Classification of Patients with Suspected Coronary Artery Disease Based on Locally Weighted Least Squares Method

Mykola Butkevych, Kseniia Bazilevych and Iryna Trofymova

National Aerospace University "Kharkiv Aviation Institute", Chkalow str. 17, Kharkiv, 61070, Ukraine

Abstract

The study explores the application of a linear regression model integrated with Locally Weighted Least Squares for diagnosing coronary artery disease. We utilized a well-established dataset, applying the locally weighted least squares method to enhance the model's sensitivity to local data variations. Our results yielded a Mean Squared Error of 0.1282 and a Mean Absolute Error of 0.2316, indicating a model with reasonable predictive accuracy. The study concludes that while the approach shows promise, it necessitates further refinement and exploration with more diverse datasets and advanced techniques. This research contributes to the evolving landscape of coronary artery disease diagnostics, aiming to improve prediction accuracy and patient outcomes.

Keywords

Coronary artery disease, machine learning, data-driven medicine, diagnostics, classification

1. Introduction

Cardiovascular diseases (CVDs) remain a significant public health challenge globally, representing the leading cause of mortality and contributing substantially to healthcare burdens [1]. Among various CVDs, coronary artery disease (CAD) is particularly noteworthy due to its high prevalence and potential to lead to severe outcomes such as heart attacks and heart failure [2]. CAD arises primarily from the buildup of plaques within coronary arteries, leading to impaired blood flow to the heart muscle [3]. The early detection and management of CAD are crucial for improving patient outcomes and reducing the risk of life-threatening events. However, accurately diagnosing CAD in its early stages remains challenging, often requiring a combination of clinical evaluation, imaging tests, and invasive procedures [4].

The advent of information technologies has brought transformative changes to the healthcare sector. Digital tools and platforms have enhanced the efficiency and quality of care delivery, enabling healthcare providers to effectively manage and analyze large volumes of patient data [5]. Integrating electronic health records, telemedicine, and mobile health applications has improved access to healthcare services and facilitated more personalized and patient-centered care [6]. Information technologies have also played a pivotal role in advancing medical research and development, driving innovations in diagnostic techniques, treatment strategies, and disease management [7].

Data-driven diagnostics represent a paradigm shift in disease detection and management [8]. Leveraging the power of big data analytics, machine learning, and artificial intelligence, this approach emphasizes utilizing vast and diverse healthcare data to uncover patterns, make predictions, and aid in clinical decision-making [9]. By analyzing data from sources such as electronic health records, imaging studies, and genomic profiles, data-driven diagnostics can offer insights beyond traditional diagnostic methods. This method promises to enhance the accuracy,

ProfIT AI 2023: 3rd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2023), November 20–22, 2023, Waterloo, Canada

nikolai.butkevych@gmail.com (M. Butkevych); ksenia.bazilevych@gmail.com (K. Bazilevych); irina.trofymova@gmail.com (I. Trofymova)

0000-0001-8189-631x (M. Butkevych); 0000-0001-5332-9545 (K. Bazilevych); 0000-0002-1537-5601 (I. Trofymova)

efficiency, and personalization of diagnostics, particularly for complex diseases where conventional approaches may fall short.

Applying modeling techniques, such as the Locally Weighted Least Squares Method, to detecting suspected CAD significantly advances cardiovascular diagnostics [10]. This approach entails creating a predictive model that can classify patients based on the likelihood of having CAD, utilizing various clinical and demographic variables. The locally weighted least squares method, known for its ability to model complex, nonlinear relationships in data, is particularly suited for handling the multifaceted nature of CAD. Applying this method makes it possible to identify subtle patterns and associations in patient data that might indicate the presence of CAD, thereby aiding in early detection and timely intervention. This novel approach underscores the potential of combining advanced analytical techniques with clinical expertise to improve the detection and management of coronary artery disease.

The aim of the paper is to develop the model of classification of patients with suspected coronary artery disease using locally weighted least squares method.

