ISSN: 2633-4828 Vol. 6 No.1, January, 2024

International Journal of Applied Engineering & Technology

AN EFFICIENT APPROACH OF HEART DISEASE DIAGNOSIS USING FEATURE EXTRACTION WITH MACHINE LEARNING MODEL

[1]Mr. H. Ramprasanth, [2]Dr. P. Nandhakumar
[1]Research Scholar, of Computer Science, Park's College (Autonomous), Tirupur
[2]Research Supervisor, Department of Computer Science, Park's College (Autonomous), Tirupur
[1]Ramprasanth1408@gmail.com and [2]drpnandhakumarphd@gmail.com

ABSTRACT

Decision support systems are used in the diagnosis of most human diseases. The selection of the most pertinent elements is what ultimately determines how effective these systems are. When there are missing values in the dataset for the different attributes, this gets more challenging. The ability of Principal Component Analysis (PCA) to handle missing attribute data is widely recognized. This work provides an approach that identifies heart disease by extracting a reduced dimensional feature subset from medical test results. The suggested approach uses Principal Component Analysis for Feature Extraction in order to extract high impact features from fresh predictions (FE-PCA). In order to minimize the dimension of the feature, PCA extracts the projection vectors that contribute the most covariance. The suggested methodology is assessed in terms of f-measure, sensitivity, specificity, and accuracy across the three datasets. To illustrate the influence of the suggested FE-PCA technique, statistical results are presented and compared to previous studies. A high precision dataset was produced by the suggested FE-PCA technique. DDoS

Keywords: - Heart Disease, Heart Disease, Feature Extraction, PCA, Chi-squire, Relief.

1 INTRODUCTION

The heart, which works like an engine or motor, is responsible for regulating blood flow throughout the body. A person's life may end due to heart failure. The World Health Organization (WHO) lists heart disease as one of the top causes of death worldwide. Surveys indicate that 56 million individuals died in 2012, with heart disease accounting for 17.5 million of those deaths. Early diagnosis, which is made possible by a battery of tests, a thorough medical history, and the patient's day-to-day activities, can help control this enormous number [1]. Until a medical specialist is present to analyze the data, the data alone will not be sufficient. Globally, there is a significant lack of medical specialists, especially in developing and third-world nations. When a patient in these nations lives in a remote location with access to basic medical tests but not to a credible expert in heart disease, the situation gets worse. Significant changes in human life styles have been brought about in large part by information technology [2]. In this sense, the medical sector has not lagged behind, and a variety of medical systems have been put forth to support medical professionals throughout the entire illness diagnosis and treatment process.

Congestive heart failure, another name for heart failure, is the result of the heart's inability to pump enough blood to meet the body's demands. Heart attacks, hypertension, obesity, vitamin deficiencies, heavy metal toxicity, smoking, alcoholism, sleep apnea, inactivity, and an unhealthful diet high in animal fats and salt are among the risk factors for heart disease. Consequently, medical professionals identify the damage that has taken place in a patient's heart and assess the patient's heart's ability to pump blood. Massive volumes of sequential data are produced by these testing methods, and it is still difficult to do effective analysis with such a vast amount of data, particularly in the beginning. In fact, a method for heart failure prediction systems with a very low error rate is crucial for medical treatments and studies [6]. Researchers will be able to help people who are at risk of developing heart disease live longer, more active lives by diagnosing and treating them early with the help of these sequential datasets. Determining the optimal answer to the issues of precise diagnosis and therapeutic delivery is the primary objective of a physical evaluation. However, there is still a significant research gap when it comes to analyzing massive data sets to forecast cardiac disease. Adequate contributions to convert or extract features to separate the class labels have not been made because the processed data has a defined feature set to describe the heart disease.

