Optimizing Predictive Models for Healthcare: A Comparison of Personal and General Datasets in Coronary Artery Disease Risk Assessment

Isaac Glenu

Abstract:

Despite the field of medicine continuing to grow, both in reach and technological advancement, Coronary Artery Disease (CAD) continues to be the leading cause of mortality in the United States [3]. CAD is a common heart condition that occurs when plaque buildup narrows the arteries, potentially leading to heart attacks and other complications that prevent the flow of blood. Assessing CAD risk is crucial yet challenging due to variability in patient data of different cardiovascular factors. Millions of people every year die of CAD, making early risk assessment a critical necessity. Machine learning has become a powerful tool in healthcare, demonstrating its potential to identify individuals at risk based on their emerging symptoms and other health factors, while providing targeted procedures to save lives. But we have realized that the effectiveness of these models depends highly on the datasets used to train them. Choosing the right training population for these models is a nuanced problem. Training a model on too specific a population risks not generalizing to a larger population, but training too large a population may risk not being personalized enough to individual patients. This paper aims to compare the efficacy of predictive models trained on more personalized datasets versus general datasets that source their data from wider and smaller populations, using data sourced from the UC Irvine Machine Learning Repository. After training models on different populations, we analyzed the results to see if there are trade-offs between accuracy and generalizability when using personal and general datasets to predict CAD. By contributing to the further optimization of predictive models for CAD, we hope to explore more effective strategies to increase CAD risk prevention.

Introduction:

Coronary Artery Disease (CAD) is a common heart condition that occurs when plaque buildup narrows the arteries, leading to heart attacks and other complications that prevent the flow of blood and is the most common cause of death in the developed world, being responsible for about one in every five deaths [3][6]. Most patients who have Coronary Artery Disease go unnoticed because many patients with CAD do not experience any symptoms before experiencing a significant first heart attack or a block in one of the main arteries/veins that connect to the heart [9]. One might have CAD for many years and not have any symptoms until one experiences a heart attack [7]. Because of this reason, early detection and risk assessment are important in improving patient outcomes and predicting CAD. In recent years, artificial intelligence and machine learning algorithms have found widespread use in autonomous driving, recommender systems in e-commerce and social media, finance, natural language comprehension, and question answering systems [8]. The use of machine learning has now reached the field of healthcare, where the use of predictive models in healthcare has gained significant attention in helping diagnose and predict different diseases, such as malaria and other bloodstream infections [11][12]. As healthcare decision-making becomes increasingly reliant on technology, there will be models that allow patients and doctors to forecast situations and better understand how predictions are made [10].

Machine learning and artificial intelligence models offer new ways to assess healthcare risk based on genetic, clinical, and lifestyle data shown in various case studies [5]. These models have the potential to revolutionize the way healthcare providers identify individuals at risk of developing CAD. Prediction models can come in different forms to optimize and conclude with accurate results that will benefit both the healthcare provider and the patient. Cardiovascular disease prediction using machine learning and deep learning has been developed and has proven to yield significant results [2]. Depending on the data that the model is given, the results could be attributed to the consensus of CAD patients, or they could be dependent on a certain individual.

The performance of predictive models can differ greatly based on how the datasets are utilized for training and validation. When we use personal/specific datasets, which are collected from specific hospitals or patient groups, the trained models may perform better on other members from that same population, but struggle to generalize to outside populations. In contrast, more general datasets that combine multiple training populations can offer diverse data samples, which allow the model to have broader applicability, but may not provide the same type of predictive precision on an individual basis.

General Literature Review:

Coronary artery disease (CAD) is hard to detect as it develops latently over many years, showing no symptoms until the disease becomes advanced or leads to a complication like a heart attack. If symptoms do appear, such as chest pain or shortness of breath, they are often mistaken for other illnesses and are usually considered too mild for any precaution to be taken. On the other hand, CAD symptoms vary widely by sex. Men who have CAD would often describe crushing chest pain, while the majority of women experience it as pressure or tightness, likely due to experiencing nausea, fatigue, and back pain. A research study on the sex differences in coronary heart disease and stroke mortality found that cardiovascular disease mortality rates are known to be lower in women than in men, and they tend to increase with age. But whether these age and sex factors have changed over recent decades or by how they have evolved in recent decades remains unclear [19]. The varying nature of this disease and the multiple factors involved in diagnosing it demand accurate prediction machine learning algorithms and models, which are essential to prevent further lives from being lost.

There have been many machine learning algorithms that have emerged as powerful tools to allow for early diagnosis and improve patient outcomes. According to a 2024 study on Predictive Modelling of Heart Disease Using Machine Learning Models, many models have been used in heart disease detection, each with distinct methodologies and levels of accuracy. The models discussed in this study are the Random Forest classifier (RF), logistic regression (LR), decision tree (DT), support vector machines (SVM), neural networks (NN), and k-nearest neighbour (KNN) [20]. Depending on the context and amount/type of data given to each of the models, different results will be found. One researcher found that the random forest classifier outperformed the support vector, decision tree, and random forest classification algorithms, with 83%, 79%, and 84 % accuracy, respectively [23]. Another study concluded that the method decision curve analysis indicated with SVM was slightly better than the logistic regression model when applied to guide clinical decision-making in the Kazakh Chinese population [24].

Another scientific inquiry evaluated the accuracy of the models in terms of their mean absolute error (MAE) and root mean squared error (RMSE) for predicting cardiovascular disease in the Hungarian

population [22]. The models predicted cardiovascular disease with less time and greater accuracy with data from the Hungarian database, with linear regression and random trees taking up the least amount of time. The m5p ML model had a prediction error of 0.2763, which was the lowest, and its MAE measure proved it. The performance of these models in the Hungarian database was assessed with the RMSE value of 0.769 and was highly recommended for accurately predicting cardiovascular diseases. The rep tree model was the most accurate for predicting cardiovascular disease, with a score of 99.81. Based on the MAE measure, naive Bayes and random tree models were found to have the lowest prediction errors, with the RMSE of 0.023, found to have the lowest prediction capability. Based on the analysis of the Statlog database, the naive Bayes model and random trees were recommended models for cardiovascular disease prediction. Utilizing the random tree algorithm was found to be faster and better for complex disease predictors [22][20]. These computer algorithms aim to bring in information and knowledge from similar patient data to predict and optimize prevention treatments for new individuals with similar factors that correspond with them individually. They use non-linear classifiers models (Machines, Random Forest, Decision Tree, and Artificial Neural Networks) that capture intricate relationships and patterns in data categorizing those instances that are not linearly separable while linear classifier models (Linear Support Vector Machine, Logistic Regression Model) are the opposite, predicting outcomes by dividing data into classes using a linear decision boundary, such as a line or a plane.

