Prediction of Cardiovascular Diseases Using Bio-Inspired MKCB Optimization Algorithm

Dr. Pavithra v
Assistant Professor, Department of Computer Science and Applications, SRM Institute of Science and Technology, Ramapuram, Chennai

Dr. Uma Shankari S
Assistant Professor, Department of Computer Science and Applications, SRM Institute of Science and Technology, Ramapuram, Chennai

Dr. Shobana R
Assistant Professor, Department of Computer Science and Applications, Auxilium College Vellore

Dr. R Shanthi
Assistant Professor, Department of Computer Science and Applications, A.M Jain College, Chennai

Abstract

Early detection and prevention of cardiovascular diseases (CVDs) are crucial for reducing mortality and morbidity associated with these illnesses. Data mining techniques can play an important role in this effort by extracting useful information from large and complex datasets, such as electronic health records (EHRs) and clinical trial data, to identify risk factors for CVDs, predict the development of these diseases, and improve the diagnosis and treatment of patients. Heart disease is one of the most common causes of mortality in the world today. While attempting to predict cardiovascular disease, clinical data analysis faces major challenges. It has been shown that it is possible to draw inferences and make predictions from the large amount of data produced by the healthcare industry with the aid of machine learning (ML). In addition, we have seen the use of ML techniques in recent developments in a number of IoT fields (IoT). Using ML to predict heart disease has only been the subject of a few studies. In this article, we propose a novel multi-Kernel optimization technique with cat boost algorithm to predict the cardiovascular disease by detecting essential features using machine learning techniques.

1. INTRODUCTION

Cardiovascular disease (CVD) encompasses a wide range of conditions affecting the heart and blood vessels, including coronary artery disease (CAD), stroke, and heart failure. The risk factors you mention, such as smoking, high blood pressure, and high cholesterol, are major contributors to the development of CVD.Early diagnosis and intervention are crucial in managing CVD, as many people with the condition may not have symptoms until they develop severe disease. For example, CAD can cause angina (chest pain) or a heart attack, but it can also be present without any symptoms. There are several non-invasive diagnostic tests

that can be used to detect heart disease, such as electrocardiography (ECG), echocardiography, and stress tests. Imaging tests, such as CT and MRI, can also be used to visualize the heart and blood vessels. To prevent CVD, it's important to address the risk factors you mentioned earlier[1]. For example, quitting smoking, maintaining a healthy diet and weight, and engaging in regular physical activity can all help to lower the risk of developing CVD. Medications, such as statins and blood pressure-lowering drugs, can also be used to control blood cholesterol and blood pressure levels.

In addition to the individual approaches, public health interventions and policies can play a critical role in reducing the burden of heart disease. This can include things like increasing access to healthy foods, promoting physical activity and reducing tobacco use, as well as policies to improve air quality, reduce poverty, and address other social determinants of health[2]. It is important to bring attention to the fact that this disease is mainly concentrated in low-income countries, and global efforts are needed to reduce the burden of heart disease in these regions through the combination of individual and population-based interventions.

Machine learning (ML) algorithms can be used to analyze large, complex medical datasets to uncover patterns and insights that would be difficult or impossible for humans to detect. By training an ML model on a large dataset of patient information, doctors and researchers can make more accurate predictions about patient outcomes and help identify risk factors for various diseases, including heart disease.

One commonly used approach in healthcare is supervised learning, where a model is trained on labelled data, meaning that the desired outcome (in this case, the presence or absence of a heart disease) is already known for each patient in the training dataset. The model can then be applied to new, unseen data to make predictions about the outcome for future patients. Another ML method is unsupervised learning, where the model looks for patterns in the data without any pre-existing labels. These methods can be used to identify subpopulations of patients with similar characteristics, or to find new biomarkers that can be used to diagnose or predict the progression of a disease. However, working with medical data can be challenging. Data collected from different sources may be in different formats, and can be incomplete or noisy. Privacy concerns also make it difficult to share and analyze patient data. There are techniques to handle that like, anonymization, aggregated data representation and differential privacy Overall, ML has the potential to revolutionize healthcare by helping doctors and researchers make more accurate predictions and personalized treatment plans for their patients[3]. But it’s important to approach these with caution as well, by taking in account data quality, legal and ethical issues.

