Enhancing Heart Disease Prediction Based on B-Spline Curves and Machine Learning Algorithms

Mohammed Abdul Ameer Jabbar[1], Abdul mohssen Jaber Almaaly[2], Mustafa Tareq Abd[3], Abdulkareem Merhej Radhi[4]

[1] Department of Intelligent Medical Systems, College of Biomedical Informatics, University of Information Technology and Communications, Baghdad, Iraq
[2] Turath University, Baghdad, Iraq
[3] Department of Computer Science, University of Technology, Baghdad, Iraq
[4]Department of Computer Science, College of Science, Al-Nahrain University, Baghdad, Iraq

E-mail address: mohammedaji@uoitc.edu.iq, Abdulmohsen.jaber@uoturath.edu.iq, mustafa.t.abd@uotechnology.edu.iq, abdulkareemradhi@gmail.com

Abstract: Predicting heart attacks, is crucial as it can save lives and reduce the personal and societal impact of cardiovascular diseases. Early detection allows for timely intervention, enabling individuals to make lifestyle changes and medical professionals to implement preventive measures. In this presented work, we propose a new model for detecting and classifying heart diseases by integrating mathematical equations related to B-spline curves with artificial intelligence methods represented by machine learning. Since this method was proposed to improve the quality of medical datasets and to treat random and outlier values, the proposed approach consists of two stages for processing medical data: In the first stage, machine learning algorithms were trained on the dataset using classical pre-processing methods with (logistic regression) and (decision tree) algorithms, and the second stage involves applying the proposed method of combining pre-processing techniques which include b-spline curves and the Random Forest algorithm on the same dataset. The proposed approach achieved excellent results through the Random Forest model, with results indicating a detection accuracy of 99.19%. This proposed method of combined strategy provides valuable insights for early intervention. This marks a significant advancement in blending mathematical interpolation with machine learning, promising enhanced prediction accuracy and practical utility in healthcare analytics.

Keywords: B-spline Curves, Machine Learning, Heart Disease Prediction, Medical Dataset Enhancing.

1. INTRODUCTION

Curve interpolation is a widely used method in various fields to create mathematical curves or functions based on sampled data points [1]. These interpolation curves are valuable for visualizing data, estimating function values, and summarizing relationships between variables in engineering and science. The accuracy of the interpolation curve is crucial for generating new data points within the range of known data. Various interpolation techniques are available to address these needs, with polynomial interpolation being particularly notable, exemplified by Lagrange and Newton interpolation [2]. Additionally, Hermite, piecewise, and spline interpolation are among the commonly used techniques, with spline interpolation being the most effective [3]. The b-spline is a special mathematical equation that is piecewise-defined by a set of polynomial. In complex cases, such as curve with discontinuous points, cusp, or turning points in the sampled data, b-spline interpolation is often preferred over polynomial interpolation. Furthermore, spline interpolation avoids Runge's phenomenon, which occurs when high-degree of polynomials cause oscillations between points. In interpolation problems, a b-spline is defined by the ordering, a Knot vector [4], and a set of control points (CP) or integrated nodes. In the context of heart attack data analysis, which often exhibits complex and intricate patterns, B-spline curves, as piecewise-defined polynomial functions, provide valuable techniques to further the representation of such datasets. Additionally, data-enhancing methods, and imbalanced data [5], [6] are employed to reveal hidden trends in the medical dataset.

Globally, heart-diseases remain the leading cause of death, with heart attacks and strokes being the primary contributors to cardiovascular-related fatalities. According to the World Health Organization (WHO), these diseases claim about 17.9 million lives annually. It is crucial to identify strategies to reduce these figures. Utilizing machine learning (ML) algorithms, previous studies have demonstrated the effectiveness of these tools in classifying and detecting heart-failure [7], [8], [9], [10], and other general diseases [11]. The decision-support system that can utilized to rapidly and inexpensively identify heart disease from medical records can be created using Machine learning techniques [12]. This type of decision-support system can help medical professionals identify the disease earlier [13]. Decision support systems can be integrated into mobile devices through the use of mobile health technologies. Mobile health technologies give patients access to real-time data collection and more effective medical care. It enhances patient monitoring without requiring a trip to the hospital [14]. In addition, the ML algorithms can accurately and efficiently predict such outcomes, helping medical professionals and patients identify early signs of heart diseases and take preventive measures. This study [15] emphasizes the need for early detection and intervention to mitigate the harmful effects of heart disease. Machine learning algorithms, known for their ability to analyze vast datasets and identify intricate patterns, offer a promising approach for predicting the onset of these diseases[16]. By applying these algorithms, medical professionals and individuals can gain insights into early signs of heart-related issues, enabling timely and proactive measures to reduce the impact of these life-threatening conditions.

This research integrates B-spline curve interpolation with machine learning algorithms to develop robust strategies for analyzing and predicting heart disease patterns. The ultimate goal is to empower healthcare professionals and individuals with the knowledge needed to take pre-emptive actions, thereby reducing the global burden of heart-related fatalities. The primary contributions of this paper encompass:

- Utilizing B-spline curve interpolation to identify the noise and outlier in the heart attack dataset.
- Assess the accuracy of B-spline curve interpolation in capturing intricate patterns in heart attack data.
- Develop robust strategies integrating B-spline interpolation and ML for early detection and prevention of heart-related issues, empowering healthcare professionals and individuals.
- Evaluate the effectiveness of combining machine learning with b-spline interpolation for predicting heart disease patterns.

The paper remaining sections are arranged as follows: A brief theoretical background of the b-spine curves and an introduction of the proposed ML algorithms that are used in this research. The list of commonly related works that concern the case study can be found in Sec2. A description of the used dataset can be found in Sec.3. The proposed methodology of ML and spline curves to enhance heart attack prediction can be seen in Sec.4. Results and discussions are included in Sec.5. Finally, Sec. 6 contains a conclusion and future directions.

A. *B-Spline curve (definition and applications)*

The distinct advantages of B-spline methods have made them very common in computer aid systems for geometry and related fields [17]. Other techniques for depicting curves and surfaces have been introduced recently [18, 19]. The more versatile Bezier Curve (BC) known as the B-Spline eliminates BC's main drawbacks, that is number of control points determines the degree of BC [20]. They are easily represented by polynomials in pieces. This makes it possible to represent a B-spline curve as an ordered set of number curves. Moreover, the curves can be improved by linear operations on control points (CP) as shown in Figure 1 [21].

