



Cross-Validated Machine Learning Stacking Ensemble Regression for CCTA-Based Coronary Artery Blockage Prediction

Ayush Vaze¹, Swarangi Gurav², Purushottam Mahajan³,
Kalyani Batwal⁴, and Omkaresh Kulkarni⁵

¹ Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology,
Pune, India, ayushvaze38367@gmail.com

² Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology,
Pune, India

³ Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology,
Pune, India

⁴ Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology,
Pune, India

⁵ Dept of Computer Engineering, Dr. D. Y. Patil Institute of Technology,
Pune, India, omkaresh.kulkarni@dypvp.edu.in

Abstract. Coronary artery disease (CAD) remains one of the top mortality causes around the world. Coronary CT angiography (CCTA) is an efficient non-invasive technique used for evaluation of coronary artery stenosis. It is nowadays used in various clinical applications, including screening and preprocedural assessment. There have been new strides in machine learning technology that sparked interest of researchers into the application of automated techniques for the analysis of cardiovascular risk factors and subsequent prediction of coronary events based on the use of CCTA data related [3, 4, 5]. However, many current solutions are based on deep learning approaches that typically require large amounts of annotated imaging data as well as high computational power which may hinder their application in practice [6, 7]. In the paper, we present a machine learning technique aimed at calculating coronary artery blockage in percentage based on the structured information about patients comprising their demographic, metabolic, and lifestyle data. A stacking ensemble regression model was established by merging a number of base learners along with a Ridge regression meta-learner, with the aim to enhance predictive stability. By utilizing the developed clinical features, the model was able to reach an R^2 score of around 0.98, thus showing high predictive quality. Then to enhance the interpretability of the developed system, the LLM (large language model) was also introduced to produce the patient-level explanations and clinical conclusions that are based on the prediction outcomes made.

Keywords: Coronary CT Angiography, Stacking Ensemble Learning, Feature Engineering, Coronary Artery Disease, Cardiovascular Risk Prediction, Machine Learning, Clinical Decision Support, Non-invasive Diagnosis.

1 Introduction

Cardiovascular diseases remain a major global health challenge, with Coronary Artery Disease (CAD) being a leading cause of mortality. The early detection of coronary stenosis is vital in preventing unfavorable cardiac events. Coronary CT Angiography (CCTA) has been recognized as an accurate technique for the assessment of coronary artery obstruction, showing promising diagnostic and prognostic performance in pre-procedural and screening studies [1], [2].

Recent advances in artificial intelligence (AI) technologies have significantly enhanced the CCTA-based stenosis detection and assessment capabilities. Recent studies on machine learning and deep learning-based stenosis detection, plaque quantification are promising [3], [4], [5], [6], [7]. However, deep learning-based imaging approaches, despite showing high diagnostic performance, require large datasets and computational power, which may not be feasible in a real-world environment.

In parallel to this, studies that incorporate imaging information with other structured clinical information such as lipid levels and metabolic risk factors have reported better cardiovascular risk prediction outcomes [8], [9]. Moreover, decision-support systems such as CarDIA-AI and multitask deep learning frameworks have emphasized the need to take the application of AI beyond diagnosis to clinical triage and risk stratification [10], [11].

Inspired by these advances in cardiovascular risk prediction with the aid of AI, the present study aims to develop an AI-based cardiac risk stratification system that predicts the percentage of blockage in coronary arteries using structured clinical information such as demographic, metabolic, and lifestyle factors. An ensemble regression method with diverse base learners and a Ridge regression meta-learner is proposed to improve the robustness of the proposed method for tabular clinical data. Furthermore, the proposed framework incorporates a large language model to enable patient-level interpretability of the proposed method [10], [11].

2 Literature Review

Boyer et al. [1] demonstrated that CCTA can reduce unnecessary invasive coronary angiography, with high sensitivity and negative predictive value. Similarly, high diagnostic accuracy (AUC 0.885) for hemodynamically significant stenoses was achieved by Langenbach et al. [2] using CT-FFR. Williams et al. [3] highlighted the growing role of AI in cardiovascular CT imaging, focusing on the application of AI in coronary segmentation, plaque analysis, CT-FFR analysis, and risk prediction. The high accuracy of AI-assisted CCTA in calcium scoring, stenosis evaluation, and dose reduction was also validated by Liao et al. [4] with AUCs of up to 0.90 in ML-assisted FFR-CT.

