IRAQI Academic Scientific Journals
العراقية المجلات الأكاديمية العلمية
University of Kut Journal
ISSN (E): 2616 - 7808 II ISSN (P): 2414 - 7419
js.uokut.edu.iq
k.u.c.j.sci@alkutcollege.edu.iq
Vol. 11 , Issue 1 , June 2026

Using K-nearest neighbor (KNN) machine learning algorithm to diagnose heart diseases in (Alkarama Teaching Hospital/ Iraq)

Hatem Abdul Hussein Ali Alquraishi [1] , Prof. Abbas Lafta Kneehr [2]

Abstract

Despite the presence of many research papers related to the prediction of heart disease through the use of artificial intelligence, the significance of heart disease and the extent of its impact on the lives of people lead to the fact that the researchers must continue to work on building the best program to predict heart disease from an early age to avoid its development, in this paper the researchers are using the K-nearest neighbor algorithm for the prediction of heart disease. The dataset used in this paper is collected from local hospitals in Iraq and was cleaned and analyzed by the researchers. The diagnosis system that is proposed in this work is to assist physicians to diagnose heart conditions by converting medical factors of the patients in to numerical representations, the simulation results show that the proposed K nearest neighbor classifier has 86% accuracy in classifying 4 medical heart conditions when using controlled databases. The researchers concluded that the most common heart conditions that are found in Iraq are the ones classified in this algorithm, which are Coronary heart disease, Congestive Heart failure, and Arrhythmias. The researchers conclude that there is a direct relationship between the amount of training data and the algorithm's accuracy and an inverse relationship between the number of neighbors K and the algorithm's accuracy.

Keywords: machine learning, K-nearest neighbor, healthcare, heart disease, diagnosis

Affiliation of Authors
[1, 2] College of Administration & Economics, University of Wasit, Iraq, kut, 52001

[1] h_alqurashi@uowasit.edu.iq
[2] alafta@uowasit.edu.iq

[1] **Corresponding Author**

Paper Info.
Published: Jun. 2026

استعمال خوارزمية أقرب الجيران (K-Nearest Neighbor, KNN) في تشخيص أمراض القلب في مستشفى الكرامة التعليمي / العراق

حاتم عبد الحسين علي القريشي [1] ، د. عباس لفته كنيهر [2]

المستخلص

على الرغم من وجود عدد كبير من الدراسات البحثية المتعلقة بالتنبؤ بأمراض القلب باستعمال تقنيات الذكاء الاصطناعي، فإن خطورة أمراض القلب واتساع نطاق تأثيرها في حياة الأفراد تفرض على الباحثين الاستمرار في تطوير أفضل النماذج القادرة على التنبؤ المبكر بهذه الأمراض بهدف الحد من تطورها. يستخدم الباحثون خوارزمية (K-Nearest Neighbor – KNN) للتنبؤ بأمراض القلب. تم جمع مجموعة البيانات المستخدمة في هذه الدراسة من مستشفيات محلية في العراق، حيث قام الباحثون بتحليلها إحصائيًا. ويهدف نظام التشخيص المقترح في هذا العمل إلى مساعدة الأطباء في تشخيص الحالات القلبية من خلال تحويل العوامل الطبية الخاصة بالمرضى إلى تمثيلات عددية قابلة للمعالجة الحاسوبية. أظهرت نتائج المحاكاة أن مصنف KNN المقترح حقق دقة بلغت 86% في تصنيف أربع حالات مرضية قلبية عند استعمال قواعد بيانات. وقد توصل الباحثون إلى أن أكثر أمراض القلب شيوعًا في العراق هي تلك التي شملها التصنيف في هذه الخوارزمية، وهي: (Coronary Heart Disease)، و (Congestive Heart Failure)، (Arrhythmias). كما خلص الباحثون إلى وجود علاقة طردية بين حجم بيانات التدريب ودقة الخوارزمية، في مقابل علاقة عكسية بين عدد الجيران K ودقة الخوارزمية.

