Classifying Patients with Myocardial Infarction and Heart Failure by Using SVM and KNN Learning Techniques

بديعة رحمن خليل
الباحثة

أ.م.د. محمد محمود فقي
جامعة السليمانية / كلية الادارة والاقتصاد

أ.م.د. سوزان صابر حيدر
جامعة السليمانية/ كلية الادارة والاقتصاد

P: ISSN : 1813-6729
E : ISSN : 2707-1359

http://doi.org/10.31272/JAE.43.2020.126.24

تأريخ أستلام البحث : 2020/9/30

مقبول للنشر بتأريخ : 2020/11/15

ABSTRACT

Cardiovascular diseases (CVD) are considered to be the leading cause of death globally and millions of people from all around the world die annually due to the different types of heart diseases. There are multiple major and minor risk factors that together contribute to developing heart disease. These risk factors include age, sex, tobacco, physical inactivity, genetics etc. Therefore, it's hard to predict heart disease in patients using conventional methods. On the other hand however, with the help of technology, it has now become easier to achieve this goal. The process begins by evaluating datasets containing patient's risk factors. Then, the evaluated datasets would be analyzed using one of the many machine learning techniques. Finally, the analyzed data would be used as a base for classifying and predicting heart disease in new patients. In this paper, we used two of the most advanced machine learning techniques Support Vector Machine (SVM) technique as well as K-Nearest Neighbor (KNN) to analyze the data that we obtained from 210 patients in Sulaimani Cardiac Hospital between (October 16th,2019 to January 9th, 2020). In conclusion, we obtained that the SVM yields more accurate results (82.6%) compared to the KNN method (73.0%).

Key Words: Cardiovascular diseases (CVD), Classification, Datasets, Support Vector machine, K-nearest neighbor, Myocardial Infarction, Heart Failure

- بحث مستل من رسالة ماجستير

مجلة الادارة والاقتصاد
العدد 126 / كانون الاول/ 2020
الصفحات : 315 - 327

1.1 INTRODUCTION

The World Health Organization (WHO) estimates the annual death from cardiovascular diseases to be 17.9 million worldwide (Cardiovascular diseases, (2020)). Cardiovascular diseases are a group of heart disorders that refers to conditions that involve narrowed or blocked blood vessels which can lead to a heart attack, chest pain (angina) or stroke. Other heart conditions, such as those that affect the heart's muscle, valves or rhythm, also are considered forms of heart disease (Heart disease - Symptoms and causes, (2020)). Heart attack is a leading cause for sudden death in both men and women. Additionally, it accounts for 85% of the deaths caused by CVDs all around the globe. Therefore, heart attack prediction has become a subject of interest to almost everyone in the world. Unfortunately, there is no single number, sign or risk factor that can be used as a sole indicator for heart attack occurrence. Different risk factors include, but are not limited to, sex, age, tobacco, physical inactivity and genetics. Some risk factors have more impact on increasing the chances for a patient to get heart attack than the others. But at the end of the day, it is the collective effect of the risk factors that increase the chance of developing cardiovascular diseases drastically which later evolve to heart attack. Analyzing hundreds and thousands of data points to obtain a relationship for predicting the likelihood of heart attack occurrence in a patient by a medical practitioner is extremely hard, if not impossible. It is very important for doctors to be able to identify the possibility of Myocardial Infarction or Heart Failure occurrence in their patients with the help of a reliable method. It is very hard for anyone to disregard the importance of such a method but before we discuss its importance, the more logical step would be to investigate the existence of a method that can be relied on for achieving this task. It is very hard, if not impossible, for a doctor or a medical practitioner to be able to analyze the risk factors for hundreds of patients by using pen and paper or traditional approaches.With the help of technology and the developed machine learning mechanisms, we can now analyze the data of thousands of patients and eventually develop a method to classify those patients based on their shared risk factors. The developed methods usually have very good accuracies. The results of these methods provide very valuable information that help doctors predict heart attack occurrence in their patients more accurately. This will definitely help doctors to take precaution measures in the early stages of heart attack development. As a result, more lives would be saved at the end of the day. There are three main types of machine learning algorithms (Supervised, Unsupervised and reinforced).

Our study aims to show the performance of two well-known machine learning algorithms and compare their results to each other. These two algorithms are SVM and KNN approaches.

1.2 Related Works

In this section, multiple studies have been presented. The studies conducted different scientific comparison between the use of machine learning techniques of SVM and KNN for classification in several applications. Most of these studies were conducted using Medical data.

