Heartcare: Predictive Analytics for Early Detection and Prevention

Daniyal Rosli[1*], Zanariah Idrus[2], Mazura Mat Din[3]

[1,2,3]*College of Computing, Informatics & Mathematics, Universiti Teknologi Mara Kedah Branch, Malaysia*

Article Information

Received: 21-11-2024
Revised: 28-11-2024
Published: 05-12-2024

Keywords
Cardiovascular disease; predictive analytics; machine learning

****Correspondence Email:***
2023375627@student.uitm.edu.my

Abstract

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, often due to late detection and prevention. Heartcare aims to leverage predictive analytics to facilitate early detection and prevention of heart diseases. By integrating machine learning algorithms such as Decision Trees, Random Forests, and Logistic Regression, Heartcare provides healthcare professionals with a powerful tool for patient health monitoring. This study focuses on developing a predictive model to assess heart disease risk using patient-specific data, such as age, sex, BMI, and lifestyle factors. The outcomes will enable healthcare professionals to make informed decisions, potentially saving lives and reducing healthcare costs.

1. Introduction

Heart disease remains a significant global health concern, accounting for approximately 31% of all deaths worldwide (Vaduganathan, 2022). Despite the advancements in medical technology and treatment options, the prevalence of heart disease continues to rise, particularly in low and middle-income countries (Qureshi,2021). Early detection and effective management of heart disease are crucial for improving patient outcomes and reducing the burden on healthcare systems (F Ali, 2020).

Traditional risk assessment tools, such as the Framingham Risk Score, have been widely used to estimate an individual's risk of developing cardiovascular disease. However, these tools often rely on a limited set of risk factors and may not capture the complex interplay of various physiological, genetic, and lifestyle factors that contribute to heart disease development (JH Law, 2023).

Recent advancements in machine learning and artificial intelligence have opened new avenues for developing more accurate and personalized prediction models. These models can integrate diverse data sources, including electronic health records, genetic information, and real-time monitoring data from wearable devices, to provide more comprehensive risk assessments (Udegbe, 2023). Several studies have demonstrated the potential of machine learning algorithms in predicting heart disease risk. For instance, (Li, 2024) developed a machine-learning model using electronic health records that outperformed traditional clinical risk scores in predicting cardiovascular events. Similarly,(D'Ancona, 2023), showed that deep learning models could accurately predict the risk of coronary artery disease using cardiac CT images.

Despite these advancements, challenges remain in developing and implementing heart disease prediction models. These include ensuring model interpretability, addressing potential biases in training data, and integrating predictive tools seamlessly into clinical workflows (Prabhod, 2023). Additionally, there is a need for models that can provide continuous risk assessment and early warnings of potential cardiac events, leveraging data from wearable devices and remote monitoring systems (Hughes, 2023).

This project aims to address these challenges by developing a comprehensive heart disease prediction model that incorporates diverse data sources, utilizes advanced machine learning techniques, and presents results through an intuitive web application. By doing so, it seeks to contribute to the ongoing efforts to improve the early detection and management of heart disease, ultimately reducing its global impact on public health.

1.1 Literature Review

Heart disease, also known as cardiovascular disease, encompasses a range of conditions affecting the heart and blood vessels, including coronary artery disease, strokes, heart attacks, and arrhythmias (Vagare, 2024). It remains the leading cause of death globally, responsible for approximately 17.9 million deaths annually, accounting for 32% of all deaths worldwide (Gaziono, 2022). Major risk factors include hypertension, high cholesterol, smoking, obesity, physical inactivity, and diabetes, along with age, gender, and genetics. Early detection and intervention are essential to mitigate severe health outcomes, yet many cases go undetected until they reach advanced stages. Predictive tools have emerged as crucial in identifying risks early by assessing factors like cholesterol levels, blood pressure, and lifestyle habits, allowing for timely and personalized management (Ullah, 2023).

The rise of predictive analytics in healthcare has really empowered the sector to adopt transformational ways of preventing and managing diseases. Predictive analytics makes use of machine learning techniques to forecast health events, such as outbreaks of diseases, patient readmissions, and chronic disease progressions, by using historical and real-time data (PM Nerkar, 2023). This data-driven approach not only improves patient care but also reduces costs and enhances operational efficiency. For heart disease, predictive models analyze data from patient medical histories, genetics, and lifestyle factors to estimate risks, offering healthcare providers tailored strategies for prevention (Freitas, 2023). Algorithms like Logistic Regression, Random Forest, and Neural Networks have proven effective in capturing complex patterns, with the potential to identify at-risk individuals and enable proactive interventions (H Yang, 2023).

