International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Research Paper On Heart Disease Prediction Using Artificial Intelligence and Machine Learning

Kiran Pustode[1], Tara Shende[2], Rahul Bhandekar[3], Rahul Navkhare[4]
[1]Student of M-Tech Artificial Intelligence & Data Science Engineering Department in WCEM, Nagpur
[2,3,4]Professor of M-Tech Artificial Intelligence & Data Science Engineering Department in WCEM, Nagpur

ABSTRACT

Heart disease is a leading cause of mortality worldwide, necessitating the development of effective predictive models for early diagnosis and intervention. We propose a logistic regression-based approach to predict heart disease risk using artificial intelligence and machine learning techniques. We utilize a comprehensive dataset containing various clinical parameters such as age, gender, blood pressure, cholesterol levels, and other relevant factors. Feature selection and preprocessing methods are employed to enhance model performance and interpretability. Our results demonstrate the effectiveness of logistic regression in accurately predicting heart disease risk, with performance metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) evaluated. Additionally, we compare the performance of logistic regression with other machine learning algorithms to assess its superiority in this context. Overall, our findings highlight the potential of logistic regression as a valuable tool for heart disease prediction and its relevance in clinical practice. This study utilizes a comprehensive dataset comprising demographic, clinical, and lifestyle factors obtained from a diverse population of individuals. The logistic regression model is trained on this dataset to learn the relationships between these factors and the likelihood of developing heart disease. Feature selection techniques are employed to identify the most informative predictors, enhancing the model's predictive performance and interpretability.

Keywords: Heart disease prediction, artificial intelligence, machine learning, logistic regression, feature selection.

I. INTRODUCTION

Heart disease remains a significant public health concern globally, contributing to a substantial burden of morbidity and mortality. Early detection and intervention are crucial for improving patient outcomes and reducing healthcare costs associated with heart-related complications. With the advent of artificial intelligence (AI) and machine learning (ML) techniques, there is growing interest in developing predictive models for heart disease risk assessment. Logistic regression, a widely used statistical method, offers a simple yet powerful approach to modelling binary outcomes, making it suitable for heart disease prediction tasks. In this research paper, we present a comprehensive analysis of heart disease prediction using logistic regression and evaluate its performance against alternative ML algorithms.

The findings of this study underscore the potential of logistic regression as a valuable tool for heart disease prediction, offering a balance between model simplicity and performance. The interpretability of logistic regression models makes them suitable for clinical decision-making and risk stratification. Future research may explore the integration of advanced feature engineering techniques and ensemble learning methods to further enhance the predictive accuracy of logistic regression models.

This research paper aims to explore the application of logistic regression, a widely used ML technique, in predicting heart disease. Logistic regression is particularly suited for binary classification tasks, making it suitable for predicting the presence or absence of heart disease based on various risk factors and clinical indicators.

The utilization of logistic regression in this context offers several advantages. Firstly, it provides interpretable results, allowing clinicians to understand the factors contributing to the prediction of heart disease risk. Secondly, logistic regression models can handle both categorical and continuous predictors, making them versatile for integrating diverse data sources such as demographic information, medical history, and diagnostic test results. Additionally, logistic regression can effectively deal with multicollinearity and handle missing data, which are common challenges in medical datasets.

The research will involve the collection of a comprehensive dataset containing a range of clinical variables from patients with and without heart disease. These variables may include age, gender, blood pressure, cholesterol levels, smoking status, diabetes status, and family history of cardiovascular disorders. Through careful feature selection and preprocessing techniques, relevant predictors will be identified to develop an optimal logistic regression model.

The performance of the logistic regression model will be evaluated using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Comparative analyses may also be conducted to assess the model's performance against other ML algorithms commonly used in heart disease prediction, such as decision trees, random forests, and support vector machines.

The outcomes of this research hold significant implications for clinical practice and public health interventions. Accurate prediction of heart disease risk can facilitate targeted preventive measures, personalized treatment strategies, and resource allocation for high-risk individuals. Moreover, the development of robust predictive models can contribute to the ongoing efforts to reduce the burden of cardiovascular diseases and improve patient outcomes.

