e-ISSN 2320 –7876 www.ijfans.org
Vol.11, Iss.12, 2022
Research Paper UGC CARE Listed (Group -I) Journal

CARDIOVASCULAR DISEASE PREDICTION USING TUNED MACHINE LEARNING TECHNIQUES

A. Sankari Karthiga[1], Dr.M.Safish Mary[2]

[1]Research Scholar
Department of Computer Science
Manonmaniam Sundaranar University
Abishekapatti, Tirunelveli,India

[2]Assistant Professor
Department of Computer Science
St.Xavier's College
Palayamkottai, Tirunelveli, India

[1]Corresponding Author Mail id:sankarikarthiga.sk@gmail.com*

ABSTRACT

Programs in healthcare will provide important role in the diagnosis for habitually accurate early identification of cardiovascular disease. Modern data mining methods may be able to help doctors draw relevant inferences from the ever expanding amount and number of medical databases. The limits of heart disease data include the choice of characteristics, the number of samples, the balance of the samples, the lack of magnitude for some aspects, etc. The main goals of this study are to reduce the number of features and improve feature selection.It was shown that feature selection algorithms were quite effective at spotting cardiovascularissues. In this study, we investigate the effects of feature selection methods on logistic regression performance, including information gain and correlation features. We also recommend Tuned Logistic Regression, a special model. Two feature selection techniques, Correlation Feature Selection and Information Gain, are analysed and contrasted with the results of the Tuned-experimental Logistic Regression. The impressive performance of Tuned Logistic Regression is evaluated using performance measures. The proposed method is demonstrated to be superior to other methods already employed in the industry when comparing the accuracy and overall result with existing algorithms.

Keywords: Cardiovascular Disease Prediction, Machine Learning Algorithms, Feature Selection Algorithms.

1. INTRODUCTION

The heart is a vital organ in the human body. It is only a pump that circulates blood throughout the body. When the body's blood circulation is poor, organs like the brain suffer.

If the heart stops beating entirely, death happens within minutes. A group of patients seeking to control heart-based diseases require accurate and effective medical identification of heart-related ailments, which is a challenging but vital duty. Since a healthy heart is essential to maintain human life, heart disease refers to a condition that affects the heart and blood vessel system and is a major source of morbidity and death in contemporary culture.

Heart disease can be easily treated if it is discovered early, which lowers the death rate. The typical method of forecasting heart illness comprises a doctor's assessment and a number of medical tests, such as an electrocardiogram (ECG), a heart MRI, and a stress test on the patient. The majority of the time, manual detection takes a long time and necessitates a specialist doctor's opinion for a significant number of patient visits. Any little symptomatic signals that the patient's heart sends to them during their usual course of action are simply inferred by them based on their knowledge.

Many fully automated heart disease classification methods are available, but they are unable to provide the expert physician with a clear indication of confirmatory No or Yes with the required levels of accuracy. To reduce the accidents based on this problem, computer - assisted specialist systems have been developed for predicting heart diseases. Therefore, it is believed that the approach that can provide good accuracy in forecasting and early detection of heart malfunctions is developed and given in this research work. This method is based on a complete literature review and practical case study-based data collecting. To present to the second opinion, the precise modalities needed to be established for detection, pattern matching, prediction, and picture processing.

1.1 Risk Elements for Heart Disease

Risk factors are the conditions or behaviours that increase a person's propensity to get a disease. They may also increase the likelihood that an existing disease may worsen.

- Irrepressible Risk Factors
- Manageable Risk Factors

The following section goes into further detail about both risk factors.

1.1.1 Irrepressible Risk Factors

Irrepressible risk variables, also referred to as non-modifiable risk factors, are those that cannot be altered in any way. The following provides definitions of the non-modifiable risk factors.

Irrepressible Risk Factor	Definition
Age	Most people who die from heart attacks over the age of 65, and heart disease typically affects men and women after menopause and after the age of 40.
Gender	Compared to premenopausal women, men have a higher risk of developing heart disease. After menopause, a woman's risk is comparable to a man's. But men and women both have a similar risk of stroke.
Family History	It includes information regarding diseases in members of the family because of the family system and interactions within the family, or because those who have a close relative who has had a heart attack may be more at risk of developing heart disease. Family history aids examination of hereditary or familial patterns and offers a ready view of issues or illnesses within the family.

1.1.2 Manageable Risk Factors

The risk factors that can be adjusted or eliminated by adopting specific measures are referred to as modifiable risk factors, also called as controllable risk factors. Below are definitions of risk factors that are drawn from the WHO.

Manageable Risk Factor	Definition
Smoking	The compounds in cigarettes that are based on nicotine encourage the development of blood clots and raise the risk of heart attacks.
Weight	Heart disease risk is directly inversely correlated with body weight. Numerous dietary guidelines should be followed in order to lower the risk of heart disease.
Cholesterol	Atherosclerosis is a type of heart disease that develops when too much blood cholesterol builds up in the artery walls.
Diabetes	High blood pressure and high cholesterol are risk factors for diabetes.It encourages damage to arterial walls and the development of blood clots
Blood Pressure	Blood pressure, which is the force of the blood on the inner walls of the blood arteries, is produced as the heart pumps blood. When a person has hypertension, their heart must work harder to pump blood throughout their body, which can harm their artery walls and lead to atherosclerosis and coronary heart disease.
Stress	Stress is characterised as a condition of psychological and physiological imbalance brought on by a discrepancy between situational demands and the person's capacity and drive to achieve those demands.
Obesity	Obesity is described as an excessive accumulation of body fat that is often 20 percent or more above a person's ideal body weight, or BMI (BMI). Obesity is associated with an increased risk of illness, disability, and mortality..