Literature has been published to showcase the recent advancement in predictive modelling in CAD healthcare and treatment. For example, one study published in the United States National Library of Medicine found that when using AI-based computer algorithms to analyze patient data, they have been able to successfully model CAD through a personalized framework. Their applications have allowed for greater accelerated data analysis and have the ability to detect genetic patterns to predict the identification of CAD [13]. The capacity to use computers to model the behaviour of the human body allows for individualized cardiac simulations targeted to particular patients, by integrating data within parameters (omics, biomarkers, clinical aspects, and genetic profiles). The next step is a diagnosis, risk stratification, decision-making, and prognosis. Then they set a feedback loop to allow the AI-driven PM algorithm to learn and enhance its effectiveness continuously [13]. So by integrating these advancements in technology into healthcare, which have allowed for a thorough evaluation of possible outcomes without putting the patient through any dangerous or life-risking operations, and providing patient information to an extent that cannot be found anywhere else. Making decisions about a patient's overall cardiovascular health provides reassurance to the doctor and the patient to proceed to the next step of action. AI modelling thrives in diverse and vast amounts of data, giving it significant ability to improve disease risk prediction.

Great success has been achieved in CAD risk assessment by using machine learning techniques for early detection. Many types of models have had varying success depending on their context and the data that they have been given to analyze. This same study, [13], found that the RF-based model outperformed the logistic regression-based model, achieving an accuracy of 85.4%. Another study, [21], achieved more accurate results with the synthetic minority over-sampling technique (SMOTE Model), which gained a 5–6% higher accuracy than the logistic regression and decision tree. But the methodology of the SMOTE model creates data points based on the distance algorithm to fully create true representations of patient data. Other ML models (Generalized linear model, Decision tree, Random forest, Support-vector machine, Neural network and K-Nearest neighbour) obtain an accuracy value higher than 0.84, which is achieved with all but the decision tree model (just below 0.80). The performance of the neural network model is outstanding, with an accuracy of 0.9303 and a recall of 0.9380 [21]. While these studies choose

their training population based on the availability of data, it is possible these performances could be improved even more if they optimized the training population for the testing population. [18]*

Recent CAD risk assessment literature has seen a significant rise in advanced computational techniques, especially within the realm of artificial intelligence (AI) [13][14]. AI can be a powerful tool for enhancing decision support systems in various applications, including medical diagnosis, by allowing the comparison of hundreds of factors that may indicate disease [4]. In terms of CAD risk assessment models, they have evolved from the traditional diagnostic methods that were used to predict patient risk. While many of them were often expensive and tedious to run, more efficient and non-invasive approaches were devised that would allow for better results and outcomes. We observed many various machine learning models that have been proposed, utilizing datasets from clinical environments to predict coronary artery blockages based on the multiple patient attributes that could be associated with CAD and the previous works that have attempted to predictively model CAD risk assessment using logistic regression, Bayesian networks, and fuzzy modeling, all of which aim to enhance predictive accuracy while minimizing the dependence on subjective expert assessments [1].

Some current machine learning models have been criticized for their limited external validity and may not perform well when applied to different demographic groups [15][16]. It has been documented that model performance can vary between patient groups due to differences in predictor levels, such as age range. This work aims to shed further light on how different training and testing demographics may influence model performance, and to clarify how CAD diagnosis models can achieve maximal efficacy in a limited data setting. Our objective is to see what type of dataset is more optimized to be more effective in training predictive models and achieve accurate as well as precise results. Are smaller, personal datasets or larger, general datasets better to assess CAD risk assessment in patients, and what are these factors that contribute to their effectiveness? We hypothesize that the impact of training a model with a more general dataset will be noticeably beneficial, compared to training separate models for demographic subsets, as the universality of a broader dataset will outweigh the specificity tradeoff.

Methods:

Data Collection:

The dataset used came from the UCI Machine Learning Repository [17]. The compilers of this dataset, Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R., collected this dataset from Cleveland, Hungary, Switzerland, and the VA Long Beach. All the names and social security numbers of the patients are removed from the database to keep the privacy and confidentiality of each patient's well-being. In essence, this dataset contains data in the healthcare field that refers to the presence of CAD in patients based on the attributes of CAD. Table 1 lists all the attributes and explains how they factor into the prediction of CAD in patients.

Attribute:	Explanation:
Sex	Sex (male/female): Studies have shown that men generally tend to be at a higher risk of coronary heart disease than women.

Age	Age: As arteries may become weaker and vulnerable to blockage over time, the risk of CAD also increases with age.
cp	Chest pain: Different types of chest pain show the severity of heart issues, which can directly correlate with CAD risk in patients
Chol	Serum Cholesterol: Cholesterol levels that remain high are key indicators of buildup in arteries that produce blockages, which are the primary causes of CAD.
Fbs	Fasting blood sugar: A major risk factor is elevated fasting blood sugar levels that could be linked to diabetes.
Exang	Exercise-Induced Angina: Physical form body exertion often signals the underlying risk of CAD, and chest pain could be a symptom of this.
Trestbps	Resting blood pressure: High resting blood pressure can damage arteries and increase CAD risk.
Oldpeak	ST depression: During physical activity, different measures of heart stress can indicate a higher value for potential CAD risk.
Restecg	Resting Electrocardiographic Results: The electrical activity of your heart, showing abnormalities in resting ECG, may point to existing heart conditions and indicate CAD.

Table 1: *Explanation for each attribute in the dataset*

To preprocess the data, we first cleaned the data from any missing, duplicate, or inconsistent values, which can lead to inaccurate results when we use the predictive models to gather results. Our next step was to normalize the data by standardizing the range and scale of numerical features in the attributes, such as blood pressure, cholesterol levels, and age. Since machine learning algorithms are sensitive to the scale of data, normalization can ensure that no single value will disproportionately influence the predictive model.

To identify whether certain features were correlated, we first created a heat map analysis on the cosine similarity across different variables to view each of their correlations. Then, we separated the data into two specific subsets, such that we could compare the performance of specifically trained models against generally trained ones.

