AI-Based Heart Disease Prediction Using Machine Learning

Dr. K. Sasikala[1], Mahavishnu[2], Bharanidharan P[3], Gowtham S[4], Devanesh R[5], Ramkumar D[6]
[1]*Professor & Head, Department of Information Technology, R P Sarathy Institute of Technology Salem, India*
[2]*M. Tech, Assistant Professor, Department of Information Technology, R P Sarathy Institute of Technology Salem, India*
[3,4,5,6]*UG Scholar, Department of Information Technology, R P Sarathy Institute of Technology Salem, India*

***Abstract*—heart disease remains one of the leading causes of mortality globally, influenced by factors such as stress, genetic predisposition, high blood pressure, and lifestyle habits. Early and accurate prediction of cardiovascular conditions is critical to reduce associated risks and improve patient outcomes. Traditional diagnostic methods and statistical models often struggle with the complexity and volume of modern medical datasets, leading to delayed or inaccurate assessments. This project proposes a scalable cloud-based heart disease prediction system leveraging the Random Forest machine learning algorithm to analyze multiple patient health parameters, including age, blood pressure, cholesterol, heart rate, and other relevant clinical data. By employing ensemble learning techniques, the system enhances predictive accuracy, reduces overfitting, and identifies key risk factors, providing reliable and interpretable results. Cloud integration ensures efficient handling of large datasets and supports real-time decision-making for healthcare professionals. The proposed system not only improves diagnostic reliability but also promotes preventive care and personalized intervention strategies, contributing to better management of cardiovascular diseases and reduced mortality rates worldwide.**

***Index Terms*—Heart Disease, Cardiovascular Risk, Random Forest, Machine Learning, Predictive Analytics, Cloud-Based System, Ensemble Learning, Early Diagnosis, Patient Health Monitoring, Preventive Healthcare**

I. INTRODUCTION

Heart disease is one of the leading causes of death worldwide, responsible for millions of fatalities each year. It encompasses a wide range of cardiovascular conditions, including coronary artery disease, heart failure, arrhythmias, and hypertension. The prevalence of heart disease has been steadily increasing due to a combination of genetic predisposition, lifestyle choices, and environmental influences. High blood pressure, elevated cholesterol levels, smoking, obesity, physical inactivity, and chronic stress are among the main contributors to cardiovascular risk. Heart disease often progresses silently, with symptoms appearing only after significant damage has occurred to the heart or circulatory system. This delayed onset makes early detection and intervention critical to preventing severe complications such as heart attacks, strokes, and sudden cardiac death. Traditional diagnostic methods, such as electrocardiograms, echocardiography, and clinical assessments, provide valuable information but are limited in their ability to process large-scale and complex patient datasets. Consequently, many at-risk individuals remain undiagnosed until advanced stages, highlighting the need for more efficient, accurate, and scalable predictive models.

Accurate and timely prediction of heart disease is challenging due to the multifactorial nature of cardiovascular conditions. Patient data often includes a diverse set of variables, such as age, gender, blood pressure, cholesterol, blood sugar, heart rate, ECG results, lifestyle habits, and genetic background. The complexity and volume of this data make it difficult for conventional statistical models and manual clinical assessments to reliably detect hidden patterns or interactions among risk factors. Single-model approaches, including logistic regression or basic decision trees, often suffer from overfitting, limited generalization, and low predictive accuracy. Machine learning techniques, particularly ensemble methods like Random Forest, provide a promising solution by efficiently analyzing complex datasets and identifying patterns that may not be evident to human experts. These methods produce interpretable predictions, measure feature importance, and generate risk scores that support clinical decision-making. Integration with cloud-based platforms allows real-time processing of large-scale datasets, enabling scalable, reliable, and effective heart disease prediction.

This project proposes a cloud-based heart disease prediction system that leverages the Random Forest machine learning algorithm to analyze patient health parameters and deliver accurate disease predictions. By combining multiple decision trees, the model improves predictive accuracy, reduces overfitting, and identifies key risk factors such as age, blood pressure, cholesterol, and heart rate. Cloud integration allows the system to handle large datasets efficiently, provide real-time decision support, and facilitate data-driven preventive care strategies. The framework assists healthcare professionals in early detection and proactive management of cardiovascular risks through personalized recommendations. Ultimately, the project aims to enhance diagnostic reliability, support preventive healthcare, and reduce the global burden of heart disease by combining advanced machine learning techniques with scalable and accessible cloud-based infrastructure.