1.2 Problem Statement

The healthcare industry collects massive amounts of healthcare data from several sources and outlets, and this data needs to be mined to elucidate hidden information for effective decision-making. Even though the majority of common illnesses and epidemics may be quickly identified from collections, there are a number of important hidden patterns of discovery and connections that are occasionally disregarded. Doctors and patients need precise information on a person's risk of getting the condition because heart disease has continued to be the top cause of mortality for the past 20 years.

The development of software to aid doctors in making decisions about heart disease in its early stages has recently been made possible thanks to computer technology and machine learning techniques. As suggested in this research, a model for such an application would be needed as a tool for practitioners to be able to classify persons with heart disease and others. The ideal model would, in fact, be capable of foretelling when the sickness would begin.

1.3 Objectives

Heart disease datasets are used to implement and test the proposed algorithms. When compared to current methods, the suggested methods perform better. To fulfil the aims of this study endeavour, the primary objectives listed below have been framed.

- To create a classification method employing database-collected datasets that quickly pinpoints the risk variables that lead to heart disease.
- To propose a machine learning technique to raise primary warning for those patients who are likely to get heart disease
- To develop a new Machine Learning based Heart Disease prediction model using Correlation based Feature Selection (CFS) with Logistic Regression Classifier, abbreviated as CFSLR model
- To introduce an efficient Heart Disease classification model using Tuned Logistic Regression, abbreviated as TLR model
- To develop an effective TLRmodel for predicting the occurrence of Heart Disease.

2. LITERATURE SURVEY

An extensive literature review is conducted to determine the state-of-the-art of expert systems that are currently being used to support working professionals. With the use of expert systems and data-driven systems for prediction and alarm systems, relevant material from a variety of journals was gathered and categorised into relevant literature. To provide a wider

perspective of the issue, the various approaches and computerised tools, algorithms employed were also taken into consideration for the literature review.

Different methods for predicting blood pressure are compared and contrasted in [1], such as the classical least squares method, ridge regression, lasso regression, ElasticNet, SVR, and the KNN algorithm. The mean absolute error evaluation index revealed that the methodologies performed better and were more accurate than their predecessors. In less than 0.1 seconds, the Gradient Boosting Decision Tree Algorithm (GBDT) can accurately predict the diastolic blood pressure of a patient, with an accuracy rate of over 64%, and the systolic blood pressure of a patient, with an accuracy rate of over 70%. The best algorithm for forecasting blood pressure, then, is the Gradient Boosting Decision Tree (GBDT). Adding new data to the system, such age, body fat, ratio, and height, can lead to better prediction results.

The authors have also presented a unique rapid conditional mutual information feature selection approach [2] to address the problem of feature selection. Using feature selection methods, classification systems can be optimised for efficiency and precision. Features were selected with the aid of feature selection algorithms and used to evaluate classifier performance. The proposed feature selection strategy (FCMIM) in combination with the Support Vector Machine classifier has been shown to be effective in experimental settings for identifying cardiovascular disorders. If you compare the proposed diagnosis system (FCMIM-SVM) to other published approaches, you'll find that it fares quite well. There is no waiting period for using the proposed approach to identify cardiovascular issues in patients.

The authors of [3] adapted a deep convolutional neural network that is suggested as part of Internet of Things architecture to address this problem and deliver a more accurate assessment of cardiovascular health (MDCNN). Vital signs are tracked using the patient's wearable computer, which can record ECGs and blood pressure readings (ECG). Depending on the nature of the issue, the incoming sensor data is labelled as "normal" or "pathological" using the MDCNN. We compare the proposed MDCNN against both traditional deep learning neural networks and logistic regression to gauge the system's performance. The outcomes demonstrate that the proposed MDCNN-based heart disease prediction system is superior to existing methods.

The Machine Learning (ML) model constructed in the author's work [4] using data from the UCI datasets was successful in predicting cardiovascular illness. The dataset is put through a battery of univariate and multivariate statistical tests to ensure its adequacy, skewness, and kurtosis, and to look for any associations between its constituent properties.

Utilizing the correlation matrix and other feature selection techniques, such as Extra Trees Classifier, features are chosen from the multiple aspects of the dataset. Hyperparameters of logistic regression algorithms are fine-tuned using grid search and random search techniques. The confusion matrix, accuracy score, precision-recall curve (PRC), and receiver operating characteristic curve are all examples of performance indicators (ROC). When compared to the other six models, the one built using the power transform, the Kernel PCA applied to the dataset, and finally the Gridsearch approach for hyperparameter management yields flawless accuracy.

According to [5], a technique dubbed MaLCaDD can be employed for reliable prediction of cardiovascular diseases (Machine Learning based Cardiovascular Disease Diagnosis). To address the imbalance and missing values in the source data, the strategy centres on the mean replacement method. Then, the Feature Importance method is applied to choose features. Predictions can be improved by combining the K-Nearest Neighbor (KNN) and Logistic Regression classifiers. The method is checked against the Cleveland, Heart Disease, and Framingham data sets, which are considered to be the industry standards. The comparison results suggest that MaLCaDD projections are better than the current best practises. This means MaLCaDD can be used to reliably and effectively diagnose cardiovascular problems at an early stage in the real world.

The authors of [6] propose a fresh method of model creation in order to solve real-world issues. GridSearchCV and other machine learning algorithms like LR, KNN, SVM, and GBC are used to predict cardiovascular illness. This system runs through a cross-validation process using five separate samples to guarantee precision. The four different strategies are compared and contrasted. The effectiveness of the models is assessed by way of tests on the Cleveland, Hungary, Swiss, Long Beach V, and UCI Kaggle datasets. The training and testing accuracies on both datasets were 100 and 99.03%, respectively, for the Extreme Gradient Boosting Classifier combined with GridSearchCV. The UCI Kaggle competition, Long Beach V, and the competitions in Hungary and Switzerland also included in this study. Using the Extreme Gradient Boosting Classifier in conjunction with GridSearchCV is made easier with the help of the supplied method, which reveals the optimum hyperparameter for measuring accuracy. The LR classifier in [7] has an accuracy of 88.6 percent, but the LR model in [7] only manages 78.56 percent. In terms of diagnostic precision, a logistic regression (LR) classifier is superior to a linear regression (LR) model for detecting cardiovascular disease.