Model Development:

For the purposes of testing more general and more specific models, in addition to the entire dataset, we create two subsets. These two subsets are divided by age groups to compare the efficacy of personal and general datasets in predicting Coronary Artery Disease (CAD). These datasets, which we refer to as the **Young Subset** and the **Old Subset**, were constructed by separating data based on age groups. The separation allows for a focused analysis of how age-specific data impacts the accuracy of predictive models. The age cutoff for the separation was 55 to make for a roughly even distribution of patient count, as shown in Table 2.

	General Dataset ()	Young Subset ()	Old Subset ()
Number of patients	297	146	151
Age Cutoff	all	< 55	55 <

Table 2: The division of the general dataset into two different subsets based on age (young/old)

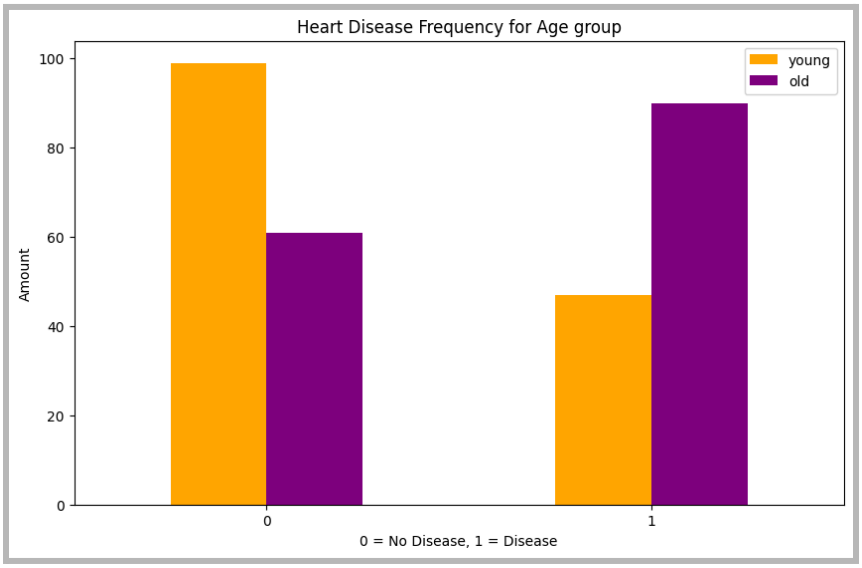


Figure 1: *Graph depicting the frequency of heart disease for each age-based subset(young/old)*
While analyzing the graph, we observed that heart disease is more common in the older population than in the young individuals, who are less likely to have heart disease. While graphing the entire population, a larger number of individuals without the disease came from the younger age group.

Model Training:
We used the scikit-learn (sklearn) library in Python to implement the three machine learning methods: K-Nearest Neighbours (KNN), Logistic Regression and Random Forest models to classify the data. Each of the models that we used has different strengths and properties that highlight the classification in predictive modelling within the healthcare domain. These characteristics are described in Table 3.

Model Testing/Evaluation Metrics:

The K-Nearest Neighbours (KNN)	The K-Nearest Neighbours (KNN) model is an algorithm that is used for classification by making predictions based on the 'K' nearest training data points. The model specifically uses proximity to create classifications and predictions about grouping individual data points. There is no type of training

	phase for this model; instead, it would just store the training data and use it for predictions. The new predicted class is typically the majority class among its neighbours. KNN is known as being simple and is a popular classification/regression classifier that is well-suited for small/medium-sized datasets.
Logistic Regression	Logistic Regression is a machine learning algorithm that is mainly used to predict the outcomes of binary classification tasks. For example, to identify if a label on a certain object or situation is binary, the model will predict that label based on the data that it is given. By graphing the model, you would find that it creates an S-shape. This specific S-shaped curve represents the model that effectively maps any data type number within the range of 0 to 1. This model is mostly used for diagnosing diseases by analyzing the presence or absence of conditions that the patient has.
Random Forest	Random forest, a common machine learning algorithm, integrates the input of data and output of multiple decision trees that lead to a single result, which one could evaluate. Using decision trees, the model handles both classification and regression situations. In terms of classifications, the output of the random forest is decided by minor decision trees, where the product would label if the object has a certain attribute or not. The Random Forest model algorithms are known to avoid and prevent the algorithm from fitting too closely to its training data.

Table 3: *Description of each Machine Learning Algorithm Models used*

Results:
We then evaluated performance outcomes of the three machine learning models used, K-Nearest Neighbours (KNN), Logistic Regression, and Random Forest, to predict Coronary Artery Disease (CAD) risk based on different datasets. Overall, we assessed the models under different conditions: training and testing on **all data**, training on **all data** and testing on **young people data**, and training on **all data** with testing on **old people data.**

By examining the results of all the models and their performances with each of these conditions, throughout age groups, we will understand which model accurately yields the most reliable predictions for CAD.

General Model:

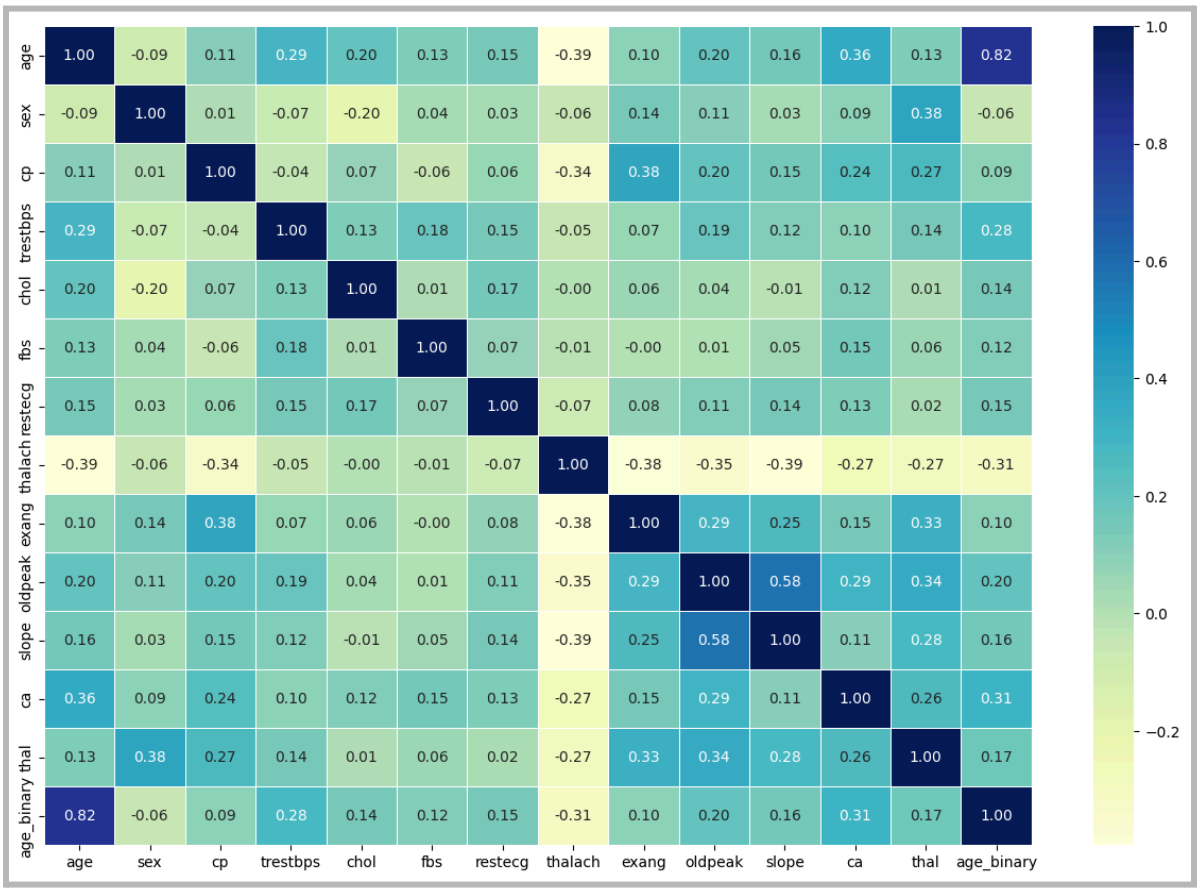


Figure 2: *Heat map depicting the correlation between independent variables in the dataset.*
This correlation heatmap shows the relationships between various features within the overall dataset used for predicting Coronary Artery Disease (CAD). The correlation values range from -1 to 1, where strong correlations (+0.50>) are observed between certain variables, such as **slope** and its binary counterpart, **oldpeak,** at 0.58, while others show moderate or weak correlations. As we evaluated this visualization, we were able to identify relationships between variables, guiding the feature selection process and ensuring that correlated features are known for the predictive modelling process.

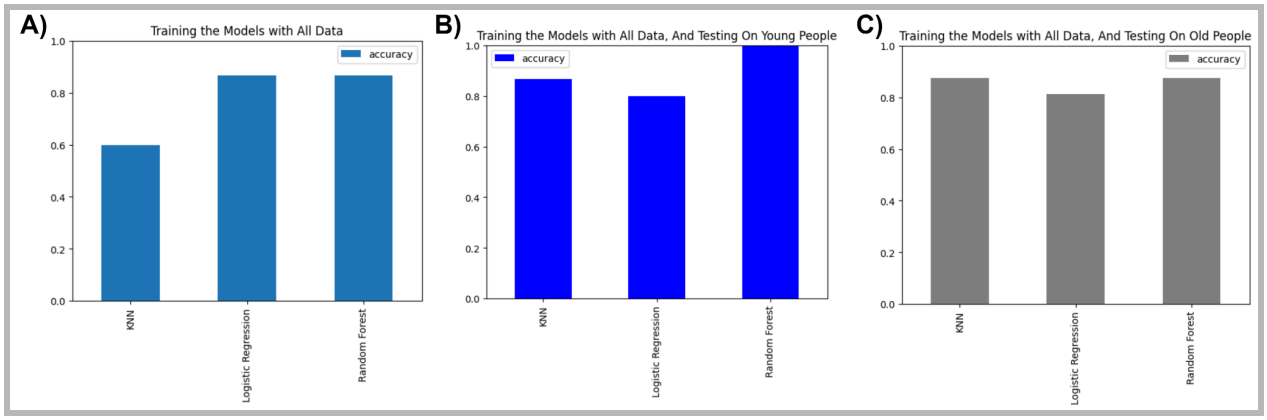


Figure 3: *Graphs that show the accuracy of three models (K-Nearest Neighbours (KNN), Logistic Regression, and Random Forest) when they are trained on* ***all data*** *and then tested under different conditions.*

To make the models, all the datasets are first set to be trained and then tested so that they are refined and ready to be tested.

Observations of results within graphs **A**, **B** and **C** of **Figure 3**:

Fig 3.A depicts the performance of models trained and tested on all available data *(Models trained on General Dataset and their results),* revealing that the KNN model has the lowest accuracy compared to the other two models (Logistic Regression and Random Forest) that show similar, high levels of accuracy. In **Fig 3.B**, the models are trained on all data but tested specifically on young individuals *(Models trained on General Dataset that are tested on Specific Datasets). There* is improvement in accuracy for all models except for the Logistic Regression model. The Random Forest model achieved a very high accuracy, outperforming the other two models, and the KNN Model also has an increased accuracy. This indicates that the Random Forest model is particularly well-suited for younger populations, as they both show enhanced precision when tested on this group.

Fig 3.C presents the models trained on all data but tested specifically on older individuals *(Models trained on General Dataset that are tested on Specific Datasets),* demonstrating that all models maintain relatively high accuracy levels. Here, the KNN and Random Forest models show similar performance, while Logistic Regression has a slightly lower accuracy rate.

When we train the data, including all ages without specification (Graph A), the model accuracy is generally the highest, but when testing on age-specific subgroups (Graphs B and C), the model itself would end up having a slight decrease in accuracy. When evaluating the results of the accuracy of each model, we see that the Random Rainforest's accuracy has the greatest average accuracy in all of the graphs, Logistic Regression is second, while KNN has the least average accuracy rate.

Specifically, the Logistic Regression model shows more promising results compared to the KNN model, since there is less difference in accuracy between all the graphs.