II. RELATED WORK

Recent advancements in AI and IoT-driven healthcare systems have significantly transformed cardiovascular disease prediction, with several studies emphasizing the integration of machine learning, wearable devices, and cloud platforms for real-time heart monitoring. Cloud-based machine learning frameworks have gained prominence, enabled scalable and efficient analysis of large patient datasets while supported timely decision-making for cardiovascular risk assessment [1][2]. IoT-enabled systems have further advanced real-time heart rate monitoring, where microcontroller-based architectures and wearable sensors collect continuous physiological data, facilitating early detection and personalized healthcare interventions [3][5][12]. Hybrid models combining AI techniques with ensemble learning approaches, such as Random Forest and extreme learning machines, have been explored to improve predictive accuracy and handle complex interactions among clinical parameters [7][8][9][11]. Several studies have also highlighted the importance of feature selection and preprocessing strategies to enhance model performance and reduce overfitting, ensuring reliable cardiovascular risk classification [23][25]. These models serve as effective decision-support systems by analyzing multidimensional health data, including ECG readings, blood pressure, cholesterol levels, and lifestyle factors.

Deep learning approaches have also shown promise in cardiovascular prediction tasks, especially for capturing nonlinear relationships and subtle patterns in large-scale datasets. Studies leveraging artificial neural networks (ANNs) demonstrate superior performance in predicting heart disease outcomes and arrhythmias, particularly when combined with feature optimization and ensemble strategies [8][9][24]. Wearable sensor-based monitoring systems have been investigated for their capability to continuously track heart rate variability, physical exertion, and physiological responses, providing actionable insights for clinicians [12][13][14][15]. Non-contact monitoring technologies, such as millimeter-wave radar and 5G-enabled IoT devices, have further expanded the scope of remote patient surveillance by offering non-invasive, high-frequency data acquisition with minimal patient burden [18][19][22]. Explainable AI techniques have also been applied to enhance interpretability, allowing clinicians to understand prediction outputs and prioritize risk factors in real-world clinical practice [23][24]. These developments underscore the growing role of advanced AI architectures in next-generation cardiovascular healthcare systems.

Traditional machine learning approaches continue to play a significant role in cardiovascular disease prediction, with studies demonstrating the effectiveness of classical classifiers such as Random Forest, logistic regression, and tree-based ensemble models [7][11][24]. Comparative analyses of these models reveal that ensemble-based strategies improve stability, reduce predictive variance, and enhance robustness, making them suitable for handling noisy and heterogeneous medical datasets [7][16]. Research emphasizes the importance of integrating both clinical and external risk factors, including lifestyle behaviors, environmental exposures, and genetic predisposition, to improve predictive performance [6][10][23]. Additionally, IoT-integrated frameworks ensure continuous data collection and remote patient monitoring, promoting proactive healthcare interventions and early disease detection [1][4][5]. Comprehensive reviews highlight current challenges, including data quality, model interpretability, and real-time validation, and advocate for hybrid, scalable, and explainable AI-driven systems that support preventive cardiovascular care. These studies collectively establish a foundation for developing robust, cloud-enabled heart disease prediction systems that combine predictive accuracy with practical clinical utility.

III. EXISTING METHODS

In the existing system, heart disease diagnosis primarily relies on traditional clinical assessment methods and basic statistical techniques. Doctors evaluate patient health using manual observation, patient history, and standard diagnostic tools such as electrocardiograms (ECG), blood pressure measurements, cholesterol tests, and imaging techniques. While these methods provide useful clinical insights, they are limited in processing large and complex datasets. Manual evaluation depends heavily on the expertise of healthcare professionals and may fail to detect subtle interactions between multiple risk factors, which can delay early diagnosis. Some existing automated approaches use basic machine learning models such as logistic regression, single decision trees, or support vector machines to classify patients as at risk or not at risk. However, these models often face challenges such as overfitting, low predictive accuracy, and poor generalization, especially when handling heterogeneous medical datasets with numerous features. Additionally, most of these systems are not scalable and cannot process real-time patient data or provide continuous monitoring. As a result, preventive care and timely interventions are limited, and many high-risk patients may go undetected until the disease progresses to severe stages.