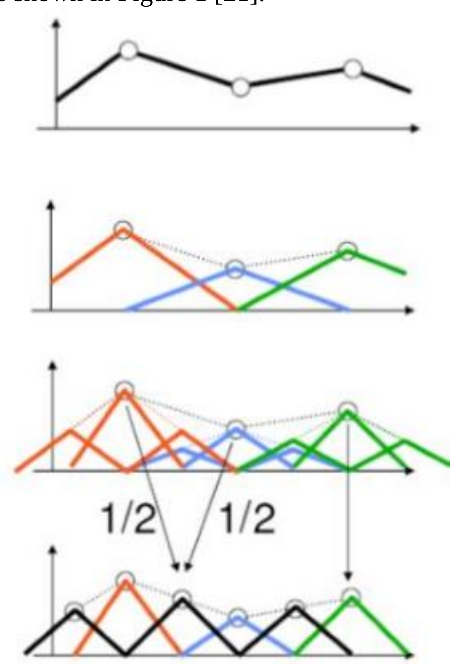


Figure 1: b-spline curve as a combination of different CP [21]

B. *Types of ML algorithms that are utilized in heart-disease prediction*

The analysis of medical datasets especially for diagnostic purposes, increasingly relies on machine-learning techniques. Modern hospitals utilize various data systems and monitoring equipment to compile extensive datasets, making it easier to capture and manage data in the digital age. This kind of data is now readily accessible and connected within the particular system, enabling the application of ML techniques for comprehensive healthcare analysis and improving the results of early prediction [22]. The ML technology can be applied to the analysis of medical data and is commonly used in specific diagnostic tasks, by entering patient records into a computer program with known and obtaining accurate diagnoses. This process creates a derived classifier that works as a decision support, which aids doctors in diagnosing new patients more quickly, accurately, and reliably. Additionally, this system is an invaluable resource for training non-specialists to diagnose patients with specific medical conditions [23].

ML techniques have become widely used artificial intelligence tools in every major field of application. Especially in the medical field and disease diagnosis, this research paper suggests improving early diagnosis of heart disease by using ML algorithms.

1) Decision Trees (DT) algorithm

The DT is an ML model utilized to make decisions based on dividing data into branches based on certain criteria [24]. This model is similar to a tree structure where the main node (the root) starts by dividing the data into sub-branches based on a certain condition (such as the value of a certain variable). This process continues until arriving at the leaves, which are the final nodes that represent final decisions or predictions [25].

The advantages of a DT algorithm include ease of understanding and visualization, as the model can be drawn in the form of a tree that people can easily understand. Decision trees can handle both categorical and numerical data [26]. However, one of the major drawbacks of decision trees is that they may be susceptible to overfitting, meaning they may be very accurate on training data but ineffective on new data. They can also be unstable, as a small change in the data can lead to large changes in the tree structure. Some examples of utilizing the DT algorithm in medical applications:

Diagnosing diseases: A decision tree can be used to analyze a patient's symptoms and medical history to diagnose diseases [27] such as diabetes [28], heart disease [29], or cancer [30]. The model can identify patterns in medical data that indicate a patient is likely to have a particular disease.

Determining a treatment plan: A decision tree can be used to determine the best treatment plan for a patient based on their characteristics such as age, gender, and medical history [31, 32]. The model can help doctors choose the most appropriate treatment for each patient.

2) Random Forest (RF) algorithm

The RF is an ML model based on the idea of using multiple DTs to make more accurate and stable predictions. Several DTs are trained independently using random samples of data, and then the results of these trees are combined to get the final result. This approach reduces the risk of over-analysis and increases model stability [33]. An RF is more accurate and stable than a single DT [34] because it takes advantage of the average of the predictions of many trees, reducing variance and skewness in the model. However, this technique is more complex and slower than a single decision tree and can be difficult to interpret when using a large number of trees [35]. Here are the examples of the RF applications in the medical field:

Cancer detection: Random forest can be used to analyze X-ray or MRI images [36] to detect cancerous tumors. The model can identify unusual patterns in images that indicate the presence of cancer [37].

Predicting surgical outcomes: RF can be used to predict surgical outcomes based on patients' previous medical data. The model can help determine the likelihood of the operation being successful or complications arising [38].

3) Logistic regrission (LR) algorithm

LR is a statistical model used to estimate the probability of a binary event (such as yes/no) occurring. This model is based on using a logistic function to convert a set of linear inputs into probabilities. These probabilities are estimated based on a set of independent variables, and the model can be used to classify the data into two different categories [39]. The advantages of LR include its simplicity and ease of interpretation. The model can determine the strength of each independent variable in the model, allowing a better understanding of the relationships between variables [40]. However, logistic regression is not suitable for non-binary classification and requires that the relationships between variables be linear. Some examples and applications in the medical field:

Heart attack prediction: LR can be used to estimate the likelihood of a heart attack [41] based on risk factors such as blood pressure, cholesterol level, and age. The model can help doctors identify patients who need urgent medical intervention.

Diagnosing infections: LR can be used to analyze the results of medical tests to diagnose infections [42] such as a urinary tract infection or pneumonia. The model can determine probabilities based on test results and symptoms.

2. RELATED WORKS

The list of related works has been presented in this section, which includes previous studies about utilizing ML algorithms to detect heart attacks by highlighting the obtained results.

- In this paper [7], an ML algorithm Support Vector Machine (SVM) was used to predict heart disease using several risk factors and make predictions about heart attacks, which is a challenging task that requires accuracy and efficiency by achieving classification accuracy at 85%.

- The paper [15] discusses various machine learning algorithms for heart attack prediction, including DT, Logistic Regression (LR), SVM, Naive Bayes (NB), Random Forest (RF), K-Nearest Neighbor (KNN), and XG Boost Classifier. SVM best results at 90% of the model accuracy.

- The author in [43] contributes several models for the detection of heart disease based on machine learning algorithms such as RF, SVM, KNN, and DT. The best results at RF reached an accuracy of 94.958%.

- In [44], the authors propose a machine learning-based model for predicting the possibility of a heart attack, with the KNN algorithm achieving an accuracy of 90.16% and recall of 87.09%.

- This paper [45], the authors discuss the use of machine learning classifiers, including Logistic Regression, to predict heart attacks with an accuracy of 82.4% and precision of 84.3%.

- In this article [46], three machine learning algorithms are implemented, DT, RF, and SVM for heart disease prediction. The best classification accuracy was achieved using SVM at 82.01% and DT at 78.98%.

- In this study [47], the authors use RF as the most compatible contender for the prediction model and give the highest performance measure among KNN and DT, and the best- accuracy in the RF model is 96.71%.

- The authors in [48], discuss the use of machine learning algorithms, such as KNN, DT, SVM, LR, Naive Bayes (NB), and RF, for analyzing heart disease using patient data. The best detection accuracy achieved in the LR model of 82%.