Despite advancements, traditional tools like the Framingham Risk Score, which estimate 10-year cardiovascular risk based on linear relationships, often fail to capture the complexity of heart disease (Gray, 2023). Recent models, including Decision Trees and Random Forests, address these limitations by uncovering non-linear relationships and improving prediction accuracy. Neural networks learn the subtlety of the pattern in the data but often require large data and substantial computational resources (Khan, 2023). Machine learning-based research on the Cleveland Heart Disease Dataset and the Framingham Heart Study promises more accurate, scalable, and pragmatic solutions for real-world application, as evident from these studies. However, several challenges remain, such as interpretability and clinical integration.

High-quality datasets and advanced algorithms are critical to improving predictive accuracy and practical application. Data preprocessing, such as cleaning and feature selection, ensures that models deliver reliable insights (Pan, 2023). While methods like Random Forests and Neural Networks have shown significant promise, the choice of algorithm often depends on factors like dataset size, feature complexity, and the trade-off between accuracy and interpretability (Misra, 2023). Models must be continuously evaluated using metrics like accuracy, precision, recall, and F1-score to optimize performance. Incorporating real-time monitoring data from wearable devices and electronic health records could further refine these models, enabling continuous risk assessment and early intervention, ultimately improving outcomes in heart disease management (Hughes, 2023).

2. Research Methods

The methodology outlines the necessary actions to achieve the goals of the project. Preliminary research, information gathering, data collection, data preparation, design and implementation, system development, and documentation are a few of these phases. Table 1 below shows the research questions and objectives for this project. This will serve as the basis for identifying the phase needed to obtain the methodology for this project.

Table 1. Research Questions and Objectives

Research Questions	**Research Objectives**
1. What variables and guidelines are essential for predicting heart disease using machine learning algorithms?	To identify the variables and guidelines used to predict heart disease
2. How can we design an effective machine-learning model that accurately predicts heart disease?	To design a machine learning model that can predict heart disease
3. What are the best metrics for evaluating heart disease prediction models, and the key requirements for developing an accessible prototype system?	To evaluate the performance of the machine learning model and develop an accessible prototype system

Figure 1 shows the research flow of the projects. The phases start with preliminary study, knowledge acquisition, data acquisition, data pre-processing, design prediction model, model training and testing, system design, system development, system development, system evaluation, and documentation

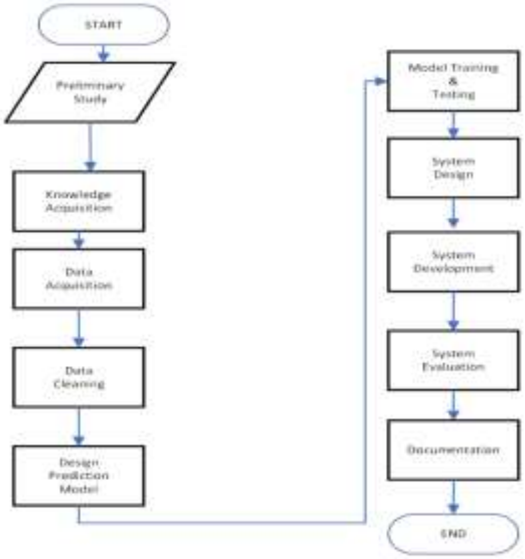
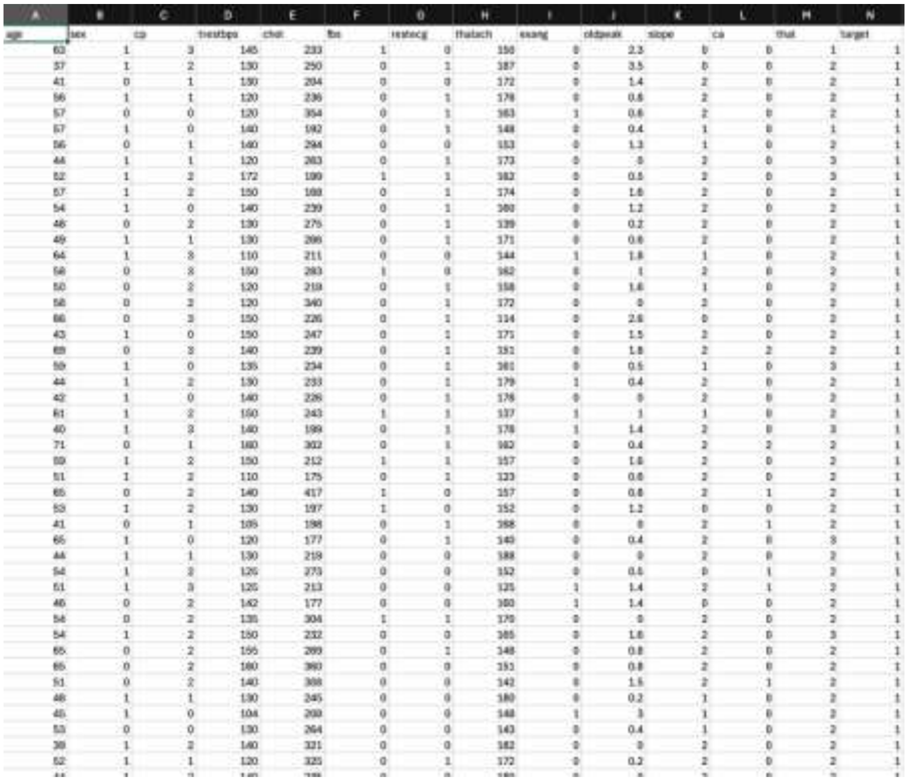