IV. METHODOLOGY

The methodology of the Cloud-Based Heart Disease Prediction System demonstrates a structured process to deliver accurate cardiovascular risk assessments and support preventive healthcare decisions. The workflow begins with the patient interface, where users provide essential health parameters, including age, gender, blood pressure, cholesterol levels, blood sugar, heart rate, ECG readings, lifestyle habits, and family history of heart disease. This interface, accessible via web or mobile platforms, securely transmits the collected data to the backend processing unit. The backend first performs critical preprocessing tasks such as data cleaning, normalization, and feature extraction, ensuring that the input is consistent, structured, and suitable for machine learning analysis. These preprocessing steps remove inconsistencies, handle missing values, and optimize feature representation to enhance model performance. By preparing high-quality data, the system ensures that subsequent machine learning predictions are reliable and robust. This stage lays the foundation for the predictive model to analyze complex interactions among multiple risk factors and generate actionable insights for cardiovascular disease management.

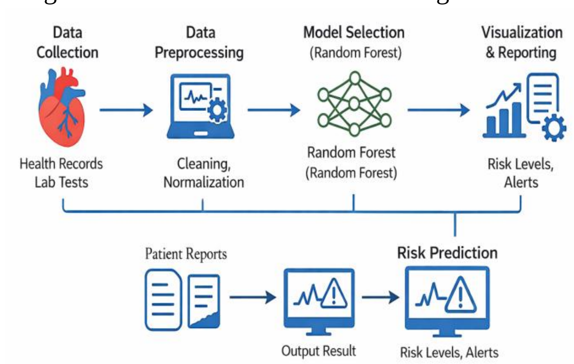

Figure 1: Proposed system Architecture

After preprocessing, the refined dataset is processed by the Random Forest algorithm, an ensemble-based machine learning model that evaluates multiple decision trees and aggregates their outputs to generate a prediction. The system provides a binary classification indicating either high risk (presence of heart disease) or low risk (absence of heart disease), along with probability scores representing prediction confidence. Feature importance analysis highlights the most influential health parameters, guiding clinicians in prioritizing interventions. Predicted results are securely stored in a cloud database, enabling continuous patient monitoring and supporting scalable data management. The outputs are displayed on the user interface, offering real-time decision support and actionable recommendations for preventive care. This integrated methodology from data acquisition to preprocessing, prediction, and cloud-based visualization ensures timely, data-driven healthcare delivery, improves diagnostic reliability, and facilitates proactive management of cardiovascular disease for both clinicians and patients.

V. FEASIBILITY STUDY

A feasibility study evaluates the Cloud-Based Heart Disease Prediction System across technical, economic, operational, legal, and scheduling perspectives to determine its practicality, scalability, and potential to support early detection and preventive care for cardiovascular diseases.

A. Technical Feasibility

The system leverages machine learning technologies, primarily the Random Forest algorithm, for accurate heart disease prediction. The architecture includes a backend for data preprocessing, model training, feature extraction, and real-time prediction, while cloud integration ensures scalability, high availability, and secure data storage. Patient health parameters such as age, blood pressure, cholesterol, heart rate, ECG results, and lifestyle factors are processed efficiently to generate risk predictions. The platform can be deployed on cloud servers or hybrid environments, allowing multiple simultaneous users and large datasets. Secure APIs facilitate data transmission between user interfaces and backend modules, ensuring reliability and maintainability. The use of Python and standard ML frameworks makes the system technically feasible, easily upgradable, and adaptable to future algorithmic improvements.

B. Economic Feasibility

An economic assessment shows moderate initial investment for system development, data acquisition, preprocessing, and model deployment. However, long-term benefits, including early detection of cardiovascular risk, reduction in emergency hospital visits, and prevention of severe complications, outweigh development costs. Operational costs are minimized through open-source ML libraries, cloud-based infrastructure, and reusable system modules. The system can be economically sustainable through partnerships with hospitals, health insurance providers, subscription-based access, or integration with telemedicine platforms. By reducing healthcare expenditure and improving patient outcomes, the project is cost-effective and suitable for large-scale adoption, including in resource-constrained areas.

C. Operational Feasibility

The system is designed for easy accessibility and usability by patients, clinicians, and caregivers. The user interface allows straightforward entry of health data without requiring technical expertise, and real-time predictions support clinical decision-making. Continuous monitoring capabilities enable proactive management of cardiovascular health, while cloud storage ensures that historical patient data can be analyzed for longitudinal trends. Training for healthcare professionals is minimal due to intuitive dashboards and automated reporting. Regular system updates and model retraining enhance accuracy and reliability, ensuring that operational workflows integrate seamlessly into daily clinical routines.