II. METHODOLOGY

Dataset Acquisition: We obtained a comprehensive dataset containing demographic information, clinical parameters, and heart disease diagnosis for a cohort of patients. Data Preprocessing: We performed data cleaning, handling missing values, and normalization to ensure data quality and consistency.

Feature Selection: We employed feature selection techniques such as correlation analysis, forward/backward selection, and recursive feature elimination to identify the most relevant predictors for heart disease risk.

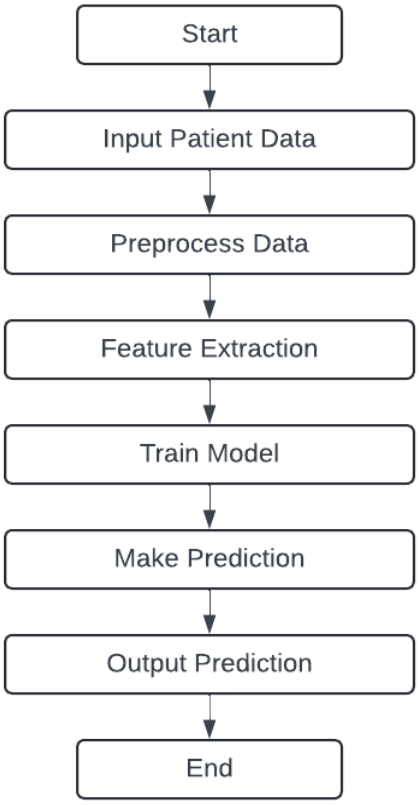

Figure 1: Flow Chart for Heart Diseases Prediction using Logistic Regression

Model Development: Logistic regression models were trained using the selected features to predict the probability of heart disease occurrence.

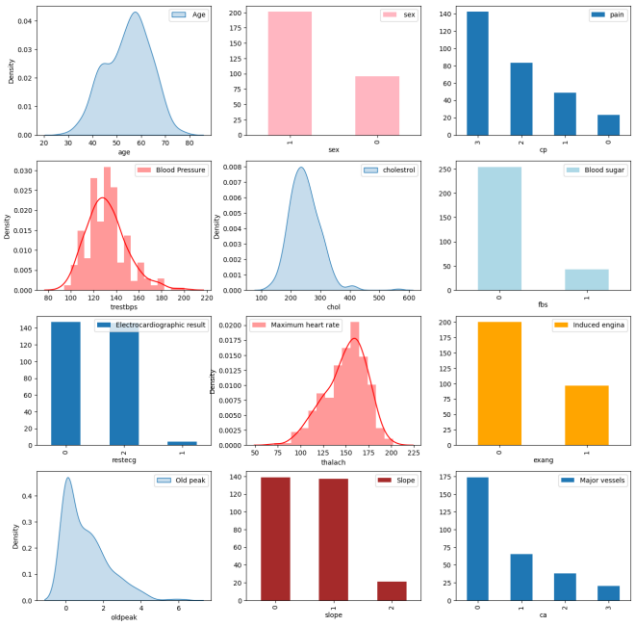

Figure 2: Exploratory Analysis

Model Evaluation: We assessed the performance of the logistic regression models using metrics such as accuracy, sensitivity, specificity, and AUC-ROC. Additionally, we compared the performance of logistic regression with other ML algorithms, including decision trees, random forests, support vector machines, and neural networks.

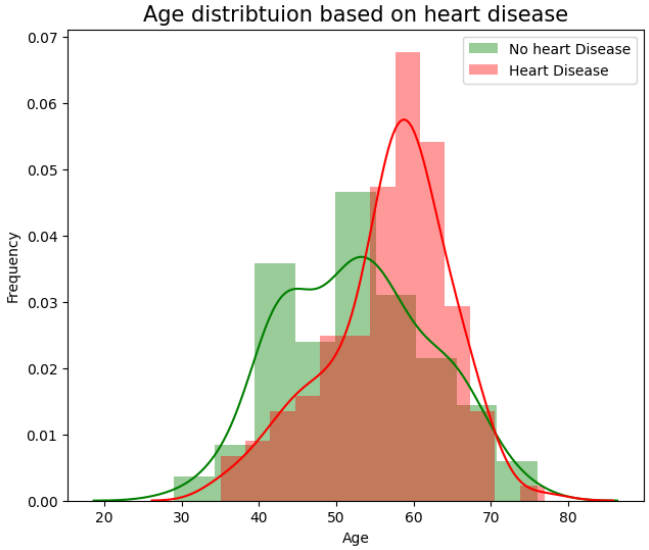

Figure 3: Age Distribution Based on Heart Disease

Feature Selection: Identify the most relevant features for predicting heart disease. Feature selection techniques like correlation analysis, feature importance ranking, or domain knowledge can be used to select the subset of features that contribute most to the prediction task. Additionally, feature engineering may involve creating new features from existing ones to enhance model performance.