The authors of [8] employed a combination of two different forms of classification to develop a scheme: the Decision Tree and the Naive Bayes. Decision Tree performed better than Naive Bayes in both of these categories. In contrast to Naive Bayes' accuracy of 89.90%, the Decision Tree's accuracy is 98.27%.

Machine learning's most popular classification methods are discussed in [9], including logistic regression, closest K-neighbour, support vector machines, decision-making tree classification, forestry random classifications, and XGBoost classification. Support vector machine algorithms yield more insightful and accurate results than the aforementioned methods.

They [10] look into Tuned K-NN and K-Nearest Neighbor. This diagnostic method was found to be accurate in determining the risk factor for cardiovascular disease. We used K-Nearest Neighbor and Tuned K-Nearest Neighbor, two variants of the K-Nearest Neighbor algorithm, to compute quantitative data features that yield positive classification results using a variety of trained classifier models for patient classification into high- and low-risk groups for normality or abnormality. Tuning the K-Nearest Neighbours algorithm yields the best prediction results at K=3, 7, and 9.

Tuned Support Vector Machines are proposed and evaluated in [11], along with a discussion of the impact that feature selection methods like information gain and correlation features can have on the performance of Support Vector Machines.

3. METHODOLOGY

Preprocessing, feature selection, and categorization are only a few of the procedures included in the suggested methodology. A classification model called Tuned Logistic Regression (TLR) is employed. This methodology starts with Heart Disease data from the UCI machine learning repository dataset. These inputs are first pre-processed, and then the proposed Tuned Logistic Regression technique is used to choose the features that matter most. The Tuned Logistic Regression classification approach is then applied to individual patient related information analysis to find the normal and abnormal classes.

3.1 Preprocessing

Preprocessing is intended to lessen undesired attributes and improve accuracy. Preprocessing involves sifting the data from a sizable database of heart patients, and it is a highly challenging process to extract some of the patient's useful information from the repository.

The majority of the time, raw medical databases will have a significant quantity of missing data. Here, pre-processing is done to separate the mathematical data from the non-numerical data in order to remove the undesired features.

3.2 Feature Selection

In order to speed up computation and increase accuracy, characteristics that are redundant or unnecessary are removed using feature selection (FS) procedures. The basic goal of feature selection is to identify a small feature subset from an intractable domain while maintaining a suitable level of accuracy in resembling the initial features. Here the features are selected by means of Correlation coefficients, Information gain feature selection methods.

3.3 Classification

A crucial use of machine learning is data mining. The target class is predicted using the classification tool, which is one of the supervised learning processes. The following section provides a description of the various machine learning techniques used for the identification and categorization of cardiovascular disease.

3.3.1 Logistic Regression (LR)

Two types of variables are modelled using a logistic regression approach. where the predictor variables can be either continuous or categorical and the dependent variable is a categorical variable, as it is in this example. This semi-parametric model does not require the data to meet the multivariate normality and equal dispersion criteria. The following form of a logistic function is utilised:

$$log\left(\frac{p}{1-p}\right) = w_0 + w_1x_1 + \cdots + w_px_p$$

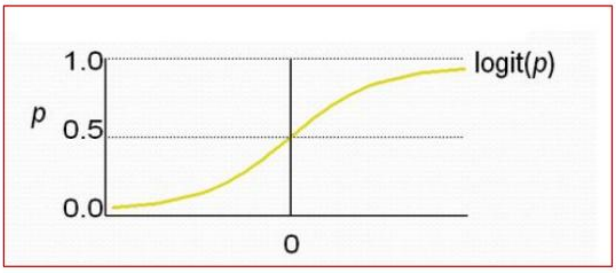

Figure 1. Logistic Function

In addition to the heuristic technique described above, the quantity (p/(1 - p)) is crucial in the interpretation of the log odds shown in the preceding Figure. Similar to a contingency table with two columns (classes) and an infinite number of rows is classification (values of x). With a finite contingency table, it is possible to empirically estimate the log-odds for each row by simply counting table counts. It requires some form of interpolation

method because there are an unlimited number of rows; logistic regression is linear interpolation for the log-odds.

3.3.2 Tuned Logistic Regression (TLR)

Tuned A prediction analysis used to describe the data is logistic regression. It is used to assess model performance indicators and select the optimum hyperparameter. Every time a machine learning algorithm is used on a particular dataset, its effectiveness is evaluated based on how well it generalises, or how it responds to brand-new, unexplored data. Changing, tuning, or tweaking specific algorithmic parameters may be necessary if the learning algorithm's performance is unsatisfactory or may be improved. Hyperparameter tuning is the act of modifying these parameters, which are referred to as "hyperparameters," in order to enhance the effectiveness of the learning algorithm. Algorithm training does not directly teach these hyperparameters. Before the data is trained, these values are fixed. They deal with factors like learning rate, which refers to how quickly the model should be able to learn, model complexity, and other factors. Every learning algorithm may have a variety of hyperparameters. An essential component of machine learning is choosing the appropriate collection of hyperparameters to achieve optimal performance.