Conforti and Guido (2005) Proposed the solution of a very critical medical decision problem (early detection of myocardial infarction) using modern and advanced learning methodologies focused on the integration of sufficient

kernels into the support vector machine structure. They were able to create very effective classifiers with strong generalization properties through the appropriate creation of a well-positioned training set. (Conforti & Guido, 2005).

Son et al. (2010) aimed at finding drug adherence predictors in HF patients. This study applied a Support Vector Machine (SVM), a machine-learning method that is useful in classifying data .The two models which best classified medication adherence in HF patients were: one with five predictors (gender, regular medication frequency, knowledge of medication, New York Heart Association [NYHA]), (Functional class, spouse), and the other with seven predictors (age, education, monthly income, ejection fraction, Mini-Mental status Examination-Korean [MMSE-K], knowledge of medication, functional class NYHA). The highest precision of detection was 77.63 percent. (Son et al, (2010)).

Yang et al. (2010) Suggested scoring model based on Vector Support System (SVM). Using Bayesian main component analysis is imputed to missing data in the clinic. Samples are categorized into three categories according to the assessment of cardiac dysfunction: the stable group (without cardiac dysfunction), the HF-prone group (at asymptomatic stages of cardiac dysfunction) and the HF group (at symptomatic stages of heart dysfunction). The model's overall accuracy in classification was 74.4 percent, with accuracies of 78.79 percent, 87.5 percent, and 65.85 percent, respectively, to classify the stable group, HF-prone group, and HF group. Compared with the reported findings in clinical practice, the model helps to improve the accuracy of HF diagnosis especially in early stage screening of HF patients. (Yang et al. (2010).

With padmavathi & Krishna (2014) Easy and efficient Magnitude Squared Coherence and Support Vector Machines dependent algorithm are provided. Detection efficiency was dependent on proper collection from MIT PTB database of Inferior Myocardial ECG signals. The total performance was reached by 99.3 per cent. Because of its simplicity and accuracy, this approach can be used to diagnose Inferior Myocardial Infarction using better results. (Padmavathi & Krishna (2014))

2.1 Materials and Methodology

The following steps have been taken to perform this paper:

a. Dataset collected from patients in Sulaimani cardiac hospital.
b. The predictable attribute has been defined as nominal, in order to classify the two types of Heart disease (Myocardial Infarction or Heart Failure).
c. The obtained data has been translated to a form which can be read and used by the Weka program.
d. Translated data has been imported into Weka program to be analyzed.
e. SMO and IBK classifiers have been chosen to analyze the data by using Support Vector machine and K-nearest neighbor algorithms respectively.
f. The data has been split into two parts (training and testing). 75% of the obtained dataset was used for achieving a training model. The remaining 25% of the data was later used for testing the training model.

2.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) algorithm is a part of the machine learning methodology that Vapnik, Boser, and Guyon first published in 1992. The Support Vector Machine learning system used a hypothetical space in a high-dimensional field in the form of linear functions, and was equipped with an

optimization theory-based algorithm **(** Anggoro & Kurnia (2020)**)**.The problem encountered in the Support Vector Machine algorithm is how to separate the two classes with a function obtained from the available training data, and the classification principle using the Support Vector Machine algorithm is simply an attempt to find the best hyperplane that acts as a separator between two groups of data in the input space (Hasibuan , et al (2017)). You will find hyperplane by calculating the margins and reaching the maximum point (Setiyorini and Asmono (2018)). Margin is the distance between the hyperplane and - class' closest pattern, where the nearest pattern is called the support vector (Fouad et al, 2019).

Very few real-world data sets are linearly separable. What makes support vector machines so exceptional is that they easily expand the simple linear structure to the case where the data set is not linearly separable. The basic concept behind this extension is to transform the input space where the data set cannot be linearly separated into a higher-dimensional space called a function space where the data can be separated linearly. Remarkably, if we carefully select these transformations, all the computations associated with the function space can be done in the input space. That is, even though we're transforming our input space to make the data linearly separable, we don't have to pay the computational costs for those transformations. The functions associated with these transformations are called kernel functions, and the kernel trick is called the method of using certain functions to switch from a linear to a nonlinear support vector machine (Lutz, (2009)).