Petch et al. [6] developed the CarDIA-AI system to guide imaging modality selection, achieving an AUC of 0.81 and improving diagnostic yield. Lu et al. [10] designed a multitask deep learning model (TabPerceiver) for risk stratification and test prediction, showing enhanced multitask performance. The work of Dreesen et al. [7]

showed that deep learning-based motion correction of images significantly improved image quality in CCTA. Shrivastava et al. [8], in a systematic review, found excellent correlations between plaque quantification with deep learning and intravascular ultrasound. Hybrid models that integrate imaging and clinical biomarkers have also been promising. Chen et al. [9] combined CCTA image features with lipid profiles using CNN models, with an AUC of 0.87. Transformer models such as TR-Net introduced by Ma et al. [5] showed high accuracy in stenosis detection on reformatted CCTA images. Moreover, a meta-analysis by Ma et al. [11] showed that the combination of radiomic and clinical features improves the prediction of MACE.

The current work fills this void by developing a stacking ensemble regression model.

3 Methodology

3.1 Overall System Framework and Study Design

The system architecture, shown in Fig. 1, is divided into two major parts: (1) Model Development Pipeline and (2) Real-Time Deployment and AI Interaction Layer. Patient information is first obtained from hospital and clinic records, followed by data engineering and feature preparation. A stacking ensemble model is then built and tested. During prediction, the user-submitted information goes through API validation, preprocessing via scaling and encoding, and feature engineering before being fed into the ensemble for blockage percentage prediction. The predicted result is then processed by a personalized recommendation engine, an LLM-based clinical explanation module, and a state-aware AI health chatbot, with the results combined into an automated PDF clinical report.

3.2 Dataset Description and Preprocessing

The dataset comprises 500 structured patient records, which have been synthetically generated and augmented based on clinically validated cardiovascular risk factors that have been identified in preventive cardiology and public health. The data generation process is guided by established clinical guidelines to ensure realistic representation of patient profiles and meaningful relationships between input features and coronary artery blockage. The choice of variables has been guided by recommendations from the Centers for Disease Control and Prevention (CDC) on the prevention of heart disease [12], diabetes and cardiovascular risk [13], control of high blood pressure [14], sleep health [15], and diabetes statistics [16]. Moreover, it is guided by the National Heart, Lung, and Blood Institute (NHLBI) healthy weight criteria [17], American Heart Association (AHA) guidelines for lessening the cardiovascular risk [18] [19], ACC/AHA guidelines for hypertension [20], and CDC standards for physical activity measurements [18].

Seventeen main variables were assessed, which includes demographic, lifestyle, clinical, and biochemical categories. The selected variables include age, smoking habit,

level of physical activity, food habits, BMI, family health history, stress level, level of drinking, sleep hours, job type, LDL, HDL, Total cholesterol, blood pressure, diabetes, resting heartbeat, and waist circumference. The CCTA_Blockage_Percentage variable shows the level of coronary stenosis expressed as a continuous rate. The relatively high R^2 value observed in this study can be attributed to the structured nature of the dataset and the strong correlation among clinically relevant features such as cholesterol levels, blood pressure, and age with coronary artery blockage.