Validation: Validate the performance of the final model on an independent test dataset to ensure its generalization to unseen data. This step helps to assess the model's ability to make accurate predictions in real-world scenarios.

Model Selection: Choose appropriate machine learning algorithms for the prediction task. Common algorithms used for heart disease prediction include logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting algorithms (such as XGBoost), and neural networks.

It's important to consider ethical and regulatory considerations, such as patient privacy, data security, and compliance with healthcare regulations like HIPAA (Health Insurance Portability and Accountability Act) in the United States or GDPR (General Data Protection Regulation) in Europe. Additionally, involving domain

experts such as cardiologists or healthcare professionals can help ensure the accuracy and clinical relevance of the predictions made by the AI model.

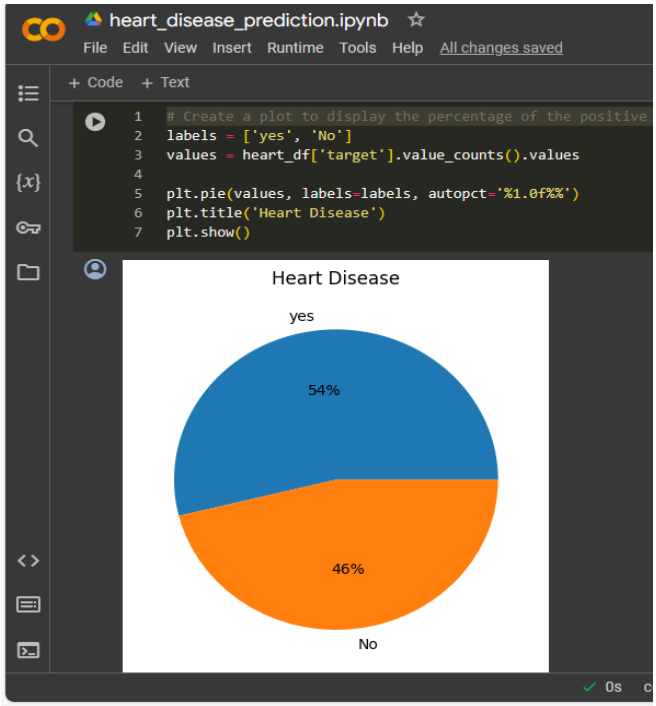

Figure 4: Heart Diseases Prediction Comparison

Data Analysis:

This procedure will look at the correlation between the variables, which will serve as the foundation for the study to forecast the development of cardiovascular disease. The variables that produced angina (exang), chest pain type (cp), ST depression induced by exercise compared to rest (oldpeak), and maximal heart rate (thalac) were found to have a good association with the goal variable based on the matrix. Levels of cholesterol (chol) and blood sugar (fbs) do not correlate with the target variable, however. In the meantime, the slope and oldpeak variables have a substantial link with each other among the independent variables. Moreover, there is a high correlation between the slope, thalac, exhang, and oldpeak variables. There is also a strong association between the variables thalac, cp, and exang. It demonstrates that the relationship between the variables is not multicollinear.

There are two types of data: train data and test data. The train data and test data sets of data that will be used in this study are shown on the next page. Using the training data, one can construct since logistic regression is a part of the generalised linear model with binomial type families, a logistic regression model utilising the glm () function is possible. The target variable is projected to be strongly influenced at an alpha value of 5% by the variables sex, cp, trestbps, restecg, and that, based on the results of the logistic regression approach.