2.3 Separable data (non-overlapping classes)

Assume that the given data are linearly separable and the line $W^T X_i + b = 0$ indicates the decision boundary between the two classes, where w represents a weight vector, b represents the bias or threshold, and x indicates the training sample. The hyperplane divides the space into two spaces (1) positive half space where the samples from the first/positive class (ω+) are located and (2) negative half space where the samples from the second/negative class (ω−) are located (Wang ,(2005)). The goal of SVM is to determine the values of w and b to orientate the hyperplane to be as far as possible from the closest samples. Moreover, SVM aims to construct the two planes, H1 and H2, as follows:

$H_1 \rightarrow W^T X_i + b = +1 \quad for\ y_i = +1$ (1)

$H_2 \rightarrow W^T X_i + b = -1 \quad for\ y_i = -1$

Where $W^T X_i + b \geq +1$ is the plane for the positive class and $W^T X_i + b \leq -1$ represents the plane for the negative class (see Fig.1). These two equations can be combined as follows:

$y_i(W^T X_i + b) - 1 \geq 0 \quad \forall i = 1,2, \ldots \ldots, N$ (2)

The SVM margin represents the sum of d1 and d2 as follows:

$$margin = d_1 + d_2 = \frac{2}{||w||}$$

Where d1 and d2 represent the distance from the first and second plane, respectively,

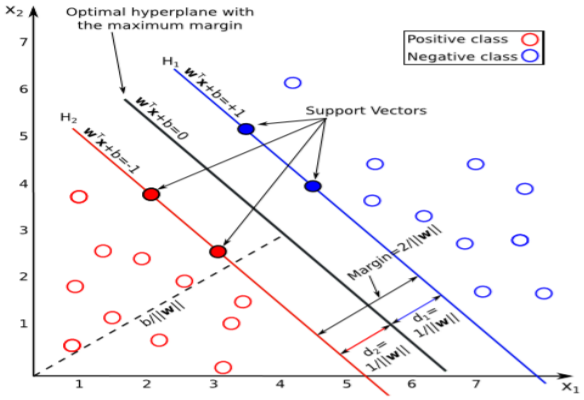

Figure (1)
An example of a binary classification problem with linear separable data using SVM

To the hyperplane and d1 =d2 as shown in Fig.1.In the SVM classifier, the margin width needs to be maximized subject to Eq. (2) as follows:

$$min\frac{1}{2}||w||^2$$
$$S.t \quad y_i(W^T X_i + b) - 1 \geq 0 \quad \forall i = 1,2,\dots\dots,N \qquad (3)$$

As reported in [12], Eq. (3) represents quadratic programming problem and it can be formalized into Lagrange formula by combining the objective function $(\min\frac{1}{2}||w||^2)$and the constraint $(yi(w^T x_i + b) - 1 \geq 0$ as follows:

$$\min L_p = \frac{||w||^2}{2} - \sum_i \alpha_i (y_i(w^T x_i + b) - 1)$$
$$= \frac{||w||^2}{2} - \sum_i \alpha_i y_i (w^T + b) + \sum_{i=1}^{N} \alpha_i \qquad (4)$$

where α_i is the Lagrange multiplier for x_i and LP indicates the primal problem. The values of w, b, and α which minimize LP in Eq. (4) are calculated, and this can be achieved by differentiating LP with respect to w and b and setting the derivatives to zero as follows:

$$\frac{\partial L_p}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (5)$$
$$\frac{\partial L_p}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (6)$$

By substituting Eqs. (5 and 6) into Eq. (4), the dual problem can be written as follows:

$$maxL_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \, y_i y_j x_i^T x_j$$
$$S.t \quad \alpha_i \geq 0 \, , \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \forall i = 1,2,\dots\dots,N \qquad (7)$$

Where LD represents the dual form of LP. Solving Eqs. (5, 6, 7) leads to determine the values of **w, b,** and **α**. In SVM, most of $\alpha_i \, s'$ are zeros; hence, sparseness is a common property of SVM. The non-zero $\alpha_i \, s'$ are corresponding to Support vectors (SVs), which are the samples closest to the separating hyperplane; thus, SVs achieved the maximum width margin.

2.4 Non-separable data (overlapping classes)

Using non-separable or overlapped data, more misclassified samples result. Thus, a slack variable $(\boldsymbol{\epsilon_i \geq 0})$ is added to relax the constraints of

linear SVM as denoted in Eq. (13), where($\boldsymbol{\epsilon_i}$) is the distance between ($\boldsymbol{x_i}$) and the corresponding margin hyperplane, and it should be minimized.