Data mining is the process of extracting useful information from large datasets and machine learning is a subset of artificial intelligence that allows systems to automatically learn and improve from the data without being explicitly programmed. These technologies have many applications in various fields including healthcare.In the medical field, data mining can be used to analyze large amounts of patient data to identify patterns and relationships that can help healthcare professionals make more informed decisions. The use of classification algorithms, in particular, can be useful in medical data mining, as they can be used to predict the class of an unseen sample based on the characteristics of the training data. This can

be used to assist in the diagnosis of diseases, predicting treatment outcomes, and identifying patients at high risk of certain conditions. It is important to note that data mining in the medical field faces certain challenges like the complexity of the data, privacy, and security issues. Therefore, the use of these algorithms should be done with care and validated with multiple datasets. Additionally, healthcare professionals should interpret the results, considering other factors such as patient demographics, medical history and lifestyle factors before making any decisions.

2. BACKGROUND

Heart disease is the leading cause of death worldwide, affecting millions of people. Medical diagnosis needs to be accurate, dependable, and supported by computers if it is to be effective in reducing diagnostic costs. Medical diagnosis needs to be accurate, dependable, and supported by computers if it is to be effective in reducing diagnostic costs. his section provides a portrayal of connected topics such machine learning and its methodologies with succinct definitions, data pre-processing, evaluation measurements, and a description of the dataset used in this study.

2.1. Machine Learning

Machine learning is a method of teaching computers to learn from data, without being explicitly programmed. It is a subset of artificial intelligence that uses statistical techniques to give systems the ability to "learn" from and make predictions or decisions without being explicitly programmed to perform a task[4]. The main goal of machine learning is to develop algorithms and models that can identify patterns in data and make predictions or decisions about new data. There are many different types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning.

2.1.1. Supervised Learning

A labelled dataset is used to train the model. It has input data and produces results. Datasets are categorised and divided into training and test sets. Our model is trained using the training dataset, while the testing dataset serves as fresh data to determine the model's accuracy. The dataset includes the output of the models.

2.1.2. Unsupervised Learning

The dataset does not contain any classified or labelled training data. Finding hidden patterns in the data is the goal. The model has been taught to create patterns.[5] For any new input dataset, it can anticipate hidden patterns with ease; nevertheless, after exploring the data, it uses the datasets to characterise the hidden patterns.

2.1.3. Reinforcement Learning

Reinforcement learning is a type of machine learning that involves training a model, called an agent, to make a sequence of decisions. The agent receives feedback in the form of rewards or penalties for its actions, and it learns to choose actions that maximize the total reward over time. This is similar to how animals and humans learn through trial and error, by receiving positive reinforcement for good actions and negative reinforcement for bad actions.

There are several algorithms used in reinforcement learning, including Q-learning, SARSA, and policy gradients. These algorithms differ in the way they represent and update the agent's knowledge of the environment, and they have been applied to a variety of tasks such as game playing, robotics, and autonomous vehicles. Reinforcement

learning is an active and promising field of research with the potential for many real-world applications.

3. Literature Survey

Health care workers can identify individuals who are at high risk of developing heart disease by using their knowledge of the risk factors linked to heart disease. Medical practitioners can diagnose cardiac illness with the aid of statistical analysis and data mining techniques. The heart and blood vessels' illnesses, such as coronary heart disease (heart attacks), cerebrovascular disease (stroke), hypertension, peripheral artery disease, rheumatic heart disease, congenital heart disease, and heart failure, have been identified using statistical analysis. The main contributors to cardiovascular disease are poor eating, cigarette use, physical inactivity, and alcohol abuse. Chest discomfort, stroke, and heart attack are the three main causes of heart disease.

Researchers Chaurasia and Pal looked at how well the heart disease database predicted heart attack risk levels. Heart disease prediction heavily relies on 11 key attributes and fundamental data mining techniques as Naive Bayes, J48 decision trees, and Bagging methods[6]. The results demonstrate that bagging strategies perform more accurately than J48 and Bayesian classification. According to the findings, the heart attack can be accurately predicted by the bagging prediction system.