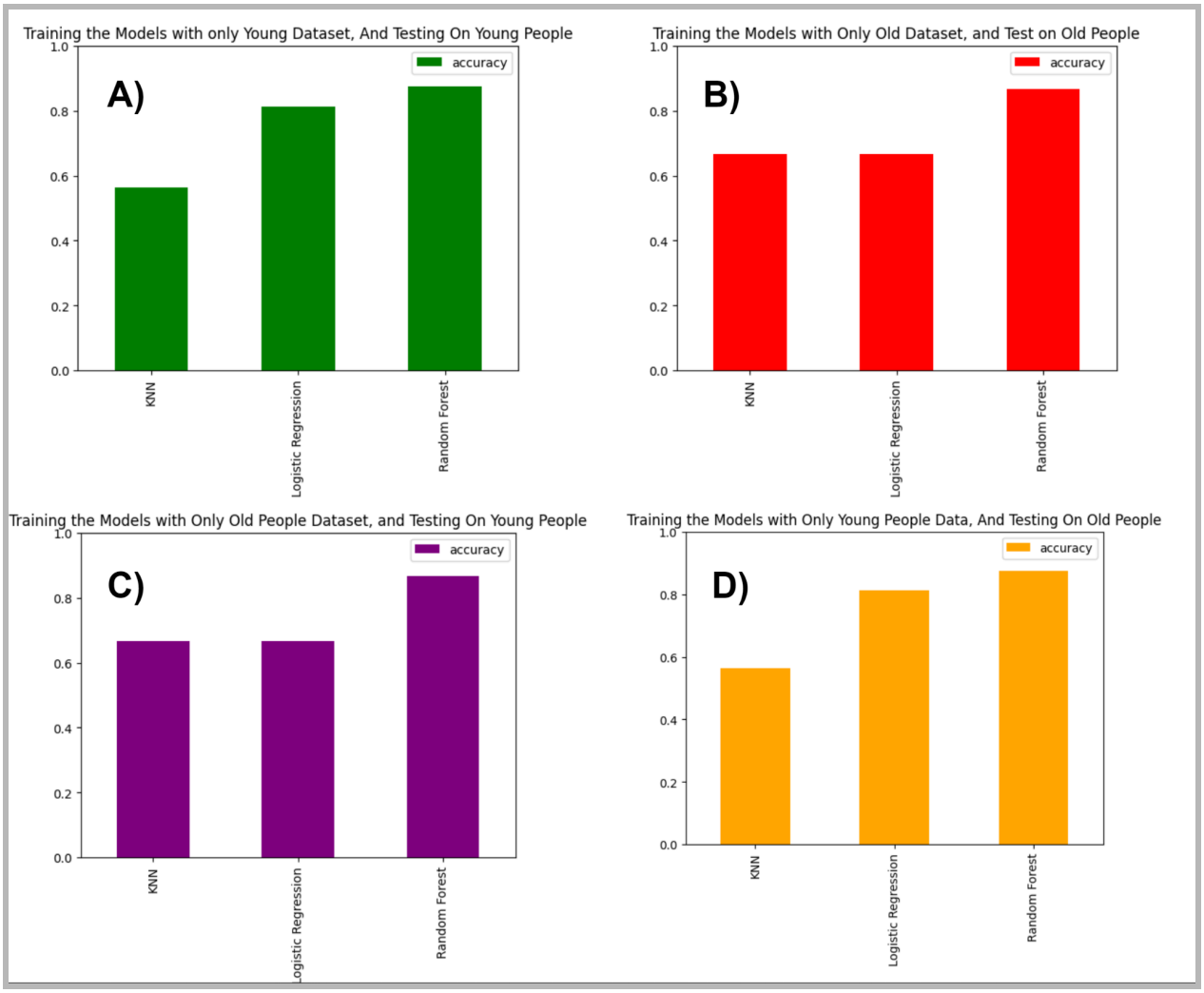
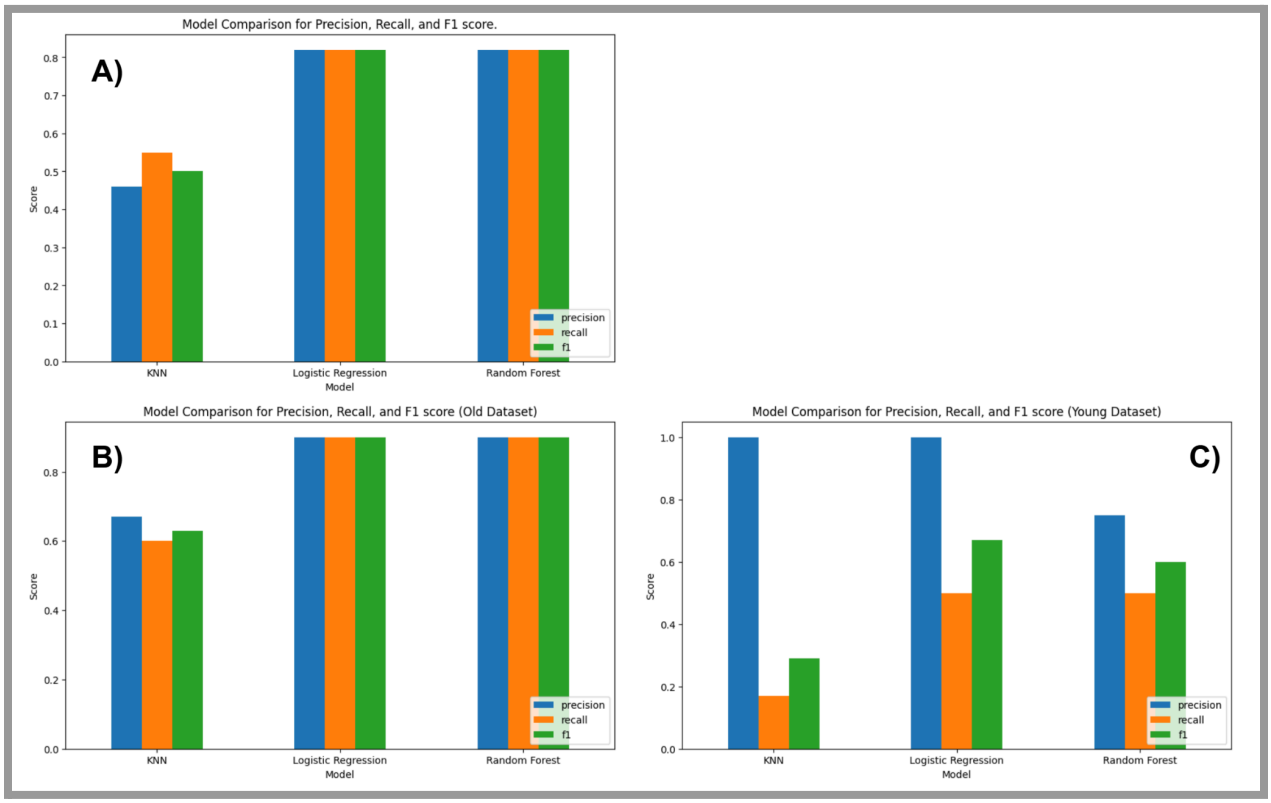