Diagnosing heart disease is considered a major undertaking that needs to be done well and quickly. The same might be automated, which would be really helpful. The world's hardest endeavor to date is still predicting heart disease. To extract features, the majority of academics have focused on data or signal gathering. However, there is not much difference between the class labels using the features that are present in the processed data. Because of this feature, supervised learning is necessary [7]. Supervised learning, however, depends on how well the training method performs, therefore sufficient accommodations must be made for handling the nonlinearities of the retrieved features. Motivated by hybrid algorithms that have been effectively applied in numerous difficult domains. As a result, an improved heart disease prediction system is required, and the people would benefit from its accurate prediction mechanism.

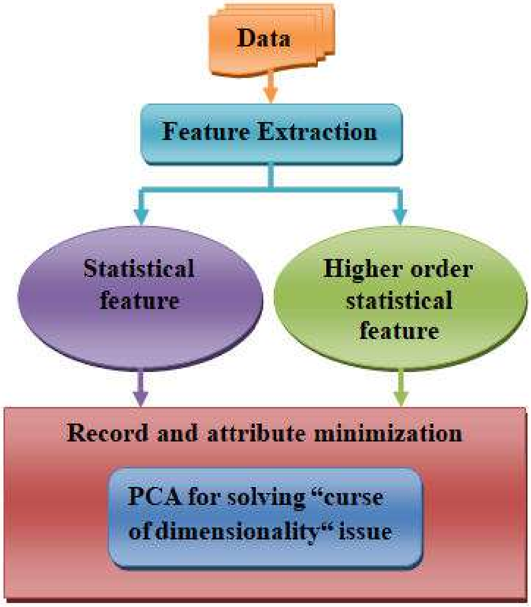
Data
Feature Extraction
Statistical feature
Higher order statistical feature
Record and attribute minimization
PCA for solving “curse of dimensionality“ issue

Figure 1. Feature Extraction

One of the design recognition approaches is Principal Component Analysis (PCA), which may be used, among other things, to explore high dimensional information that is easy to understand by merely glancing at the massive amount of data. Prior to plotting and interpreting the results, the high element of the information must be reduced to a low measurement for information analysis. A few basic plots—the scoring plot and stacking plot, in particular—benefit from the addition of significant data thanks to PCA. In the subject of examination, a lot of information is hard to deconstruct [9]. The highly connected informative indexes are related to one another according to the PCA computation. The relationship between information containing elements as sections and perceptions or tests as lines is made clear by PCA's numerical computation in linear algebra. Reducing the vast number of interconnected elements to a manageable amount is the aim of the PCA computation. Principal components are the interrelated elements in this statement. The principal method involves building a network with the greatest quantity of data in the first two sections, and then using a 2-dimensional visualization in MATLAB programming to explore the data.

It is challenging to comprehend the mathematics of Principal Component Analysis (PCA) without a strong basis in linear algebra. Students' comprehension was significantly enhanced by visualizing the transformation between characteristics and principal components, according to research conducted by Data Science at General Assembly in San Francisco [8]. Four main components make up principal component analysis (PCA), a method for reducing dimensionality: component selection based on explained variance, Eigen-decomposition, feature covariance, and principal component transformation.

1.1. Heart Disease

One of the hallmarks of heart disease, or coronary heart disease (CHD), is the BuildUp of fat in the blood vessels that feed the heart muscles. Heart patients are not diagnosed until the blockage is greater than 70%, and heart disease can start as early as age 18. These obstructions develop gradually until pressure builds up to the point where the membrane covering the obstruction bursts. If blood clots from the chemicals released by a ruptured membrane combine with the blood, heart disease may result [10].

The elements that cause obstruction are known as risk factors. We categorize these risk variables as either changeable or non-modifiable. Heredity, age, and gender are risk factors that cannot be changed. Heart disease will always be caused by these risk factors because they are unchangeable. Risk factors that are modifiable are ones that we can change with our actions. Adaptable risk variables consist of,

1) Food related,

2) Habit related,

3) Stress related,

4) Bio chemical and miscellaneous risk factors.