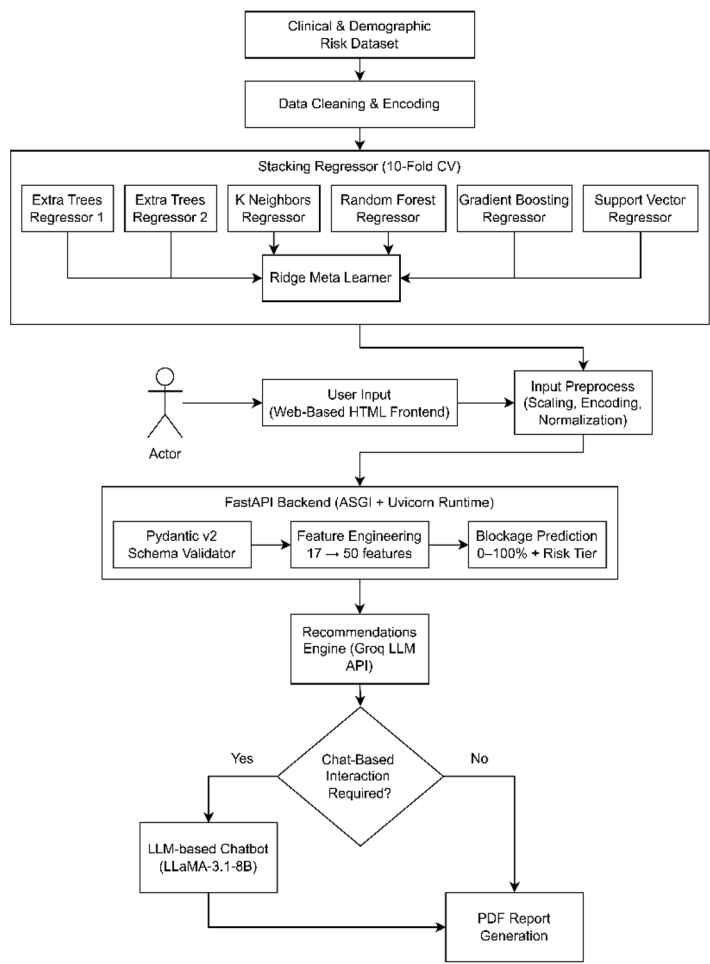


Fig. 1. System Architecture

3.3 Feature Engineering and Problem Formulation

To improve the robustness of predictions, domain knowledge-driven feature engineering was used to expand the 17 core variables into a set of 49 engineered features. The categories of transformations involved lipid ratios like the LDL/HDL ratio and Total Cholesterol/HDL ratio, which are known to be established atherogenic risk factors. Interaction terms involving anthropometric variables, namely the product term BMI $\times$ Waist Circumference, were also developed to quantify joint risk due to adiposity. Combinations of metabolic-hemodynamic factors quantifying joint effects of blood pressure and lipid variables were also included, together with lifestyle and clinical interaction terms to link behavioral risk factors with clinical markers of risk. Polynomial transformations of core continuous variables were also used to quantify nonlinear relationships in the feature space. The learning problem is stated as a supervised regression problem:

$$f : X \in \mathbb{R}^{49} \rightarrow y \in \mathbb{R} \quad (1)$$

where X is the 49-dimensional engineered feature vector, and y is the predicted CCTA blockage percentage. The objective function seeks to minimize the mean squared error while maximizing the R^2 coefficient of determination on the training and validation sets.

3.4 Model Development and Stacking Ensemble Design

Multiple predictive architectures were evaluated, including standalone regressors and neural network configurations. The final model adopted is a Stacking Ensemble Regressor, constructed using a two-layer architecture consisting of diverse base learners and a regularized meta-learner.

Five heterogeneous base estimators were incorporated as the first layer. The ExtraTrees Regressor exploits extreme randomization to reduce variance across high-dimensional feature spaces. The Random Forest Regressor provides bootstrap aggregation with decorrelated tree ensembles for stable prediction. The Gradient Boosting Regressor sequentially minimizes residual error through additive tree expansion. The K-Nearest Neighbors Regressor identifies local similarities in feature space and non-linear proximity relationships. Support Vector Regression, finally, uses kernel-based margin maximization for robust nonlinear regression. Ridge Regression was chosen as the meta-model to perform L2-regularized aggregation of base learner predictions. The final ensemble output is computed as:

$$\hat{Y} = \sum^I w^I f^I(X) \quad (2)$$

where $f^I(X)$ is the prediction made by the i^{th} base model, and w^I is the weight associated with it, learned using Ridge optimization on the predictions of the held-out fold. The optimal weight vector is obtained by minimizing the L2-regularized objective function:

$$w^* = \arg \min_w (\|y - Zw\|^2 + \alpha \|w\|^2) \quad (3)$$

where $Z \in \mathbb{R}^{(n \times K)}$ denotes the matrix of out-of-fold predictions from the K base learners, w is the meta-learner weight vector, and α is the regularization parameter controlling the bias–variance tradeoff. This stacking representation has a better bias–variance tradeoff than individual regressors, as expected by the performance characteristics of ensemble learning methods as discussed in [5]