Figure 4: Graphs depicting *accuracy of models* (KNN, Random Forest And Logistic Regression): *when the model is trained on the young dataset and tested on young people (A), trained on the old dataset and tested on old people (B), trained on the old dataset and tested on young people (C), trained on the young dataset and tested on old people (D).*

Observations of results within graphs **A**, **B, C** and **D** of **Figure 4**:

In **Fig 4.A**, the Random Forest model has the highest accuracy among the three models. We can conclude that the Logistic Regression, though it has higher accuracy than the KNN model, still has a lower accuracy than the Random Forest model when trained on the young dataset and tested on young people. When the models are trained on the old dataset and tested on older individuals, **Fig 4.B**, the Logistic Regression and KNN models show similar accuracy levels. However, the Random Forest model outperforms both of these models, achieving higher accuracy. This suggests that, when working with older data, the Random Forest model is better equipped to handle the nuances of this demographic, providing more reliable results than the other models.

Fig 4.C shows the performance of the models trained on the old dataset but tested on younger individuals, which yields results similar to Graph B. Here again, the Logistic Regression and KNN models exhibit the same level of accuracy, while the Random Forest model maintains its higher performance, surpassing both the other models. This consistency in the Random Forest model's superiority indicates its robustness across different age groups, even when trained on older data. **Fig 4.D** presents the models trained on the young dataset and tested on older individuals; the results mirror those seen in Graph A. Both Logistic Regression and Random Forest models show strong performance, with Random Forest maintaining a slight edge over Logistic Regression, while KNN continues to lag in accuracy. This graph reinforces the idea that the Random Forest model tends to outperform the others, regardless of whether it is tested on younger or older populations.

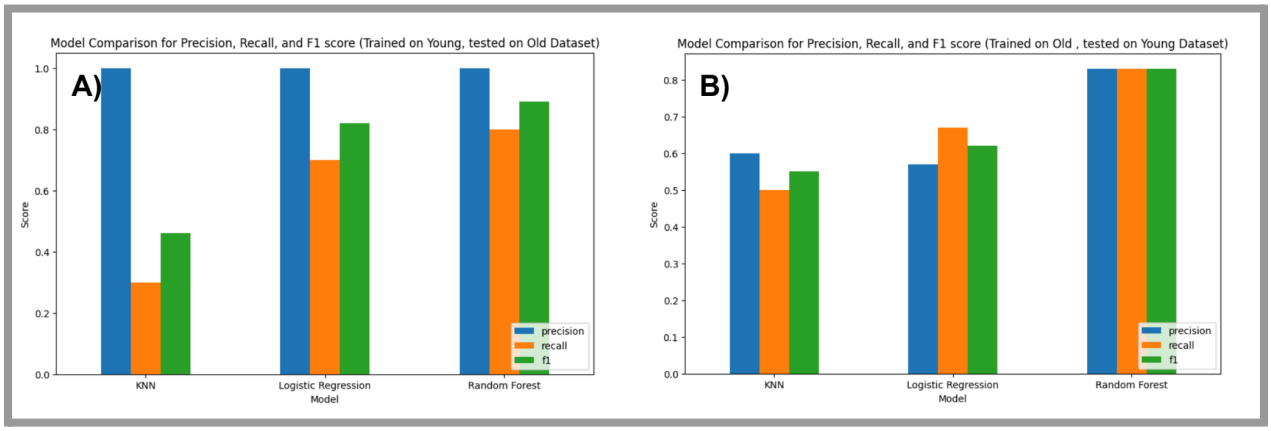


- **Figure 5:** *Graphs that compare the Precision, Recall and F1 score for each of three models (K-Nearest Neighbours (KNN), Logistic Regression, and Random Forest) when they are trained*

Fig 5.A compares the general models (KNN, Logistic Regression and Random Forest) performance in terms of Precision, Recall and F1 score.

Fig 5.B aims to compare the models and their performance in terms of precision, recall, and F1 score according to the Old Dataset.

Fig 5.C showcases the performance of models based on specific datasets, their performance in terms of precision, recall, and F1 score according to the Young Dataset.



- **Figure 6:**
- A. Comparing the models (KNN, Logistic Regression and Random Rainforest) and their performance in terms of Precision, Recall and F1 score, trained on the Old Dataset and tested on the Young Dataset.
- B. Comparing the models (KNN, Logistic Regression and Random Rainforest) and their performance in terms of Precision, Recall and F1 score, trained on the Young Dataset and tested on the Old Dataset.

Each of the bar graphs compares the performance of three machine learning models (K-Nearest Neighbours (KNN), Logistic Regression, and Random Forest) across precision, recall, and F1 score. In all the figures, the KNN model exhibits lower scores in all metrics compared to the other models, indicating its limited performance in accurately predicting CAD in patients. By observing the figures, both Logistic Regression and Random Forest show similar precision, recall, and F1 scores, reflecting their ability to balance prediction accuracy and generalizability through all of the different datasets and their conditions. Although the Logistic Regression does show promising results, Random Forest slightly edges out in all three metrics, making it the most effective model for CAD prediction in this analysis.

Discussion:

Interpretation of Results:

From the start of our exploration we divided the dataset into young and old, with the age of 55, to make for an even distribution of patient count. We first modelled the number of people who have CAD in both young and old datasets, then compared the results in a graph. We found that the older population tends to have higher rates of CAD patients. The predictability of these models was assessed by separating the dataset into young and old. Through training and assessing the ML Model's performance between the datasets, we have found that there is a clear trade-off between models trained on personalized (age-specific) datasets and those trained on general (all-inclusive) datasets in predicting heart disease. We have concluded that personalized datasets, represented by training and testing on either young or old age groups separately, tend to have higher prediction accuracy when applied to the same demographic. An example of this is when the training and testing on young data shows about a 0.9 accuracy for the

Random Forest Model, with Logistic Regression around 0.8, and KNN slightly below an accuracy of 0.6. In the same way, for the old data, Random Forest reaches 0.85, while KNN and Logistic Regression reach around 0.65. This finding implies that catering models to a specific age range could have the possibility of creating nuanced solutions since the older the age, the higher the correlations to certain factors that pertain to CAD, like thalassemia rate (thal) at 0.37 or exercise-induced angina (exang) at -0.39 in the heatmap model.