An efficient decision support system ought to be created in order to counter the threat posed by heart disease. In this research, an enhanced principal component analysis (PCA) based feature reduction method is proposed. This technique maximizes the prediction of cardiac illness and addresses the shortcomings of the conventional Principal Component Analysis (PCA) approach [11].

For data reduction, the proposed method outperforms the standard PCA and FAST methods. The proposed method makes a compelling case for dimensionality reduction.

The remainder of this article is organised as follows. Section 2 examines the various approaches taken by the researchers to the problem statement. Section 3 investigates existing methodologies with various approaches or solutions to the problem statement. Section 3: Explain the system approach used in the proposed scheme for solving the problem statement. Section 4 compares the proposed FE-PCA approach's performance to existing methods. Section 5 concludes by summarising the proposed work.

2. RELATED WORK

Heart disease's performance qualities and accessibility are influenced by several feature extraction techniques. It is impossible to identify one algorithm that performs better than another when analyzing the research that has been published on data mining algorithms and cardiovascular disease risk prediction because each method has pros and cons of its own [12]. The planned study is focused on medical diagnostics; however, choosing the methodology will take additional work.

Ghosh, P., et al., (2021) analyze about the Cardiovascular disease [3]. It has become one of the world's major causes of death. Accurate and timely diagnosis is of crucial importance. We constructed an intelligent diagnostic framework for prediction of heart disease, using the Cleveland Heart disease dataset. We have used three machine learning approaches, Decision Tree (DT), K- Nearest Neighbor (KNN), and Random Forest (RF) in combination with different sets of features. We have applied the three techniques to the full set of features, to a set of ten features selected by “Pearson's Correlation” technique and to a set of six features selected by the Relief algorithm. Results were evaluated based on accuracy, precision, sensitivity, and several other indices. The best results were obtained with the combination of the RF classifier and the features selected by Relief achieving an accuracy of 98.36%. This could even further be improved by employing a 5-fold Cross Validation (CV) approach, resulting in an accuracy of 99.337%.

Brik, Y., Djerioui, M., & Attallah, B. (2021) addressed the problem of the leading cause of death in the world according to the World Health Organization (WHO) [4]. Authors proposed an efficient medical decision support system based on twin support vector machines (Twin-SVM) for heart disease diagnosing with binary target. Unlike conventional support vector machines (SVM) that finds only one optimal hyperplane for separating the data points of first class from those of second class, which causes inaccurate decision, Twin-SVM finds two non-parallel hyper-planes so that each one is closer to the first class and is as far from the second class as possible. Furthermore, a comparison between the proposed method and several well-known classifiers as well as the state-of-the-art methods has been performed. The obtained results proved that our proposed method based on Twin-SVM technique gives promising performances better than the state-of-the-art. This improvement can seriously reduce time, materials, and labor in healthcare services while increasing the final decision accuracy.

For the purpose of diagnosing heart disease patients, Elhoseny, M., et al., (2021) non-invasive medical procedures based on machine learning (ML) methods provide reliable HD diagnosis and efficient prediction of HD conditions [5]. However, the existing models of automated ML-based HD diagnostic methods cannot satisfy clinical evaluation criteria because of their inability to recognize anomalies in extracted symptoms represented as classification features from patients with HD. Authors proposed an automated heart disease diagnosis (AHDD) system that integrates a binary convolutional neural network (CNN) with a new multi-agent feature wrapper (MAFW) model. The MAFW model consists of four software agents that operate a genetic algorithm (GA), a support vector machine (SVM), and Naïve Bayes (NB). The agents instruct the GA to perform a global search on HD features and adjust the weights of SVM and BN during initial classification. A final tuning to CNN is then performed to ensure that the best set of features is included in HD idenfication. The CNN consists of five layers that categorize patients as healthy or with HD according to the analysis of optimized HD features. Authors evaluated the classification performance of the proposed AHDD system via 12 common ML techniques and conventional CNN models cross-validation technique and by assessing six evaluation criteria. The AHDD system achieves the highest accuracy of 90.1%, whereas the other ML and conventional CNN models attain only 72.3%–83.8% accuracy on average.