In the medical field, data mining techniques are used to predict cardiac disease utilising a variety of prediction approaches. Colombet et al looked at CART and ANN algorithms for heart disease prediction. They perform better on the examination and succeed. In the areas directly connected to this study, there is a wealth of related work. The introduction of ANN has enabled the medical industry to forecast events with the highest degree of accuracy. The back propagation multilayer perception (MLP) of ANN is employed to foretell cardiac disease.

Patterns are found using NN, DT, Support Vector Machines SVM, and Naive Bayes on data of heart disease patients gathered from the UCI laboratory. Using these algorithms, the outcomes are contrasted for effectiveness and accuracy[7]. Results for the F-measure obtained by the proposed hybrid technique are 86.8%, placing it in the same league as the other methods already in use. The medical business produces a lot of data, but it hasn't always been put to good use. With the new methods described here, heart disease can be more easily and effectively predicted, and costs can be reduced. The many research methodologies addressed in this study for the prediction and categorization of heart disease using ML and deep learning (DL) techniques are extremely accurate in determining the efficacy of these methods.

A Hnin Wint Khaing proposed an effective method for heart attack risk level prediction using the heart disease database in 2011. First, the K-means clustering technique is used to cluster the heart illness database in order to retrieve the information pertinent to heart attacks. Through its k parameter, this technique enables controlling the number of fragments. The MAFIA (Maximal Frequent Item set Algorithm) algorithm is then used to extract the frequent patterns from the retrieved data that are pertinent to heart disease. For the purpose of accurate heart attack prediction, the machine learning algorithm is trained using the chosen significant patterns. To demonstrate the severity of a heart attack using a decision tree, they used the ID3 algorithm as the training

algorithm. Results indicated that the intended prediction was accurate.

The particle swarm with multi-Kernel optimization methodology is a new method we present in this paper. This study's major goal is to increase the performance accuracy of heart disease prediction. There has been numerous research that have produced limitations on feature selection for algorithmic use. The suggested solution, in contrast, makes use of every function without any feature selection limitations. Here, we conduct tests to determine the characteristics of a machine learning system with an improved technique[8]. The findings of the experiment demonstrate that our suggested hybrid strategy has a higher ability to predict heart disease than existing methods.

The rest of the paper is structured as follows: Part II discusses heart-related works, available approaches, and current methods. In Part III, we also give an overview. Part IV describes MKCB Data pre-processing, which is followed by feature selection, classification modeling, and performance measure. The experimental setup and the algorithms utilized are provided in Section V.

4. Proposed Methodology

Different approaches can be used to choose features. In this experiment, the best features are chosen using a combination of Multi Kernel Optimization Random Search with CATBOOST algorithm. The features chosen by this method are used in the machine learning model to determine the model's accuracy. The study reveals that the features chosen using this methodology have increased accuracy.

4.1. OVERVIEW OF THE PROPOSED METHOD

Ensembles of models are groups of models whose combined predictions or outputs result in a single output. Support Vector Machines (SVMs) are a type of kernel-based method that can be used when selecting the base learners for ensemble methods like layered generalisation. According to Joachims et al. [9], merging two kernels is advantageous if they both employ different data instances as support vectors and produce performance that is close to the same. Multiple kernels must be taken into account, as demonstrated by recent improvements in SVMs and other kernel-methods. As a result, there is flexibility and a reflection of the fact that various, heterogeneous data sources are frequently used in ordinary learning issues.

The reasoning is identical to merging different classifiers: it is preferable to have a set and let an algorithm handle the selection or combination step rather of picking just one kernel function. In two ways, MKL can be helpful

1. Since a kernel defines how similar two instances are, selecting a particular kernel may introduce bias because different kernels correspond to different notions of similarity.

2. We can use a mix of a kernel set or import a learning mechanism to select the best kernel to prevent this. A better answer can be found by giving a learner a selection of kernels.

3. It is possible for different kernels to use input from various representations. Combining kernels can be done to combine multiple information sources because different kernels may have different measures of similarity.

A multi-kernel algorithm is a machine learning method that enhances the performance of a model by combining numerous kernels, or similarity functions[10]. These kernels, which are used to compare the similarity between pairs of data points, can be any form of function, including linear, polynomial, or radial

basis functions. A multi-kernel approach can capture a wider range of correlations between data points and increase the model's accuracy by merging numerous kernels.