الكلمات المفتاحية: تعلّم الآلة، خوارزمية أقرب الجيران، الرعاية الصحية، أمراض القلب ، التشخيص

انتساب الباحثين
[1، 2، 3] كلية الادارة والاقتصاد، جامعة واسط، العراق، الكوت، 52001

[1] h_alqurashi@uowasit.edu.iq
[2] alafta@uowasit.edu.iq

[1] **المؤلف المراسل**

معلومات البحث
تأريخ النشر : حزيران 2026

Introduction

Heart diseases are the leading cause of death worldwide, The Iraqi Ministry of Health in their annual report, reported 140,621 deaths in 2019, 28.83% of those deaths were due to heart

disease, while cancer related deaths were only 9.33% and road traffic accidents accounted for 4.9%. Motivated by the worldwide increase in mortality of heart disease patients each year, and the availability of gigantic volumes of healthcare data, which can be used to extract useful knowledge, to diagnose heart diseases with high accuracy to help physicians make informed decisions when treating patients. Researchers are increasingly adopting machine learning in diagnosing heart diseases [1], the researchers wanted to examine the accuracy of machine learning methods in diagnosing heart diseases to determine the most suitable way for diagnosing [2]. The researcher chose K-Nearest Neighbors (KNN) Algorithm as the algorithm of choice to be tested as a machine learning algorithm. The researchers chose KNN for multiple reasons one of the reasons is its ease of use compared to other algorithms, the other reason is its effectiveness based on multiple studies which examined the effectiveness of KNN against other machine learning algorithms most concluded that KNN was the most effective algorithm in diagnosing heart disease. The researchers used a dataset collected from local Iraqi hospitals that consists of 201 patients.

Research Problem

Though machine learning approaches have broadly found their application in the diagnosis of heart disease, most works available depend on publicly available datasets and do not depict local population characteristics. In this view of the Iraqi healthcare scenario, data-driven diagnostic systems based on locally collected clinical data are scarce for such models to apply in the real clinical environment. Besides, choosing a suitable algorithm among various options with reasonable accuracies that are simple and interpretable for medical use is another challenge. Therefore, this study frames and tests the problem of building a K-Nearest Neighbor (KNN)-based diagnostic model for heart diseases using actual patient data obtained from Alkarama Teaching Hospital in Iraq.

Significance of the Study

This study is practically and scientifically important. In practice, the proposed system will help physicians in diagnosing common heart diseases by providing another decision support based on clinical data about the patient. On a scientific note, This study contributes to the existing literature by evaluating the performance of the KNN algorithm through a dataset collected locally rather than standardized international datasets. Other findings include how training data size relates to the number of neighbors (K) and classification accuracy, which can guide further research as well as system development in similar healthcare environments.

Related work

[3]: conducted a comparative study of five machine-learning algorithms for predicting heart disease using the UCI dataset. The algorithms chosen by the researchers were K-nearest neighbor (KNN), Support Vector Machine Algorithm (SVM), decision tree, logistic regression, gradient descent. And the results showed that the algorithm with the highest accuracy was KNN with an accuracy of 85.7%. And the lowest algorithm for diagnosing heart diseases was gradient descent which had an accuracy of 73.63%.

[4]: did a correlation analysis using Pearson correlation to find the effective data in heart failure when using K-nearest neighbor (KNN) as a

machine learning algorithm for diagnosing heart diseases, the researchers concluded that highly correlated features significantly affected the performance of KNN, and the KNN model had a high accuracy of 97.07% in diagnosing heart failure using a dataset from Kaggle data repository .

[5]: published a study comparing two supervised machine learning algorithms which are K-Nearest Neighbor (KNN) and Random Forest in diagnosing heart diseases, the experimental results obtained by the researchers showed that KNN had the highest accuracy of 86.885% compared to random forest which had an accuracy of 81.967%.

[6]: published a detailed study looking into the ability of using KNN in diagnosing Coronavirus Disease 2019 their results showed that the algorithm achieved high accuracy and stable outcomes compared to other machine learning algorithms they tested in diagnosing Covid-19.

Background

Machine learning is a part of artificial intelligence that has algorithmic approaches which allow machines to solve problems without explicit computer programming [7]. Machine learning earned its role in medicine as it gave physicians a way to optimize clinical care of patients with chronic diseases and allowed them to make informed decisions in diagnosing and treating those patients [8]. Figure 1 shows the amount of research articles that included machine learning in some way or another into the medical field. The trend shows an exponential growth for those scientific articles especially in the fields of diagnostics and drug discovery. Artificial intelligence is expected to save the United States over 150 billion US dollars in healthcare spending by the year 2026 [9].