Fig. 1 Research Methodology

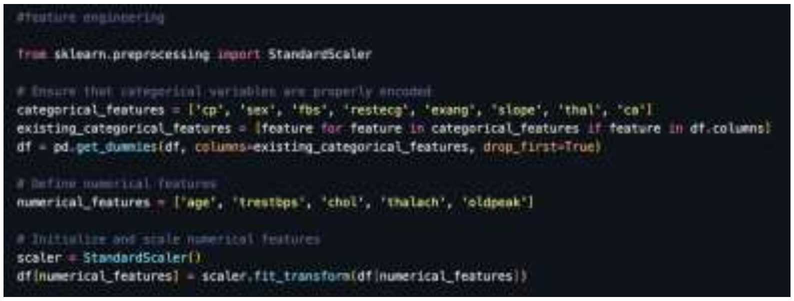
The research is structured, starting with data collection to develop a heart disease prediction system. In this regard, 304 anonymized patient records containing key health indicators such as age, sex, cholesterol levels, and chest pain type were sourced from reliable repositories like Kaggle and the UCI Machine Learning Repository. These datasets provide the foundation for training and evaluating machine learning models.



A	B	C	D	E	F	G	H	I	J	K	L	M	N
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1
40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
53	1	2	130	197	1	0	152	0	1.2	0	0	2	1
41	0	1	105	198	0	1	168	0	0	2	1	2	1
65	1	0	120	177	0	1	140	0	0.4	2	0	3	1
44	1	1	130	219	0	0	188	0	0	2	0	2	1
54	1	2	125	273	0	0	152	0	0.5	0	1	2	1
51	1	3	125	213	0	0	125	1	1.4	2	1	2	1
46	0	2	142	177	0	0	160	1	1.4	0	0	2	1
54	0	2	135	304	1	1	170	0	0	2	0	2	1
54	1	2	150	232	0	0	165	0	1.6	2	0	3	1
65	0	2	155	269	0	1	148	0	0.8	2	0	2	1
65	0	2	160	360	0	0	151	0	0.8	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
48	1	1	130	245	0	0	180	0	0.2	1	0	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
53	0	0	130	264	0	0	143	0	0.4	1	0	2	1
39	1	2	140	321	0	0	182	0	0	2	0	2	1
52	1	1	120	325	0	1	172	0	0.2	2	0	2	1

Fig. 2 Dataset of Heart Disease

Preprocessing of data was done by cleaning the data, handling missing values, normalizing continuous features, encoding categorical variables, and outlier detection. The dataset was split into training and testing to validate the performance of the models for accurate predictions. This step is important in letting machine learning algorithms learn effectively. Figure 3 shows a snippet of Python code used for data pre-processing.



```
#feature engineering

from sklearn.preprocessing import StandardScaler

# Ensure that categorical variables are properly encoded
categorical_features = ['cp', 'sex', 'fbs', 'restecg', 'exang', 'slope', 'thal', 'ca']
existing_categorical_features = [feature for feature in categorical_features if feature in df.columns]
df = pd.get_dummies(df, columns=existing_categorical_features, drop_first=True)

# Define numerical features
numerical_features = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']

# Initialize and scale numerical features
scaler = StandardScaler()
df[numerical_features] = scaler.fit_transform(df[numerical_features])
```

