SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 OF THE UGC ACT, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

CFD: Cardiovascular disease prediction through Feature engineering and Data mining

122003023 - Anandalakshmi R
122003050 - Bhargavi T
122003096 - Jaisaiarun P Srinivasan

Guide Name:
Dr. A. Joy Christy

Index

1. Abstract
2. Literature survey
3. Methodology
4. Results
5. Conclusion
6. Future Works
7. Reference

Abstract

- Cardiovascular diseases (CVDs) are one of the important causes of death worldwide.
- The amount of data in the healthcare industry is huge.
- Data mining turns the large collection of raw healthcare data into information that can help to make informed decision and prediction
- Data mining techniques are used be operate upon the massive set of data collected from the healthcare sector to produce impeccable results in the prediction of CVDs
- It is important to select the correct combination of significant features that can improve the performance of the prediction models.
- This paper focuses on identifying significant features and data mining techniques that will improve the accuracy of predicting CVDs.
- The prediction models are developed using different combination of features, and seven classification techniques.
- Classification techniques are k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote.
- *Base Paper :* **Identification of significant features and data mining techniques in predicting heart disease**
- *Authors:* **Mohammad Shafenoor Amina, Yin Kia Chiama, Kasturi Dewi Varathan**

Literature survey

TITLE	AUTHORS	MODEL	WORK
Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease.	Paul, A.K., Shill, P.C., Rabin, M.R.I., Akhand, M.A.H.	Neural Network with Fuzzy.	Genetic algorithm based fuzzy decision support system for predicting the risk level of heart disease.
A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data.	Verma, L., Srivastava, S., Negi, P.C.	C4.5 - A Decision Tree induction algorithm.	CAD diagnosis using correlation based feature subset (CFS) selection with particle swarm optimization (PSO) search method.
An Extreme Learning Machine (ELM) Predictor for Electric Arc Furnaces' vi Characteristics.	Ismaeel, S., Miri, A., Sadeghian, A., Chourishi, D.	Extreme Learning Machine (ELM). It is a single, hidden Layer Feed-forward neural Network (SLFN).	ELM time series prediction strategy to estimate the current and voltage behaviour of an Electric Arc Furnace (EAF).
Feature selection using Artificial Bee Colony for cardiovascular disease classification.	Subanya, B., Rajalaxmi, R.R.	SVM (Support Vector Machine).	Swarm intelligence based Artificial Bee Colony (ABC) algorithm is used to find the best features in the disease identification.

TITLE	AUTHORS	MODEL	WORK
Feature analysis of coronary artery heart disease data sets.	El-Bialy, R., Salamay, M.A., Karam, O.H., Khalifa, M.E.	Decision Tree.	Integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease.
Computational intelligence for heart disease diagnosis: a medical knowledge driven approach.	Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P.	Naïve Bayes.	Detection of heart disease with help of medical knowledge driven feature selection process (MFS).
Early prediction of heart diseases using data mining techniques.	Chaurasia, V., Pal, S.	CART (Classification and Regression Tree).	Prediction models for heart disease survivability using data mining algorithms.
Heart disease classification using neural network and feature selection.	Khemphila, A., Boonjing, V.	Neural Network with Genetic Algorithm.	Classify the presence of heart disease with reduced number of attributes.
Using decision tree for diagnosing heart disease patients.	Shouman, M., Turner, T., Stocker, R.	Decision Tree with Gain Ratio.	Investigates applying a range of techniques to different types of Decision Trees seeking better performance in heart disease diagnosis.