3.5 Model Validation and Performance Evaluation

Robustness of the model was tested using 10-fold cross-validation, a stratification of the data into ten non-overlapping subsets to test generalization on different subsets of patients. Results were measured using the Coefficient of Determination (R^2), the Root Mean Squared Error (RMSE), and the Mean Absolute Error (MAE). The Coefficient of Determination (R^2), measuring the proportion of variance explained by the model, is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

where y_i is the true blockage percentage, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of observed values. The proposed Stacking Ensemble Regressor achieved an R^2 of 0.9807, an RMSE of 4.57, and an MAE of 3.52. These findings suggest that about 98% of the variance in CCTA blockage percentage is explained by the model.

3.6 AI Integration

The LLM Response Generator integrates a Large Language Model (LLaMA 3, 8B parameters) with structured prompt templates to produce patient-specific responses including a risk explanation statement in layman terms, evidence-based lifestyle recommendations, and a clinically structured summary that is interpretable by physicians. Finally, the PDF Report Generator is used to automatically generate clinical reports using ReportLab, and LLM responses into a structured clinical report that can be downloaded.

4 Result and Discussion

4.1 Result

Comparative R^2 Performance. The comparative R^2 performance of all the models considered for evaluation is represented in Fig. 2. The stacking ensemble method produced the best model with the highest coefficient of determination, with an R^2 value of 0.9807, which implies that the model was able to explain about 98% of the variance in the coronary artery blockage percentage. The Extra Trees and KNN with R^2 values of 0.9799 and 0.9771, respectively. The SVR (RBF) model had an R^2 value

of 0.9613, while the XGBoost with an R^2 of 0.9362. The stacking ensemble model performed better than all the individual models.

Mean Absolute Error (MAE) Analysis. Fig. 2 shows the comparison of MAE among models. The stacking ensemble model had the lowest MAE of 3.52, which indicates the smallest average absolute deviation of predicted values from actual blockage percentages. The Extra Trees model had an MAE of 3.64, followed by the KNN model with an MAE of 3.89. The SVR (RBF) and XGBoost models had relatively higher errors of 5.05 and 4.76, respectively. The lower MAE of the proposed ensemble model indicates better precision in stenosis percentage prediction, which is critical in risk assessment.