- The researchers in [49] propose several models based on ML algorithms, which include (NB, RF, DT, and KNN) to predict heart disease by utilizing a heart-disease dataset. The best prediction accuracy was achieved in the KNN classifier at 90.78.

- In [23], the authors also used the heart disease dataset and proposed several solutions for heart attack prediction analysis based on ML algorithms, the best classification results were achieved utilizing the KNN model at a classification accuracy of 94.1%.

- The researchers [50] proposed a hybrid method utilizing ML algorithms, RF with a linear model in order to optimize heart disease prediction. The accuracy level was achieved in this experiment at 88.7%.

- The author demonstrates in this review [51] paper to show the latest trends of ML techniques and methods that utilized recently in the task of heart disease prediction.

- In this paper [52], the authors proposed a heart failure system based on machine learning algorithms and feature selection methods utilizing the UCI dataset. The best-achieved classification results in SVM and DT, and the accuracy at 92.3% and 93.19 respectively.

- In [53], the authors combine recent works in a systematic review, that employs methodology to identify the difficulties arising from unbalanced data in heart disease classification and prediction.

- Authors in [54], their suggested method predicted the presence of heart disease using an RF. Relevant attributes were chosen through feature selection using the chi-square method. Compared to the decision tree, the authors' approach yielded higher accuracy.

- In [55], this study refers to developing a neural network-optimized genetic algorithm system to predict heart disease. It performed more accurately than the traditional neural network.

- In [56], authors have created a group of prediction models for diagnosing heart diseases. The ensemble model was built using gradient boosting, random forest, and extreme gradient boosting classifiers.

3. HEART ATTACK DATASET DESCRIPTION

The heart attack dataset published by IEEE data port in [57], was used in this analysis and contains various attributes such as age, cholesterol levels, blood pressure, and other health indicators. Contains 313 samples and 13 attributes classified as (0,1) normal and heart attack patients. Table 1 shows the heart attack dataset attributes and their description.

TABLE 1. HEART ATTACK DATASET

Features/attribute	Full-text description and data values
Age	The age of the patient in years
Sex	Sex of patient (1, 0) as (M/F)
CP	Ch-pain valued at (0,1,2,3) for t-angina, at-angina, non-anginal pains, and asymptomatic: respectively
trtbps	Bloud-pressure
chol	Cholesterol
fbs	Sugar of fasting blood > 120 mg/dl (0= false, 1=true)
restecg	Resting electrocardiographic result (0=normal ,1=abnormal ,2= hypertrophy)
thalachh	Maximum heart rate
exng	do Exercise (yes=1, no=0)
oldpeak	Exercise related to rest
slp	The slop of peek (0,1,2) as up-slop, flat, down-slop
CAA	Number of major vessels (0-3)
Thal	Thala-ssemia (1,2,3)
output	Target values (1= heat disease,0= normal)

4. PROPOSED METHOD

The proposed methodology of applying B-spline interpolation curves is to enhance the heart attack dataset by determining the outliers and noise in the data samples. This technique improves the proposed machine learning model to classify and detect results accurately. The proposed system includes three major phases: the pre-processing is the first phase which includes data integration and gathering, applying a b-spline curve to the attributes to find the relation between features. Furthermore, implementing a b-spline curve to highlight the outliers in the samples by drawing the curve between two attributes and for all datasets, this technique is effective for better smoothing the original dataset. The second phase includes preparing the smoothed dataset by splitting it into training, validation, and testing sets in order to build the proposed model utilizing decision trees, random forest, and logistic regression which are the common ML algorithms that are used to detect heart attack disease. The final phase involves evaluating the proposed model according to the classification metrics including accuracy, Mean Square Error (MSE), precision, and recall by computing the confusion matrix on the testing set, Moreover, comparing the proposed method with previous related works. The following subsection contains details about each phase. Figure 2 illustrates the proposed method block diagram.

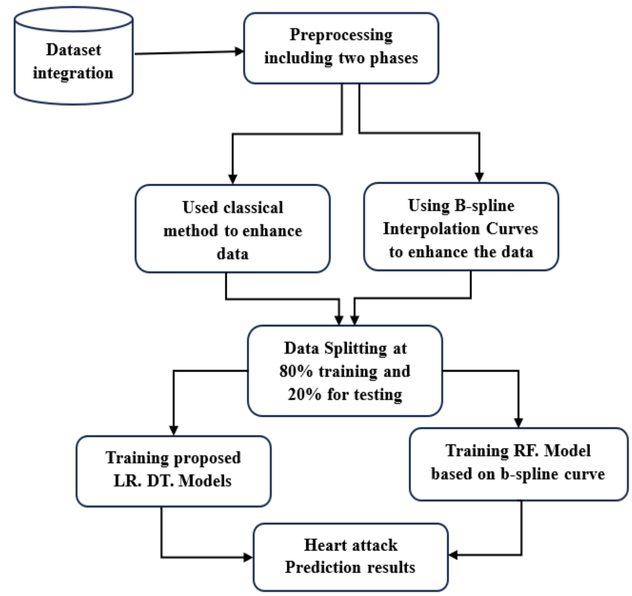


Figure 2 The block diagram of the proposed method

The proposed pre-processing steps include data integration, b-spline algorithm implementation in Python, and data normalization techniques. In these steps, the data will be ready for splitting into two sets (training and testing) in order to start training the proposed ML models.

A. Implementing B-spline Curve Interpolation:

The equations define the curve fitting the points in general where denoted in [58], [59]. For a given sequence of knots, a unique spline is given by:

$$p(t) = \sum_{i=0}^{n} Ni, k(t)\, Pi\,, \tag{1}$$

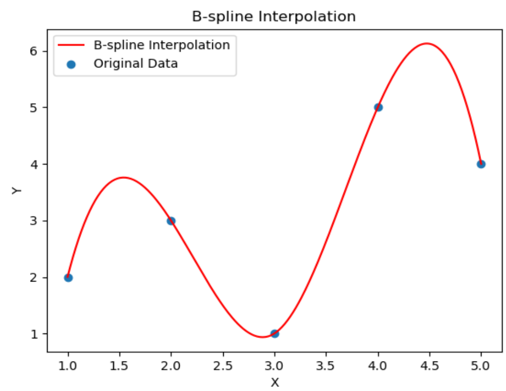
starting with the b-spline of degree (p=0), which is means, utilizes a constant polynomials.

$$B_{j,1}(t) = \begin{cases} 1 & if \quad t_i \in [\, u_j\,,\ u_{j+1} \\ 0 & \quad otherwise \end{cases} \tag{2}$$