$w^T x_i + b \geq +1 - \epsilon_i \quad for\ y_i = +1\}$ (8)

$w^T x_i + b \leq -1 + \epsilon_i \quad for\ y_i = -1\ \}$

If $0 \leq \epsilon_i \leq 1$, then the sample is in between the margin and the correct side of the hyperplane and this means that the sample is correctly classified. If $\epsilon_i > 1$; hence,$y_i(w^T x_i + b) \geq 1 - \epsilon_i$; thus, the decision function $(w^T x_i + b)$and the class label (yi) have different signs which indicate that the sample (xi) is misclassified. The objective function of SVM after adding ϵ_i will be as follows:

$$min\ \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\epsilon_i$$

$$s.t.\quad y_i(w^T x_i + b) - 1 + \epsilon_i \geq 0 \qquad \forall i = 1,2,\ldots..,N \qquad (9)$$

Where C is the penalty parameter and it controls the trade-off between the size of the margin and the slack variable penalty. Equation (9) is formalized in to Lagrange formula as follows:

$$L_p = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\epsilon_i - \sum_{i=1}^{N}\alpha_i\,[y_i(w^T x_i + b) - 1 + \epsilon_i] - \sum_{i=1}^{N}\mu_i\,\epsilon_i \qquad (10)$$

Where μi ≥ 0 is the Lagrange multipliers to enforce the positivity of ϵ_i . By differentiating LP with respect to w, b and $\epsilon_i$and setting the derivatives to zero as in Eqs. (5, 6, 11):

$$\frac{\partial L_P}{\partial \epsilon_I} = 0 \rightarrow C = \alpha_i + \mu_i \qquad (11)$$

From Eq.(11),it can be remarked that αi is limited by the upper-bound C. Moreover, SVs with αi =C lie outside the margin or on the margin boundary.

2.5 Nonlinear separable data

If the data are nonlinearly separable, kernel functions can be used to transform the data into a higher-dimensional space using a nonlinear function (φ), where the data can be linearly separable. The kernel function is defined as the dot product of nonlinear functions as follows:

$$\boldsymbol{K(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j)}$$

and the objective function of SVM will be as follows:

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\epsilon_i \qquad (12)$$

$$s.t.\, y_i(w^T \emptyset(x_i) + b) - 1 + \epsilon_i \geq 0 \quad \forall i = 1,2,\ldots\ldots.N$$

In SVM, the most widely used kernel functions are:

– Linear kernel $K(x_i, x_j) = \langle x_i, x_j \rangle$.

– Radial basis function (RBF) kernel $(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$.

– Polynomial kernel of degree d, $K(X_i, X_j) = (\langle X_i, X_j \rangle)^d$ (Tharwat, (2019)).

2.6 K-Nearest Neighbor

K-NN is an instance based or lazy method of learning used in data mining classification. Because of sample-based learning, K-NN is also known as a lazy learner when training examples are present; K-NN learns from example, and builds a model. K-NN is an easy and good classificatory. To train our model we apply Heart disease dataset classifier K-NN. We got 38 instances properly classified, and 14 instances incorrectly classified. K-NN classifies instances in close range. In terms of the Euclidean distance measure, closeness is. The distance to the Euclidean is between two points i.e. X and Y given in the equation below **(13)**

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (13)$$

Where:
Yi = Data Samples
Xi = Testing Data
i = Data Variable
D = Distance
n = Data Dimension

We have repeatedly applied the K-NN classifier, i.e. we run it in multiple iterations by adjusting the value of 'K' until the right accuracy is achieved. To measure the precision, sensitivity and specificity, the confusion matrix is used. In Table 4 the uncertainty matrix for the classifier K-NN is shown. The K Nearest Neighbors' Algorithm (K-NN) in pattern recognition is a non-parametric approach used for classification and regression. Learning is supervised. The 'K' in KNN algorithm is taken into account by the number of nearest neighbors (Lutz, (2009)).