According to a review of the literature, few adaptive strategies for finding the heart disease using the feature extraction and principle component analysis have been proposed.

3 EXISTING METHODOLOGY

3.1. Chi squared

All that the Chi-squared feature evaluation does is show how important each original feature is. The user can choose which features to keep and which to remove based on this. In Chi-squared feature selection, the relevance of a feature is ascertained using the Chi-squared test statistic between the feature and the target class [13]. Equation (1) is used to compute the Chi-squared statistic, where observed denotes the actual number of class observations and expected denotes the number of class observations that would be anticipated in the absence of any association between the feature and the class. Chi squared needs discrediting numerical characteristics prior to calculation because the sum is computed across each feature value.

$$X^2 = \sum\left((observed - expected)^2/(expected)\right) \qquad (equ. 1)$$

A high Chi-squared test score indicates that the feature and the target class are unlikely to be independent, and thus the feature should be retained in our new dataset.

3.2. ReliefF

Based on variations in feature and class values across neighboring instances, the ReliefF calculates feature scores. ReliefF reduces the feature's score when a group of nearby examples share the same class value but differ in the feature's value. Alternatively, ReliefF increases the feature's score [14] if neighboring instances have varying feature values and class values. To get an overall score for each feature, this method is repeated for a set of sampled cases and their closest neighbors. An equation such as this one can be used to determine the rank of each attribute:

$$R = \Sigma\left((X - Miss)^2 - (X - Hit)^2\right) \qquad (equ.2)$$

Where X represents the feature value of a random sample, Miss represents the feature value of a nearest neighbour with the opposite class value of X, and Hit represents the feature value of a nearest neighbour with the same class value as X.

4 THE PROPOSED MODEL

Due to the redundant and inconsistent data in this pre-processed dataset, there is more room for data storage and search. We have to eliminate any unnecessary and duplicated data in order to attain accurate classification [21]. The dimensionality reduction approach is used to compress high dimensional data to smaller dimensional data, subject to certain limitations. When extracting features, the most pertinent feature—the key feature—is extracted using principal component analysis.

4.1. Feature extraction

An essential technique in data mining and pattern recognition is feature extraction. In order to accomplish the goal of dimensionality reduction, it extracts the significant feature subset from the original dates using a series of criteria to minimize machine training time and space complexity [15]. Feature extraction converts the input data into a set of features, and the new reduced representation includes most of the pertinent data from the original datasets.

Feature extraction is the process of converting unprocessed data into numerical features that may be processed while keeping the original data set's information intact. Reducing the quantity of unnecessary data in a data source is made easier by feature extraction. Lastly, data reduction speeds up the learning and generalization phases of the machine learning process and enables the model to be constructed with less computational effort.

4.2. Principal component analysis

Principal component analysis is a statistical method that preserves the variability in the data set while reducing the information of a large set of associated variables into a small number of variables (referred to as "principal components"). The variables in the data set are combined linearly to create the principle components, and weights are selected to guarantee that the principal components are uncorrelated with one another [16]. Every component is ordered so that the first few components account for the most of the variability, and each provides new information to the data set.

Suppose we have a random vector population X, where

$$X = (x1, x2, \dots xn)^T \qquad (equ.3)$$

And the mean of that population is denoted by,

$$\mu_x = E(X) \qquad (equ.4)$$

And the covariance matrix of the same data set is

$$C_X = E\{(X - \mu_x)(X - \mu_x)^T\} \qquad (equ.5)$$

By definition, the covariance matrix is symmetric. By determining the eigenvalues and eigenvectors of a symmetric covariance matrix, we can compute an orthogonal basis. The first eigenvector represents the direction of the data's highest variance, and by arranging the eigenvectors in descending order of eigenvalues (biggest first), an ordered orthogonal basis can be produced. This enables us to determine which directions in which the data set has the highest concentrations of energy.