Regularization parameter (λ). The "penalty" term that is added to the cost function contains a constant called the regularisation parameter (). Regularization is the process of including this penalty in the cost function. L1 and L2 regularisation are the two types. In the formula for the punishment, they vary.

$$L1 = \lambda \sum_{j=1}^{k} |\theta_j|$$

$$L2 = \lambda \sum_{j=1}^{k} \theta_j^2$$

The cost function in a linear regression is the sum of squared errors. It becomes L2 regularised after the addition of the term:

$$cost = \sum_{i=1}^{m} (y^i - h(x)^i)^2$$

$$cost_{regularization} = \sum_{i=1}^{m} (y^i - h(x)^i)^2 + \lambda \sum_{j=1}^{k} \theta_j^2$$

The binary cross entropy, or log loss, function serves as the cost function in tuned logistic regression. It becomes L2 regularised when a term is added:

$$cost = -\sum_{i=1}^{m} y^i log(h(x^i)) + (1-y^i) log(1-h(x^i))$$

$$cost_{regularization} = -\sum_{i=1}^{m} y^i log(h(x^i)) + (1-y^i) log(1-h(x^i)) + \lambda \sum_{j=1}^{k} \theta_j^2$$

The regularisation parameter is closely related to the Logistic Regression Classifier's parameter C, which is inversely proportional to C=1/ λ.

LogisticRegression(C=1000.0, random_state=0)

A model is designed to determine a weight for each feature during training. Each weight corresponds to a theta vector value. It encourages the model to move weights closer to 0 for some features because there is now a penalty for having a weight for a feature. Therefore, regularisation reduces a model's complexity to prevent overfitting.

In this study, the regularisation parameter C is used. C has the value 1/λ. The trade-off between enabling the model to become as complicated as it wants and attempting to keep it simple is controlled by lambda (λ). For instance, if λis extremely low or 0, the model will have the ability to overfit by giving each parameter's weights large values. The model will, however, tend to underfit if the value of is increased because the model will then be oversimplified.

The reverse is true for parameter C. We raise the regularisation strength for low values of C, which leads to the creation of straightforward models that underfit the data. For high values of C, we reduce the efficacy of regularisation, allowing the model to become more complex and overfit the data.

To make decisions with a computer-based model at a low computation cost, it is highly recommended to use an automated medical diagnosis tool.

<table>
<tr><th colspan="3">Proposed Model</th></tr>
<tr><td rowspan="2">Dataset : Heart Disease (UCI)
Features : 14
Instances : 303
Classes : 2
Present Samples : 44.50%
Absent Samples : 55.50%</td><td colspan="2">Proposed Method:
1. Correlation Feature Selection with LR (CFSLR)
2. Information Gain Feature Selection with LR (IGFSLR)
3. Tuned Logistic Regression (TLR)</td></tr>
<tr><td>Compared Methods
• LR
• LR+Information Gain
• LR+Correlation
• TLR</td><td>Performance Measures
• Accuracy
• Precision
• Recall
• F1-score
• Time complexity</td></tr>
</table>

In this study, a classifier for tuned logistic regression was used. In this phase, Tuned LR is integrated to create an effective classification model. Testing of the proposed classifier model revealed that it outperformed the compared approaches in several performance metrics.

Cardiologists can use cardiovascular disease prediction to help in diagnosis. The proposed study includes a number of procedures, including preprocessing, feature selection based on Correlation Feature Selection (CFS), Information Gain Feature Selection (IGFS), and Tuned Logistic Regression (TLR) techniques, to name just a few. In the classification model, these techniques are employed. The method is used to classify both typical and atypical heart disease situations.

4. PERFORMANCE VALIDATION PHASE

This section evaluates how well the system using Python to implement it performs the proposed heart disease prediction approach. The experiment is run on an i5 processor with 4GB RAM. Here, the suggested methodology is tested experimentally using the Cleveland dataset from the UCI machine learning library. A brief description of the datasets can be found in the section below: which can be downloaded from http://archive.ics.uci.edu/ml/datasets/statlog+ The Heart-Statlog dataset consists of two classes, 14 features, and 303 instances, with 55.50% of the sample being made up of absent instances and 44.50% of the sample being made up of present instances.

The Python working platform is used to accomplish the suggested methodology. The following list of evaluation criteria was employed in this study project: The dataset contains a total of 313 records with 15 attributes, which are divided into two sets with a training set percentage of 60% and a testing set percentage of 40%..

In that instance, an ideal matrix was put together with the goal of accurately predicting the precise state of the cardiovascular stoics (patients):

Patients with no cardiovascular defects are denoted by B, while stoics with cardiovascular complications are denoted by A.

Table 1. Confusion Matrix of A and B

Model	**A (patient withheartdisease)**	**B (patientwithNoheartdisease)**
A (patient withheartdisease)	TruePositive	False Negative
B (patientwithNOheartdisease)	False Positive	TrueNegative

The above grid named confusion matrix encloses data about actual and predict

classificationscompleted by a classification system.

The data enclosed in the matrix are estimated so that thefunctioning of such systems is known.

True Positive (TP) - unhealthy people correctly identified as unhealthy

True Negative (TN) - healthy people correctly identified as healthy

False Positive (FP) - unhealthy people incorrectly identified as unhealthy

False Negative (FN) - healthy people incorrectly identified as healthy

ThedatasetisdownloadedfromtheUCIrepositorythatcontains303recordswith15attributes.The attributes are age, sex, chest pain, cholesterol, and so on.

