable of making predictions on unseen data.

The evaluation of the model on a test dataset revealed encouraging results, with respectable accuracy and precision rates. However, further improvements are needed to enhance the model's recall rate and ensure comprehensive coverage of all instances of heart disease.

The additional K-fold cross-validation experiment conducted in this study further strengthens the evaluation of the PGM-based model's performance and generalization capabilities. By partitioning the dataset into multiple folds and iteratively training and testing the model on different combinations, a more robust assessment of the model's behavior is obtained. The variability in performance metrics across folds underscores the importance of comprehensive evaluation techniques and highlights potential areas for improvement, such as feature engineering or data augmentation.

Overall, the PGM-based approach holds significant potential for revolutionizing heart disease diagnosis and improving patient outcomes. By leveraging the power of probabilistic modeling and incorporating domain knowledge and data-driven insights, this approach can augment clinical decision-making and facilitate more accurate and timely diagnoses of heart disease.

Future research directions may involve refining the model architecture, incorporating additional data sources, and conducting extensive validation studies in diverse clinical settings. With continued advancements in machine learning and healthcare analytics, the PGM-based approach to heart disease prediction stands poised to make a meaningful impact on clinical practice and patient care.

References

[Janosi and Detrano, 1988] Janosi, Andras, S. W. P. M. and Detrano, R. (1988). Heart Disease. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52P4X.