In this scenario, each base MKL learner of this ensemble model needs to handle fewer kernels than the total number of kernels that need to be handled by employing a single MKL model. It is typically used to combine two ensemble approaches. For particular regression issues, Wolpert et al. [15] offer a number of methods that stacking can be utilised in conjunction with the bootstrap procedure to further enhance bagging's performance. For classification, Kai et al. [16] also describe strategies for combining bagging and stacking.

Contributions:

1. Hybrid kernel learning method with catboost algorithm in this study Research results show that both techniques can be used to improve the accuracy of the model

2. The results also demonstrate that MKCB can handle more instances with the same number of kernels or combine more kernels with the same number of instances than MKL even without using parallel computation. This makes MKCB flexible for use in practical applications.

The CatBoost algorithm uses the gradient method to strengthen decision trees. It can handle categorical features and outperforms earlier gradient boosting approaches. CatBoost employs permutation, one-hot-max-size encoding, greedy algorithms at new tree splits, and target-based statistics to convert labels into numbers. The algorithm incorporates the following

steps: 1. Randomly permuting the dataset;

Step 2: changing the numerical values for the categorical variables as well as the labels to integer numbers.

Step 3: RFC constructs trees via the bagging method, bootstrap feature selection, and random feature distribution. Additionally, it combines many decision trees.

1. The top ensemble algorithm is known to be gradient boosting. The loss function is optimised via the gradient boosting algorithm using the gradient descent method.

2. The algorithm is iterative, and these are the

Initialise the first simple algorithm $b0$

$$s = -\nabla F = \begin{pmatrix} -L'_z(y_1, a_{n-1}(x_1)), \\ \dots \\ -L'_z(y_\ell, a_{n-1}(x_\ell)) \end{pmatrix}.$$

(1)

3. On each iteration we make a shift vector s = (s1,..sl). si — the values of the algorithm bN(xi) = si on a training sample

$$b_n(x) = \operatorname{argmin}_b \frac{1}{\ell} \sum_{i=1}^{\ell} (b(x_i) - s_i)^2,$$

(2)

4. Finally, we add the algorithm bN to the ensemble

$$a_n(x) = \sum_{m=1}^{n} b_m(x)$$

(3)

5. Then the linear combination of the kernel can be written as follows

$$K(x, x') = \sum dmKm(x, x')\, M\, m{=}1 \quad (4)$$

$$dm \geq 0, \sum dm = 1\ M\ m=1 \quad (5)$$

6. Set input X, target Y and penalty value C

7. Calculate matrix K that is dot product vector input due to the multiple kernel function

8. Get support vector, data that satisfy $\leqslant$
$$\alpha i \leqslant C \rightarrow |f(xi)|=1 \quad (7)$$

9. Get bias that is $b=yi-wxi$.

10. M is number of kernel function. If the optimal hyperplane of single kernel is obtained by solving the following optimization problem:

$$\min 1\ 2\ \|w\|\ 2 + C\sum \xi i\ n\ i=1 \quad (8)$$

4.2. DATASET DESCRIPTION

The data set has 14 attributes, and during data pre-processing, some records with missing values are deleted from the data set in order to develop a binary classifier that will identify whether the patient has heart disease or not. Following data pre-processing, feature selection techniques assist in lowering the number of irrelevant features.

The data set is displayed in 4.1 displays the dataset's range values. The data were taken from the UCI Repository. The last attribute is the goal attribute, which indicates whether the patient has heart disease. The first 13 attributes are input features. In order to process the records in various categorization models and assess the outcomes, the dataset is further divided into training and testing in an 80:20 ratio.