Table 2. The Detailed Dataset Attributes Description and Distribution

SL.NO	SYMBOL	DESCRIPTION	TYPE	DATA RANGE
1	Age	Subject Age in years	Numeric	[29, 77]
2	Sex	Subject gender	Binary	1 = male 0 = female
3	Tresbps	Resting blood pressure in mmHg	Numeric	[94, 200]
4	Cp	Chest pain type	Nominal	0= typical angina 1= atypical angina 2=non-anginal pain 3 =asymptomatic
5	Chol	Serum cholesterol in mg/dl	Numeric	[126, 564]
6	Fbs	Fasting blood sugar with value >120mg/dl	Binary	0 – false 1 – true
7	Restecg	Resting electrographic results	Nominal	0 = normal 1=having ST-T wave abnormality 2 = showing probable or definite left ventricular Hypertrophy
8	Thalach	Maximum heart rate	Numeric	[71, 202]
9	Old peak	ST depression induced by exercise relative to rest	Numeric	[0, 6.2]
10	Exang	Exercise angina induced	Binary	0 = no 1 = yes
11	Ca	No of major vessels colored by	Nominal	0-3 value

		fluoroscopy		
12	Slope	Slope of the peak exercise ST segment	Nominal	1=up-sloping 2=flat 3=downsloping
13	Thal	Defect type	Nominal	3-normal 6=fixed defect 7=reversible defect
14	Obes	Obesity	Binary	1=yes 0=no
15	Num	Diagnosis disease of heart	Binary	0-no 1-yes

Precision

The inputs are accurately estimated by applying precision to the correlation that exists between sequential events. When data is retrieved automatically, for instance, the goal is to assign an ID that may or may not be significant to the search.

Accuracy

Here, accuracy is defined as the total number of correctly categorised samples divided by the total number of input samples. It is possible to quantify by utilising a variety of ambiguous datasets of varying classes and processing each and every positive and negative term that can be conceived of. Accuracy estimation in ML methods is crucial to making realistic decisions because it helps keep costs down while making fewer mistakes. False-positive diabetes diagnoses, for instance, are part of medical DSSs and drive up exam costs and patient anxiety.

Sensitivity and specificity

The FNs and FPs are eliminated during the estimate of sensitivity and specificity. Only a detector with a sensitivity that has been optimised will ever be truly one of a kind. Let's pretend that in clinical DSSs we either discover plenty of healthy persons or lots of folks who aren't healthy. Maximum sensitivity is achieved when people with the disease are efficiently grouped into distinct categories. Similarly, if healthy people are included in the analysis of sick people, then becomes highly specific. Therefore, sensitivity or recall can be thought of as the ratio of the total number of TPs to the total number of sick people in the population. To put it more precisely the ratio of the total number of TNs to the total number of healthy people in the population.

In this chapter, we covered the four different approaches to research that have been shown to be effective in predicting cardiovascular disease. In this section, we laid out the entire layered structure of the system and described how it works. The evaluation parameters

and dataset used to analyse this chapter's work have been discussed. The methods proposed for predicting cardiovascular events will be described in depth in subsequent chapters.

Performance Tables and graphs

The proposed model of TLR performance with graphs below

Accuracy wise performance

Table 3. Accuracy Wise Performance

Iteration	Accuracy %				
	TKNN	TSVM	CFS LR	IGFS LR	TLR
1	**84.43**	86.81	90.23	90.65	93.66
2	81.46	79.12	90.21	89.95	**97.26**
3	83.81	81.32	90.56	89.03	94.03
4	81.6	84.62	**90.72**	90.55	95.83
5	82.98	**89.01**	90.57	91.13	96.41
6	82.01	80.22	90.25	89.47	96.57
7	79.55	80.22	89.03	90.03	93.05
8	83.78	87.91	90.62	**91.37**	96.94
9	83.76	76.92	89.18	90.79	93.71
10	79.53	78.02	89.85	91.35	94.42

The above table shows the performance of the all algorithms accuracy. The overall accuracy of the TKNNis upto84.43%, Accuracy of TSVM is 89.01%, accuracy of the CFS LR is 90.72%, Accuracy of the IGFS LR is 91.37% and Accuracy of TLR is 97.26%.

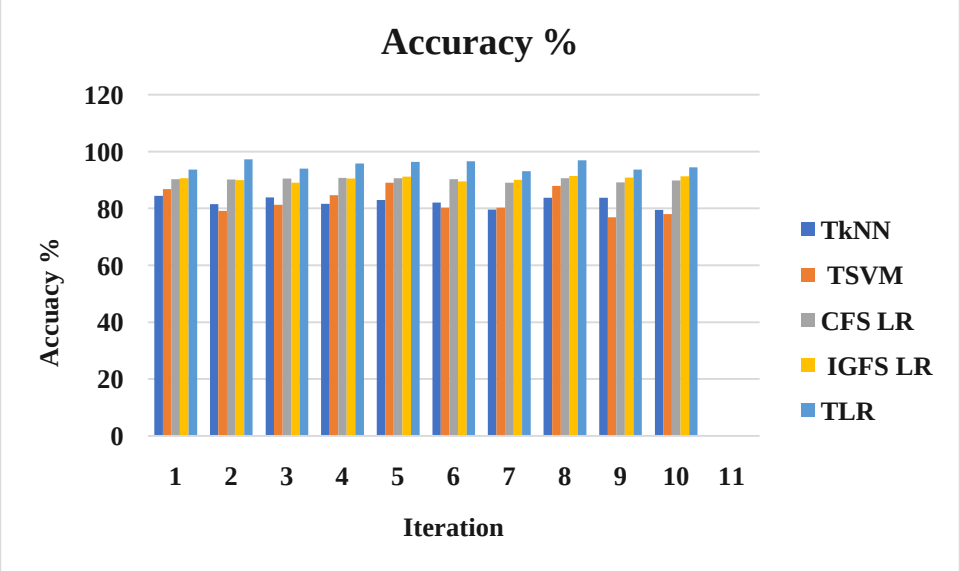

Figure 2 Performance of Accuracy Metric

The above chart depicts the accuracy of the TKNN, TSVM, CFSLR, IGFS LRand TLR

Performance Metrics of Precision

Table4. Performance Metrics of Precision

Iteration	Precision%				
	TKNN	TSVM	CFS LR	IGFS LR	TLR
1	86.46	83.93	89.93	90.59	92.18
2	84.33	75.47	89.24	91.09	91.72
3	85.63	78.57	**89.96**	90.11	91.81
4	82.25	83.33	89.95	**91.93**	91.1
5	82.68	86.79	89.6	90.68	**92.79**
6	86.45	77.59	89.77	91.45	92.32
7	83.51	75.93	89.85	90.28	92.15
8	84.08	**97.04**	89.05	90.17	92.53
9	85.12	83.33	89.29	91.07	92.45
10	**85.99**	69.49	89.94	91.54	91.21

The above table shows the precision value of all the algorithmscompared. The overall precision of the TKNN is upto 85.99%, precision of TSVM is 97.04%,precisionof the CFS LR is 89.96%, precisionof the IGFS LR is 91.93% and precisionof TLR is 92.79%.