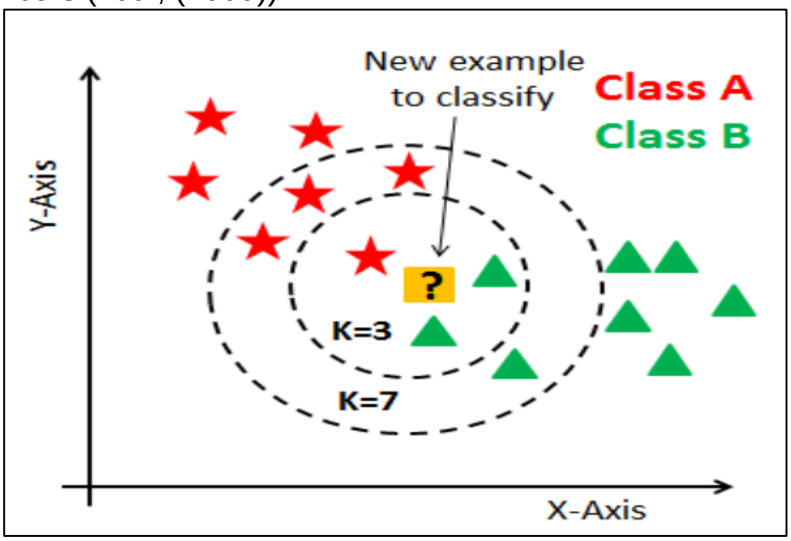

Figure (2)
shows how KNN classification can differ based on the k-value. If (K=3) the result would be green [Class B] and if (K = 7) the result for the example data would yield red [Class A]

In order to classify a dataset with the K-Nearest Neighbor algorithm we followed the steps below:
First, we determined the parameter k (number of nearest neighbors). We tested multiple k-values and eventually chose K=7 as it gave highest accuracy compared to (K= 3, 5, 7). Next, we chose the distance between data points to be equal to (1/d). The Weka program sorted the distances from high value to low value. It also determined the closest distance to the K order. Finally, we obtained the classified test data based on the training datasets mentioned in the confusion matrix that was provided by the weka program.

The K value is suggested to be odd, and more than one. The value of K is fine, based on the data number. The higher the K-value, the lower the classification noise effect. The tool used in evaluating distance is the Euclidean distance tool. The K-Nearest Neighbor algorithm has the advantage of being immune to training data which contains a lot of noise and will be

sufficient if the training data is high. It is possible to measure distance neighbors using the Euclidean distance as in equation 13

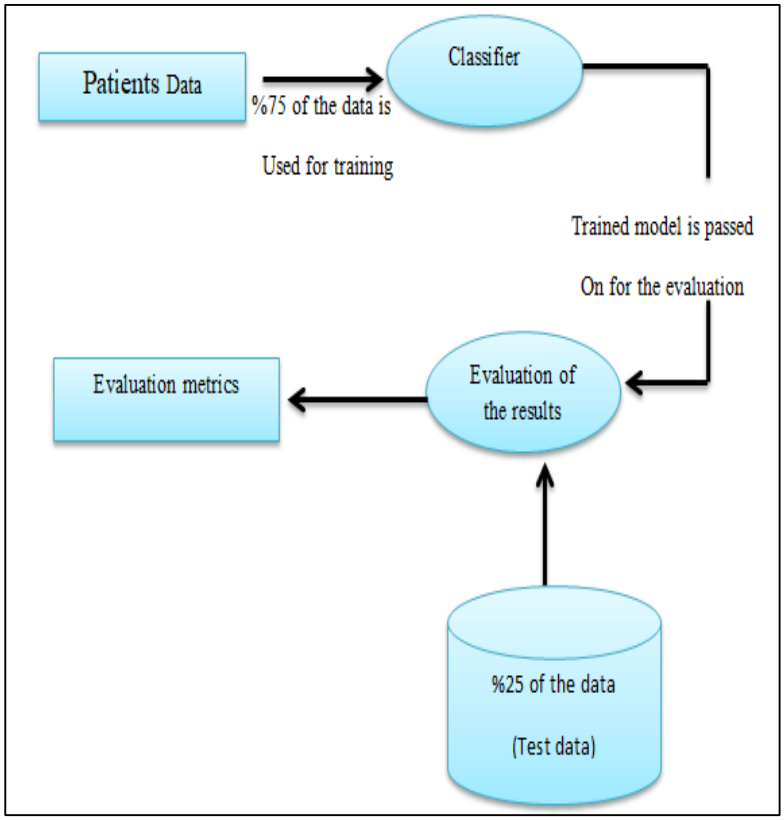

Figure (3)
Flow Diagram of Classifier

2.7 Weka

Weka is simply a little bird and it finds only in New Zealand's islands, but in this case weka is a toolkit for data mining. It is a workbench is an acronym for Waikato Environment for Knowledge Analysis, and has been developed at Waikato University (Ranga & Rohila (2018)). The Weka software version which was used in this research was 3.9.3. Weka is software which is open source. Data mining is performed using a group of Weka-embedded machine learning algorithms. All the tools that are needed for preprocessing data are implemented in weka. Weka supports segregated Comma (.csv) and (.arff) file formats.