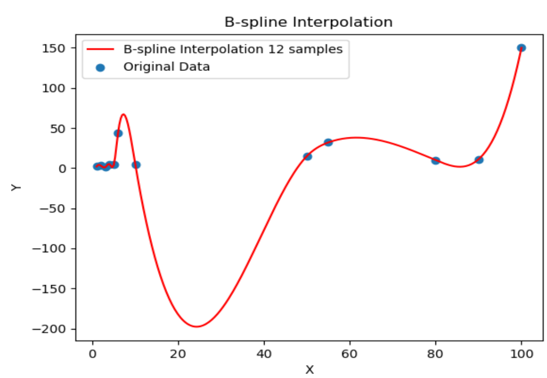
The highest degree at (*P*+1) b-splines are defined by recursion.

$$B_{i,m+1}(t) = \frac{t - u_j}{u_{j+m} - u_j} B_{j,m}(t) + \frac{u_{j+m+1} - t}{u_{j+m+1} - u_{j+1}} B_{j+1,\ m}(t) \tag{3}$$

The implementation in Python 3.7 shows good results at t = 3 and control points at (2-4), first, we tested the curve implementation on randomly generated points at (5, 12) points as shown in figure 3: where (a, b) illustrate the curve at random points (5, 12 points) respectively.



(a)



(b)

Figure 3 (a), shows the b-spline curve in random data generated 5 points, (b) the B-spline curve in random data generated 12 points.

After designing the function of the proposed curve in Python, the curve was fitted into the most valued attributes in the heart attack dataset in order to enhance the data by identifying the outlier values. There are three common techniques after detecting the outliers in the dataset: the first technique is to compute the mean of the attribute column values and replace the outlier. The second technique is to replace the outlier values with (0). The third technique is simply to remove the entire rows of outliers. In the age attribute, we cannot set it as 0, hence, we use the mean instead. The noise in other attributes that contain binary values like (0,1) also cannot utilize the mean or set it as 0 because of its effects on the model performance. So, it must be removed.

B. Data normalization:

This technique is used to reduce the variation in the dataset by converting the data points into form between (0 – 1). The standardization method was used for normalization, employing an effective technique that utilizes both the mean and standard deviation.

C. The Proposed Heart Attack Detection Models

The proposed method integrates B-spline data smoothing with three powerful machine learning algorithms: Decision Tree, Random Forest, and Logistic Regression, each trained independently. The smoothed data enhances the input for these machine-learning models. Decision Trees offer transparency, Random Forest reduces overfitting, and Logistic Regression provides probabilistic insights. This combination aims to optimize the analysis of heart disease attributes, enabling early detection and actionable insights for healthcare professionals and individuals. The dataset was split into three sets 80% for training and 10% for validation sets to ensure the robustness of the proposed models. The training set is utilized to train the proposed models. Subsequently, the validation set helps fine-tune the model parameters, optimizing its performance. Finally, the model is tested on a separate 10% testing set to evaluate its predictive accuracy and generalization capabilities.

D. Evaluation of the Proposed Models

Evaluating model performance is a critical aspect of the proposed methodology. After building and training the model by integrating B-spline data smoothing with machine learning algorithms like Decision Tree, Random Forest, and Logistic Regression, a rigorous assessment is conducted. The classification metrics [60] such as precision (P), recall (R), accuracy (acc), MSE, and F1 score are utilized to quantify the classification models accuracy and effectiveness in predicting heart disease patterns. The formulas for these classification metrics are as follows:

$$acc. = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fn} + \mathrm{fp}} \tag{4}$$

$$P = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \tag{5}$$

$$R = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \tag{6}$$

$$E.rate = \frac{\mathrm{fn} + \mathrm{fp}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fn} + \mathrm{fp}} \tag{7}$$

$$F1\ score = \frac{2(\mathrm{P.R})}{\mathrm{P} + \mathrm{R}} \tag{8}$$

The model is tested on a separate dataset to ensure its ability to generalize beyond the training data. This thorough evaluation aims to provide insights into the model's reliability and suitability for early detection of heart-related issues. Comparisons with related works in the field serve as valuable benchmarks, highlighting the novel contributions and advancements of the proposed approach.

5. RESULTS AND DISCUSSION

B-spline curves can provide a clearer representation of how specific attributes evolve across the dataset. At the pre-processing phase, Figure 4 shows the visualizing of age against chol heart attack risk may reveal non-linear patterns that are not immediately evident in the original data.

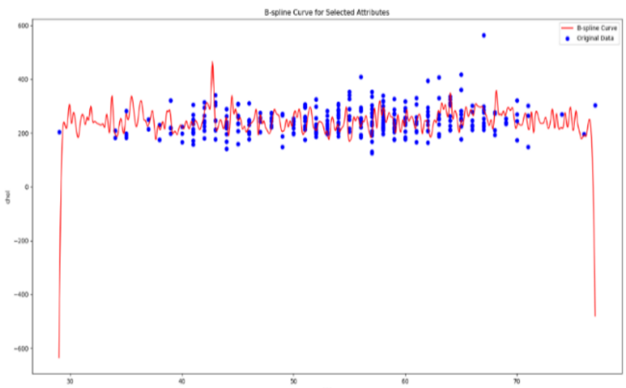
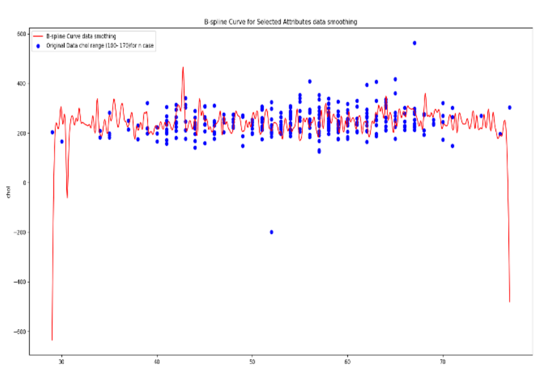
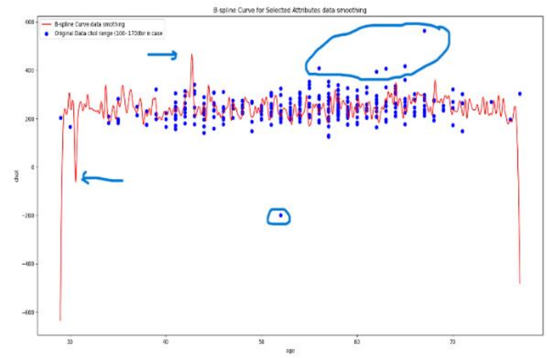


Figure 4. Shows the proposed curve of age and chol attributes

The B-spline can be more affected in the values of the attributes such as (age, chol, and blood pressure) in order to smooth heart attack data points and start to build the machine learning model. This improvement in medical data lies in the small size and also a small number of attributes. Practically, at the control point = 5: figure 5 (a), (b) can easily indicate the noise in this dataset at the blue point that lies over the range of the original data.