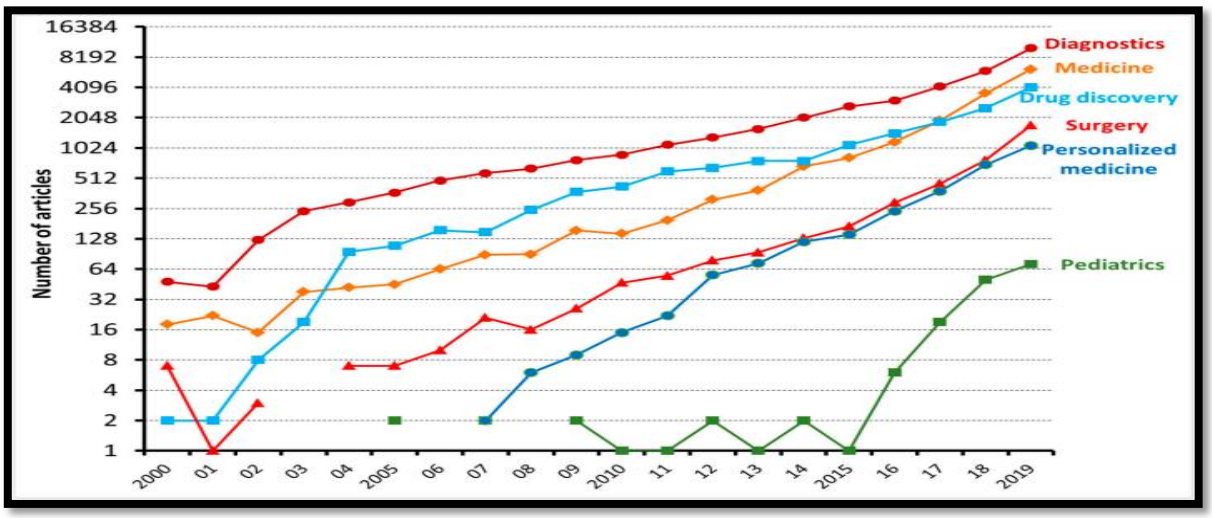

Figure (1): Number of articles, reviews, and editorials [7].

The way artificial intelligence learns to interpret data is by using multiple layers of nonlinear processing units. This way it can make predictions or classify records. In this way, artificial intelligence can analyze electronic medical records that have information that physicians normally skip or do not give attention to and that algorithm makes connections with those patterns from millions of similar cases. Most hospitals possess a vast amount of valuable data that can be used to diagnose, treat, and discover new drugs and the

only way to harvest that power is by using artificial intelligence.

K-Nearest Neighbors (KNN)

Is a non-parametric machine learning method developed originally by Evelyn Fix and Joseph Hodges in 1951. In spite of its ease of use, KNN continues to perform effectively with large training datasets. It operates based on the fundamental assumption underlying all available predictions, which is the fact that observations with similar attributes usually have similar outcomes. KNN predicts a value for each new observation based on the mean of its K which is the number of neighbors that are calculated in the training set. If we take a hypothetical example thinking about the fact that there is an infinite amount of data, thus any observation will have thousands of neighbors that are very close to each other in terms of attributes or characteristics so the outcome will provide the best possible prediction that can be provided [10].

K-Nearest Neighbors is an uncomplicated, genuine data mining method that is sometimes referred to as memory-based classifier because the training data need to be stored in the memory when the classification is being done [11]. When training KNN with continuous data a mathematical formula is needed to calculate the distance between its neighbors there are 8 major distance families and those eight have a total of 54 ways to calculate distance.

Dataset

A large number of relevant inputs must be considered during the diagnosis, in order to diagnose a patient with heart disease with high accuracy. Physicians usually depend on all recorded symptoms, patient medical history, medical examination and laboratory results, as well as the physician's own experience .