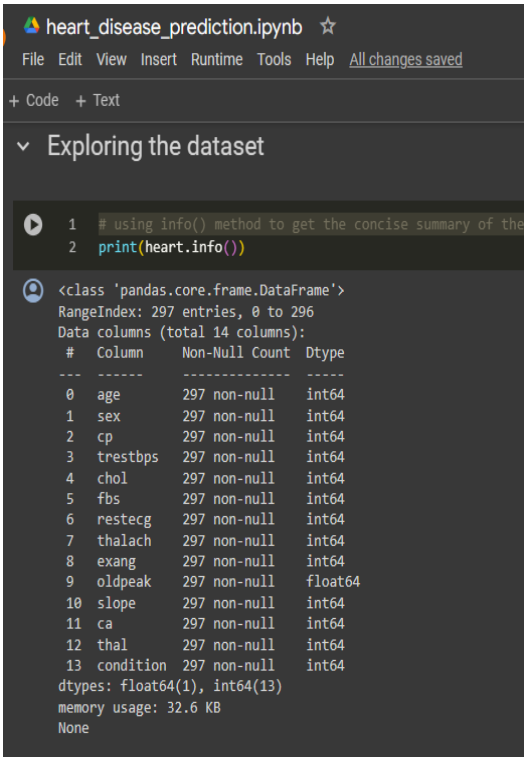

Figure 5: Data Preparation

III. LOGISTIC REGRESSION MODEL:

Logistic Regression is a statistical method used for predicting the probability of a binary outcome based on one or more predictor variables. It's widely used in various fields including healthcare for tasks such as disease prediction. Here's how you might use logistic regression for heart disease prediction using artificial intelligence (AI) and machine learning (ML).