Dataset: The dataset used in this paper has been collected from patients in the Sulaimani cardiac hospital between October 16th,2019 to January 9th, 2020. This dataset consists of total 210 patient records. Each row represents one patient record. The record contains 12 attributes out of which one is the predictable attribute called Y whose value indicates the type of heart disease (either myocardial infraction or heart failure). The remaining 11 attributes are used in the predication part of the algorithm. All the 12 attributes are categorical attributes. The following table demonstrates the dataset used in this research paper.

Table (1)
Dataset Description

No	Variable	Attribute Name	Attribute Description	Values
1	X1	Age	Age of the person	No particular range(Real)
2	X2	Sex	Gender of the person(Binary value)	Female = 0 Male = 1
3	X3	BP	Blood pressure in mmHg	No particular range(Real)
4	X4	Fbs	Fasting blood sugar(Binary value)	Fasting blood sugar > 120 mg/dl True =1 and False = 0
5	X5	ECG	Electrocardiographic (Binary value)	Normal = 0 Abnormal = 1
6	X6	CP	Chest pain (Binary value)	No = 0 Yes = 1
7	X7	SOB	Shortness of Breath (Binary value)	No = 0 Yes = 1
8	X8	PAL	Palpitation (Binary value)	No = 0 Yes = 1
9	X9	COU	Cough (Binary value)	No = 0 Yes = 1
10	X10	SMO	Smoking (Binary value)	No = 0 Yes = 1
11	X11	Med	Medication (Nominal value)	Tab = 1 Stent placement = 2
12	Y	Class =Type	Result (Target variable) Yi = +1 A Yi = -1 B	Myocardial Infraction = A Heart Failure = B

The objective of this paper is to conduct parametric analysis of the obtained dataset using two different machine learning algorithms. In order to classify our data, we used Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) techniques. Weka program has been used in this study for generating confusion matrices in order to check the effectiveness of the evaluated techniques

3 Results and Discussion

3.1 Measures for Performance Evaluation

Related classifier output is measured using confusion matrix. Confusion Matrix stores the real and expected class information in tabular form (estimated by the classifier) (Tomar, & Agarwal (2014)). As shown in Table (2).

Table (2)
Confusion Matrix

Actual Class	Predicted Class	
	MI	HF
MI	True Positive (TP)	False Negative (FN)
HF	False Positive (FP)	True Negative (TN)

This paper tests the efficiency of the technique proposed using precision, specificity, sensitivity and geometric mean. The right prediction in proportion to the total number of predictions made by a classifier decides its accuracy which is formulated as:

$$Accuracy(MI, HF) = \frac{(TP+TN)}{(TP+FP+FN+TN)} * 100\% \quad (14)$$

Where,
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

3.2 Experimental Results and Discussion

The experiment was conducted with the use of SVM and KNN classification methods to compare the results obtained. Weka version 3.9.3 is used to analyze Intel ® core (TM) i5 system data with 2.50GHz processor clock speed with 4 GB memory. First the dataset is converted into (csv) file format. Then the result attribute is translated to nominal in weka preprocessed. This processed dataset is divided into 75 percent to be used for model training and the remaining 25 percent is used for model testing.

3.3 Analysis with SVM

Table (3)
Confusion Matrix SVM for testing 25% of the dataset

SVM	Predicted Class		
Actual Class	MI	HF	SUM
MI	25 (TP)	2 (FN)	27
HF	7 (FP)	18 (TN)	25
SUM	32	20	52

The above confusion matrix provides multiple different details about the results obtained from the evaluated method. As we mentioned earlier, 75% of the preprocessed dataset has been used as training set and the remaining 25% as testing. There are 52 testing instances or information from 52 patients has been used to test the evaluated method. 32 patients were actually diagnosed for having Myocardial Infarction (MI) and the remaining 20 patients were diagnosed for having Heart failure. The SVM classifier was able to accurately classify the disease for 43 of the patients (TP and TN) and failed to classify the condition of 9 patients (FP and FN) correctly.

3.4 Analysis with KNN

Table (4)
Confusion Matrix KNN = 7 for testing 25% of the dataset

KNN	Predicted Class		
Actual Class	MI	HF	SUM
MI	26(TP)	8(FN)	34
HF	6(FP)	12(TN)	18
SUM	32	20	52

The above table shows that the conditions for only 38 patients have been classified correctly by this model (TP and TN). On the other hand, however, the actual condition for 14 patients has been classified mistakenly (FP and FN).

