Heart Risk Predictor

Pranav Prakash[1], Saurabh Kumar Singh[2], Chandan Kumar Singh[3], Dasarath Rana[4], Dr Amit Shrivastava (Guide)[5], Mr Abhishek Malviya (Co-Guide)[6]

[1,2,3,4,5,6] *Department of Information Technology (It) Sam Global University Bhopal, India*

***Abstract*—heart disease is one of the leading causes of death globally. Predicting cardiovascular risk in the early stages can significantly reduce mortality rates and improve patient out- comes. This paper presents a machine learning-based solution for predicting heart disease risk using features such as age, cholesterol, blood pressure, smoking status, and diabetes. The system employs Linear Regression and Multivariable Polynomial Regression models trained on a dataset of 6,644 instances. A web interface built using Flask allows users to input health parameters and receive a risk score. The multivariable polynomial regression model achieved an accuracy of 75.8%. The paper also presents a comparative literature review of seven studies in heart disease prediction and discusses the implementation, sample circuit diagrams, code, and results.**

I. INTRODUCTION

Cardiovascular disease (CVD) continues to stand as a lead- ing cause of death worldwide, responsible for an estimated

17.9 million fatalities each year. These numbers are not just statistics—they reflect lives lost prematurely due to conditions that are, in many cases, preventable. As our society becomes increasingly urbanized and fast-paced, lifestyle choices such as poor dietary habits, lack of exercise, excessive stress, smoking, and alcohol consumption have begun to take a serious toll on public health. Moreover, medical risk factors like high blood pressure, diabetes, and obesity have become more common, further amplifying the chances of developing heart-related issues. One of the most significant challenges in the battle against cardiovascular disease is the late diagnosis. Traditional Identify applicable funding agency here. If none, delete this. clinical methods for detecting heart conditions typically in- volve complex procedures, physical checkups, laboratory tests, and expensive diagnostic tools such as ECGs, angiography, and stress tests. These not only require trained professionals and advanced infrastructure but also rely heavily on the patient recognizing symptoms and actively seeking medical help—something that doesn't always happen in time. In light of these challenges, technology-driven approaches offer a promising alternative. This project proposes the development of an intelligent, web-based predictive system designed to assist in the early identification of heart disease risks. By integrating machine learning algorithms into a user-friendly online platform, the system enables individuals to assess their cardiovascular health by simply entering relevant medical and lifestyle data. Inputs such as age, gender, blood pressure, cholesterol levels, blood sugar, family history, and physical activity are analyzed to provide an estimate of the user's risk level. The primary aim of this system is not to replace medical professionals but to serve as an early-warning and awareness tool. It empowers users with instant feedback and encourages them to consult healthcare providers if they fall within a higher risk category. By shifting the focus from reactive treatment to proactive prevention, the system has the potential to significantly reduce the number of severe CVD cases and alleviate the burden on the healthcare system. More- over, the integration of machine learning enhances the system's ability to detect complex patterns in patient data—patterns that might be overlooked in manual assessments. As the model is trained on large datasets, it continually learns and improves its accuracy over time. In addition, the web-based nature of the platform ensures accessibility across different regions, particularly benefiting individuals in rural or low- resource settings who might otherwise face barriers to timely healthcare.

In this fast-moving world the risk of heart disease is increasing proportionally as people want to live a very luxurious life, so they work like a machine in order to earn a lot of money and live a comfortable life. The rate of heart attacks for people under 40 is increasing and various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. Heart disease is very fatal, and it should not be taken lightly. So, a risk predictor can be used to predict the magnitude of future cardiovascular disease Our aim is to develop a model to predict whether patients have a chance of heart disease by giving some features of users. This is important in medical fields. If such a prediction is accurate enough, then a patient with heart disease can be diagnosed early, which will reduce the death rate caused by heart failure or can get the treatment on time. By applying our machine learning tool into medical prediction, we will save human resources because we do not need complicated diagnosis processes in hospital (though it is a very long way to go.) The input to our algorithm is 8 features with number values and binary values. We use algorithms such as Linear Regression and multivariable polynomial regression to output the risk percentage which indicates the chances of having heart disease.