This would lead to improved precision, recall, and F1 scores within that group (e.g., young-trained models on young data show F1 scores up to 0.8 for Random Forest). We realized that, though this personalization comes at the cost of generalizability in terms of data, it was revealed that cross-testing the data drops in the performance of models trained on young data but tested on old data, at an accuracy of 0.55 for KNN. The Logistic Regression and Random Forest, in this case, maintain around an accuracy score of 0.80-0.85. On the other hand, old-trained models tested on young data dropped to about 0.7 for Logistic Regression, while KNN had a slight increase at 0.65, with Random Forest at 0.9. While graphing Precision-recall graphs, we found further indication of this, with young-trained models on old data exhibiting low recall (around 0.2 KNN score) and overall imbalanced F1 scores. We have deduced that failure to generalize across age-related physiological differences was shown by the frequency graph, where disease prevalence is usually higher in older people rather than in younger groups.

Models that were trained on general datasets tend to demonstrate strengths of balanced performance and broader application to general populations. The accuracy rates overall for these models would be Random Forest and Logistic Regression at about 0.85, and KNN at approximately 0.65. When tested on subgroups, they maintain accuracies of 0.9 for Random Forest on the young dataset, 0.8 on the old dataset, with precision-recall metrics showing consistent F1 scores around 0.8 across models. The strength of these models stems from capturing diverse correlations in the heatmap, such as links between chest pain (cp) and thalach (-0.34) or oldpeak and slope (-0.58) correlations that prompt modelling from there. But there are weaknesses to this, including a slightly lower peak accuracy compared to personalized models on matched data and potential overlooking of subgroup-specific biases, as seen in marginally lower performance on old test data, Logistic Regression at 0.75.

Implications for Healthcare:

These findings have significant insights into the further use of personalized medicine in terms of general public health predictions. We found that specific-personalized models could enable more accurate risk assessments for individuals, for example, tailored interventions that could be modified for the patient such as targeted screenings for older patients where disease frequency is higher. This aligns with the precision aspect of healthcare prediction, potentially reducing false negatives in high-risk groups, therefore improving recall in old datasets where the Random Forest excels. For public healthcare, general dataset models offer scalable tools for population-level predictions, supporting broad understandings of CAD and providing reliable factors across demographics.

However, we found that over-reliance on personalized models risks worsening health inconsistencies if data from underrepresented groups, like the young people group, where the disease is sparse, leading to biased predictions that are not necessarily true for general groups. Integrating both approaches could enhance hybrid systems, such as AI-driven apps that switch between general baselines and personalized refinements based on patient age.

Several limitations result from this matter. Data quality issues may arise from the dataset's potential biases, such as overrepresentation of certain features (e.g., a binary age split might oversimplify continuous age effects, as shown by the 0.82 correlation between age and binary_age). Model overfitting is evident in personalized training, where high in-group accuracy contrasts with poor cross-group performance, suggesting models memorize subgroup patterns rather than learning generalizable features. Biases in dataset selection, like unequal disease distribution (young skewed toward no disease, old toward disease), could inflate accuracies in matched scenarios while masking real-world variability. Additionally, the small sample sizes implied by frequency counts (~150 young, ~150 old) may limit statistical power, and the absence of external validation datasets raises questions about real-world applicability.

The major limit is bias. Predictive models are statistically or computationally trained on the preprocessed data, with the overall aim of open access optimising the quality of the predictions. All parts of this process are vulnerable to bias and other forms of error, for instance, due to missing data, how it is managed, and the representativeness of the training data. Hence, the need for model validation. In a field like healthcare, it is important to test all the possibilities to gain a deeper understanding of the relationships between the patients and the data.

Conclusion:

This study reveals that models trained on personalized (age-specific) datasets outperform general models in matched scenarios, with Random Forest consistently leading (up to 0.9 accuracy on young-young splits), but suffer from reduced generalizability in cross-age testing (drops to 0.55-0.85). General dataset models provide stable, high accuracies (~0.85 for Random Forest and Logistic Regression) across all subgroup tests, though they may not capture fine-grained subgroup nuances. Precision, recall, and F1 scores mirror this, with personalized models showing imbalances in cross-testing (e.g., low recall on mismatched data), while general models maintain equilibrium. The correlation heatmap and frequency graph underscore age as a key differentiator, with disease more prevalent in older groups and features like oldpeak and slope highly interlinked.

We have concluded that simpler models like KNN may struggle with complex datasets, more advanced models such as Random Forest and Logistic Regression provide more robust and reliable performance, while Random Forest demonstrates a slight advantage. These results don’t generalize outside the population, in practice, as broadly as possible. Based on the models' results, we conclude from our figures that specific models (that focus on personal datasets) often achieve higher accuracy for predicting Coronary Artery Disease (CAD) in patients, as they are trained on features from a smaller population.

An AI model is only as good as the data used to teach it. Our study indicates and proves that personal datasets are best for high accuracy within a specific group, while general datasets are safer and more reliable for a broader range of people. We find that Random Forest is the most suitable algorithm, outperforming Logistic Regression and KNN because it can better handle the complex connections between health factors. However, the major danger is applying a model trained on a specific group (like young people) to a different group (like older people), which can lead to poor performance and potential misdiagnoses because the health patterns don't match. Therefore, while specific models are excellent for their intended targets, general models are better overall because they prevent these errors and work

reliably for the general public at a given level, but discretion must be involved when assessing the status of individual patients.

Recommendations and Future Findings:

For healthcare practitioners, we recommend the use of general dataset models for initial population screenings or when patient data spans demographics, prioritizing Random Forest for its superior accuracy and balance. One would switch to personalized models for more unique individual cases within specific age groups to maximize precision, but always make sure with cross-testing to avoid any mistakes. Researchers should focus on general models for scalable studies but integrate personalization for subgroup analyses, ensuring diverse datasets to minimize biases. In resource-limited settings, Logistic Regression offers a computationally efficient alternative with comparable performance to more complex models like Random Forest.