Fig. 3 Data pre-processing using Jupyter Notebook and Python

The algorithms used in model development to predict heart disease outcomes include Logistic Regression, Decision Trees, and Random Forests. The models were trained using the pre-processed dataset, with techniques such as cross-validation and hyperparameter tuning employed to optimize metrics, including accuracy, precision, recall, and F1-score. The best model selected was based on its evaluation results for deployment. As shown in Figure 4, a model comparison between 3 algorithms shows that Random Forest has the highest accuracy out of the 3 algorithms.

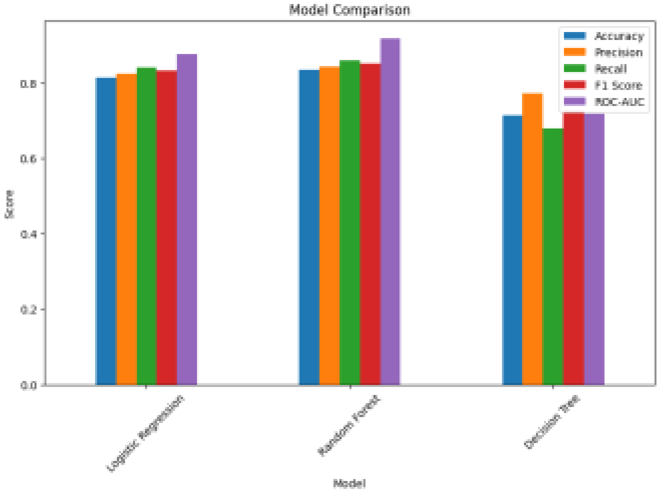


Fig. 4 Model comparison of the chosen algorithms

It includes the integration of the trained model into the web application, both for the front end with Next.js and the back end with Django, which allows the user to input health data for immediate predictions in a very user-friendly format. Much attention was paid to the security and privacy concerns regarding the processing of sensitive data. The system has been fully tested for the correctness of the predictions and for providing a seamless user experience. Figure 5 shows a web application design for Heartcare using Figma first before it is integrated in Next.js and Django.

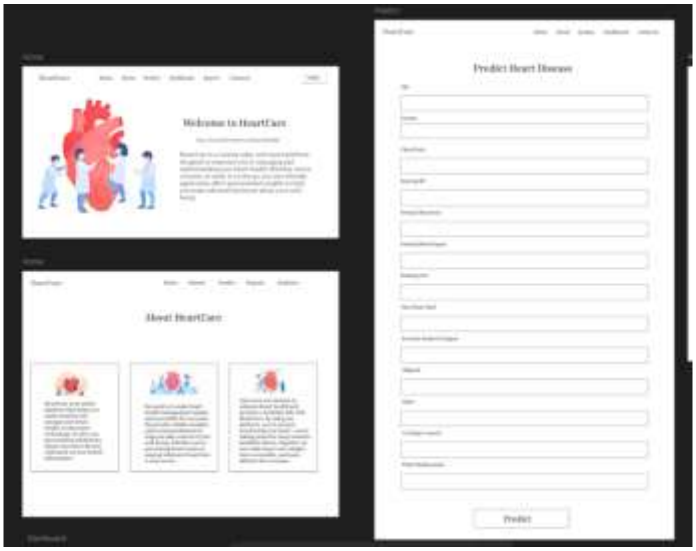


Fig. 5 Web application design for Heartcare

The final system evaluation measured the performance of both the model and the application using quantitative metrics, such as accuracy and qualitative user feedback. These evaluations led to iterative improvements in usability and accuracy. Thorough documentation was developed to support both end-users and developers, which will ensure clarity of understanding and support future updates.