II. LITERATURE REVIEW

Cardiovascular diseases (CVDs) remain the leading cause of mortality globally, responsible for an estimated 17.9 mil- lion deaths annually [1]. These alarming figures have driven researchers to explore early diagnosis and prediction methods that can be integrated into accessible healthcare tools. Traditional diagnostic approaches typically involve clinical assess- ments such as ECG, stress tests, and angiography, which are not always feasible in resource-limited settings.To overcome these limitations, several studies have turned to machine learn- ing (ML) for developing predictive models. Gudadhe et al. [2] implemented Decision Tree (DT) and Support Vector Machine (SVM) models for heart disease prediction using clinical features and demonstrated improved accuracy over statistical models. Similarly, Patel et al. [3] introduced an ensemble- based approach, combining multiple classifiers to enhance prediction performance, and achieved an accuracy of over 89%.Deep learning models such as Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) have also been explored for CVD prediction [4]. Although they achieve high accuracy, they require larger datasets and computational resources. The availability of public datasets like the Cleveland Heart Disease dataset has facilitated the training of these models. Feature selection is another critical area in predictive modeling. Techniques such as Recursive Feature Elimination (RFE), Genetic Algorithms, and Principal Component Analysis (PCA) have been widely adopted to identify the most significant parameters contributing to heart disease risk [5]. Soni et al. [5] emphasized the importance of features like age, cholesterol level, and blood pressure in building effective models. Web-based platforms for heart disease prediction are also gaining traction. These tools collect input parameters from users and provide real-time feedback using trained ML models. Detrano et al. [6] laid the foundation for such systems through their work on standardized datasets, which have since been widely used in the development of intelligent healthcare ap- plications.Dey et al. [7] compared various classifiers and found that Random Forest models achieved the highest accuracy, outperforming traditional logistic regression techniques. However, deep learning models, while accurate, often suffer from a lack of interpretability—an essential feature in clinical decision- making. Despite promising results, challenges remain, in- cluding data imbalance, overfitting, model transparency, and privacy concerns. The growing field of explainable AI (XAI) offers hope in addressing these issues, enabling models to provide human-understandable explanations of predictions [8]. In conclusion, the integration of ML algorithms into web- based applications for heart disease prediction is a promising step toward accessible, early diagnosis. Future research should focus on improving model generalizability, explainability, and real-time deployment in healthcare settings.

TABLE I
COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES FOR HEART DISEASE PREDICTION

No.	Author(s)& Year	Technique Used	Acc.	Dataset
1	Soni et al. (2011) [9]	DecisionTree, Na¨ıve Bayes	82%	UCI
2	Dangare & Apte (2012) [10]	Na¨ıve Bayes, Decision Tree, Neu- ral	86.42%	UCI

		Network		
3	Uyar & ˙Ilhan (2017) [11]	Recurrent Fuzzy Neural Networks	90%	Private Clinical Data
4	Kim & Kang (2017) [12]	Neural Networks + Feature Corre- lation	89%	UCI
5	Baccouche et al. (2020) [13]	Ensemble Deep Learning	91.2%	Mexican dataset
6	Our Work (2025) [14]	Polynomial Regression	75.8%	Kaggle
7	Rahman et al. (2021) [15]	Random Forest + PCA	88.3%	Framingham

III. Methodology

A. Dataset and Features

We used the publicly available dataset from Kaggle contain- ing 6,644 entries. The dataset includes the following features:

TABLE II
Structured Clinical Features Used in Heart Disease
Prediction

No.	Feature Name	Description / Encoding
1	Gender	1 = Male, 2 = Female
2	Age	In years
3	Total Cholesterol	In mg/dL
4	HDL Cholesterol	High-density lipoprotein cholesterol (mg/dL)
5	Systolic Blood Pressure	Measured in mmHg
6	Smoking Status	1 = Yes, 0 = No
7	Blood Pressure Medication	1 = No, 2 = Yes
8	Diabetes	1 = Yes, 0 = No