The Cleveland database from UCI Machine Learning Repository includes sufficient and carefully chosen attributes for heart disease diagnosis, thus, Data were collected from 201 patients in Alkarama Teaching Hospital was based on those medical attributes with the cardiologist diagnosis of them as either being normal, having coronary heart disease, having arrhythmias, or having congestive heart failure. However, the researchers had to make some modifications to the attributes with the help of a cardiologist, due to the difficulty of obtaining them. The new dataset with the modified attributes is laid out in Table (1) below.

Table (1): The 14 attributes and their data type

1	**Age**	**Continuous Data**	
2	Sex	Female “0”	Male “1”
3	Smoking	Yes “0”	No “1”
4	Previous family history of HD	Yes “0”	No “1”
5	Previous attack of angina	Yes (positive) “0”	No (negative) “1”
6	Fasting blood sugar if > 120 mg/dl	Yes “0”	No “1”
7	Serum cholesterol in mg/dl.	Normal “0”	Abnormal “1”
8	Systolic blood pressure	Continuous Data	

9	Diastolic blood pressure	Continuous Data			
10	HR: heart rate achieved.	Continuous Data			
11	Chest pain	Typical angina “0”	Atypical angina “1”	Non-angina pain “2”	Asymptomatic “3”
12	Electrocardiographic results	Normal “0”	ST-T wave abnormality "1"	probable or definite left ventricular hypertrophy "2"	
13	The angina is exercise induce	Yes“0”	No“1”		
14	Echo finding for hypokinesia	Yes “0”	No “1”		

Some of the gathered descriptive attributes were changed into their respective binary forms such as the sex, smoking, family history of heart disease etc... By this way "Yes" was coded as 0 and "No" was coded as 1. For the case of the chest pain attribute, it needed more than two values as it had four descriptive values which are: Typical angina, Atypical angina, Non-angina pain, Asymptomatic so It was changed to numbers ranging from zero to three. Meanwhile, some of the attributes were continuous, a binary formula will not work on such attributes such as age, blood pressure and heart rate. A problem arises when these continuous values are retained without altering them which is the bias of large numbers in the data such as the age and the heart rate which reaches a maximum value of 145 compared to more important attributes such as chest pain which only maxes at 3. For this reason, those attributes were normalized with Z-Score normalization [12]. The Z-Score normalization process was made using SPSS, and the results (Descriptive statistics) are shown in Table (2).

Table (2): The Distributive’s statistics

Continuous attributes	N	Minimum	Maximum	Mean	Std. Deviation
Age	201	31	95	58.95	14.957
systolic blood pressure	201	40	200	120.95	23.272
diastolic blood pressure	201	30	100	73.11	13.496
heart rate achieved	201	32	145	88.53	16.896

As seen in the Table (2), age for example has a minimum value of 31 and a maximum value of 95 with a mean of 58.95 and a standard deviation of 14.957. Keeping the data without normalization will grant a bigger weight to age than it should, compared to an attribute like smoking which is a binary variable. After data normalization, the age had a max of 2.41020 and a min of -1.86868. Systolic blood pressure had a max of 3.39694 and a min of -3.47817, diastolic blood pressure had a max of 1.99217 and a min of -3.19470, and heart rate had a max of 3.34198 and a min of -3.34581 as shown in Table (3).

Table (3): z-score attributes

Continuous attributes	Min	Max
Age	-1.86868	2.41020
systolic blood pressure	-3.47817	3.39694
diastolic blood pressure	-3.19470	1.99217
heart rate achieved	-3.34581	3.34198

Despite the presence of many research papers related to the prediction of heart disease through the use of artificial intelligence, the significance of heart disease and the extent of its impact on the lives of people lead to the fact that the researchers must continue to work on building the best program to predict heart disease from an early age to avoid its development, in this part of the project, we use the K-nearest neighbor algorithm for the prediction of heart disease. The testing and prediction operation is laid out in the following flowchart as shown in Figure (2).