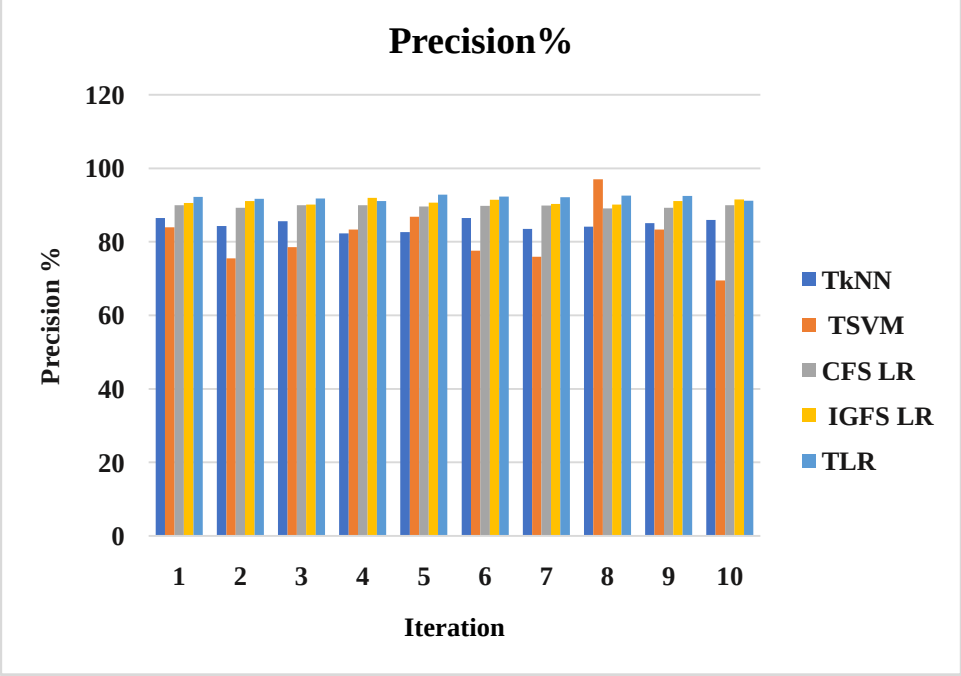

Figure 3. Performance Of Precision Metric

The above chart depicts the precision performance of the TKNN, TSVM, CFSLR, IGFS LR, and TLR.

Performance Metrics Recall

Table 5.Performance Metrics Recall

Iteration	Recall %				
	TKNN	TSVM	CFS LR	IGFS LR	TLR
1	90.22	94	**91.86**	93.65	95.56
2	88.51	86.96	91.13	93.07	92.96
3	90.44	89.8	91.85	94.98	93.54
4	**90.69**	86.96	91.82	93.17	94.65
5	86.11	93.88	91.69	94.76	**95.01**
6	82.21	90	91.7	94.6	94.17
7	85.44	89.13	91.23	94.46	95.04
8	86.21	92.16	91.79	**94.89**	92.84
9	88.64	78.95	91.59	93.74	95.36
10	87.61	**95.35**	91.73	93.99	95.52

The above table shows the performance of the recall in all the algorithms. The overall recall of the TKNN is up to 84.43%, recallof TSVM is 95.35% recallof the CFS LR is 91.86%, recallof the IGFS LR is 94.89%and recall of TLR is 95.01%.

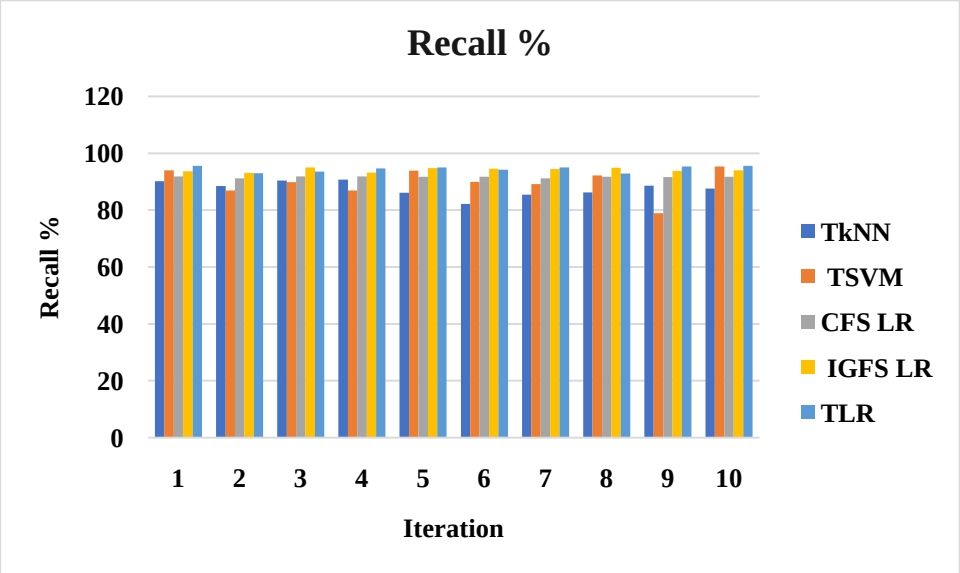

Figure 4. Performance Metrics of Recall

The above chart delineates the recall performance of the TKNN, TSVM, CFSLR, IGFS LRand TLR.