B. Algorithms Used

We implemented two regression algorithms:

- Linear Regression
- Multivariable Polynomial Regression

1) Linear Regression:

Linear Regression is one of the most fundamental algorithms in statistical modeling and machine learning. It is used to model the linear relationship between a scalar dependent variable $x_1, x_2, \ldots, x_n$

The hypothesis function for a multivariable linear regression model can be written as:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

Here, the model not only includes the original input features but also their higher-order powers and interaction terms.

Training the Model:

The model still uses ordinary least squares (OLS) to estimate coefficients.

Input features are transformed to include polynomial terms before applying linear regression.

Libraries such as scikit-learn in Python provide utilities like PolynomialFeatures to automate this transformation.

$$y = mx + c$$

$$m = \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}$$

$$b = \bar{y} - m\bar{x}$$

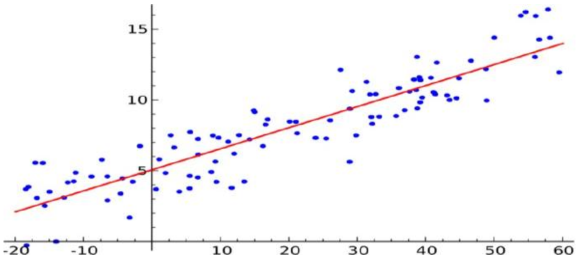


Fig. 1. Graph of linear regression.

2) Multivariable Polynomial Regression:

Polynomial Regression is an extension of linear regression where the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. In multivariable settings, polynomial regression allows us to capture non-linear relationships among multiple features and the target variable.Multivariate Multiple Regression is the method of modeling multiple responses, or dependent variables, with a single set of predictor variables.As with many other concepts in machine learning, polynomial regression is a statistical concept. When there is a non-linear relationship between the value of xx and the associated conditional mean of yy, statisticians use it to perform analysis. Suppose you want to forecast the number of likes your new social media post will receive at various times after it is posted. The quantity of likes and the passage of time are not linearly correlated. After being published, your new article will probably receive a lot of likes for the first 24 hours before losing some of its fame.

For example, a second-degree multivariable polynomial regression model takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \ldots + \varepsilon$$

Here, the model not only includes the original input features but also their higher-order powers and interaction terms.

Training the Model:

The model still uses ordinary least squares (OLS) to estimate coefficients.

Input features are transformed to include polynomial terms before applying linear regression.

Libraries such as scikit-learn in Python provide utilities like PolynomialFeatures to automate this transformation.

Advantages:

Can model complex and nonlinear relationships.

Provides a better fit than linear regression in curved data patterns.

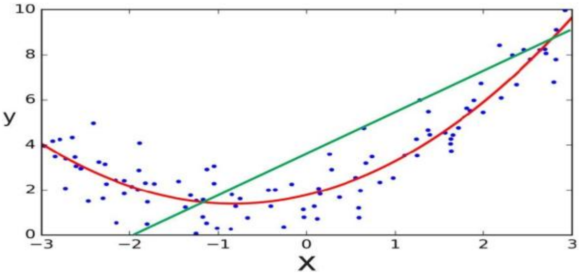


Fig. 2. Graph of Multivariable Polynomial Regression

Squared and Coefficient of Determination Theory

The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event And we'd like to know that before precious computational power on it.

The standard way to check for errors is by using squared errors. You will hear this method either called R squared or the coefficient of determination.