2. Current research analysis

The current research landscape in the domain of CAD diagnosis has increasingly gravitated towards integrating advanced analytical techniques with traditional clinical methods. This synergy is driven by the growing recognition that conventional diagnostic approaches, while effective, may only partially capture the complexity and heterogeneity inherent in CAD. Recent studies have demonstrated a keen interest in exploring data-driven models, particularly those employing machine learning and statistical analysis, to enhance the precision and predictive power of CAD diagnostics. These investigations typically focus on using patient data, encompassing clinical parameters, imaging results, and biochemical markers, to develop algorithms capable of identifying patterns indicative of CAD with greater accuracy and efficiency than traditional methods alone. The thrust of this research underscores a paradigm shift towards more personalized and preemptive healthcare strategies, underlining the significance of early and accurate detection of CAD for better patient outcomes.

The research [11] focuses on evaluating a hybrid system combining Genetic Algorithms (GAs), Biogeography-Based Optimization (BBO), and Particle Swarm Optimization (PSO) with neural networks for heart disease diagnosis. The study employs the Z-Alizadeh Sani dataset, which contains 303 records. The model utilizes the top 14 weight features from the dataset, determined through trial and error, noting that increasing the number of features did not enhance prediction accuracy. The performance of the developed models, GAsBBO-MLPNNs, BBO-MLPNNs, and PSO-MLPNNs, is contingent on several factors, including the number of iterations, the number of neurons in hidden layers, population size, and the activation function of the hidden layers. The research employs a tenfold cross-validation method for evaluating the models, using 90 percent of the dataset for training and the remaining 10 percent for testing. Performance metrics such as overall accuracy, F-score, confusion matrix, Sensitivity (Recall), and Specificity are used to assess the effectiveness of the proposed models.

The paper [12] introduces a novel wrapper feature selection method for diagnosing coronary artery disease, utilizing Grey Wolf Optimization (GWO) and Support Vector Machine (SVM) classifier. The study proposes a two-stage approach to address the challenge of large datasets in medical diagnostics, which often contain redundant and irrelevant features. Initially, GWO is employed for efficient feature selection in the disease identification dataset, aiming to enhance the relevance and quality of the data used. Subsequently, the fitness function of GWO is evaluated using an SVM classifier. The methodology is validated using the Cleveland Heart disease dataset, with the results demonstrating that the proposed GWO-SVM method surpasses current approaches, achieving 89.83% accuracy, 93% sensitivity, and 91% specificity. This research highlights the potential of integrating advanced optimization techniques with machine learning classifiers to improve critical illnesses like coronary artery disease diagnosis.

The paper [13] proposes a new classification system for coronary artery abnormalities (CAAs) following Kawasaki Disease, using only coronary artery z-scores. This system was developed after reviewing echocardiograms from 1990 to 2007 of patients with a history of Kawasaki Disease. The study focuses on refining the classification of CAAs, arguing that current methods underestimate their severity. By analyzing z-scores and their distribution, the study suggests an optimized definition of CAA sizes. This research is pivotal in accurately identifying and classifying CAAs in Kawasaki Disease, which is crucial for effective management and prognosis.

The study [14] employs Chi-squared Automatic Interaction Detection (CHAID) to explore the interrelation of significant risk factors in the development of CAD. It includes a retrospective analysis of 1381 patients who underwent coronary angiography at a cardiology clinic between January 1999 and February 2003. The research assesses various factors such as sex, age, diabetes, hypercholesterolemia, hypertension, smoking status, family history of CAD, and body mass index. The findings highlight sex and age as primary risk factors, with diabetes mellitus being notably significant in certain age groups. For older females, hypercholesterolemia emerges as a key predictor. The study concludes by ranking the risk factors in order of their classification importance for CAD.

The paper [15] presents a novel method for assessing the severity of CAD from electronic health records. It utilizes a recurrent capsule network model to extract semantic relations from coronary arteriography texts, primarily using Chinese datasets. The model's performance is validated on data collected from Shanghai Shuguang Hospital, showcasing high accuracy and efficiency in CAD severity classification. This approach represents a significant advancement in using deep learning techniques for the automated analysis of CAD severity, demonstrating the potential for improved diagnostic methods in healthcare.