(a)



(b)

Figure 5:(a) on the left side shows the noise detection using a cubic B-spline curve at degree 3, while **Figure 5**:(b) on the right side shows how the cubic spline can easily indicate outliers in the heart attack dataset

At CP = 3, figure 6 illustrates the curve of the attribute (age, chol) as (X, Y) which shows another indicator before handling with outliers' points. Figure 7 shows the curve after handling noise and outliers.

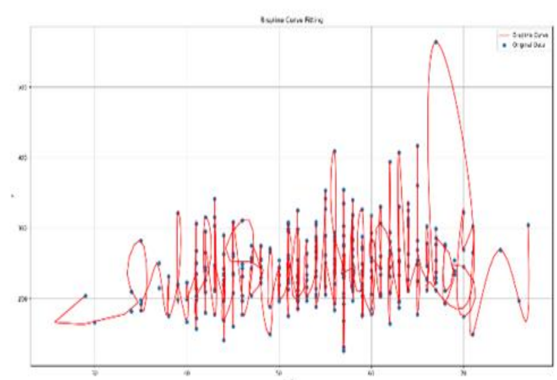


Figure 6. The b-spline curve proposed an indicator before processing the noise and outlier at CP=3.

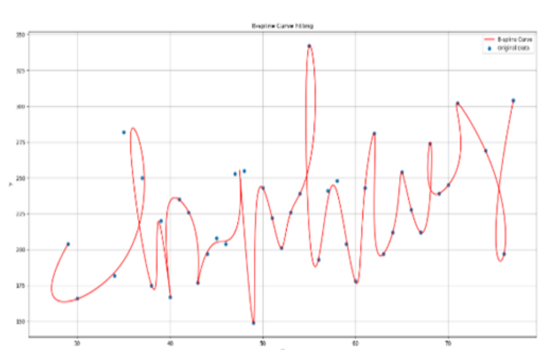
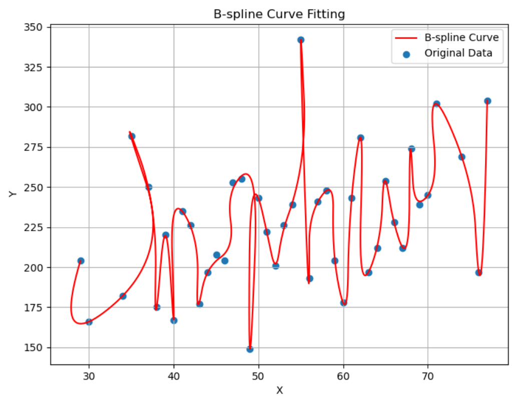
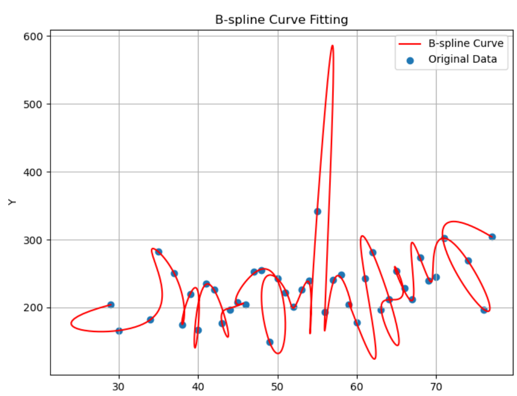


Figure 7. The b-spline curve proposed indicator after processing the noise and outlier at CP=3.

In order to achieve the best fitting point in the proposed curve, we tested on other CPs such as (2, 4) for further improvement, and the results are shown in Figure 8 (a, b).



(a)



(b)

Figure 8. (a) cubic b-spline curve at CP = 2 in, while (b) cubic b-spline curve at CP=4 both curves after handling with noise and outlier.

The curve at CP= 2 shows the best results of fitting points after handling noise and outliers. The obtained results from this experiment of the proposed hybrid technique (combining B-spline curve as data-smoothing and ML algorithms) in the medical dataset, opened the door to investigate other fields such as big data. It aroused curiosity about the effect of this curve with big data and whether it can benefit from handling and detecting noise and outliers. For this reason, we took a sample of large data related to a group of attacks on a specific network. This data is characterized by its large size, amounting to more than 3 million records. After performing the same previous experiment, we found that the curve was too complex to be a good indicator for detecting outliers and noise when the dataset is very large. Figure 9 shows the implanting b-spline curve on the big dataset.

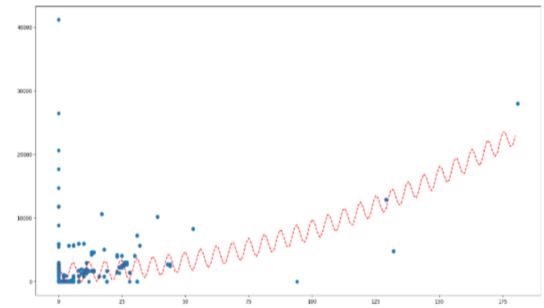


Figure 9. Cubic B-spline curve at CP = 5 in the large distrusted dataset

The suggested machine learning model's results demonstrate how well it can predict patterns of heart disease. The model exhibits high sensitivity for early detection and high precision in identifying positive cases. Its accuracy is further supported by its ability to predict continuous variables with minimal error. Overall, the findings highlight the promising potential of the integrated approach and provide valuable insights for taking preventative action against heart-related issues. Table 2 illustrates the classification metrics of the proposed model, evaluated on a testing set comprising 125 data points also figure 10 shows the comparison of these metrics for both phases.

TABLE 2. THE CLASSIFICATION METRICS OF THE PROPOSED MODELS

ML model	**Accuracy**	**Precision (P)**	**Recall (R)**	**F1 score**	**Error rate**
LR-model	0.7258	0.75	0.8182	0.7826	0.2742
DT- model	0.7581	0.8333	0.7692	0.80	0.2419
RF-model & b-spline curve	**0.9919**	**0.9844**	**1.0**	**0.9921**	**0.0081**