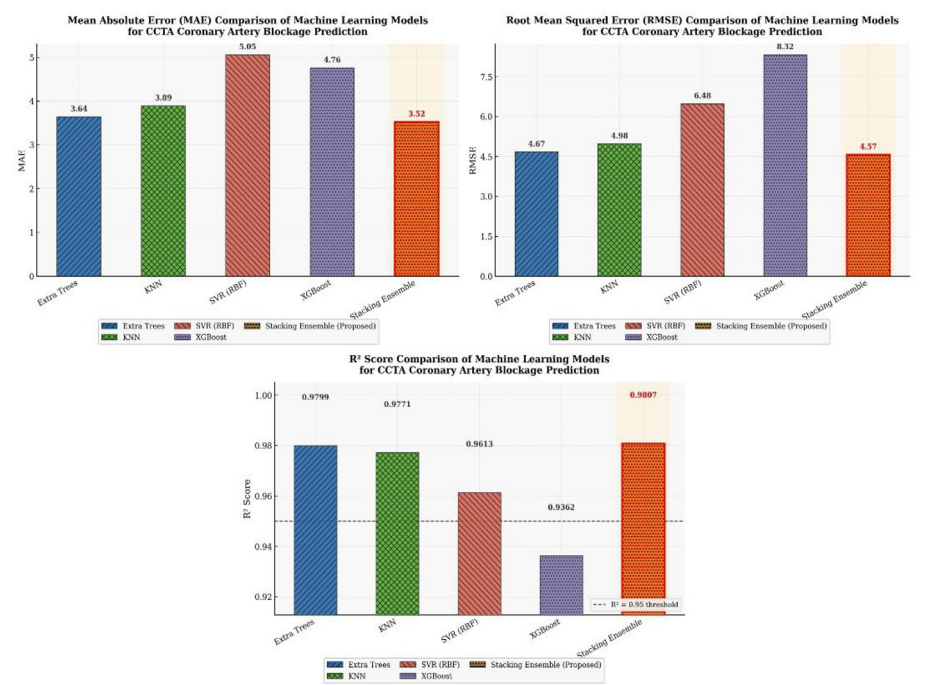


Fig. 2. . Comparative Evaluation of Machine Learning Models Using R^2 , MAE, and RMSE for CCTA Coronary Artery Blockage Prediction

Root Mean Squared Error (RMSE) Evaluation. The comparison of the RMSE values in Fig. 2 further confirms the robustness of the proposed model. The stacking ensemble method had the lowest RMSE of 4.57, followed closely by Extra Trees with an RMSE of 4.67 and KNN with an RMSE of 4.98. The SVR and XGBoost methods had relatively higher RMSE values of 6.48 and 8.32, respectively. Since the RMSE

gives more weight to larger errors than the MAE, the lower RMSE of the stacking ensemble model further confirms its better ability to control the predictions. The comparative performance of all tested regression models is presented in Table I.

Table 1. Comparative Performance of Regression Model

Model	R ² Score	MAE	RMSE
Stacking Ensemble (Proposed)	0.9807	3.52	4.57
Extra Trees	0.9799	3.64	4.67
K-Nearest Neighbors (KNN)	0.9771	3.89	4.98
Support Vector Regression (RBF)	0.9613	5.05	6.48
XGBoost	0.9362	4.76	8.32

4.2 Discussion

Impact of Heterogeneous Ensemble Learning. KNN captures local similarities since SVR can capture non-linear associations between clinical variables and incidence of blockage. Tree based classifiers specialize in complex patterns among predictors. Ridge regression takes an ensemble approach of merging predictions made by different learners using L2 regularization in order to reach the best combination for predictions. This enables the method to cancel bias and variance saving both at the same time while performing better compared to methods like XGBoost which cannot achieve such balance. Such high R² value in this research is due to the correlation between clinical variables themselves.

Comparison with Existing CCTA-Based AI Studies. These studies confirmed this research. According to Lu et al. [10], AUC reached 0.76 in the study of CAD in 14,021 people. Ma et al. [5] managed to gain high accuracy of stenosis detection by transformer model. However, apart from images, no other structured data were used. In the same way, Chen et al. [9] showed proper classification by using plaques and lipids via CNN. Nevertheless, it seems that the problem persists due to one-mode learning approach.

Clinical Deployment and Practical Implications. The inference system developed with the help of FastAPI allows to make predictions and automatically classify the risk level into one of the three categories: low (up to 20%), medium (from 21% to 50%), and high (more than 50%). Such innovations simplify the process of non-invasive diagnostics in cardiology and reduce the burden of working on interpreting results for physicians. Furthermore, with the help of the LLaMA-based chatbot, it is now easier and faster for health workers to get an understanding of the obtained results.

Limitations and Future Directions. Despite the promising results that have been obtained, there are some limitations that should be mentioned. First of all, the sample size of 500 observations, including the artificial ones, is quite low. Future work will include the validation of the model on several centers, use of SHAP analysis, calibration tests, and cohort study.

5 Conclusion

The study showcases a model of machine learning that predicts the degree of the blockage of blood vessels using the information obtained from patients who underwent the procedure of coronary CT angiography. A stacking ensemble regression model was utilized which was composed of a range of various types of base learners and the ridge model which primarily served as a meta-learner to increase predictive power in the study. Thanks to the model it was possible to discover interesting findings regarding the correlations among different factors that should be considered among patients including those being related to their lipid levels, blood pressure, metabolic status and lifestyle. The model reached R^2 value of over 0.9807 that confirms the excellent level of predictability achieved in this study.

6 Declarations

Financial Support: None Competing Interests: The authors declare that they have no competing interests. Ethical Issues: Not applicable Data availability: The data used in this paper are artificially generated. Author Contribution: All authors were involved in the conception, design, and execution of this research.