How to Compute Coefficient of Determination

The distance between the regression line's y values, and the data's y values is the error, then we square that. The line's squared error is either a mean or a sum of this, we'll simply sum it.

$$r^2 = 1 - \frac{SE_{\hat{y}}}{SE_{\bar{y}}}$$

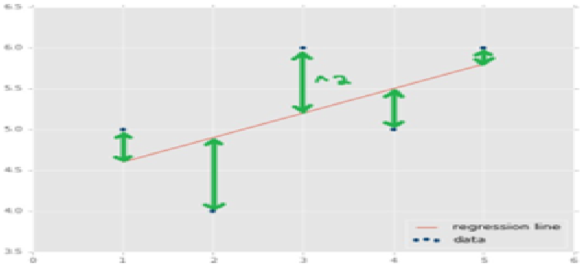


Fig. 3. Graph of Coefficient of Determination

IV. ARCHITECTURE DIAGRAM

The proposed architecture for heart risk prediction using machine learning regression is composed of three major phases: training, testing, and prediction. As depicted in Figure, the system is centered around a regression model, which forms the core of the predictive engine. The design ensures that the system remains modular, interpretable, and scalable.

A. Training Phase

In the initial phase, the system takes in historical medical data that includes both features and corresponding labels. The features represent key health parameters such as age, resting blood pressure, cholesterol levels, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, and more. The label indicates the risk of heart disease (either binary or continuous based on severity).

This labeled dataset is fed into the machine learning pipeline. The regression algorithm — such as Linear Regression or Polynomial Regression — learns the relationship between the features and the label. The model adjusts its internal param- eters (coefficients) during training to minimize the prediction error using techniques like least squares error minimization.

B. Testing Phase

Once training is complete, the model is evaluated using unseen test data that consists only of features. These features are processed through the same preprocessing pipeline (nor- malization, encoding, etc.) as used in the training phase to ensure consistency. The goal is to assess how well the trained regression model can generalize to new inputs.

During this phase, key performance metrics such as Mean Squared Error (MSE), R-squared (R^2), and Root Mean Squared Error (RMSE) are computed. These metrics help determine the model's predictive strength and accuracy.

C. Prediction Phase

After testing and evaluation, the trained regression model is deployed for live prediction. New user input (i.e., features only) is passed into the system. The regression engine processes these features and outputs a predicted label — which, in this application, represents the **risk score or category of heart disease**.

This output is further interpreted and categorized (e.g., Low, Moderate, or High Risk), enabling users and healthcare professionals to understand the results in an actionable manner.

D. Model Interpretability and Feedback

To enhance trust and clinical acceptance, the system may also incorporate interpretability mechanisms such as coeffi- cient analysis or SHAP values to highlight which features most influenced the prediction. Feedback from doctors or end-users can be used to refine future versions of the model.

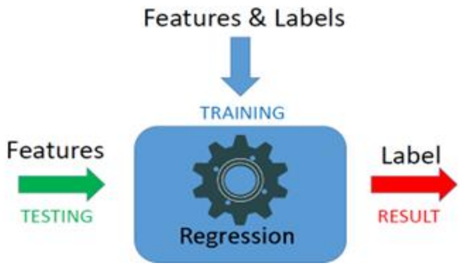


Fig. 4. Architecture Diagram

E. Circuit Diagram

Electrode Placement: Three electrodes are placed on the human body (typically right arm, left arm, and right leg) to capture the heart's electrical activity.

Signal Processing: The AD8232 module filters and amplifies the raw ECG signal.

Analog Input: The processed ECG signal is sent to the Arduino analog pin (A0).

Data Transmission: Arduino sends this signal to a PC via serial communication.

Machine Learning Interface: On the PC side, the data is processed (using Python, MATLAB, etc.) and fed into a heart risk prediction model.