ATTRIBUTES	DESCRIPTION	TYPE
Age	Patient's age in completed years	Numeric
Sex	Patient's Gender (male represented as 1 and female represented as 0)	Nominal
Cp	The type of chest pain categorized into 4 values:1. typical angina, 2.non-anginal pain 3. Asymptotic	Numeric
Chol	Serum cholesterol in mb/dl	Numeric
Trestbps	Level of blood pressure at resting mode (in mm/Hg at the time of admitting in the hospital)	Numeric
FBS	Blood sugar level on fasting >120mg/dl; represented as 1 in case of true and 0 in case of false	Nominal
Resting	Results of electrocardiogram while at rest are represented in 3 distinct values: Normal state is represented as value 0. abnormality in ST wave as value 1. (Which may include inversion of T-wave and /or depression or elevation of ST of >0.05mv) and any probability or certainty of 1.V HUPERTOPHY NY Estes criteria as value 2	Nominal
Oldpeak	Exercise -induced ST depression in comparison with the state of rest.	Nominal
Slope	Exercise -induced ST depression in comparison with the state of rest	Numeric
Cs	Fluoroscopy coloured major vessels numbered from 0 to 3	Nominal
Thal	Status of the heart illustrated through three distinctly numbered values. Normal numbered as	Numeric

	3. fixed defect as 6 and reversible defect as 7	
Num	Heart diseases diagnosis represented in 5 values. With 0 indicating total absence and 1 to 4 representing the presented in different degrees	Nominal
Exang	Angina induced by exercise (0 depending 'no 'and 1 depending 'yes')	Nominal

5. RESULTS AND DISCUSSION

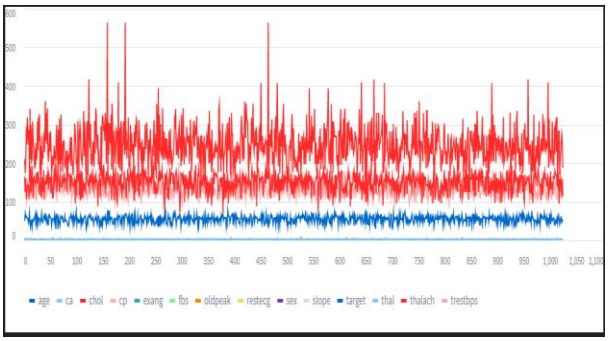
5.1. Experimental setup for evaluation

The dataset is taken from UCI RESPOSITORY, discusses the dataset's range value. The dataset contains the expected output values 0 and 1, where 0 denotes the absence of cardiac disease and 1 denotes its existence. Without using any reduction techniques, 13 features were chosen in the initial step. The technique-MKCB 9 features are recognized as relevant features in the second phase with the aid of feature reduction, and assessment parameters are compared for all Classifier models.

Phase 1

The feature importance score for each feature identified by Svm with MKL are shown in Figure 5.1.

Fig 5.1. Feature Selection Based on MKL



The significant features selected based on the feature importance of Multi Kernel Optimization Technique with cat boost model.

Fig 5.2. Significant Feature Selected by AdaBoost

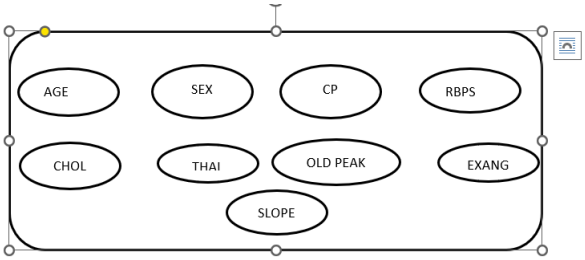


Fig 5.3. Feature Selection Based on Feature Correlation

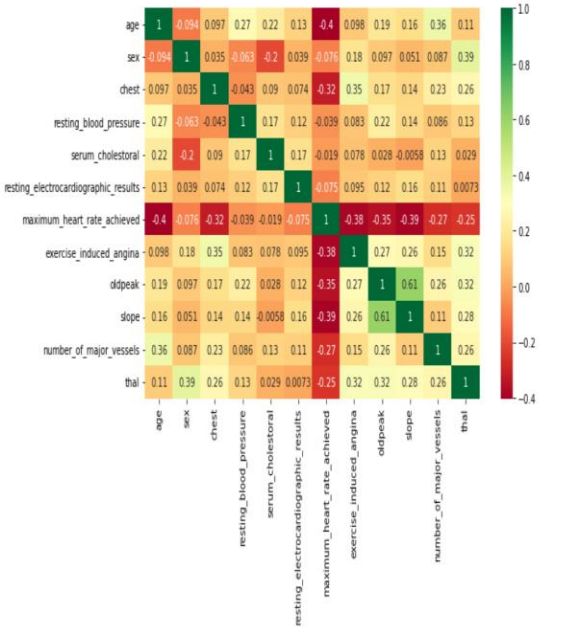
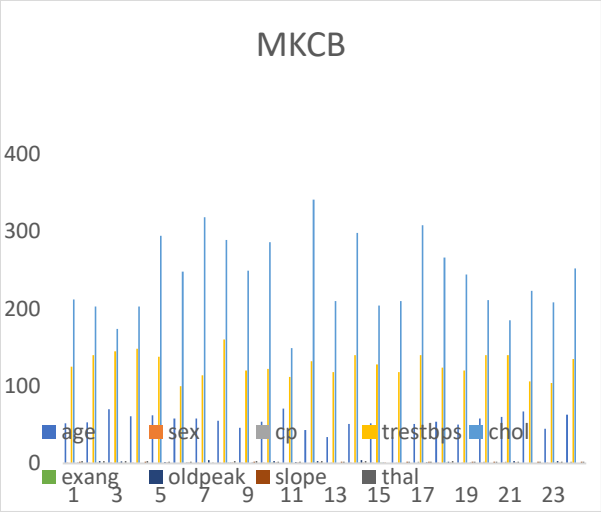