Performance Metrics of F1-Score

Table 6. Performance Metrics of F1-Score

Iteration	F1-Score %				
	TKNN	TSVM	CFS LR	IGFS LR	TLR
1	80.26	88.68	89.21	88.04	92.4
2	87.79	80.81	87.95	90.2	92.52
3	84.27	80.81	88.92	88.6	91.59
4	80.14	85.11	88.36	89.27	92.32
5	85.27	**90.2**	87.9	**90.66**	92.64
6	88.19	83.33	88.65	90.92	92.93
7	82.31	82	88.59	88.77	**92.95**
8	**88.97**	89.52	89.4	88.15	92.23
9	86.02	81.08	**89.91**	89.99	91.72
10	85.36	80.39	88.27	90.36	92.2

The above table shows the performance measureF1-scorein all algorithms. The overall F1-score of the TKNN is upto 88.97%, F1-score of TSVM is 90.2%,F1-score of the CFS LR is 89.91%, F1-score of the IGFS LR is 90.66% and F1-score of TLR is 92.95%.

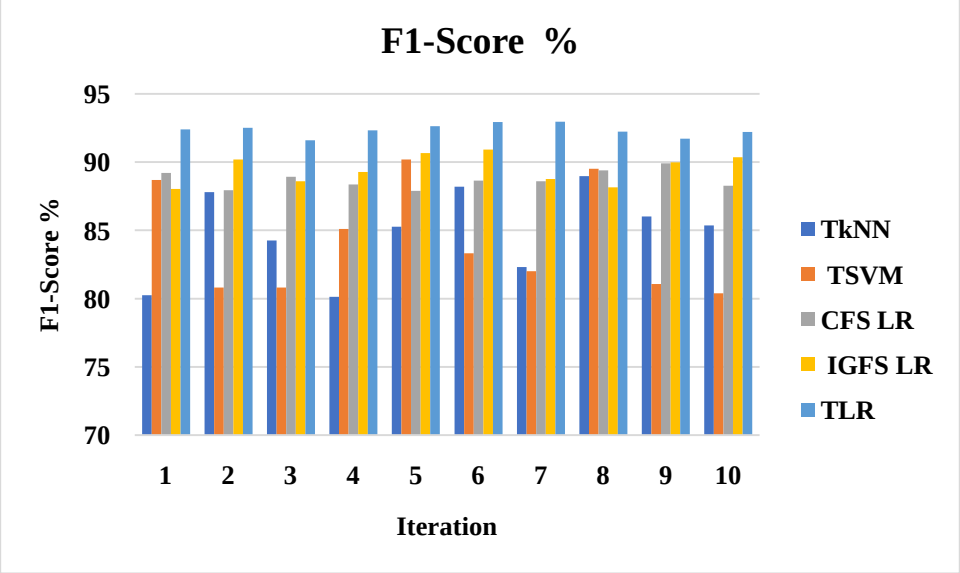

Figure 5. Performance of F1- Score Metrics

The above chart delineates the F1-score performance of the TKNN, TSVM, CFSLR, IGFS LR, and TLR.

Time Complexity

Table 7. Time Complexity

Iteration	Time Taken in seconds				
	TKNN	TSVM	CFS LR	IGFS LR	TLR
1	1.261	0.8681	0.508	**0.189**	0.329
2	1.627	0.7912	0.635	0.399	0.552
3	1.574	0.8131	0.441	0.582	0.745
4	1.79	0.8461	0.209	0.909	0.487
5	1.385	0.8901	**0.134**	0.401	0.603
6	0.972	0.8021	0.616	0.255	0.462
7	**0.909**	0.8021	0.761	0.578	0.706
8	1.171	0.8791	0.624	0.677	0.761
9	1.236	**0.7692**	0.935	0.822	**0.011**
10	1.172	0.7802	0.673	0.873	0.351
Average	**1.3097**	**0.82413**	**0.5536**	**0.5685**	**0.5007**

The above tabledelineates the time complexity of the TKNN, TSVM, CFSLR, IGFS LR, and TLR.

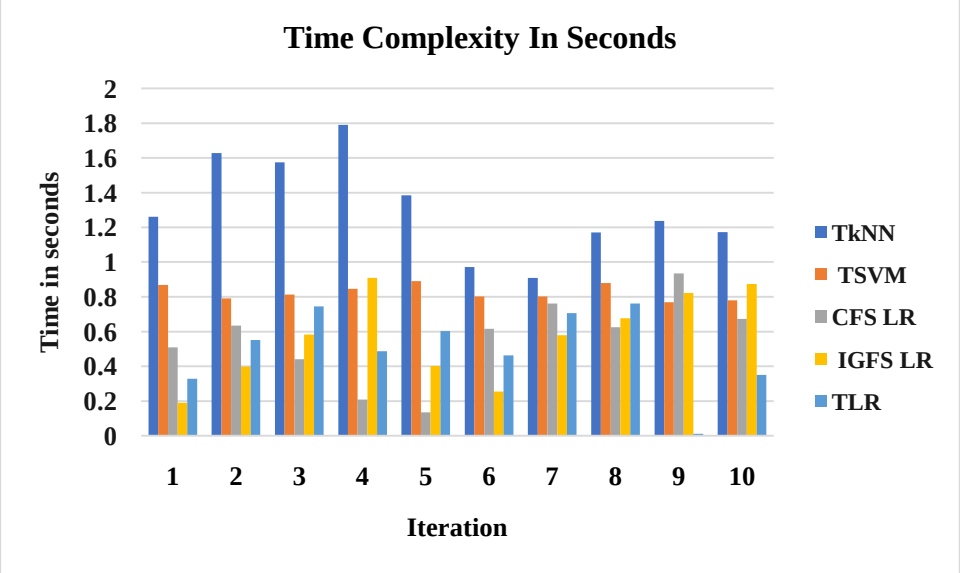

Figure 6.Time Complexity in Seconds

The above chart delineates the time complexity of the TKNN, TSVM, CFSLR, IGFS LR, and TLR.