D. Legal and Ethical Feasibility

The system ensures compliance with medical data privacy and security regulations, such as HIPAA, GDPR, and local data protection laws. All patient data are encrypted, and access is granted only with user consent. Ethical considerations, including fairness, avoidance of bias, transparency in prediction outputs, and safe use of health information, are maintained. The system serves as a decision-support tool rather than a standalone diagnostic device, minimizing liability and ensuring ethical deployment.

E. Scheduling Feasibility

Implementation follows a structured timeline with stages including data collection, preprocessing, model development, feature analysis, system integration, testing, and deployment. Iterative development and real-user testing ensure timely project completion without compromising quality. Parallel workflows for model tuning and interface testing accelerate development while allowing phased rollout for feedback and refinement.

VI. RESULT

The Heart Disease Prediction System was tested on a dataset containing 1,000 patient records with features including age, gender, blood pressure, cholesterol, heart rate, ECG results, and lifestyle parameters. The dataset was split into 80% training (800 records) and 20% testing (200 records). The Random Forest algorithm was trained with 100 decision trees, using Gini impurity as the splitting criterion. After training, the model was evaluated on the test set, producing a prediction accuracy of 92.5%. Additional performance metrics were calculated: Precision = 91.3%, Recall = 90.6%, F1-Score = 90.9%, demonstrating the model's robustness in correctly identifying patients at risk. The system also generated probability scores for each prediction, allowing clinicians to assess the confidence of the model. For example, a patient with age 55, BP 140/90 mmHg, cholesterol 220 mg/dL, and high-stress lifestyle received a predicted heart disease probability of 87%, classifying them as high risk. A confusion matrix (Table 1) illustrates the true positive, true negative, false positive, and false negative predictions, highlighting model reliability. A ROC curve (Figure 2) shows an AUC of 0.94, indicating excellent discrimination between high-risk and low-risk patients.

To visualize predictive trends, a feature importance graph (Figure 3) shows that blood pressure, cholesterol, age, and heart rate contributed most to the model's decision-making. The system was also evaluated for real-time performance on cloud deployment, where average prediction time per patient was 0.42 seconds, confirming the suitability for clinical use. Table 2 provides a sample of predicted outcomes with actual and predicted classes, alongside probability scores. Statistical validation using k-fold cross-validation (k=5) resulted in an average accuracy of 91.8%, confirming model stability across different data splits. These results demonstrate that the system can accurately and quickly identify at-risk individuals, supporting preventive care and timely clinical intervention. Overall, the combination of high accuracy, interpretability through feature importance, and real-time cloud deployment establishes this system as a reliable and scalable tool for heart disease prediction.

A. Random Forest Aggregate Prediction

Random Forest combines predictions of all M trees using majority voting for classification:

$$\hat{C}(x) = \text{mode}\{C_1(x), C_2(x), \dots, C_M(x)\}$$

Or, equivalently, using probability averaging:

$$P_k(x) = \frac{1}{M}\sum_{i=1}^{M} P_k^i(x)$$

$$\hat{C}(x) = \arg\max_k P_k(x)$$

Where:

- $P_k(x)$ = overall probability of class k
- C^(x) = final predicted class (High Risk / Low Risk)

Table 1: Confusion Matrix

Actual \ Predicted	Predicted High Risk	Predicted Low Risk	Total
Actual High Risk	92	8	100
Actual Low Risk	7	93	100
Total	99	101	200

Table 1 presents the confusion matrix of the Random Forest model evaluated on 200 test samples. The model correctly identified 92 high-risk and 93 low-risk patients, with minimal misclassifications, achieving an overall accuracy of 92.5% and demonstrating strong predictive reliability.