We find that future research should explore the use of hybrid models which aim to combine both the general and personalized datasets, using an ensemble of methods or tools to make weighting age-specific or factor-based, for instance, boosting thalach for young or exang for old based on correlations. By further expanding to a larger, more diverse dataset, we could incorporate longitudinal data or other demographics, which have great potential to address inherent biases. We have realized that these models could be attributed to other diseases, like diabetes or cancer, where age plays a pivotal role in determining diseases. Incorporating advanced techniques like transfer learning or federated learning could mitigate overfitting while preserving privacy in personalized data.

References:

[1] N. A. Setiawan, P. A. Venkatachalam, and A. F. M. Hani, "Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System," in Proceedings of the International Conference on Man-Machine Systems (ICoMMS), Batu Ferringhi, Penang, Malaysia, 11-13 October 2009.

[2] Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. Frontiers in Medicine, 10, 1150933.

[3] CDC. (2024, April 29). Heart Disease Facts. Heart Disease. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[4] Ong Ly, C., Unnikrishnan, B., Tadic, T. et al. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. npj Digit. Med. 7, 124 (2024). https://doi.org/10.1038/s41746-024-01118-4

[5] Behar, J. A., Levy, J., & Celi, L. A. (2023). Generalization in medical AI: a perspective on developing scalable models. ArXiv.org. https://arxiv.org/abs/2311.05418

[6] American Heart Association. (2024, January 10). Coronary Artery Disease - Coronary Heart Disease. Www.heart.org. https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease/coronary-artery-disease

[7] Cleveland Clinic. “Coronary Artery Disease.” Cleveland Clinic, Cleveland Clinic, 2023, my.clevelandclinic.org/health/diseases/16898-coronary-artery-disease.

[8] Yang CC. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. J Healthc Inform Res. 2022 Feb 11;6(2):228-239. doi: 10.1007/s41666-022-00114-1. PMID: 35194568; PMCID: PMC8832418.

[9] Mayo Clinic, "Coronary artery disease: Symptoms & causes," Mayo Clinic, [Online]. Available: https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613

[10]Markham S. Patient perspective on predictive models in healthcare: translation into practice, ethical implications and limitations? BMJ Health Care Inform. 2025 Jan 16;32(1):e101153. doi: 10.1136/bmjhci-2024-101153. PMID: 39824519; PMCID: PMC11751774.

[11]Al Meslamani AZ, Sobrino I, de la Fuente J. Machine learning in infectious diseases: potential applications and limitations. Ann Med. 2024 Dec;56(1):2362869. doi: 10.1080/07853890.2024.2362869. Epub 2024 Jun 10. PMID: 38853633; PMCID: PMC11168216.

[12] Nahian, J. A., Masum, A. K. M., Abujar, S., & Mia, M. J. (2022). Common human diseases prediction using machine learning based on survey data. arXiv preprint arXiv:2209.10750.

[13]Singh M, Kumar A, Khanna NN, Laird JR, Nicolaides A, Faa G, Johri AM, Mantella LE, Fernandes JFE, Teji JS, Singh N, Fouda MM, Singh R, Sharma A, Kitas G, Rathore V, Singh IM, Tadepalli K, Al-Maini M, Isenovic ER, Chaturvedi S, Garg D, Paraskevas KI, Mikhailidis DP, Viswanathan V, Kalra MK, Ruzsa Z, Saba L, Laine AF, Bhatt DL, Suri JS. Artificial intelligence for cardiovascular disease risk assessment in personalised framework: a scoping review. EClinicalMedicine. 2024 May 27;73:102660. doi: 10.1016/j.eclinm.2024.102660. PMID: 38846068; PMCID: PMC11154124.

[14]Patel SJ, Yousuf S, Padala JV, Reddy S, Saraf P, Nooh A, Fernandez Gutierrez LMA, Abdirahman AH, Tanveer R, Rai M. Advancements in Artificial Intelligence for Precision Diagnosis and Treatment of Myocardial Infarction: A Comprehensive Review of Clinical Trials and Randomized Controlled Trials. Cureus. 2024 May 11;16(5):e60119. doi: 10.7759/cureus.60119. PMID: 38864061; PMCID: PMC11164835.

[15] Chava L Ramspek, KJ Jager, FW Dekker, C Zoccali, M van Diepen. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021 Jan;14(1):49–58. doi:10.1093/ckj/sfaa188.

[16] Van Calster B, Steyerberg EW, Wynants L, et al. There is no such thing as a validated prediction model. *BMC Med.* 2023;21:70. doi:10.1186/s12916-023-02779-w.

[17] Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [dataset]. *UCI Machine Learning Repository.* 1989. doi:10.24432/C52P4X.

[18] Akella A, Akella S. Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. Future Sci OA. 2021 Mar 29;7(6):FSO698. doi: 10.2144/fsoa-2020-0206. PMID: 34046201; PMCID: PMC8147740.

[19] Bots SH, Peters SAE, Woodward M. Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. BMJ Global Health. 2017;2:e000298. https://doi.org/10.1136/bmjgh-2017-000298

[20] Olatunji A, Abdul-Yekeen AM. Predictive modelling of heart disease using machine learning models [preprint]. 2024 Jan 02. Available from: https://www.researchgate.net/publication/377079313_Predictive_Modelling_of_Heart_Disease_Using_Machine_Learning_Models

[21] Akella A, Akella S. Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. Future Sci OA. 2021 Mar 29;7(6):FSO698. doi: 10.2144/fsoa-2020-0206. PMID: 34046201; PMCID: PMC8147740.

[22]Nadakinamani RG, Reyana A, Kautish S, Vibith AS, Gupta Y, Abdelwahab SF, Mohamed AW. Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques. Comput Intell Neurosci. 2022 Jan 11;2022:2973324. doi: 10.1155/2022/2973324. Retraction in: Comput

Intell Neurosci. 2023 Aug 16;2023:9815067. doi: 10.1155/2023/9815067. PMID: 35069715; PMCID: PMC8767405.

[23]Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics, 2, 100016. https://doi.org/10.1016/j.health.2022.100016

[24]Jiang Y, Zhang X, Ma R, Wang X, Liu J, Keerman M, Yan Y, Ma J, Song Y, Zhang J, He J, Guo S, Guo H. Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China. Clin Epidemiol. 2021 Jun 9;13:417-428. doi: 10.2147/CLEP.S313343. PMID: 34135637; PMCID: PMC8200454.

—----------