Overall Performance

Table 8. Overall Performance

Performance Metric %	TKNN	TSVM	CFS LR	IGFS LR	Proposed TLR
Accuracy	84.43	89.01	90.72	91.37	**97.26**
Precision	85.99	87.04	89.96	91.93	**92.79**
Recall	90.69	95.35	91.86	94.98	**95.56**
F1-score	88.97	90.2	89.91	90.92	**92.95**

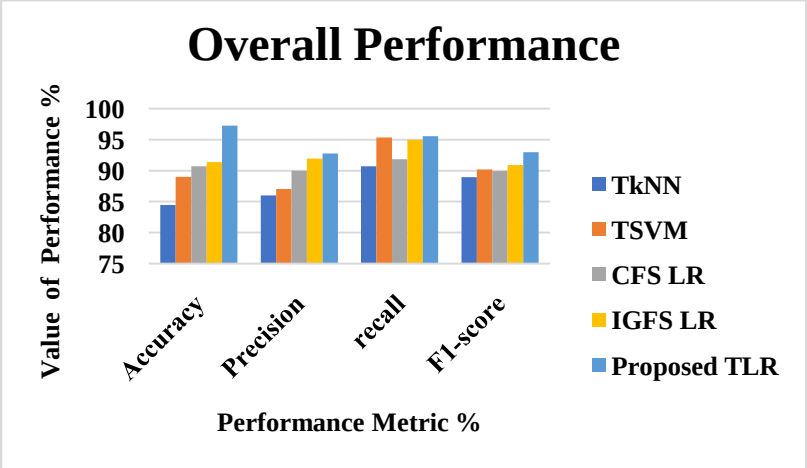

Figure 7. Overall Performance

The above table depicts that Tuned Logistic Regression got 97.26% accuracy with lesser time compared with TunedSupport Vector Machine. Thus, the experiment has justified the use of the proposed Tuned Logistic Regression algorithm.In this paper a medical data classification using Tuned K-Nearest Neighbour (TKNN), Tuned Support Vector Machine (TSVM), and TunedLogistic Regression (TLR) is compared. The experimental result analysis using UCI dataset is done and the performance is analyzed in terms of sensitivity, specificity, accuracy, precision, recall and F-1 Score measured.

5. CONCLUSION

The study's main objective is to more precisely estimate patients' risk of getting heart disease using patient information that was extracted from enormous databases. According to this point of view, the goal of the study is to create a smart model for predicting heart disease. In this study, the TKNN, TSVM, and TLR classification models were introduced for the accurate prediction of heart disease. Information about heart illness is among the exam's components, which are taken from the dataset of the UCI machine learning repository.A variety of assessment criteria are used to determine the efficiency of the proposed method, which demonstrates that it is more effective at preventing heart disease. The suggested Tuned Logistic Regression (TLR) is substantially more accurate than the TKNN, TSVM, CFS LR, and IGFS LR models, with an accuracy of 97.26%. Using the overall performance standards of tuned machine learning models, Tuned Logistic Regression is evaluated for its high performance and improved accuracy.

REFERNECES

[1] B. Zhang, J. Ren, Y. Cheng, B. Wang and Z. Wei, "Health Data Driven on Continuous Blood Pressure Prediction Based on Gradient Boosting Decision Tree Algorithm," in *IEEE Access*, vol. 7, pp. 32423-32433, 2019, doi: 10.1109/ACCESS.2019.2902217.

[2] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[3] M. A. Khan, "An IoT Framework for Heart Disease Prediction Based on MDCNN Classifier," in *IEEE Access*, vol. 8, pp. 34717-34727, 2020, doi: 10.1109/ACCESS.2020.2974687.

[4] SateeshAmbesange*et al*., "Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques,"in *IEEE Access*,978-1-7281-6823-4/20/$31.00 c2020 IEEE, 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4).

[5] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[6] Ahmad, GhulabNabi, et al. "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV." *IEEE Access*, vol. 10, 2022, pp. 80151–73. *DOI.org (Crossref)*, https://doi.org/10.1109/ACCESS.2022.3165792.

[7] B. Manoj Kumar,et al. "Accuracy Analysis of Heart Disease Prediction Using Logistic Regression in Comparison with the Linear Regression Algorithm." *Journal of Pharmaceutical Negative Results*, vol. 13, no. SO4, Jan. 2022. *DOI.org (Crossref)*, https://doi.org/10.47750/pnr.2022.13.S04.199..

[8] A. Sankari Karthiga, Dr. M. Safish Mary. Early Prediction of Heart Disease Using Decision Tree Algorithm. M.Phil Scholar, Mother Teresa Women's University International Journal of Advanced Research in Basic Engineering Sciences andTechnology. 2017; Vol.3 Issue.3

[9] A. Sankari Karthiga, Dr. M. Safish Mary, "Cardiovascular Disease Detection Using Machine Learning Techniques", International Journal of Mechanical Engineering, ISSN: 0974-5823, Vol. 7 (Special Issue, Jan.-Feb. 2022).

[10] Karthiga, A.S. and Mary, M.S. 2022. "A predictive analysis of heart diseases using machine learning algorithms". International journal of health sciences. 6, S3 (Apr. 2022), 3108–3125. Doi:https://doi.org/10.53730/ijhs.v6nS3.6309.

[11] A. Sankari Karthiga, Dr. M. Safish Mary, "Prediction of Cardiovascular Disease at Early Stage Using Feature Selection Based Tuned-Support Vector Machine. Jundishapur Journal of Microbiology.Vol. 15 No. 1 (2022).