The analyzed papers collectively illustrate the dynamic and innovative landscape of research in CAD diagnostics. They underscore a shift towards integrating advanced computational methods, such as machine learning, neural networks, and optimization algorithms, with traditional clinical approaches. This convergence has led to more accurate and efficient diagnostic tools, reflecting a trend toward personalized medicine and improved patient outcomes. The studies, ranging from feature selection methodologies to severity classification models, demonstrate the potential of these advanced techniques in enhancing the precision of CAD diagnostics. This evolving field continues to offer promising avenues for research, potentially impacting the future of cardiovascular healthcare significantly.

3. Materials and methods

In the field of medical diagnostics, the classification of diseases such as CAD is a crucial task [16]. Our methodological approach in this study employs the Locally Weighted Least Squares (LWLS) technique to train a linear regression model specifically for CAD diagnosis. This involves selecting a representative patient dataset, applying LWLS to train the model, and using this model to predict CAD in new patient data. The performance of this model is critically evaluated against known outcomes to ensure its accuracy and reliability, with iterative improvements made based on these assessments. This approach aims to offer a nuanced and effective tool for CAD diagnosis.

Linear regression is a statistical approach to modeling the relationship between a dependent variable and one or more independent variables [17]. The dependent variable in the context of CAD diagnosis is the likelihood of CAD presence. In contrast, the independent variables are various patient data points, such as age, cholesterol levels, and blood pressure. The general form of the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon, \tag{1}$$

where y represents the predicted outcome (e.g., the probability of CAD); β_0 is the y-intercept, β_1, β_2, ..., β_n are coefficients that represent the impact of each independent variable x_1, x_2, ..., x_n on the predicted outcome, and ϵ is the error term, accounting for the deviation of the predictions from the actual values.

LWLS is an enhancement of linear regression, which applies a weighting scheme to the data points [18]. Each data point gets a weight based on its proximity to the point where the prediction is made. The closer the data point to the prediction point, the higher its weight. This is expressed mathematically as:

$$\min_{\beta} \sum_{i=1}^{n} \omega_i \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in})\right)^2, \tag{2}$$

where ω_i is is the weight assigned to the *i*-th data point. The weights are typically assigned based on a function of the distance between the data points and the point of prediction, with common choices being Gaussian or exponential functions.

The integration of linear regression with LWLS for CAD diagnosis involves using the weighted least squares approach within the linear regression framework. This combination tailors the model to local data variations, making it more adaptable and sensitive to the specific characteristics of the CAD dataset. Doing so aims to increase the accuracy of CAD predictions, acknowledging that the importance and influence of specific risk factors can vary across different patient groups.

For our study we have used the Framingham Heart Study dataset publicly available on Kaggle [19]. The dataset is a collection of data from the renowned Framingham Heart Study focused on identifying common factors or characteristics contributing to cardiovascular disease. The dataset contains various patient information, such as age, sex, blood pressure, cholesterol levels, and smoking status. It is widely used in medical research, particularly in studies related to heart disease and its risk factors. This dataset is valuable for developing predictive models and conducting epidemiological studies in cardiovascular health. The characteristic of the dataset is presented in Table 1.

Table 1
The dataset characteristic

Attribute	Scale type	Range
Sex	Boolean	0, 1
Age	Metric	32..70
Education	Metric	1..4
CurrentSmoker	Boolean	0, 1
CigsPerDay	Metric	0..43
BPMeds	Boolean	0, 1
PrevalentStroke	Boolean	0, 1
PrevalentHyp	Boolean	0, 1
Diabetes	Boolean	0, 1
TotChol	Metric	143..696
SysBP	Metric	83,5..295
DiaBP	Metric	48..143
BMI	Metric	15,54..38,53
HeartRate	Metric	50..110
Glucose	Metric	45..268
TenYearCHD	Boolean	0, 1

4. Results

The training process of a model is a meticulous and multifaceted procedure. It starts with data preparation, which involves thoroughly cleaning and organizing the dataset, addressing any missing values, and standardizing variables on different scales. Special attention is given to the nature of each variable, as the dataset comprises a blend of categorical (Boolean) and continuous (Metric) variables.