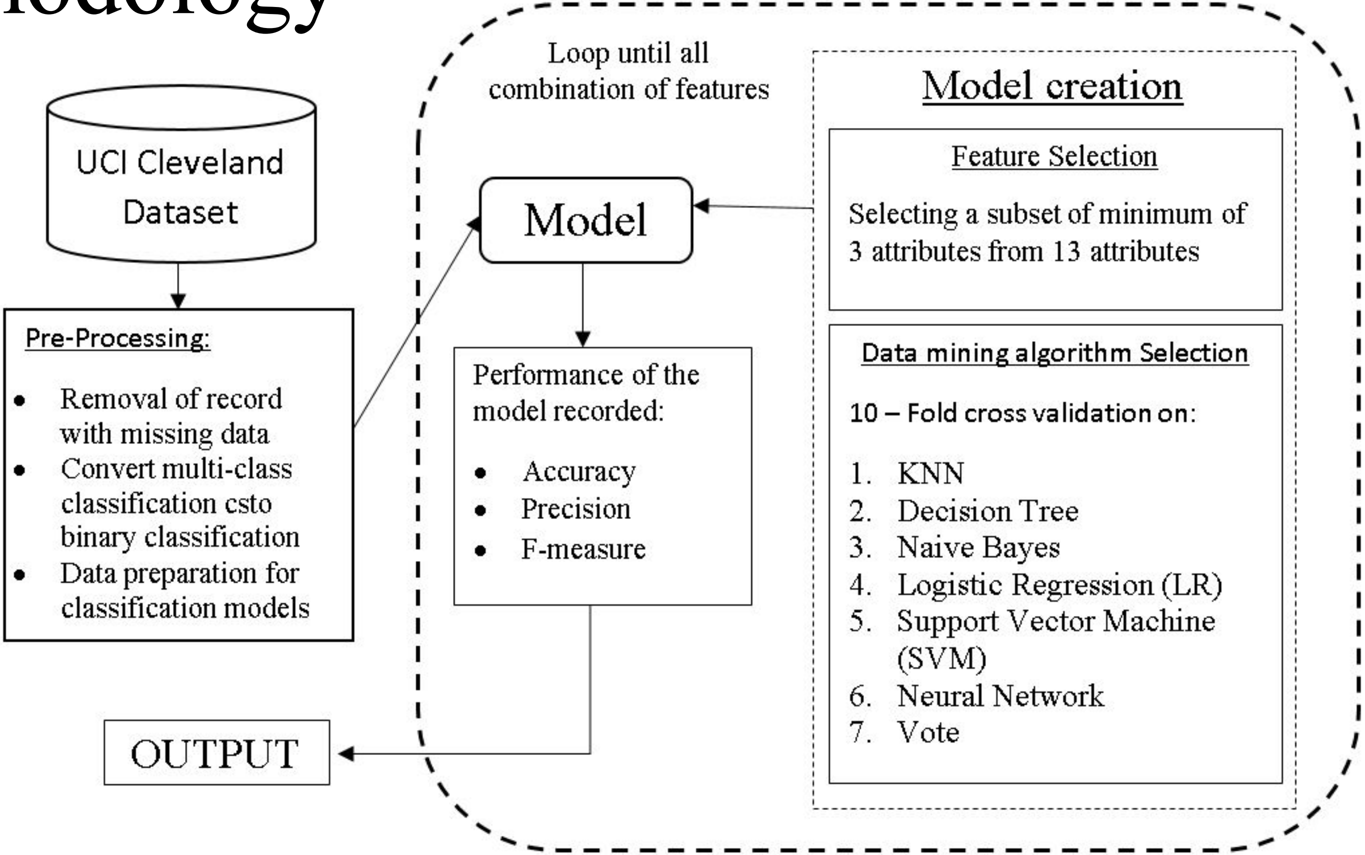
Methodology

- **Dataset :** UCI Cleveland Dataset
- **Set of data mining techniques :** k-NN, Decision Tree, Naive Bayes, Logistic Regression (LR), Support Vector Machine (SVM), Neural Network and Vote(i.e. a hybrid technique with Naïve Bayes and Logistic Regression).

1. Starts with the pre-processing on the dataset :
   1.1. Removal of record with missing data
   1.2. Convert multi-class classification to binary classification
   1.3. Data preparation for classification models
2. Followed by Model creation :
   2.1. different combination of attributes is selected from the entire set of attributes.
   2.2. A data mining technique is selected
   2.3. the model created, based on the selected attributes and data mining technique
3. The performance of the model created in K-fold cross validation is recorded
   3.1. To identify the significant features accuracy, precision and f-measure were used.
   3.2. To identify best data mining technique for models, the accuracy and precision measures were found.
4. Steps 3 and 4 are iterated as a subset containing minimum 3 features are chosen and is applied to model.
5. By analyzing the results, the significant features and data mining technique that have significant impacts in creating the best performing models are identified to predict heart disease.

- An analysis on how many times an attribute was selected in the model that has performed with the highest accuracy, precision and F-measure was done to select the significant features.

Methodology



UCI Cleveland Dataset

Pre-Processing:

- Removal of record with missing data
- Convert multi-class classification csto binary classification
- Data preparation for classification models

Loop until all combination of features

Model

Performance of the model recorded:

- Accuracy
- Precision
- F-measure

OUTPUT

Model creation

Feature Selection

Selecting a subset of minimum of 3 attributes from 13 attributes

Data mining algorithm Selection

10 – Fold cross validation on:

1. KNN
2. Decision Tree
3. Naive Bayes
4. Logistic Regression (LR)
5. Support Vector Machine (SVM)
6. Neural Network
7. Vote

Methodology

- Evaluation is conducted to validate the findings of the significant features and data mining techniques identified
- **Evaluation Dataset :** UCI Statlog Dataset

1. Starts with the pre-processing on the dataset :
   - 1.1. Removal of record with missing data
   - 1.2. Converting the validation dataset features to the format of dataset.
2. Followed by feature and model selection
   - 2.1. Significant features found are selected from the evaluation dataset
   - 2.2. Top 3 classification models are selected.
3. The performance of the modes in K-fold cross validation is recorded
   - 3.1. Average of accuracy of K-fold cross validation

- The accuracy of the models using all the features and using the significant features is compared.