Table (5)
The classification accuracy, sensitivity and specificity of proposed model

Classifier	Sensitivity	Specificity	Accuracy	Correctly classified	Incorrectly classified
SVM	92.59	72	82.69	43	9
K-NN = 7	76.47	66.67	73.08	38	14

$$\text{Accuracy (MI, HF)} = \frac{TP+TN}{(TP+FP+FN+TN)} * 100 \quad (15)$$

$$\text{Accuracy} = \frac{25+18}{25+7+2+18} * 100 = 82.6923\%$$

$$\text{Error rate (MI, HF)} = \frac{FP+FN}{TP+FP+TN+FN} * 100 \quad (16)$$

$$\text{Error rate} = \frac{7+2}{25+7+18+2} * 100 = 17.3077$$

$$\text{Sensitivity (Recall)} = \frac{TP}{(TP+FN)} * 100 \quad (17)$$

$$\text{Sensitivity (Recall)} = \frac{25}{25+2} * 100 = 92.59\%$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} * 100 \quad (18)$$

$$\text{Specificity} = \frac{18}{(18+7)} * 100 = 72\%$$

The same procedures are applied to KNN as in equations (15), (16), (17), (18) and the results are shown in table (5)

$$\text{Accuracy (MI, HF)} = \frac{26+12}{26+6+8+12} * 100 = 73.08\%$$

$$\text{Error rate (MI, HF)} = \frac{6+8}{26+6+12+8} * 100 = 26.92\%$$

$$\text{Sensitivity (Recall) (MI, HF)} = \frac{26}{26+8} * 100 = 76.47\%$$

$$\text{Specificity (MI, HF)} = \frac{12}{12+6} * 100 = 66.67\%$$

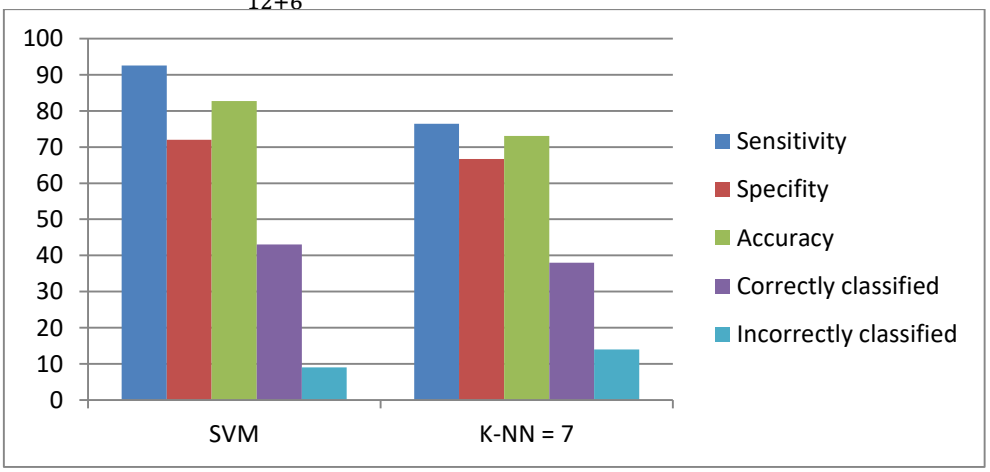

Figure (4)
illustrates the difference in accuracy, sensitivity and specify of the proposed models.

In this paper, we compare the effectiveness of the proposed models based on their achieved accuracies. Among the analyzed data, we obtained

that the SVM gives the highest accuracy compared to the KNN models. The dataset we used in this research was non-linear. The SVM model gave 82.69% accuracy. On the other hand, however, the KNN models did not classify the instances like the SVM approach. The highest accuracy we obtained from the evaluated KNN models was 73.08% (K=7).

4 CONCLUSION AND RECOMMENDATION

4.1 CONCLUSION

In this paper, we have discussed the parametric analysis of Myocardial Infarction (MI) and heart failure (HF) prediction using the Suleiman's cardiac hospital dataset. Our obtained data contained risk factors for 210 patients. Two different classification techniques have been described and their valuation metrics were calculated. According to the results achieved from this paper, Support Vector Machine classifier works better and gives more accurate results compared to the KNN classifier. The accuracy obtained from SVM model was 82.69% but the highest value for KNN model under the same conditions was 73.08%. It can be observed that due to the smaller number of instances in the dataset the accuracy percentages and other prominent metrics are not very high in our paper.

5 4.2 RECOMMENDATION

I recommend that the future researchers use two other machine learning algorithms like Naive Bayes and Decision tree for the same health conditions with similar risk factors and variables. Additionally, they have to compare the accuracy of the four algorithms.