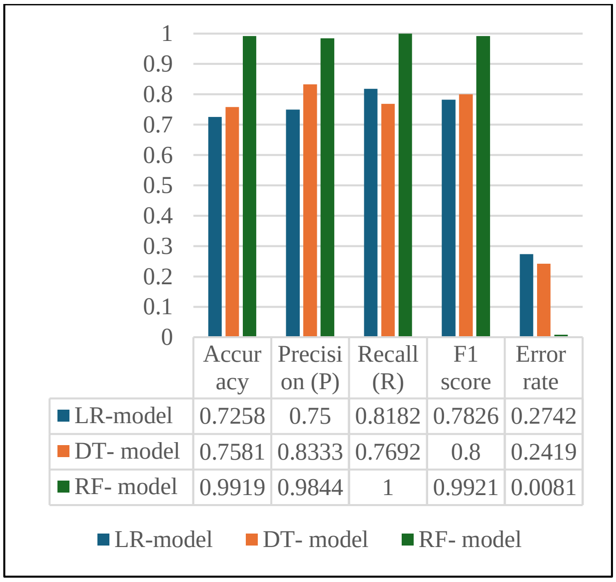


Figure 10. The comparison of prediction results for both proposed phases

The Random Forest model utilized in this study stands out for seamlessly integrating the proposed combination of B-spline curve interpolation and machine learning (ML) techniques. This innovative approach harnesses the power of B-spline curves, specifically tailored to handle the complex and intricate patterns within the heart attack dataset. In contrast, other models such as Logistic Regression and Decision Tree were trained directly on the heart attack dataset without incorporating B-spline interpolation. By leveraging the combined strength of B-spline techniques and ML algorithms, the Random Forest model aims to enhance the accuracy and interpretability of predictions, presenting a unique strategy to address the challenges posed by heart disease data analysis. Table 4 illustrates the comparison results with previous related works.

TABLE 3. THE COMPARISON RESULTS OF THIS WORK AND PREVIOUS RELATED WORKS

Authors	ML Models	Best obtained Accuracy %
In [7]	SVM	85
In [15]	DT, LR, SVM, NB, RF, KNN, XG Boost	SVM (90)
In [43]	RF, SVM, KNN, DT	RF (94.96)
In [44]	KNN	90.16
In [45]	LR	82.4
In [46]	SVM, DT	SVM (82.71)
In [47]	RF, DT, KNN	RF (96.71)
In [48]	LR, SVM, RF, NB, KNN	LR (82)
In [49]	KNN, RF, DT, NB	KNN 90.78
In [23]	KNN and others ML models	KNN 94.1
In [50]	Hybrid RF with a linear model	88.7
In [52]	Feature selection method with DT, SVM	DT 93.19
This work	**RF (based on B-spline)**	**99.19**

6. CONCLUSION AND FUTURE WORKS

This research presents a promising method for predicting heart disease patterns by combining B-spline curve interpolation with machine learning, demonstrated using the algorithm of the Random Forest to create a robust model. The obtained classification results achieved a detecting accuracy of 99.19%, this combined strategy outperforms existing models, and the other detection methods of heart diseases were presented in comparison between the proposed method and related works. The integration of B-spline interpolation enhances precision, capturing intricate patterns in the heart attack dataset. The results not only underscore the model's superiority but also highlight its potential for early detection and intervention in heart attacks.

This combination strategy opens doors for researchers to use mathematical curves in terms of text data enhancement, which is a good indicator for outliers and noise data specifically in medical diagnosis and computer-aided diagnosis systems which are shown in the presented results of this research.

The future research, directions include exploring diverse interpolation techniques, scalability to different medical datasets, and optimizing model parameters for enhanced predictive accuracy.

Another important idea for future direction is that, using b-spline curves to transform the text datasets into images. By transforming each row in the dataset to an image that contains the CP as attributes which are drawing the curve. This technique may be used to train the model based on a convolutional neural network with transformed images instead of text data samples, this proposed method for future work could be resistant to adversarial attacks, as we know that such attacks can poison the model during the training or the testing processes.

Acknowledgment

This work was supported by the University of Technology, College of Computer Science.

References

[1] L. Li, "Application of cubic b-spline curve in computer-aided animation design," *Computer-Aided Design and Applications*, vol. 18, no. S1, pp. 43-52, 2020, doi: 10.14733/cadaps.2021.S1.43-52.

[2] M. Ammad and A. Ramli, "Cubic B-spline curve interpolation with arbitrary derivatives on its data points," in *2019 23rd International Conference in Information Visualization–Part II*, 2019: IEEE, pp. 156-159, doi: 10.1109/IV-2.2019.00038. [Online]. Available: https://ieeexplore.ieee.org/document/8812013/

[3] H. Haron, A. Rehman, D. Adi, S. Lim, and T. Saba, "Parameterization Method on B ‐ Spline Curve," *Mathematical Problems in Engineering*, vol. 2012, no. 1, p. 640472, 2012, doi: 10.1155/2012/640472.

[4] R. Yeh, Y. S. Nashed, T. Peterka, and X. Tricoche, "Fast automatic knot placement method for accurate B-spline curve fitting," *Computer-aided design*, vol. 128, p. 102905, 2020, doi: 10.1016/j.cad.2020.102905.

[5] D. S. Saleh and M. S. Othman, "Exploring the Challenges of Diagnosing Thyroid Disease with Imbalanced Data and Machine Learning: A Systematic Literature Review," *Baghdad Science Journal*, 2023, doi: 10.21123/bsj.2023.8544.

[6] R. Sethi, M. Mehrotra, D. Sethi, and G. Mehrotra, "Medical Image Augmentation Framework for Resolving Chest X-Ray Data Imbalance," *International Journal of Computing and Digital Systems*, vol. 12, no. 1, pp. 1161-1170, 2022, doi: 10.12785/ijcds/120192.

[7] S. Krishnan and S. Geetha, "Prediction of heart disease using machine learning algorithms," in *2019 1st international conference on innovations in information and communication technology (ICIICT)*, 2019: IEEE, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[8] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87-97, 2021, doi: 10.1007/s12553-020-00505-7.

[9] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, vol. 6, no. 4, pp. 854-873, 2018, doi: 10.4236/wjet.2018.64057.

[10] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *2018 4th International conference on computing communication and automation (ICCCA)*, 2018: IEEE, pp. 1-4.

[11] V. Kamra, P. Kumar, and M. Mohammadian, "Diagnosis support system for general diseases by implementing a novel machine learning based classifier," *International Journal of Computing and Digital Systems*, 2021, doi: 10.12785/ijcds/100168.

[12] P. Rani, R. Kumar, N. M. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263-275, 2021, doi: 10.1007/S40860-021-00133-6.

[13] A. J. S. T. Varun Sapra, "Two-Phase Heart Disease Diagnosis System Using Deep Learning," *International Journal of Control and Automation*, vol. 12, no. 5, pp. 558 - 573, 12/31 2019. [Online]. Available: http://sersc.org/journals/index.php/IJCA/article/view/2690.

[14] N. Vithanwattana, G. Mapp, and C. George, "Developing a comprehensive information security framework for mHealth: a detailed analysis," *Journal of Reliable Intelligent Environments*, vol. 3, pp. 21-39, 2017, doi: 10.1007/s40860-017-0038-x.