4.3. Feature Extraction with PCA

Principle Component Analysis (PCA) is a popular feature extraction technique in data research. Using PCA, one can project data into a new subspace of equal or less dimensions by identifying the eigenvectors of a covariance matrix with the highest eigenvalues. In real life, PCA creates a new dataset with (ideally) less than n features from an n-feature matrix. To put it another way, it does so by generating a new, smaller set of variables that manage to capture a sizable amount of the data present in the original features [17]. Nevertheless, as explaining the idea of PCA is a task best left to others, the purpose of this tutorial is to show PCA in action.

Principal Components Analysis (PCA) analysis here, selected PCA because, when applied to connected characteristics, it yields good results. Since we are working with test qualities for the detection of heart disease, PCA was our choice. It looks for patterns in the data set and evaluates how each attribute differs and how similar it is. It is an effective data analysis tool. The UCI repository's data collection on heart illness was chosen. The initial set of data and its mean are selected. A covariance matrix is constructed. The Eigen vectors and Eigen values are subsequently chosen using the covariance matrix. The eigenvector with the highest Eigen value is the main element of the data set on heart disease. It shows how the data attributes are most strongly related to one another. The Eigen values have an ascending order. The most important data are selected, and the least important ones are eliminated or thrown away. By doing this, data with higher dimensions are reduced to data with smaller dimensions [18]. A sample data set on heart disease is selected from 500 data sets. In order to diagnose cardiac disease, several aspects are taken into account.

5. PERFORMANCE EVALUATION

The suggested hybrid model is divided into three stages: preprocessing was done in the first research phase, and AKSVM was attained in the second. In the third phase, the genetic optimization technique is put into practice. In the last stage of the suggested model, the classification of heart disease is predicted [19]. By computing the conventional metrics of accuracy, precision, sensitivity, specificity, and F-measure, the three approaches' executions were evaluated using the Confusion Matrix predictive classification table.

To distinguish the instances of different classes the confusion matrix acts as a valuable tool for evaluating the algorithms. It reveals the amount of correct and wrong predictions prepared by the model compared with the actual categorizations in the dataset. True Positive (TP) and True Negative (TN) are helpful to identify, when the algorithm is generating the actual data [20]. The parameter for the evaluation measure such as False Positive (FP) and False Negative (FN) are used to know, when the classifier is producing the faulty information.

- **Accuracy:** It is the proportion between the quantity of right predictions and complete number of predications.

$acc=TP+TNTP+TN+FP+FN$ (equ. 6)

- **Precision:** It is the proportion between the quantity of right positives and the quantity of true positives in addition to the quantity of false positives.

$(p)=TPTP+FP$ (equ. 7)

- **Recall:** It is the proportion between the quantity of right positives and the quantity of true positives in addition to the quantity of false negatives.

$recall=TPTP+FN$ (equ. 8)

- **F-score:** It is known as the consonant mean of precision and review.

$acc=112(1p+1r)=2prp+r$ (equ. 9)

.

5. CONCLUSION

This paper presents a heart disease diagnosis methodology that, in the first step, extracts a feature subset using principal component analysis (PCA). The selection of the main components is accomplished through parallel analysis. Three UCI datasets are used: Cleveland, Hungarian, and Swiss. For Cleveland, Hungarian, and Switzerland, the proposed feature extraction using principal component analysis (FE-PCA) methodology produced feature subsets with dimensions reduced by 70%, 62%, and 70%, respectively. FE-PCA is used for extraction, resulting in the classification of suspected heart disease instances into heart disease patient and normal subject classes. As evaluation metrics, accuracy, sensitivity, and specificity are used. In comparison to the existing technique, the proposed FE-PCA technique performed well across all three metrics. Experiment results are also presented, and their statistics increased our confidence in the proposed technique.

6. REFERENCES

[1] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680.

[2] Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1), 23-69.