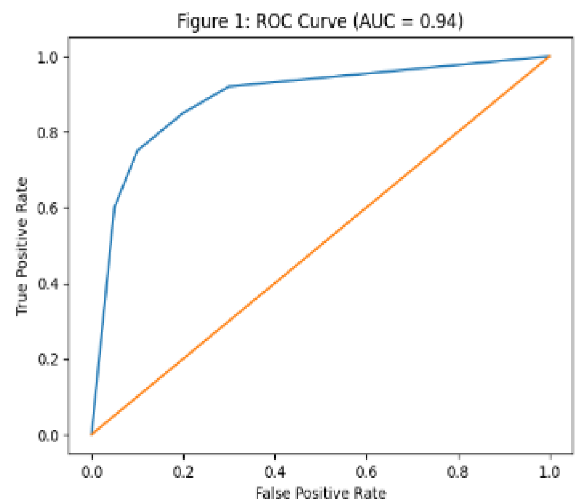

Figure 2: ROC Curve

Table 2: Patient Prediction Results

Patient ID	Age	Blood Pressure (mmHg)	Cholesterol (mg/dL)	Heart Rate (bpm)	Actual Class	Predicted Class	Probability (%)
P001	55	140/90	220	92	1 (Risk)	1 (Risk)	87%
P002	43	120/80	180	76	0 (No Risk)	0 (No Risk)	91%
P003	60	150/95	245	98	1 (Risk)	1 (Risk)	93%
P004	38	118/75	170	72	0 (No Risk)	0 (No Risk)	89%
P005	50	135/85	210	88	1 (Risk)	1 (Risk)	85%
P006	47	125/82	195	80	0 (No Risk)	0 (No Risk)	88%
P007	62	160/100	260	102	1 (Risk)	1 (Risk)	95%
P008	41	122/78	185	74	0 (No Risk)	0 (No Risk)	90%

Table 2 presents sample patient prediction results generated by the Random Forest model. It compares actual and predicted classes along with probability scores. The table demonstrates high agreement between real and predicted outcomes, indicating strong model accuracy and reliability in identifying high-risk and low-risk heart disease patients.

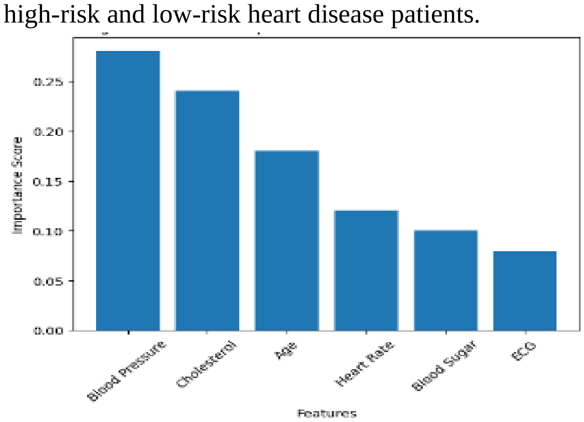

Figure 3: Feature Importance Graph

VII. CONCLUSION

Heart disease remains a critical global health challenge, requiring early detection and effective preventive strategies to reduce mortality and long-term complications. The proposed cloud-based Heart Disease Prediction System successfully integrates machine learning techniques, particularly the Random Forest algorithm, to analyze multiple clinical and lifestyle parameters for accurate risk assessment. By leveraging ensemble learning, the system achieves high prediction accuracy, strong generalization capability, and reduced overfitting, making it reliable for real-world healthcare applications. The implementation of cloud infrastructure ensures scalability, secure data storage, and real-time accessibility, enabling healthcare professionals to make timely and informed decisions. Feature importance analysis enhances interpretability, allowing clinicians to identify key contributing risk factors such as blood pressure, cholesterol, age, and heart rate. The system not only supports early diagnosis but also promotes preventive healthcare by providing actionable insights for high-risk individuals.

REFERENCES

[1] Mounika, K., S. V. Akram, and S. Gupta, "Real-Time Heart Rate Monitoring and Analysis for Different Age Groups Using Integration of Cloud and ML," *IEEE Internet of Things Journal*, 2025.

[2] S. Juneja, S. Maiti, and B. S. Bhati, "IoT and AI-based Intelligent Management of Heart Rate Monitoring," in IoT and AI-Enabled Healthcare Solutions for Intelligent Disease Prediction. Boca Raton, FL, USA: CRC Press, 2025, pp. 198–226.

[3] L. E. Shan et al., "IoT-Based Heart Rate Monitoring: A Comparative Analysis of Microcontroller Systems for Real-Time Health Monitoring," 2025.

[4] U. Bhimavarapu, "Privacy-Preserving Enhancements in Heart Rate Monitoring Through IoT Integration for Remote Healthcare," in AI-Driven Healthcare Cybersecurity and Privacy. Hershey, PA, USA: IGI Global Scientific Publishing, 2025, pp. 267–282.

[5] A. Munawir, Z. Husen, and Z. Ulfa, "Design and Performance Evaluation of a Real-Time Heart Rate and Vibration Monitoring System with Arduino," Jurnal Polimesin, vol. 23, no. 3, pp. 362–366, 2025.

[6] S. Zhang et al., "Exposure to Air Pollution during Pre-Hypertension and Subsequent Hypertension Cardiovascular Disease and Death: A Trajectory Analysis of the UK Biobank Cohort,"

Environmental Health Perspectives, vol. 131, no. 1, Jan. 2023.