Selecting the most relevant features is a critical step that impacts the model's effectiveness. This selection is based on the characteristics of the data and the goal of predicting the TenYearCHD outcome.

A logistic regression model is typically chosen for its appropriateness in handling binary outcomes and its ability to provide probabilities. The training phase involves fitting this model to the selected features using a designated portion of the dataset.

Following training, the model undergoes a rigorous evaluation to assess its accuracy, precision, recall, and other relevant metrics. This evaluation is crucial in determining the model's effectiveness in predicting coronary heart disease.

If the initial results are unsatisfactory, the model undergoes tuning, where hyperparameters are adjusted to optimize performance. Cross-validation techniques are employed to ensure the model's robustness and generalizability.

This process is inherently iterative, requiring continuous refinement and adjustments based on the performance metrics and validation outcomes. This rigorous methodology ensures the development of a reliable and accurate predictive model.

The results are presented in Table 2.

Table 2
Modeling results

Metric	Value
Mean Squared Error	0.1282
Mean Absolute Error	0.2316

The results indicating a Mean Squared Error (MSE) of 0.1282 and a Mean Absolute Error (MAE) of 0.2316 are pretty revealing. The MSE, being relatively low, suggests that the model's predictions are generally close to the actual values. This is indicative of a model with a good fit to the data. However, it is essential to consider the complexity of the dataset and the nature of CAD prediction, where even small errors can be clinically significant.

The MAE provides a straightforward interpretation of the average magnitude of prediction errors without considering their direction. An MAE of 0.2316, in the context of CAD prediction, can be considered moderate. It highlights that while the model has predictive validity, there is still room for improvement, especially in reducing false positives and false negatives, which are critical in medical diagnostics.

These results should be contextualized within the study's limitations, including the dataset's representativeness and the model's generalizability to other populations. Future work could focus on incorporating more diverse datasets, exploring more complex models, or integrating additional relevant features that could enhance the model's predictive accuracy. Additionally, the implications of these findings for clinical practice should be cautiously interpreted, considering the balance between the benefits of early detection and the risks of over-diagnosis.

5. Conclusions

In the conclusion of this study, we underscore the criticality of enhancing CAD diagnostics in light of its increasing prevalence. The methodological novelty introduced by combining linear regression with Locally Weighted Least Squares has shown significant promise. Our findings, reflecting a balance of accuracy and areas for improvement, indicate a substantial stride in CAD diagnostic approaches. Future research directions aim to enrich the model's robustness through diverse and comprehensive datasets and explore the integration of more sophisticated techniques, possibly encompassing artificial intelligence. Such advancements could revolutionize CAD diagnostics, contributing immensely to predictive medicine and improving patient outcomes in cardiovascular health. This research serves as a stepping stone towards more accurate, efficient, and non-invasive diagnostic methods, aligning with the goal of enhancing healthcare delivery and patient care in cardiovascular diseases.

Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic "Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management".