[15] K. Wang *et al.*, "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP," *Computers in biology and medicine*, vol. 137, p. 104813, 2021, doi: 10.1016/j.compbiomed.2021.104813.

[16] H. Abubaker, F. Muchtar, A. R. Khairuddin, A. N. A. Nuar, Z. M. Yunos, and C. Salimun, "Exploring Important Factors in Predicting Heart Disease Based on Ensemble-Extra Feature Selection Approach," *Baghdad Science Journal*, vol. 21, no. 2 (SI), pp. 0812-0812, 2024, doi: 10.21123/bsj.2024.9711.

[17] H. Hang, X. Yao, Q. Li, and M. Artiles, "Cubic B ‐ Spline Curves with Shape Parameter and Their Applications," *Mathematical Problems in Engineering*, vol. 2017, no. 1, p. 3962617, 2017, doi: 10.1155/2017/3962617.

[18] E. Evans, M. Scott, X. Li, and D. Thomas, "Hierarchical T-splines: Analysis-suitability, Bézier extraction, and application as an adaptive basis for isogeometric analysis," *Computer Methods in Applied Mechanics and Engineering*,

vol. 284, pp. 1-20, 2015, doi: 10.1016/j.cma.2014.05.019.

[19] L. Xumin, X. Weixiang, G. Yong, and S. Yuanyuan, "Hyperbolic polynomial uniform B-spline curves and surfaces with shape parameter," *Graphical Models*, vol. 72, no. 1, pp. 1-6, 2010, doi: 10.1016/j.gmod.2009.10.001.

[20] H. Prautzsch, W. Boehm, and M. Paluszny, *Bézier and B-spline techniques*. Springer, 2002.

[21] A. Chaudhuri, "B-splines," *arXiv preprint arXiv:2108.06617*, 2021, doi: 10.12785/ijcds/120192.

[22] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, and B. Alshraideh, "Enhancing heart attack prediction with machine learning: A study at jordan university hospital," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 5080332, 2024, doi: 10.1155/2024/5080332.

[23] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational intelligence and neuroscience*, vol. 2021, no. 1, p. 8387680, 2021, doi: 10.22452/mjcs.sp2022no1.10.

[24] W. A. Belson, "Matching and prediction on the principle of biological classification," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 8, no. 2, pp. 65-75, 1959, doi: 10.2307/2985543.

[25] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[26] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381-386, 2020, doi: 10.21275/ART20203995.

[27] G. R. Banu, "A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease," *International Journal of Computer Sciences and Engineering*, vol. 4, no. 11, pp. 64-70, 2016. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/51273936/22-IJCSE-01866-libre.pdf?1484024803=&response-content-disposition=inline%3B+filename%3DA_Role_of_decision_Tree_classification_d.pdf&Expires=1720317429&Signature=Zp4jGO9zA~Vd1auWtrQuhGHJEQcaC4uuylPAU9si00bI~tGfXMnLgYc~PDUIO713ZFunaKTIYA1sC14m9sy7s~BST6BUtWXAaBSoh-j5UVIsAlYTwXuRsa~OW4itPNeRWl2rObTnFSe~C0sodq87YpBL9a-xnIl59tPjXTe3R3~rD9vQNYOHULw9V4xznezTBInQnB8UHACbQ0F9aB3JGNucVgL5ueI3qIpk0aCVic0v3qjqcUgIi6sawOCzccy5~ncL0DMzkFILudHEwlZUn6N~wjCHRG5tUU2cwYZLtAAKPv3wgHbzlv4TiHTLbjk7bzTDp1T80R25DqHYajkn8g__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

[28] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, and A. Tkachenko, "Classification and Prediction of Diabetes Disease using Decision Tree Method," in *IT&AS*, 2021, pp. 163-172. [Online]. Available: https://ceur-ws.org/Vol-2824/paper16.pdf. [Online]. Available: https://ceur-ws.org/Vol-2824/paper16.pdf

[29] M. Shouman, T. Turner, and R. Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients," *AusDM*, vol. 11, pp. 23-30, 2011. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/8174511/Using_decision_trees_in_diagnosing_heart_disease_patients-libre.pdf?1390854454=&response-content-disposition=inline%3B+filename%3DUsing_Decision_Tree_for_Diagnosing_Heart.pdf&Expires=1720317660&Signature=TpnIY1XUFrkPVqlAQTX5z-BkKq~XQX7kNeXfuJQxVaT7hBtqb5zSlRe8PVtXDNZ-Dl~wK7hWYS~qu~4lpndN2phdAPdkGhEvaLqFI0XNAAvroKNPSOvuFLOXUSxlwuTyv~nyqQbJeQ0VsuEDZEGhYwBfCBYb5jH0OAKX3BE9xqBLbXn4D8bJ0rQ~FP2BIF-bfdrcSayo7ci785W5BCPOJxhO4eL~vVB2dbN6nBWQ2MUNg8dxq4A7b15pm9~miBQhwS2cxZiR9Euje1LyrJbGTT8oFEJc~3I7BwAfEEmTdetcMBBt~d1ovh7YqLE2nK0RljLnAAkQMHQimPipFzvk6w__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

[30] J. Keerthika, D. Sruthi, D. Swathi, S. Swetha, and R. Vinupriya, "Diagnosis of breast cancer using decision tree data mining technique," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2021, vol. 1: IEEE, pp. 1530-1535, doi: 10.1155/S1110724303210032.

[31] C. W. Olanow, R. L. Watts, and W. C. Koller, "An algorithm (decision tree) for the management of Parkinson's disease (2001): treatment guidelines," *Neurology*, vol. 56, no. suppl_5, pp. S1-S88, 2001, doi: 10.1212/WNL.56.suppl_5.S1.

[32] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," in *2012 Japan-Egypt Conference on Electronics, Communications and Computers*, 2012: IEEE, pp. 173-177, doi: 10.1109/JEC-ECC.2012.6186978.