[3] Ghosh, P., Azam, S., Karim, A., Jonkman, M., & Hasan, M. Z. (2021, May). Use of efficient machine learning techniques in the identification of patients with heart diseases. In 2021 the 5th International Conference on Information System and Data Mining (pp. 14-20).

[4] Brik, Y., Djerioui, M., & Attallah, B. (2021). Efficient heart disease diagnosis based on twin support vector machine. Diagnostyka, 22(3), 3-11..

[5] Elhoseny, M., Mohammed, M. A., Mostafa, S. A., Abdulkareem, K. H., Maashi, M. S., Garcia-Zapirain, B., ... & Maashi, M. S. (2021). A New Multi-Agent Feature Wrapper Machine Learning Approach for Heart Disease Diagnosis. Computers, Materials & Continua, 67(1).

[6] Gárate-Escamila, A. K., El Hassani, A. H., & Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19, 100330.

[7] Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. Computer Methods and Programs in Biomedicine, 195, 105635.

[8] Sun, S. (2021). Segmentation-based adaptive feature extraction combined with Mahalanobis distance classification criterion for heart sound diagnostic system. IEEE Sensors Journal, 21(9), 11009-11022.

[9] Gupta, A., Arora, H. S., Kumar, R., & Raman, B. (2021). DMHZ: a decision support system based on machine computational design for heart disease diagnosis using z-alizadeh sani dataset. In 2021 International Conference on Information Networking (ICOIN) (pp. 818-823). IEEE.

[10] Putra, L. S. A., Isnanto, R. R., Triwiyatno, A., & Gunawan, V. A. (2018). Identification of Heart Disease With Iridology Using Backpropagation Neural Network. In 2018 2nd Borneo International Conference on Applied Mathematics and Engineering (BICAME) (pp. 138-142). IEEE.

[11] Sun, S., Wang, H., Cheng, C., Chang, Z., & Huang, D. (2017). PCA-based heart sound feature generation for ventricular septal defect discrimination. In 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP) (pp. 128-133). IEEE.

[12] Suseendran, G., Zaman, N., Thyagaraj, M., & Bathla, R. K. (2019). Heart Disease Prediction and Analysis using PCO, LBP and Neural Networks. In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 457-460). IEEE.

[13] Kumar, P. R., Ravichandran, S., & Narayana, S. (2021). Parametric Analysis on Heart Disease Prediction using Ensemble based Classification. In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-13). IEEE.

[14] Sonawane, R., & Patil, H. D. (2022). Prediction of Heart Disease by Optimized Distance and Density-Based Clustering. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 1001-1008). IEEE.

[15] Ambesange, S., Vijayalaxmi, A., Sridevi, S., & Yashoda, B. S. (2020). Multiple heart diseases prediction using logistic regression with ensemble and hyper parameter tuning techniques. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4) (pp. 827-832). IEEE.

[16] Chandra, R., Kapil, M., & Sharma, A. (2021). Comparative Analysis of Machine Learning Techniques with Principal Component Analysis on Kidney and Heart Disease. In 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1965-1973). IEEE.

[17] Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. Physica A: Statistical Mechanics and its Applications, 482, 796-807.

[18] Rehman, A., Khan, A., Ali, M. A., Khan, M. U., Khan, S. U., & Ali, L. (2020). Performance analysis of PCA, sparse PCA, kernel PCA and incremental PCA algorithms for heart failure prediction. In 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-5). IEEE.

[19] Taguchi, Y. H., & Taguchi, Y. (2020). Applications of PCA based unsupervised FE to bioinformatics. Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based Approach, 119-211.

[20] Lakshmi, G., & Sujatha, P. (2023). Early Detection and Classification of Heart Diseases by Employing IFCMML and 2L-C Model with I-GA Machine Learning Methods. Indian Journal of Science and Technology, 16(15), 1107-1117.

[21] Hole, K. R., & Anand, D. (2023). AMVAFEx: Design of a Multispectral Data Representation Engine for Classification of EEG Signals via Ensemble Models. International Journal of Intelligent Engineering & Systems, 16(6).