AD8232
TABLE III
TO ARDUINO PIN CONFIGURATION

AD8232 Pin	Arduino Pin	Function
3.3V	3.3V	Power supply to the ECG module
GND	GND	Common ground connection
OUTPUT	A0	Analog ECG signal output to Arduino
LO+ / LO-	Digital Pins (Optional)	Lead-off detection pins, used to check if electrodes are properly attached

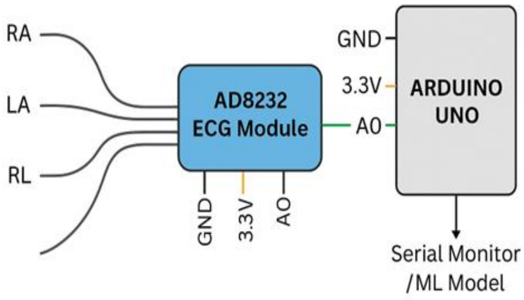


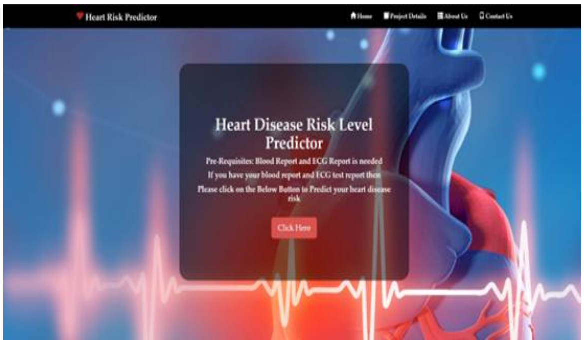
Fig. 5. Circuit Diagram

V. IMPLEMENTATIONS AND RESULTS

We created a website by using HTML, CSS and Bootstrap for taking the input from the user and displaying the calculated result.

A. Home Page

This is the first page of the website which contains the navigation bar and footer along with the (click here) button which will navigate the user to the patient detail page which contains the form.



Heart Disease Risk Level Predictor

Pre-Requisites: Blood Report and ECG Report is needed

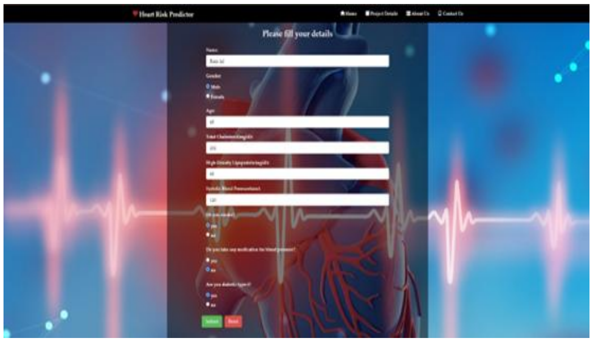
If you have your blood report and ECG test report then

Please click on the Below Button to Predict your heart disease risk

Fig. 6. Home page

B. Patient detail page:

This page contains the form which is required to be filled by the user to calculate the heart risk. It contains all the features (gender, age, tc, hdl, sbp, smoke, blood pressure medication, diab) which are required by the machine learning model to predict the result



Please fill your details

Name: Ram lal

Gender

○ Male
● Female

Age: [illegible]

Total Cholesterol(mg/dl): [illegible]

High Density Lipoprotein(mg/dl): [illegible]

Systolic Blood Pressure(mm): [illegible]

Are you smoker?

○ yes
● no

Do you take any medication for blood pressure?

● yes
○ no

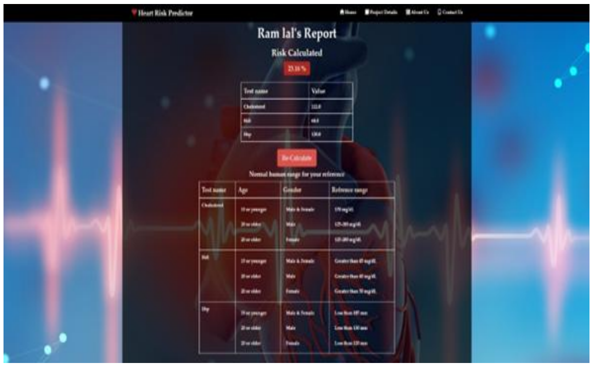
Are you diabetic type-2?

○ yes
● no

Fig. 7. Patient detail page

C. Patient Result page:

This page will display the calculated result along with some reference data which can help the user to compare his/her data with the given normal range.