[33] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in

International conference on intelligent data communication technologies and internet of things (ICICI) 2018, 2019: Springer, pp. 758-763, doi: 10.1007/978-3-030-03146-6_86.
[34] S. Angra and S. Ahuja, "Machine learning and its applications: A review," in *2017 international conference on big data analytics and computational intelligence (ICBDAC)*, 2017: IEEE, pp. 57-60, doi: 10.1109/ICBDACI.2017.8070809.
[35] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," *Baltic Journal of Modern Computing*, vol. 5, no. 2, p. 221, 2017, doi: 10.22364/bjmc.2017.5.2.04.
[36] N. Lay *et al.*, "Detection of prostate cancer in multiparametric MRI using random forest with instance weighting," *Journal of Medical Imaging*, vol. 4, no. 2, pp. 024506-024506, 2017, doi: 10.1117/1.JMI.4.2.024506.
[37] A. Murugan, S. A. H. Nair, and K. S. Kumar, "Detection of skin cancer using SVM, random forest and kNN classifiers," *Journal of medical systems*, vol. 43, no. 8, p. 269, 2019, doi: 10.1007/s10916-019-1400-8.
[38] Z. G. Merali, C. D. Witiw, J. H. Badhiwala, J. R. Wilson, and M. G. Fehlings, "Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy," *PloS one*, vol. 14, no. 4, p. e0215133, 2019, doi: 10.1371/journal.pone.0215133.
[39] T. Rymarczyk, E. Kozłowski, G. Kłosowski, and K. Niderla, "Logistic regression for machine learning in process tomography," *Sensors*, vol. 19, no. 15, p. 3400, 2019, doi: 10.3390/s19153400.
[40] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *Journal of clinical epidemiology*, vol. 110, pp. 12-22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.
[41] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, pp. 685-693, 2018, doi: 10.1007/s00521-016-2604-1.
[42] A. Naess, S. S. Nilssen, R. Mo, G. E. Eide, and H. Sjursen, "Role of neutrophil to lymphocyte and monocyte to lymphocyte ratios in the diagnosis of bacterial infection in patients with fever," *Infection*, vol. 45, pp. 299-307, 2017.
[43] M. A. Kadhim and A. M. Radhi, "Heart disease classification using optimized Machine learning algorithms," *Iraqi Journal For Computer Science and Mathematics*, vol. 4, no. 2, pp. 31-42, 2023, doi: 10.52866/ijcsm.2023.02.02.004.
[44] M. Rizwan, S. Arshad, H. Aijaz, R. A. Khan, and M. Z. U. Haque, "Heart attack prediction using machine learning approach," in *2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)*, 2022: IEEE, pp. 1-8, doi: 10.1109/INTELLECT55495.2022.9969395. [Online]. Available: https://ieeexplore.ieee.org/document/9969395/
[45] J. S. Rose, P. M. Bruntha, S. Selvadass, and M. Rajath, "Heart Attack Prediction using Machine Learning Techniques," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2023, vol. 1: IEEE, pp. 210-213, doi: 10.1109/ICACCS57279.2023.10113045. [Online]. Available: https://ieeexplore.ieee.org/document/10113045/
[46] C. Sharma, V. Batra, M. Manchanda, D. Goyal, and P. Gupta, "Analytic Approach for Heart Disease Prediction using Supervised Machine Learning Algorithms," *Turkish Online Journal of Qualitative Inquiry*, vol. 12, no. 8, 2021. [Online]. Available: https://openurl.ebsco.com/EPDB%3Agcd%3A9%3A16065673/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A160601216&crl=c.
[47] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of coronary heart disease using supervised machine learning algorithms," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 2019: IEEE, pp. 367-372, doi: 10.1109/TENCON.2019.8929434. [Online]. Available: https://ieeexplore.ieee.org/document/8929434/
[48] B. K. Mengiste, H. K. Tripathy, and J. K. Rout, "Analysis and prediction of cardiovascular disease using machine learning techniques," in *International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy*, 2020: Springer, pp. 133-141, doi: 10.1007/978-981-15-8685-9_13. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-8685-9_13
[49] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques,"

SN Computer Science, vol. 1, no. 6, p. 345, 2020, doi: 10.1007/s42979-020-00365-y.

[50] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[51] M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, and P. Singh, "Latest trends on heart disease prediction using machine learning and image fusion," *Materials today: proceedings*, vol. 37, pp. 3213-3218, 2021, doi: 10.1016/j.matpr.2020.09.078.

[52] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019, doi: 10.14569/IJACSA.2019.0100637.

[53] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligence in Medicine*, vol. 128, p. 102289, 2022, doi: 10.1016/j.artmed.2022.102289.

[54] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Prediction of heart disease using random forest and feature subset selection," in *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015) held in Kochi, India during December 16-18, 2015*, 2016: Springer, pp. 187-196, doi: 10.1007/978-3-319-28031-8_16. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-28031-8_16

[55] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," *Computer methods and programs in biomedicine*, vol. 141, pp. 19-26, 2017, doi: 10.1016/j.cmpb.2017.01.004.

[56] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two‐tier classifier ensemble," *BioMed Research International*, vol. 2020, no. 1, p. 9816142, 2020, doi: 10.1155/2020/9816142.

[57] M. A. Sufian, "heart.csv," *IEEE Dataport*, September 28 2023, doi: 10.21227/ha3p-pa05.

[58] J. Manning, "Continuity conditions for spline curves," *The computer journal*, vol. 17, no. 2, pp. 181-186, 1974, doi: 10.1093/comjnl/17.2.181.

[59] A. M. S. Rahma, A. Abdulla, and A. AbdulHossen, "Modified Method for Generating B-Spline Curves of Degree Three and Their Controlling," *Eng. & Tech. Journal*, vol. 30, no. 1, pp. 109-125, 2012. [Online]. Available: https://www.iasj.net/iasj/download/06d6ba6d5be2a76a.

[60] Ž. Vujović, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599-606, 2021, doi: 10.14569/IJACSA.2021.0120670

Dr. Mohammed Abdul Ameer Jabbar received a B.Sc Degree in computer science from the University of Digla, M.Sc. Degree in computer science from Al-Nahrain University College of Science, Ph.D. University of Technology. His work at UOITC as a Faculty member and his research interests are in Machine Learning, Pattern Recognition, Neural Networks, Artificial Intelligence, Computer Vision, Image Recognition, Image Segmentation, Data Clustering, Data Mining, and Deep Learning.

Dr. Abdul Mohssen Jaber Almaaly Professor currently works at Al-Turath University and is a specialist in Applied Mathematics, Statistics, Analysis, Curves, and Mathematical Models.

Dr. Mustafa Tareq Abd received the B.Sc. degree in computer science from ALTurath University College, Iraq, in 2012, the M.Sc. degree in network technology from the National University of Malaysia (UKM), Malaysia, in 2016, and the Ph.D. degree in computer science from the Centre of Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM) in 2021. His research interests include the Internet of Things, data mining, clustering algorithms, evolving data streams, computer networks, and Telecommunications.

Dr. Abdulkareem Merhej Radhi Professor at Al-Nahrain University has research interests in Data Mining and Knowledge Discovery, Machine Learning, Text Mining, Natural Language Processing, Pattern Recognition, Classification, Data Clustering, Neural Networks and Artificial Intelligence, Feature Extraction, Feature Selection, Prediction.