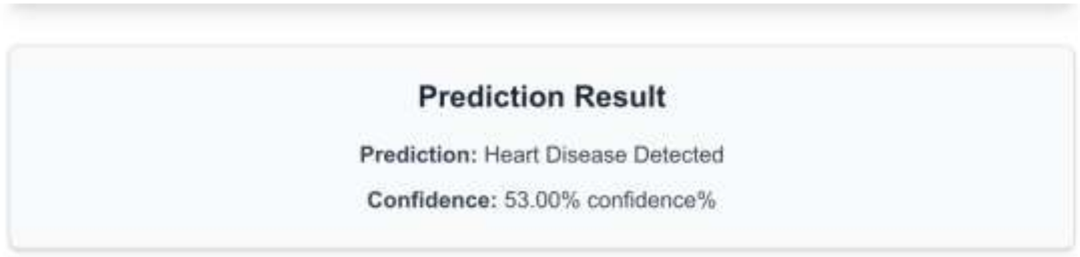
3. Result and Discussion

The HeartCare system was able to accomplish the main objective of predicting heart disease risk with good accuracy. Different machine learning models were trained, evaluated, and compared on a dataset of 304 records. Among all, the Random Forest algorithm was found to be the best-performing model with an accuracy of 83.52%, precision of 84.31%, recall of 86%, and F1-score of 85.15%. Other models, such as Logistic Regression and Decision Trees, have shown moderate performances of 81.31% and 71.42%, respectively. Below is a Table 2 to showcase the model's performance for Heartcare.

Table 2. Model Performance

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.813187	0.823529	0.84	0.831683	0.876585
Random Forest	0.835165	0.843137	0.86	0.851485	0.918537
Decision Tree	0.714286	0.772727	0.68	0.723404	0.718049

The web application of the system, built using Next.js for the front end and Django for the back end, integrates the trained machine learning model seamlessly. It allows users to input health data such as age, cholesterol levels, and chest pain type and get real-time predictions regarding the risk of heart disease. The prediction results indicate the risk involved, such as "Heart Disease Detected" or "No Heart Disease Detected," along with a confidence score for healthcare providers and patients to make appropriate decisions. Figure 6 shows the prediction result after inputting all the necessary details for Heartcare.



Prediction Result

Prediction: Heart Disease Detected

Confidence: 53.00% confidence%

Fig. 6 Prediction Result of Heartcare

4. Conclusions

HeartCare epitomizes how machine learning can be leveraged to enhance the diagnosis and treatment of heart ailments early. The system provides a user-friendly web application that can make very accurate and actionable predictions using a high-performing Random Forest model integrated into it. This approach facilitates timely interventions and enhances decision-making by healthcare providers. While the system shows a promising result, future development, such as the addition of real-time data from wearable devices and ensuring diverse dataset representation, will further enhance the reliability and applicability of the system. HeartCare therefore marks one of the positive steps toward using technology in improving health outcomes and tackling this global challenge in heart disease.

5. References

Vaduganathan, M., Mensah, G. A., Turco, J. V., Fuster, V., & Roth, G. A. (2022). The global burden of cardiovascular diseases and risk: a compass for future health. *Journal of the American College of Cardiology*, *80*(25), 2361-2371.

Qureshi, N. Q., Mufarrih, S. H., Bloomfield, G. S., Tariq, W., Almas, A., Mokdad, A. H., ... & Samad, Z. (2021). Disparities in cardiovascular research output and disease outcomes among high-, middle-and low-income countries–an analysis of global cardiovascular publications over the last decade (2008–2017). *Global Heart*, *16*(1).

Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208-222.

Law, J. H., Sultan, N., Finer, S., & Fudge, N. (2023). Advancing the communication of genetic risk for cardiometabolic diseases: a critical interpretive synthesis. *BMC medicine*, *21*(1), 432.

Udegbe, F. C., Nwankwo, E. I., Igwama, G. T., & Olaboye, J. A. (2023). Real-time data integration in diagnostic devices for predictive modeling of infectious disease outbreaks. *Computer Science & IT Research Journal*, *4*(3).

Li, C., Liu, X., Shen, P., Sun, Y., Zhou, T., Chen, W., ... & Gao, P. (2024). Improving cardiovascular risk prediction through machine learning modeling of irregularly repeated electronic health records. *European Heart Journal-Digital Health*, *5*(1), 30-40.

D'Ancona, G., Massussi, M., Savardi, M., Signoroni, A., Di Bacco, L., Farina, D., ... & Benussi, S. (2023). Deep learning to detect significant coronary artery disease from plain chest radiographs AI4CAD. *International Journal of Cardiology*, *370*, 435-441.

Prabhod, K. J. (2023). Integrating Large Language Models for Enhanced Clinical Decision Support Systems in Modern Healthcare. *Journal of Machine Learning for Healthcare Decision Support*, *3*(1), 18-62.