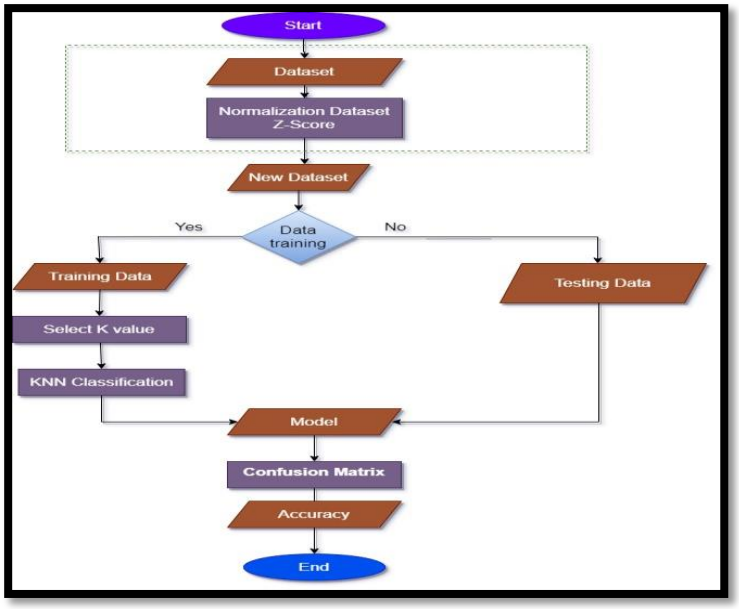

Figure (2): the KNN testing and prediction operation

The use of K-Nearest Neighbor for heart disease diagnosis

The collected dataset consists of 201 subjects, each having 14 parameters and divided into four categories of diagnosis which are coronary heart disease, congestive heart failure, arrhythmias and normal, the data were collected locally and are therefore representative of the local population, the approach for the diagnosis is the use of the K-nearest neighbor algorithm for classification.

The data were divided multiple times; each time, they were split into two groups which are training and testing. The first test used a training group which had 75% of the full dataset and the second group had the remaining 25%, a special code was used to randomly distribute the patients in the dataset depending on the required percentages for example put randomly 75% of the data into group A and put the rest in group B. Filling the groups in this way will eliminate some of the bias which arise from the researcher choosing patients that

make more sense to the algorithm and removing patients with asymptomatic symptoms. Then the process starts with modeling the algorithm. One of the features of using MATLAB is that it offers the function "fitcknn" which allows us to use the K nearest neighbor algorithm and control its parameters easily. The algorithm was trained by dividing the training data which has 15 parameters into two sections the first being the parameters which are the first 14 columns and the second being the target which is the number of the class or the disease which would be the 15th column. After that we set the algorithm's parameters which are the number of neighbors and the distance measurement type. There is no standard formula for selecting the optimal number of neighbors or K thus, the experimental method was used by the researchers. In this test a K value from one to ten was tested on the same data groups and the accuracy was measured each time as laid out in Table (4).

Table (4): the test with 75% training data and 25% testing data

75% Training	25% Testing	K value	Accuracy
151	**50**	k=1	86
		k=2	78
		k=3	66
		k=4	56
		k=5	48
		k=6	44
		k=7	36
		k=8	30
		k=9	28
		k=10	28

As seen from the results the algorithm's accuracy decreased every time the K value increased like when K was one the algorithm's accuracy was the highest at 86% and at K equals nine which means taking nine of closest neighbors to the test value the algorithm's accuracy became 28% a sharp decrease due to the number of neighbors.

The second test had an equal split putting 50% of the data into training and the other 50% into testing the algorithm and in the same time testing the value of K which ranged from one to ten as shown in Table (5).

Table (5): the test with 50% training data and 50% testing data

50% Training	50% Testing	K value	Accuracy
100	**101**	k=1	83.1683
		k=2	71.2871
		k=3	60.3960
		k=4	49.5050

		k=5	38.6139
		k=6	34.6535
		k=7	29.7030
		k=8	25.7426
		k=9	22.7723
		k=10	19.8020

In this experiment a decrease in the algorithm's accuracy was seen and it will become a common pattern in the upcoming experiments as well. When we had 100 training samples the accuracy was 83.1% compared to when we had 151 training samples the accuracy was 86%. Every time the training data is reduced the algorithm's accuracy suffers. And The K values results was similar to the first experiment. Every time the number of neighbors increased the algorithm's accuracy decreased .

The third test put the highest percentage into the testing group which had 75% of the data and the training group had the remaining 25% As shown in Table (6).