References

[1] E. Goldsborough, N. Osuji, and M. J. Blaha, "Assessment of Cardiovascular Disease Risk: A 2022 Update," *Endocrinology and Metabolism Clinics*, vol. 51, no. 3, pp. 483–509, Sep. 2022, doi: 10.1016/j.ecl.2022.02.005.
[2] A. Kr. Malakar, D. Choudhury, B. Halder, P. Paul, A. Uddin, and S. Chakraborty, "A Review on Coronary Artery disease, Its Risk factors, and Therapeutics," *Journal of Cellular Physiology*, vol. 234, no. 10, pp. 16812–16823, Feb. 2019, doi: 10.1002/jcp.28350.
[3] A. Lala and A. S. Desai, "The Role of Coronary Artery Disease in Heart Failure," *Heart Failure Clinics*, vol. 10, no. 2, pp. 353–365, Apr. 2014, doi: 10.1016/j.hfc.2013.10.002.
[4] P. G. Steg and G. Ducrocq, "Future of the Prevention and Treatment of Coronary Artery Disease," *Circulation Journal: Official Journal of the Japanese Circulation Society*, vol. 80, no. 5, pp. 1067–1072, Apr. 2016, doi: 10.1253/circj.CJ-16-0266.
[5] M. Mazorchuck, et al., "Web-Application Development for Tasks of Prediction in Medical Domain," *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 5–7, Sep. 2018, doi: 10.1109/stc-csit.2018.8526684.
[6] K. Bazilevych, M. Butkevych, and N. Dotsenko, "Cardiac Studies Diagnostic Data Informative Features Investigation based on Cumulative Frequency Analysis," *CEUR Workshop Proceedings*, vol. 3348, pp. 84–89, 2022.
[7] D. Chumachenko, "On Intelligent Multiagent Approach to Viral Hepatitis B Epidemic Processes Simulation," *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pp. 415–419, Aug. 2018, doi: 10.1109/dsmp.2018.8478602.
[8] N. Dotsenko, et al., "Modeling of the Processes of Stakeholder Involvement in Command Management in a Multi-Project Environment," *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, pp. 29–32, Sep. 2018, doi: 10.1109/stc-csit.2018.8526613.
[9] S. Yakovlev *et al.*, "The Concept of Developing a Decision Support System for the Epidemic Morbidity Control," *CEUR Workshop Proceedings*, vol. 2753, pp. 265–274, 2020.
[10] M. Bevilacqua and F. Marini, "Local classification: Locally weighted–partial least squares-discriminant analysis (LW–PLS-DA)," *Analytica Chimica Acta*, vol. 838, pp. 20–30, Aug. 2014, doi: 10.1016/j.aca.2014.05.057.
[11] M. I. Dwaikat and M. Awad, "Hybrid Model for Coronary Artery Disease Classification Based on Neural Networks and Evolutionary Algorithms," *Journal of Information Science and Engineering*, vol. 38, no. 5, pp. 1001–1020, 2022.
[12] Q. Al-Tashi, H. Rais, and S. Jadid, "Feature Selection Method Based on Grey Wolf Optimization for Coronary Artery Disease Classification," *Advances in intelligent systems and computing*, vol. 843, pp. 257–266, Sep. 2018, doi: 10.1007/978-3-319-99007-1_25.

[13] C. Manlhiot, K. Millar, F. Golding, and B. W. McCrindle, “Improved Classification of Coronary Artery Abnormalities Based Only on Coronary Artery z-Scores After Kawasaki Disease,” *Pediatric Cardiology*, vol. 31, no. 2, pp. 242–249, Dec. 2009, doi: 10.1007/s00246-009-9599-7.

[14] T. Mevlüt, K. Imran, and K. Turhan, “Analysis of intervariable relationships between major risk factors in the development of coronary artery disease: A classification tree approach,” *Anadolu Kardiyoloji Dergisi*, vol. 7, no. 2, pp. 140–145, 2007.

[15] Q. Wang, J. Qiu, Y. Zhou, T. Ruan, D. Gao, and J. Gao, “Automatic Severity Classification of Coronary Artery Disease via Recurrent Capsule Network,” *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pp. 1587–1594, Dec. 2019, doi: 10.1109/bibm.2018.8621136.

[16] V. P. Mashtalir, V. V. Shlyakhov, and S. V. Yakovlev, “Group Structures on Quotient Sets in Classification Problems,” *Cybernetics and Systems Analysis*, vol. 50, no. 4, pp. 507–518, Jul. 2014, doi: 10.1007/s10559-014-9639-z.

[17] T. Masuda *et al.*, “Development and Validation of Generalized Linear Regression Models to Predict Vessel Enhancement on Coronary CT Angiography,” *Korean Journal of Radiology*, vol. 19, no. 6, pp. 1021–1021, Jan. 2018, doi: 10.3348/kjr.2018.19.6.1021.

[18] Y. Xie, “Fault monitoring based on locally weighted probabilistic kernel partial least square for nonlinear time-varying processes,” *Journal of Chemometrics*, vol. 33, no. 12, Dec. 2019, doi: 10.1002/cem.3196.

[19] A. Bhardwaj, “Framingham heart disease dataset,” *Kaggle*, 2021. https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset (accessed Sep. 01, 2023).