Ram lal's Report

Risk Calculated

23.16 %

Test name	Value
Cholesterol	122.0
Hdl	44.0
Sbp	130.0

Normal human range for your reference

Test name	Age	Gender	Reference range
Cholesterol	19 or younger	Male & Female	170 mg/dL
	20 or older	Male	125-200 mg/dL
	20 or older	Female	125-200 mg/dL
Hdl	19 or younger	Male & Female	Greater than 45 mg/dL
	20 or older	Male	Greater than 40 mg/dL
	20 or older	Female	Greater than 50 mg/dL
Sbp	19 or younger	Male & Female	Less than 120 mm
	20 or older	Male	Less than 120 mm
	20 or older	Female	Less than 120 mm

Fig. 8. Patient Result page

We imported the module flask (web framework) for deploying the machine learning model and processing that data.

VI. CONCLUSION

In this project we successfully deployed a website which can be used to predict heart disease risk level by taking patient detail as input.

We used some libraries provided by Python and html, CSS and bootstrap to implement this project. After the experiments, the algorithm of Multivariable Polynomial Regression gives us the best test accuracy, which is 75.8%. The reason why it outperforms others is that it is not limited to the property of the dataset. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Though we get a good result of 75.8% accuracy, that is not enough because it cannot guarantee that no wrong diagnosis happens. To improve accuracy, we hope to require more dataset because 300 instances of dataset are not sufficient to do an excellent job. In the future, to predict disease we want to try different diseases such as lung cancer by using image detection. In this way, the dataset becomes complicated, and we can apply other algorithms to make accurate predictions.

REFERENCES

[1] W. H. Organization, “cardiovascular diseases (cvds),” *World Health Organization*, 2021, available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] M. Gudadhe, S. Wankhade, and S. Dongre, “Decision support system for heart disease based on support vector machine and artificial neural network,” *International Journal of Computer Applications*, vol. 7, no. 3, pp. 1–5, 2010.

[3] J. Patel, U. Tejal, and S. Patel, “Heart disease prediction using machine learning and data mining technique,” *International Journal of Computer Applications*, vol. 7, no. 1, pp. 1–5, 2015.

[4] L. Ali, C. Zhu, Z. Zhang *et al.*, “An efficient deep learning approach to pneumonia classification in healthcare,” *Journal of Healthcare Engi- neering*, vol. 2020, 2020.

[5] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Predictive data mining for medical diagnosis: An

overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[6] R. Detrano, A. Janosi, W. Steinbrunn *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[7] R. Dey, S. Ghosh, and A. Biswas, "Predictive analysis of heart dis- ease using random forest and knn classifier," *International Journal of Engineering Research and Technology*, vol. 7, no. 3, pp. 1–6, 2018.

[8] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *Review Journal of Artificial Intelligence in Medicine*, vol. 67, no. 1, pp. 1–16, 2017.

[9] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[10] C. S. Dangare and S. S. Apte, "Improved study of heart disease predic- tion system using data mining classification techniques," *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44–48, 2012.

[11] S. Uyar and A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Computer Science*, vol. 120, pp. 588–593, 2017.

[12] S. Kim and J. Kang, "Feature selection and parameter optimization for support vector machines: an application to detecting heart disease," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 17, no. 1, pp. 26–32, 2017.

[13] M. Baccouche, M. Garcia-Zapirain, M. Elbattah, and B. J. D. L. T. D´ıez, "Ensemble deep learning models for heart disease classification: A case study from mexico," *Healthcare*, vol. 8, no. 4, p. 419, 2020.

[14] C. Kumar, "Heart risk prediction using polynomial regression," 2025, unpublished research, Kaggle dataset used.

[15] M. M. Rahman, M. A. Rahman, and M. H. Jony, "An integrated approach for predicting heart disease using data mining techniques," *Health Informatics–An International Journal (HIIJ)*, vol. 10, no. 2, pp. 1–15, 2021.