Table (6): the test with 25% training data and 75% testing data

25% Training	75% Testing	K value	Accuracy
50	**151**	k=1	78.1457
		k=2	60.9272
		k=3	49.0066
		k=4	37.0861
		k=5	33.1126
		k=6	28.4768
		k=7	24.5033
		k=8	18.5430
		k=9	17.2185
		k=10	13.9073

The results were similar to those of the previous tests the accuracy suffers with the decrease in the amount of training data and it suffers with the increase of the number of neighbors for example the highest accuracy recorded was 78.1% when K was equal to one a 7.9% decrease in accuracy compared to the first test .

The fourth and final test gave much higher weight into testing which was 80% of the whole dataset and gave only 20% to training as shown in Table (7).

Table (7): the test with 20% training data and 80% testing data

20% Training	80% Testing	K value	Accuracy
40	**161**	k=1	73.9130
		k=2	50.3106
		k=3	40.9938
		k=4	31.0559
		k=5	29.1925
		k=6	24.8447
		k=7	21.7391
		k=8	14.9068
		k=9	10.5590
		k=10	8.0745

This test showed the worst-case scenario where the highest weight was given to testing not training the accuracy had a sharp decrease to 8% when the number of neighbors was 10 which is the lowest recorded accuracy in all four tests.

The algorithm's accuracy was calculated for each test following this method which is: After the training is completed, the researchers built a new matrix based on the number of classes they had which means a 5x5 matrix then the MATLAB's Predict function is used to check whether the algorithm correctly predicted the classes or not and it's done by putting the model that have been trained and comparing it with column number 15 which is the class or disease type then the results are saved in the new 5x5 matrix. For example equation (1).

$$\boldsymbol{R} = \begin{bmatrix} \mathbf{26} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{14} & \mathbf{3} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Ans} = \mathbf{86} \qquad (\mathbf{1})$$

The following matrix which is the results of the first experiment the rows here represent the classes, and the columns represent what the algorithm has predicted. It predicted that there are 26 subjects belong to class number 1 without any wrong predictions and it predicted 14 subjects belong to class number 2 and had a mismatch of 4 subjects putting them in class 2 wrongfully and so on. Calculating the Main diagonal and divide it by the number of patients. This results in the algorithm's accuracy as shown in equation (2).

$$\mathbf{algorithm's\ accuracy} = \left(\frac{\mathbf{sum(diag(R))}}{\boldsymbol{n}}\right) * \mathbf{100} \qquad (\mathbf{2})$$

The prediction algorithm

Now that the training is complete and the best value of K was chosen the final phase of the algorithm can begin which is building the diagnosis algorithm. With the use of Excel's Kutools add-in the full dataset was stripped from the last column which is the diagnosis column and then the data was split into 201 spreadsheets each sheet contains one row of data each row has the attributes of one patient and then each spreadsheet

was saved as a (.xlsx) file. The reason behind this is that the data can be fed into the testing algorithm one patient at a time and at the same time that data is stripped from the diagnosis column so it's considered foreign to the algorithm. All this can be swapped with an easy-to-use user interface that allows physicians to enter the data easily and by the press of one button it can do all the above automatically but this work is out of the scope of this research as it requires backend and front-end programming experience.

Here an individual with the following attribute was fed into the algorithm as shown in Figure (3).

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Untitled													
VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14
Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
0.5382	1	0	1	1	1	0	0.3891	0.5102	0.2644	0	1	0	0

Figure (3): patient data that is stripped from the diagnosis column

The algorithm output was: as shown in Figure (4).

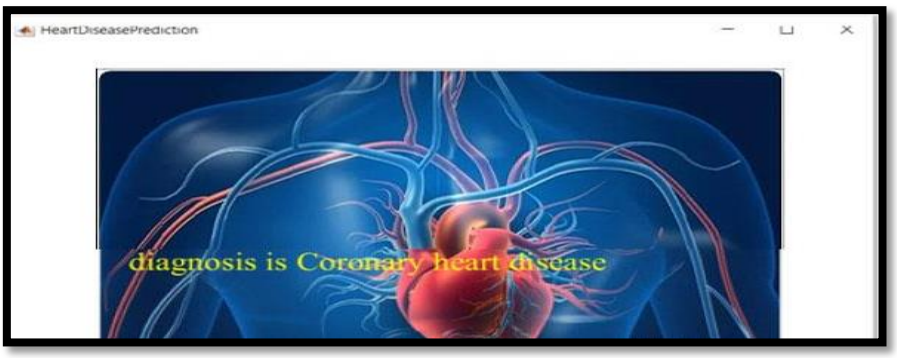

Figure (4): patient diagnosis with coronary heart disease

By comparing the algorithm's diagnosis with the physician's diagnosis of the same patient as shown in Figure (5).