References

1. J. Boyer, A. Bartoli and . P. Deharo, "Feasibility of Non-Invasive Coronary Artery Disease Screening," 2023.
2. M. C. Langenbach, I. L. Langenbach, B. Foldyna and V. Mauri, "Advanced CT measures of coronary artery disease with intermediate stenosis in patients with severe aortic valve stenosis".
3. C. M. Williams, R. J. Weir-McCall and A. L. Baldassarre, "Artificial Intelligence and Machine Learning for Cardiovascular Computed Tomography (CCT): A White Paper of the Society of Cardiovascular Computed Tomography (SCCT)," *Journal of Cardiovascular Computed Tomography*, 2024.
4. J. Liao, L. Huang, M. Qu, B. Chen and G. Wang, "Artificial Intelligence in Coronary CT Angiography: Current Status and Future Prospects," 2022.
5. Y. Ma, M. Li and H. Wu, "The Machine Learning Models in Major Cardiovascular Adverse Events Prediction Based on Coronary Computed Tomography Angiography: Systematic Review," *JOURNAL OF MEDICAL INTERNET RESEARCH* , 2025.

6. J. Petch, J. P. T. Bortesi, T. Sheth and M. Natarajan, "Coronary Computed Tomographic Angiography to Optimize the Diagnostic Yield of Invasive Angiography for Low-Risk Patients Screened With Artificial Intelligence: Protocol for the CarDIA-AI Randomized Controlled Trial," 2025.
7. H. J. H. Dreesen, C. Stroszczynski and M. M. Lell, "Optimizing Coronary Computed Tomography Angiography Using a Novel Deep Learning Based Algorithm," *Journal of Imaging Informatics in Medicine*, 2024.
8. P. Shrivastava, S. Kashikar, H. P. Parihar, P. Kasat, P. Bhangale and P. Shrivastava, "A systematic review on deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction," *European Journal of Radiology Open*, 2025.
9. J. Chen, S. Song and R. Zhuo , "Prediction model based on artificial intelligence for identifying risk of coronary atherosclerotic heart disease in computed tomography," *Journal of Radiation Research and Applied Sciences* , 2024.
10. J. Lu, M. Bennamoun, J. Stewart, J. K. Eshraghian, Y. Liu, B. Chow, M. F. Sanfilippo and . G. Dwivedi, "Multitask Deep Learning for Accurate Risk Stratification and Prediction of Next Steps for Coronary CT Angiography Patients," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, 2023.
11. X. Ma, G. Luo, W. Wang and K. Wang, "Transformer Network for Significant Stenosis Detection in CCTA of Coronary Arteries," 2021.
12. "Centers for Disease Control and Prevention (CDC), Preventing Heart Disease," [Online]. Available: <https://www.cdc.gov/heart-disease/prevention/index.html>.
13. "National Heart, Lung, and Blood Institute (NHLBI/NIH), Heart-Healthy Living: Healthy Weight," [Online]. Available: <https://www.nhlbi.nih.gov/health/heart-healthy-living/healthy-weight>.
14. "CDC, Diabetes and Your Heart," [Online]. Available: <https://www.cdc.gov/diabetes/diabetes-complications/diabetes-and-your-heart.html>.
15. "CDC, Sleep About Sleep," [Online]. Available: https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html.
16. "CDC, National Diabetes Statistics Report," [Online]. Available: <https://www.cdc.gov/diabetes/data/statistics-report/index.html>.
17. "American College of Cardiology (ACC)/AHA, New High Blood Pressure Guideline," [Online]. Available: <https://www.acc.org/About-ACC/Press-Releases/2025/08/13/20/03/New-high-blood-pressure-guideline-emphasizes-prevention-early-treatment-to-reduce-CVD-risk>.
18. "CDC, Physical Activity Data," [Online]. Available: <https://www.cdc.gov/physicalactivity/data/index.html>.
19. "CDC, Blood Pressure Basics," [Online]. Available: <https://www.cdc.gov/bloodpressure/>.
20. "AHA, Reduce Your Cardiovascular Risk," [Online]. Available: <https://www.heart.org/en/health-topics/reduce-your-risk>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