Hughes, A., Shandhi, M. M. H., Master, H., Dunn, J., & Brittain, E. (2023). Wearable devices in cardiovascular medicine. Circulation Research, 132(5), 652-670.

Vagare, R. D., Ubale, H. H., & Atar, A. A. (2024). CARDIOVASCULAR DISEASE-AN OVERVIEW.

Gaziano, T. A. (2022). Cardiovascular diseases worldwide. *Public Health Approach Cardiovasc. Dis. Prev. Manag*, *1*, 8-18.

Bevan, G., Pandey, A., Griggs, S., Dalton, J. E., Zidar, D., Patel, S., ... & Al-Kindi, S. (2023). Neighborhood-level social vulnerability and prevalence of cardiovascular risk factors and coronary heart disease. *Current problems in cardiology*, *48*(8), 101182.

Ullah, M., Hamayun, S., Wahab, A., Khan, S. U., Rehman, M. U., Haq, Z. U., ... & Naeem, M. (2023). Smart technologies are used as smart tools in the management of cardiovascular disease and their future perspective. *Current Problems in Cardiology*, *48*(11), 101922.

Islam, M. N., Raiyan, K. R., Mitra, S., Mannan, M. R., Tasnim, T., Putul, A. O., & Mandol, A. B. (2023). Prediction: an IoT and machine learning-based system to predict the risk level of cardiovascular diseases. *BMC Health Services Research*, *23*(1), 171.

Freitas, A. T. (2023). Data-Driven Approaches in Healthcare: Challenges and Emerging Trends. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, 65-80.

Yang, H., Luo, Y. M., Ma, C. Y., Zhang, T. Y., Zhou, T., Ren, X. L., ... & Lin, H. (2023). A gender-specific risk assessment of coronary heart disease based on physical examination data. *NPJ digital medicine*, *6*(1), 136.

Del Giorgio Solfa, F., & Simonato, F. R. (2023). Big Data Analytics in Healthcare: exploring the role of Machine Learning in Predicting patient outcomes and improving Healthcare Delivery. *International Journal of Computations, Information and Manufacturing (IJCIM)*, *3*.

Mann, A., Cleveland, B., Bumblauskas, D., & Kaparthi, S. (2024). Reducing Hospital Readmission Risk Using Predictive Analytics. *INFORMS Journal on Applied Analytics*.

MacKay, C., Klement, W., Vanberkel, P., Lamond, N., Urquhart, R., & Rigby, M. (2023). A framework for implementing machine learning in healthcare based on the concepts of preconditions and postconditions. *Healthcare Analytics*, *3*, 100155.

Ajegbile, M. D., Olaboye, J. A., Maha, C. C., & Tamunobarafiri, G. (2024). Integrating business analytics in healthcare: Enhancing patient outcomes through data-driven decision making.

Liu, W., Laranjo, L., Klimis, H., Chiang, J., Yue, J., Marschner, S., ... & Chow, C. K. (2023). Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. *European Heart Journal-Quality of Care and Clinical Outcomes*, *9*(4), 310-322.

Gray, M. P., Berman, Y., Bottà, G., Grieve, S. M., Ho, A., Hu, J., ... & Rogers, C. (2023). Incorporating a polygenic risk score-triaged coronary calcium score into cardiovascular disease examinations to identify subclinical coronary artery disease (ESCALATE): protocol for a prospective, nonrandomized implementation trial. *American Heart Journal*, *264*, 163-173.

Sadar, U., Agarwal, P., Parveen, S., Jain, S., & Obaid, A. J. (2023, December). Heart disease prediction using machine learning techniques. In *International Conference on Data Science, Machine Learning and Applications* (pp. 551-560). Singapore: Springer Nature Singapore.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. Materials today: proceedings, 80, 3782-3785.

Thomas, A., Ryan, C. P., Caspi, A., Liu, Z., Moffitt, T. E., Sugden, K., ... & Gu, Y. (2024). Diet, pace of biological aging, and risk of dementia in the Framingham Heart Study. *Annals of Neurology*, *95*(6), 1069-1079.

Pan, B., Hirota, K., Jia, Z., & Dai, Y. (2023). A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 126866.

Misra, P. K., Kumar, N., Misra, A., & Khang, A. (2023). Heart disease prediction using logistic regression and random forest classifier. In *Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem* (pp. 83-112). CRC Press.