Fig 5.4. Significant Feature Selected by Linear Correlation



CONCLUSION

In this study, a bioinspired algorithm for cardiovascular disease prediction was proposed. The accuracy of the outcome of the feature selection will be impacted by the correlation bias problem in SVM. Large sets of related attributes from a synthetic dataset analysis dataset were utilized to assess the algorithms. It was demonstrated that the MKCB worked. Also, a comprehensive and effective implementation of the suggested method was provided. By using the ensemble method, the proposed algorithm's stability can be increased.

REFERENCES

Felker, G. M., & Mann, D. L. (2019). Heart Failure: A Companion to Braunwald's Heart Disease E-Book. Elsevier Health Sciences.

World Health Organization. (2016). NCD Mortality and Morbidity. [Online]. Available: https://www.who.int/gho/ncd/mortality_morbidity/en/.

Manoj Raman, Vijay Kumar Sharma.(2017).Classification Utility & Procedures for Recognition of Heart Disease: A Review", International Journal of Scientific Research in Science and Technology , pp.383-387. (IJSRST).

Savla, A., Israni, N., Dhawan, P., Mandholia, A., Bhadada, H., & Bhardwaj, S. (2015, March). Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture. In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-7). IEEE.

Banu, N. S., & Swamy, S. (2016, December). Prediction of heart disease at early stage using data mining and big data analytics: A survey. In 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) (pp. 256-261). IEEE.

O. Gualdrón, J. Brezmes, E. Llobet, A. Amari, X. Vilanova, B. Bouchikhi, X. Correig,Variable selection for support vector machine based multisensor systems, Sens.Actuators B: Chem. 122 (2007) 259–268.[6] M. Pardo, G. Sberveglieri, Random forests and nearest shrunken centroids forthe classification of sensor array data, Sens. Actuators B: Chem. 131 (2008)93–99

J.H. Cho, P.U. Kurup, Decision tree approach for classification and dimension-ality reduction of electronic nose data, Sens. Actuators B: Chem. 160 (2011)542–548.[8] R. Kaur, R. Kumar, A. Gulati, C. Ghanshyam, P. Kapur, A.P. Bhondekar, Enhancingelectronic nose performance: a novel feature selection approach using dynamicsocial impact theory and moving window time slicing for classification of Kan-gra orthodox black tea (Camellia

sinensis (L.) o. kuntze), Sens. Actuators B: Chem.166 (2012) 309–319

S. Marco, A. Gutiérrez-Gálvez, Signal and data processing for machine olfactionand chemical sensing: a review, IEEE Sens. J. 12 (2012) 3189–3214.[10] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classificationusing support vector machines, Mach. Learn. 46 (2002) 389–422.

Arjunan, S., Pothula, S.: A survey on unequal clustering protocols in wireless sensor networks. J. King Saud Univ.-Comput. Inf. Sci. 31, 304–317 (2017)

Arjunan, S., Sujatha, P.: Lifetime maximization of wireless sensor network using fuzzy based unequal clustering and ACO based routing hybrid protocol. Appl. Intell. 48(8), 2229–2246 (2018)