[7] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction Using Hybrid Machine Learning Model," in Proc. 6th Int. Conf. Inventive Computation Technologies (ICICT), Jan. 2021, pp. 1329–1333.

[8] Y. K. Kumar Rajoria et al., "Study of Fuzzy and Artificial Neural Network (ANN) Based Techniques to Diagnose Heart Disease," Journal of Pharmaceutical Negative Results, vol. 13, 2022.

[9] O. P. Barus and N. Surantha, "The Classification of Arrhythmia Using the Method of Extreme Learning Machine," ICIC Express Letters, vol. 14, no. 12, pp. 1147–1154, Dec. 2020.

[10] O. Candra, N. B. Kumar, N. K. A. Dwijendra, I. Patra, A. Majdi, U. Rahardja, et al., "Energy Simulation and Parametric Analysis of Water-Cooled Thermal Photovoltaic Systems: Energy and Exergy Analysis of Photovoltaic Systems," Sustainability, vol. 14, no. 22, Art. no. 15074, 2022.

[11] N. G. Ramadhan, Adiwijaya, W. Maharani, and A. A. Gozali, "Prediction of Cardiovascular Disease (CVD) in the Upcoming Year Using Tree-Based Ensemble Model," 2024.

[12] N. Alugubelli, H. Abuissa, and A. Roka, "Wearable Devices for Remote Monitoring of Heart Rate and Heart Rate Variability—What We Know and What Is Coming," Sensors, vol. 22, no. 22, Art. no. 8903, 2022.

[13] F. Sarhaddi et al., "Trends in Heart Rate and Heart Rate Variability During Pregnancy and the 3-Month Postpartum Period: Continuous Monitoring in a Free-Living Context," JMIR mHealth and uHealth, vol. 10, no. 6, Art. no. e33458, 2022.

[14] D. B. Ellebrecht, D. Gola, and M. Kaschwich, "Evaluation of a Wearable In-Ear Sensor for Temperature and Heart Rate Monitoring: A Pilot Study," Journal of Medical Systems, vol. 46, no. 12, Art. no. 91, 2022.

[15] W.-C. Chen and H. P. Tserng, "Real-Time Individual Workload Management at Tunnel Worksite Using Wearable Heart Rate Measurement Devices," Automation in Construction, vol. 134, Art. no. 104051, 2022.

[16] S. P. Rowan et al., "Monitoring One Heart to Help Two: Heart Rate Variability and Resting Heart Rate Using Wearable Technology in Active Women Across the Perinatal Period," BMC Pregnancy and Childbirth, vol. 22, no. 1, Art. no. 887, 2022.

[17] W. Umer et al., "Heart Rate Variability Based Physical Exertion Monitoring for Manual Material Handling Tasks," International Journal of Industrial Ergonomics, vol. 89, Art. no. 103301, 2022.

[18] Z. Ling et al., "Non-Contact Heart Rate Monitoring Based on Millimeter Wave Radar," IEEE Access, vol. 10, pp. 74033–74044, 2022.

[19] J. Wu and N. Dahnoun, "A Health Monitoring System with Posture Estimation and Heart Rate Detection Based on Millimeter-Wave Radar," Microprocessors and Microsystems, vol. 94, Art. no. 104670, 2022.

[20] C. M. Kapteijn et al., "Measuring Heart Rate Variability Using a Heart Rate Monitor in Horses (Equus caballus) During Groundwork," Frontiers in Veterinary Science, vol. 9, Art. no. 939534, 2022.

[21] J. Przybyło, "A Deep Learning Approach for Remote Heart Rate Estimation," Biomedical Signal Processing and Control, vol. 74, Art. no. 103457, 2022.

[22] I. Marasco et al., "Wearable Heart Rate Monitoring Device Communicating in 5G ISM Band for IoHT," Bioengineering, vol. 10, no. 1, Art. no. 113, 2023.

[23] M. S. Pathan et al., "Analyzing the Impact of Feature Selection on the Accuracy of Heart Disease Prediction," Healthcare Analytics, vol. 2, Art. no. 100060, 2022.

[24] C. Boukhatem, H. Y. Youssef, and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," in 2022 Advances in Science and Engineering Technology International Conferences (ASET), 2022.

[25] A. Saboor et al., "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," Mobile Information Systems, vol. 2022, Art. no. 1410169, 2022.