ZAge	Sex	Smoking	FHHD	PERANG	FBS	CHOL	ZBPH	ZBPL	ZHR	ChestPain	ECG	Exang	HYP	Heartdiseasediagnose
0.53819	1	0	1	1	1	0	0.38908	0.51021	0.26441	0	1	0	0	2

Figure (5): physician's diagnosis of the same patient

The patient has coronary heart disease which means that the algorithm's prediction was correct.

In this test the patient with the following attributes was imported to the algorithm as shown in Figure (6).

VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14
Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
0.7388	0	1	1	0	1	0	-0.9000	-0.9718	-0.6825	0	0	0	1

Figure (6): patient data that is stripped from the diagnosis column

The algorithm prediction was as shown in Figure (7).

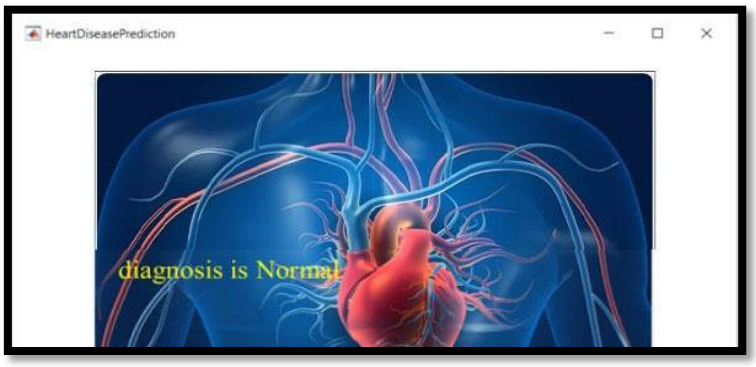

Figure (7): patient diagnosis with normal

Looking at the physician's diagnosis of the same patient below as shown in Figure (8).

ZAge	Sex	Smoking	FHHD	PERANG	FBS	CHOL	ZBPH	ZBPL	ZHR	ChestPain	ECG	Exang	HYP	Heartdiseasediagnose
0.73876	0	1	1	0	1	0	-0.90001	-0.97175	-0.68253	0	0	0	1	1

Figure (8): physician's diagnosis of the same patient

The algorithm has successfully predicted that the patient is normal .

Another example is a patient with the following attributes was entered to the prediction algorithm as shown in Figure (9).

VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14
Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
0.8725	1	1	1	0	0	1	0.3891	-0.9718	-0.5642	2	1	0	0

Figure (9): patient data that is stripped from the diagnosis column

The algorithm diagnosed this patient with arrhythmias as shown in Figure (10).

Figure (10): patient diagnosis with arrhythmias

Matching the results of the algorithm with the physician diagnosis below, showed that the algorithm had successfully diagnosed this patient as shown in Figure (11).

ZAge	Sex	Smoking	FHHD	PERANG	FBS	CHOL	ZBPH	ZBPL	ZHR	ChestPain	ECG	Exang	HYP	Heartdiseasediagnose
0.87247	1	1	1	0	0	1	0.38908	-0.97175	-0.56416	2	1	0	0	4

Figure (11): physician's diagnosis of the same patient

One more example is a patient with the attributes below as shown in Figure (12).

VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14
Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
-0.5984	1	1	1	0	1	1	-0.4703	-0.9718	-1.6887	2	1	0	1

Figure (12): patient data that is stripped from the diagnosis column

The algorithm diagnosed the patient with congestive heart failure as shown in Figure (13).

Figure (13): patient diagnosis with congestive heart failure

Pairing the algorithm's diagnosis of the patient with the physician's diagnosis below showed that the algorithm had predicted the disease successfully as shown in Figure (14).

ZAge	Sex	Smoking	FHHD	PERANG	FBS	CHOL	ZBPH	ZBPL	ZHR	ChestPain	ECG	Exang	HYP	Heartdiseasediagnose
-0.59839	1	1	1	0	1	1	-0.47031	-0.97175	-1.68866	2	1	0	1	3

Figure (14): physician's diagnosis of the same patient

One last example is a patient with the following attributes was fed into the algorithm as shown in Figure (15).

VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11	VarName12	VarName13	VarName14
Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
-1.8687	1	0	1	1	0	0	-0.4703	0.5102	-0.6825	0	0	1	1

Figure (15): patient data that is stripped from the diagnosis column

And it resulted in: as shown in Figure (16).

Figure (16): patient diagnosis with coronary heart disease

The algorithm predicted that the patient has coronary heart disease comparing it with the physician's diagnosis below as shown in Figure (17).

ZAge	Sex	Smoking	FHHD	PERANG	FBS	CHOL	ZBPH	ZBPL	ZHR	ChestPain	ECG	Exang	HYP	Heartdiseasediagnose
-1.86868	1	0	1	1	0	0	-0.47031	0.51021	-0.68253	0	0	1	1	3

Figure (17): physician's diagnosis of the same patient

The physician diagnosed this patient with congestive heart failure This indicates that the algorithm's prediction was incorrect and that is due to the fact that the algorithm's accuracy achieved in the training process was 86% and that means there is an error rate of 14% and figure number (17) shows that.

Conclusion

1- The diagnostic system proposed in this study is intended to assist physicians to diagnose heart conditions by converting medical factors of the patients into numerical representations, The simulation results showed that the proposed K nearest neighbor classifier has 86% accuracy in classifying 4 medical heart conditions when using controlled databases.

2- Most of the studies that explored this subject along with this work concluded that using the K-nearest neighbor algorithm (with 14 parameters) is one of the best options when it comes to heart disease classification and this work proved that by achieving 86% accuracy while classifying diseases into four categories.

3- The most common heart conditions that are found in Iraq are the ones classified in this algorithm, which are Coronary heart disease, Congestive Heart failure, and Arrhythmias and that is based on the data that has been collected locally from hospitals.

4- There is a direct relationship between the amount of training data and the algorithm's accuracy and an inverse relationship between the number of neighbors K and the algorithm's accuracy.

Recommendations

1- Obtaining a larger dataset with additional parameters so that the classification of heart conditions can be improved to a higher

accuracy or increasing the number of subjects in the dataset in the training phase.

2- In terms of classification accuracy, use other machine learning techniques and compare which technique is better.

3- Using different K-nearest neighbor algorithm parameters in hope to achieve a higher accuracy.

4- Programming an easy-to-use user interface that does most of the back-end work to make the entering and prediction of data seamless to physicians.

References

[1] Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–1219.

[2] Anggoro DA, Supriyanti W. Improving accuracy by applying Z-score normalization in linear regression and polynomial regression model for real estate data. Int J Emerg Trends Eng Res. 2019;7(11):549–555.

[3] Arghandabi H, Shams P. A comparative study of machine learning algorithms for the prediction of heart disease. Int J Res Appl Sci Eng Technol. 2020;8(12):677–683.

[4] Assegie TA, Nair PS. The performance of different machine learning models on diabetes prediction. Int J Sci Technol Res. 2020;9(01).

[5] Garg A, Sharma B, Khan R. Heart disease prediction using machine learning techniques. In: IOP Conference Series: Materials Science and Engineering. Vol 1022. IOP Publishing; 2021. p. 012046.

[6] Ye H, Wu P, Zhu T, Xiao Z, Zhang X, Zheng L, et al. Diagnosing coronavirus disease 2019 (COVID-19): efficient Harris Hawks-inspired fuzzy K-nearest neighbor prediction methods. IEEE Access. 2021;9:17787–17802.

[7] Rajula HSR, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina. 2020;56(9):455.

[8] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255–260.

[9] Stanfill MH, Marc DT. Health information management: implications of artificial intelligence on healthcare data and information management. Yearb Med Inform. 2019;28(01):56–64.

[10] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med. 2001;23(1):89–109.

[11] Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Waltham (MA): Morgan Kaufmann; 2012.

[12] Imron MA, Prasetyo B. Improving algorithm accuracy k-nearest neighbor using Z-score normalization and particle swarm optimization to predict customer churn. J Soft Comput Explor. 2020;1(1):56–62.