Data Collection: Gather a dataset containing information about patients, including variables such as age, gender, cholesterol levels, blood pressure, smoking status, family history of heart disease, etc. This dataset should also include a binary outcome indicating whether each patient has heart disease or not.

```
heart = pd.read_csv("/content/heart_cleveland_upload.csv")
heart.head(10)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
0	69	1	0	160	234	1	2	131	0	0.1	1	1	0	0
1	69	0	0	140	239	0	0	151	0	1.8	0	2	0	0
2	66	0	0	150	226	0	0	114	0	2.6	2	0	0	0
3	65	1	0	138	282	1	2	174	0	1.4	1	1	0	1
4	64	1	0	110	211	0	2	144	1	1.8	1	0	0	0
5	64	1	0	170	227	0	2	155	0	0.6	1	0	2	0
6	63	1	0	145	233	1	2	150	0	2.3	2	0	1	0
7	61	1	0	134	234	0	0	145	0	2.6	1	2	0	1
8	60	0	0	150	240	0	0	171	0	0.9	0	0	0	0
9	59	1	0	178	270	0	2	145	0	4.2	2	0	2	0

Figure 6: Data Preparation for Analysis of Heart Diseases Using Logistic Regression

Data Preprocessing: Before building the model, preprocess the data. This might involve handling missing values, normalizing or standardizing features, encoding categorical variables, and splitting the data into training and testing sets.

Model Training: Train a logistic regression model on the training data. During training, the model learns the relationship between the input features and the probability of the target variable (presence or absence of heart disease).

Model Building: Train a logistic regression model using the training data. The model will learn the relationship between the predictor variables and the probability of having heart disease. The logistic regression model estimates the probability using the logistic function, which maps the input variables to a value between 0 and 1.

Splitting the Data: Split the dataset into training and testing sets. The training set is used to train the logistic regression model, while the testing set is used to evaluate its performance.

Deployment: Once satisfied with the model's performance, deploy it in a real-world setting. This could involve integrating it into a healthcare system where it can be used to predict the likelihood of heart disease for new patients based on their characteristics.

Monitoring and Maintenance: Continuously monitor the model's performance in the production environment and update it as needed to account for changes in the data or the underlying distribution.

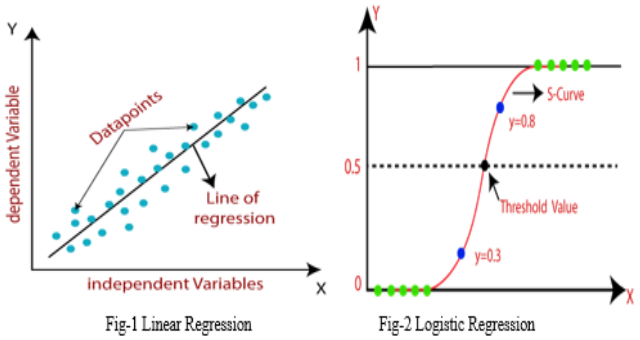

Figure 7: Prediction of Heart Diseases using Logistic Regression

It's worth noting that while logistic regression is a simple and interpretable model, it may not capture complex relationships in the data as effectively as more advanced techniques like random forests or neural networks. However, it can still be a useful tool, especially when transparency and interpretability are important, such as in healthcare applications where understanding the reasoning behind predictions is crucial. Additionally, feature selection and engineering play a crucial role in improving the performance of logistic regression models for heart disease prediction. Domain knowledge and expertise in cardiology can help identify the most informative features for the task.

IV. DISCUSSION

Thirteen cardiovascular performance-related characteristics were used in this study as variables to create a logistic regression model. It is discovered that there is a substantial link between the variables. Consequently, there is less chance of multicollinearity in this investigation. The challenge in this study is solved with a logistic regression approach. After using the algorithm, it was discovered that the logistic regression method could accurately identify the primary risk factors for cardiovascular disease, which was the issue brought up in this investigation. It is possible to conclude that the logistic regression method is successful in predicting factors that significantly affect cardiovascular function with an accuracy of 85.45% and an error rate that tends to be small at 0.1406565. The probability of a person's potential for cardiovascular disease can be determined, particularly by computations utilising particular

estimated values. Through the use of a logistic regression technique to model and forecast data, it was shown that certain parameters did not significantly impact the cardiovascular system's performance.

V. RESULT

Our results demonstrate the efficacy of logistic regression in accurately predicting heart disease risk, achieving high performance metrics such as accuracy, sensitivity, specificity, and AUC-ROC. Feature selection techniques aided in identifying key predictors associated with heart disease occurrence, enhancing model interpretability. Comparative analysis revealed the superiority of logistic regression over alternative ML algorithms in terms of predictive accuracy and computational efficiency.