Results

- The highest accuracy, the highest precision and the highest f-measure achieved by each data mining technique and the combination of features used in the model is identified.
- Nine significant feature found: “sex”, “cp”, “fbs”, “restecg”, “exang”, “oldpeak”, “slope”, “ca” and “thal”.
- The three tables show the performance of the best classification techniques with performance metrics

Performance Metric	Data Mining Algorithm	Features	Score
Accuracy	SVM	Age, sex, cp, chol, fbs, exang, oldpeak, slope, ca	86.87%
Precision	Decision Tree and K-NN	Sex, restecg, exang	95%
F-Measure	SVM	Age, sex, cp, chol, fbs, exang, oldpeak, slope, ca	88.22%

- Vote, Naïve Bayes and SVM are the top three techniques for all of the 8100 combinations by getting an average of 78.20%, 78.20% and 78.15% in accuracy.
- The average values of precision indicate that Vote, Naïve Bayes and SVM are the top three techniques
- The top three techniques that have achieved the highest average F-measure are LR, SVM and Naïve Bayes

Results

- An analysis on how many times an attribute was selected in the model that has performed with the highest accuracy, precision and F-measure was done.
- Among all of the 13 attributes, **‘sex’** is the attribute with the highest number of total occurrence. This indicates that this attribute is the most significant attribute.
- Nine features are identified as significant features in heart prediction: **“sex”, “cp”, “fbs”, “restecg”, “exang”, “oldpeak”, “slope”, “ca” and “thal”**.
- During the evaluation phase the accuracy of results achieved by the prediction models developed using the 9 significant features and top three classification

Number of features	**Vote**	**Naive Bayes**	**SVM**
With 13 features	86.30%	84.07%	82.22%
9 significant features	87.41%	84.81%	85.19%

- The accuracy of prediction models developed using the 9 significant features is better than the models developed using all 13 attributes.

Conclusion

- Unlike existing methods this study focuses on discovering the significant features and data mining techniques to improve the accuracy of predicting cardiac diseases.
- The Cleveland dataset from UCI machine learning repository was used identify the significant features and the top performing data mining techniques.
- The performance metrics for all combinations of features and classification models are obtained.
- The Best Significant features and classification models were validated using USI Statlog Heart Disease Dataset.
- This research identifies 9 significant features and 3 data mining techniques that predicts cardiac disease with high accuracy.
- This study compares the accuracy of proposed model and existing models from comparison we can conclude that proposed model predicts heart disease with high accuracy.

Future Work

- This study can be upgraded by conducting the same experiment on a large scale dataset.
- Proposed model uses hybrid technique called vote, it’s a combination of Naïve Bayes and Logistic Regression.
- In the same way we can test alternative combinations of data mining techniques to predict cardiac disease. Besides, new feature selection methods can be used to gain a broader perspective on significant features and enhance accuracy in prediction of cardiac disease.

Reference

- Paul, A.K., Shill, P.C., Rabin, M.R.I., Akhand, M.A.H., 2016. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. (ICIEV). In: 5th International Conference on Informatics, Electronics and Vision. IEEE, pp. 145–150.
- Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., Wang, Q., 2017. A hybrid classification system for heart disease diagnosis based on the RFRS method. Comput. Math. Methods Med.
- Ismaeel, S., Miri, A., Sadeghian, A., Chourishi, D., 2015. An Extreme Learning Machine (ELM) Predictor for Electric Arc Furnaces' vi Characteristics. IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), New York, pp. 329–334.
- Subanya, B., Rajalaxmi, R.R., 2014. Feature selection using Artificial Bee Colony for cardiovascular disease classification. International Conference on Electronics and Communication Systems (ICECS), pp. 1–6.
- El-Bialy, R., Salamay, M.A., Karam, O.H., Khalifa, M.E., 2015. Feature analysis of coronary artery heart disease data sets. Procedia Comput. Sci. 65, 459–46
- Nahar, J., Imam, T., Tickle, K.S., Chen, Y.P.P., 2013. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst. Appl. 40 (1), 96–104.
- Chaurasia, V., Pal, S., 2013. Early prediction of heart diseases using data mining techniques. Carib. J. SciTech. 1, 208–217.
- Khemphila, A., Boonjing, V., 2011. Heart disease classification using neural network and feature selection. In: 21st International Conference on Systems Engineering (ICSEng). IEEE, Las Vegas, pp. 406–409.
- Shouman, M., Turner, T., Stocker, R., 2011. Using decision tree for diagnosing heart disease patients. Proceedings of the Ninth Australasian Data Mining Conference (AusDM’11), Darlinghurst, Australia, pp. 23–30.