REFERENCES

1. Anggoro, D. A., & Kurnia, N. D. (2020). Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease. International Journal, 8(5).
2. A. Fouad, H. M. Moftah and H. A. Hefny (2019), MRI Brain cancer diagnosis approach using gabor filter and support vector machine, vol. 7, no. 11, pp. 2–6.
3. Cardiovascular diseases (2020). Retrieved 21 August 2020, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
4. Conforti, D., & Guido, R. (2005). Kernel-based support vector machine classifiers for early detection of myocardial infarction. *Optimization Methods and Software*, *20*(2-3), 401-413.
5. C. A.Hasibuan, M. A. Mukid and A. Prahutama (2017), Klasifikasi diagnosa penyakit demam berdarah dengue (dbd) menggunakan support vector machine (svm) berbasis gui matlab, vol. 6, pp. 171–180.
6. Heart disease - Symptoms and causes. (2020). Retrieved 21 August 2020, from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118
7. Lutz H.Hamel (2009) Knowledge Discovery with Support Vector Machines (Wiley Series on Methods and Applications Data Mining).
8. Padmavathi, K., & Krishna, K. S. R. (2014, November). Myocardial infarction detection using magnitude squared coherence and support vector machine. In *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)* (pp. 382-385). IEEE.
9. Ranga, V., & Rohila, D. (2018). Parametric Analysis of Heart Attack Prediction Using Machine Learning Techniques. *INTERNATIONAL JOURNAL OF GRID AND DISTRIBUTED COMPUTING*, *11*(4), 37-48.
10. Son, Y. J., Kim, H. G., Kim, E. H., Choi, S., & Lee, S. K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, *16*(4), 253-259.
11. T. Setiyorini and R. T. Asmono (2018), Komparasi metode neural network , support vector machine dan linear regression pada estimasi kuat tekan, vol. 15, no. 1, pp. 51–56.

12. Tharwat, A. (2019). Parameter investigation of support vector machine classifier with kernel functions. Knowledge and Information Systems, 61(3), 1269-1302.
13. Tomar, D., & Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. International Journal of Bio-Science and Bio-Technology, 6(2), 69-82.
14. Wang L (2005), Support vector machines: theory and applications, vol 177. Springer, Berlin
15. Yang, G., Ren, Y., Pan, Q., Ning, G., Gong, S., Cai, G. ... & Yan, J. (2010, October). A heart failure diagnosis model based on support vector machine. In *2010 3rd International Conference on Biomedical Engineering and Informatics* (Vol. 3, pp. 1105-1108). IEEE.
https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

تَصنيف مرضى النوبة القلبية وعجزِ القلب بأستخدام تقنيات Support Vector machine(SVM) و K-nearest neighbor learning (KNN)

المستخلص:

تعتبر أمراض القلب والأوعية الدموية في صدارة أسباب الوفيات في جميع أنحاء العالم، ويموت الملايين من الأشخاص في جميع أنحاء العالم سنويًا بسبب أنواع مختلفة من أمراض القلب. هناك العديد من عوامل الخطرة الرئيسية والثانوية التي تساهم معًا في الإصابة بأمراض القلب فمثلا العمر والجنس والتدخين وقلة النشاط البدني وعلم الوراثة وغيرها. لذلك من الصعب التنبؤ بأمراض القلب لدى المرضى باستخدام الطرق التقليدية. لكن من ناحية أخرى، وبأستخدام التكنولوجيات الحديثة، أصبح الآن من السهل تحقيق هذا الهدف، وذلك بتقييم مجموعة من البيانات التي تحتوي على عوامل الخطر لدى المريض. بعد ذلك، ومن ثم تحليل تلك البيانات التي سبق تقييمها باستخدام إحدى تقنيات التعلم الآلي. وأخيرًا ، سيتم استخدام البيانات التي تم تحليلها كأساس ولتصنيف أمراض القلب و التنبؤ بها للمرضى. وفي هذا البحث، استخدمنا طريقتين من أكثر تقنيات التعلم الآلي تقدمًا وهما تقنية Support Vector Machine (SVM) بالإضافة إلى K-Nearest Neighbour (KNN) لتحليل البيانات التي تم الحصول عليها من مركز الأمراض القلبيه في مستشفى السليمانية لـ 210 مريض للفتره (من 16 أكتوبر 2019 إلى 9 يناير 2020). وقد أظهرت النتائج التي حصلنا عليها بأن تقنية SVM تعطي نتائج أكثر دقة بنسبة (82.6٪) مقارنة بطريقة KNN والتي تكون دقة نتائجها بنسبة (73.0٪).

.....
.....
.....