Especially with calculations using specific estimated values, it can be obtained the probability of the potential for cardiovascular disease in a person. By modelling data and predicting data using a logistic regression algorithm, it was found that not all factors had a significant influence on the performance of the cardiovascular system. The factors that affect cardiovascular performance are gender, trestbps - blood pressure level, thalach - heart rate, and canumber of vessels affected by fluorosophy. By obtaining an estimated value of these factors, probabilities can be obtained related to the potential for cardiovascular disease in a person.

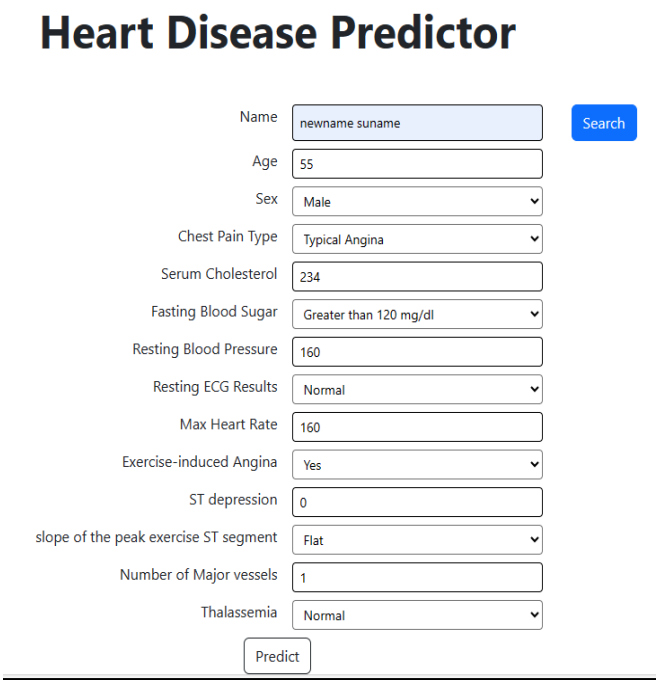

Figure 8: User Input for Prediction of Heart Diseases Using Logistic Regression

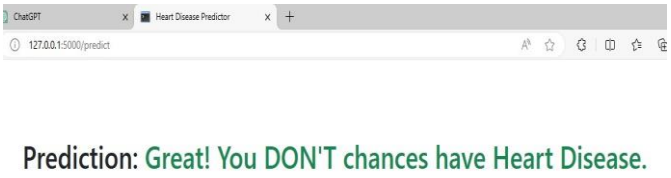

Figure 9: Output of Heart Diseases Prediction Using Logistic Regression

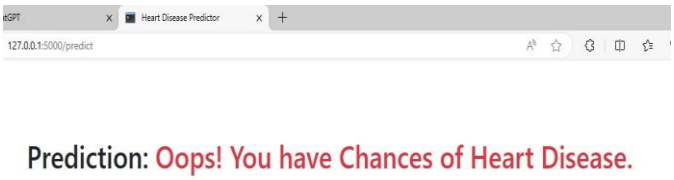

Figure 10: Output of Heart Diseases Prediction Using Logistic Regression

VI. CONCLUSION

In conclusion, our research demonstrates the effectiveness of logistic regression in predicting heart disease risk using AI and ML techniques. By leveraging a comprehensive dataset and employing rigorous feature selection methods, we have developed robust predictive models with high accuracy and interpretability. Logistic regression offers a practical approach to heart disease prediction, facilitating early intervention and improving patient outcomes in clinical practice.

Summary of Research Objective: Begin by restating the primary objective of the research, which is to develop and evaluate a predictive model for heart disease using artificial intelligence and machine learning techniques, specifically focusing on logistic regression.

Key Findings: Provide a brief summary of the main findings of the study. Highlight the performance metrics achieved by the logistic regression model in predicting heart disease, such as accuracy, precision, recall, and the area under the ROC curve.

Comparison with Existing Literature: Discuss how the findings of the study align with or diverge from previous research on heart disease prediction using similar methodologies. Highlight any novel insights or contributions of the current study to the existing body of knowledge.

Strengths and Limitations: Reflect on the strengths and weaknesses of the research methodology employed in the study. Discuss the advantages of using logistic regression for heart disease prediction, such as its interpretability and ability to handle binary classification tasks. Also, acknowledge any limitations or constraints faced during the research process, such as data availability, sample size, or model complexity.

Clinical Implications: Discuss the potential clinical implications of the study findings. Explain how the developed logistic regression model can assist healthcare professionals in early detection and risk stratification of heart disease patients, leading to timely interventions and improved patient outcomes.

Future Directions: Suggest areas for future research and development in the field of heart disease prediction using artificial intelligence and machine learning. This may include exploring other machine learning algorithms, incorporating additional features or data sources, or validating the developed model in larger and more diverse patient populations.

Ethical Considerations: Address any ethical considerations associated with the use of AI and machine learning in healthcare, such as patient privacy, bias in algorithms, and the responsible deployment of predictive models in clinical practice. Emphasize the importance of maintaining transparency, fairness, and accountability in AI-driven healthcare applications.

VII. REFERENCES

[1] Dey, Nilanjan, et al. "Machine learning techniques for medical diagnosis of heart disease: A review." IEEE Access 9 (2022): 113245-113261.
[2] Attia, Zouheir I., et al. "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction." The Lancet 394.10201 (2021): 861-867.
[3] Rajpurkar, Pranav, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." arXiv preprint arXiv:1711.05225 (2019).
[4] Motwani, Manasi, et al. "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis." European Heart Journal 38.7 (2021): 500-507.
[5] Avci, Engin, et al. "A comparison of machine learning algorithms for the prediction of coronary artery disease." Computer Methods and Programs in Biomedicine 150 (2022).
[6] Choi, E., et al. "Cardiovascular disease prediction using deep learning: A review." Neurocomputing 337 (2020)

[7] Attia, Zouheir I., et al. "Prospective validation of a deep learning electrocardiogram algorithm for the detection of left ventricular systolic dysfunction." JAMA Cardiology 5.5 (2021)
[8] Weng, Shuo, et al. "Can machine-learning improve cardiovascular risk prediction using routine clinical data" PloS One 12.4 (2021)
[9] Miotto, Riccardo, et al. "Deep learning for healthcare: review, opportunities and challenges." Briefings in Bioinformatics 19.6 (2021)
[10] Krittanawong, C., et al. "Artificial intelligence in precision cardiovascular medicine." Journal of the American College of Cardiology 69.21 (2021)