OKAN ÜNİVERSİTESİ

Heart disease prediction using machine learning techniques: a survey

A paper reviewed by ***Mr. Chahyaandida Ishaya***

Student ID: 253307014

Program Code: 3307010

Instructor: Assoc. Prof. Dr. Evrim Guler

This report is prepared as part of the coursework for **AIE504 – Machine Learning**, Department of Computer Engineering, **MSc in Artificial Intelligence Engineering**, Istanbul Okan University. The aim of the report is to review the article *"Heart disease prediction using machine learning techniques: a survey"* by V. V. Ramalingam, Ayantan Dandapath, and M. Karthik Raja. The review focuses on the study's originality, objectives, methodology, findings, practical relevance, and its contribution to the broader field of machine learning in healthcare.

Summary

This paper reviews the literature on the application of machine learning algorithms for heart disease prediction using supervised classifiers like Naive Bayes, Support Vector Machines (SVM), K Nearest Neighbour (KNN), Decision Trees (DT), Random Forest (RF), and ensembles. It is also used for the discussion of dimensionality reduction techniques that include feature extraction techniques like Principal Component Analysis (PCA) or correlation-based feature selection and feature selection techniques like chi-square. These authors compile results of various studies on the application of machine learning techniques for heart disease prediction, concluding that SVM, RF, and ensembles work better, while DT tends to overfit.

Citation details

Title of the paper:
Heart disease prediction using machine learning techniques: a survey.

Author(s):
V. V. Ramalingam, Ayantan Dandapath, and M. Karthik Raja.

Journal/conference:
International Journal of Engineering Technology (IJET).

Year of publication:
2018 (Volume 7, Issue 2.8, pages 684–687).

Research Problem or Objective

The major issue that is being tackled is about building trustworthy, precise, and viable systems for the prediction of Cardiovascular Diseases using machine learning since heart diseases are one of the dominant causes of death in both developed and developing nations, and the data in the medical field is large and noisy. It is clear that the aim of the paper is not about developing any new prediction system, but it is more about surveying various prediction systems and techniques with emphasis on finding the effectiveness of various algorithms and techniques for dimensionality reduction that have been successfully used for prediction of heart diseases. A related task is finding the flaws in the current approaches related to dealing with high dimensional data and overfitting.

Methodology Used

Nature of the study

This is a narrative study, not an implementation study, in that it discusses the results of previous experiments/contributions, summarizing or synthesizing the outcomes instead

of training models to achieve a task. It chooses previous works that involved using machine learning techniques for cardiovascular or heart disease prediction tasks that use data like the UCI Cleveland heart datasets.

Dimensionality reduction:

This paper begins with a presentation of dimensionality reduction as an important step in heart disease prediction tasks. There are basically two approaches:

- Feature extraction: Techniques such as Principal Component Analysis (PCA) are used to transform the original features into a set of new features that are uncorrelated, capturing the variance. These features are then used as input for classifiers. References cited make use of PCA for reducing dimensionality before the process of classification. This helps in improving efficiency.
- Feature selection: Techniques such as correlation-based feature selection with Best First Search or chi-square statistics could select a set of features that are thought to be of great significance for the prediction task among the original features. These techniques were used to select significant features from heart disease databases before applying models such as Random Forest or hybrid models.

Algorithms and techniques surveyed

Next, the authors group the results of the survey by type of algorithm.

- Naive Bayes: A probabilistic classifier with a simplifying assumption of feature independence given the class. Some papers report accuracies of low to mid 80% on the Cleveland data set, especially when using techniques such as SVM RFE (Recursive Feature Elimination) or gain ratio for selecting the features.
- Support Vector Machine (SVM): Supervised learning technique with the objective of maximizing the margin hyperplane that divides classes in the feature space. This survey refers to various instances where SVM performs accurately with a high percentage of about 98.9% on a hospital database, mid 80% on UCI databases, accompanied by good F-measure values, especially with kernel or boosting techniques.
- K Nearest Neighbour (KNN): This is a non-parametric approach that assigns to a data point a label given by the majority of its k nearest neighbours. In the studies mentioned, accuracies have ranged from 70% for more demanding tasks to over 80% for certain heart disease problems, with k, distance measures, or even metaheuristics like Ant Colony Optimization playing a role in determining the classification performance.
- Decision Trees (DT): These are tree-based classifiers that divide the data using either information gain or the entropy of the data. According to the survey, basic decision trees tend to perform poorly with accuracy rates of around 43%, whereas

modified versions such as J48 or alternative decision trees along with PCA or feature selection can achieve rates of over 90%.

- Random Forest (RF): A combination of decision trees developed using random data samples and random feature subsets. The authors point out the importance of research that reports significant improvement in Random Forest accuracy over individual decision trees (with Cleveland data, for example, approximately 91.6%, whereas for other heart data it is about 97-97.7%).
- Ensemble models: A combination of multiple classifiers (SVM, KNN, ANN; or Naive Bayes, Decision Tree, SVM) with a voting mechanism or other techniques. It is mentioned in the survey that the accuracies of certain models reached high values (e.g., about 94.1% for CHD prediction models, even about 98% for other heart-related issues like syncope).

General operational principles

At a more conceptual level, it is clear that the systems investigated in these surveys all operate on a similar pattern: gathering or preprocessing medical data, dimensionality reduction through feature extraction or selection, followed by the use of individual or combined supervised learning models for predicting the presence or absence of heart disease. A number of systems employ cross-validation techniques for their evaluation.

Key Findings or Conclusion

This study ends with the conclusion that machine learning algorithms possess great potential for predicting cardiovascular diseases effectively if combined with dimensionality reduction techniques correctly. It is concluded that:

- SVM, Random Forest, and ensembles tend to provide the best results for heart diseases on different datasets with high levels of accuracy and F-measure.
- Basic decision trees tend to work poorly because of overfitting, whereas tree-based models with the use of PCA, feature selection, or alternation (e.g., alternating decision trees) tend to work very well.
- Naive Bayes classifiers are computationally efficient and yield reasonable results, although marginally inferior to the best results yielded by SVMs or ensembles.

The authors strongly point out that while the current situation with regard to heart disease prediction using various systems is fairly good, some issues still remain with regard to dealing with high dimensional data in order to reduce the chances of overfitting.

Critical Analysis

Strengths

- Focused and relevant topic: It is important to note that the study focuses on a relevant topic since it deals with predicting cardiovascular disease, which is a significant problem in the medical field.

- Coverage of key ML techniques: It comprehensively describes principal supervised learning techniques such as Naive Bayes, SVM, K-NN, Decision Trees, Random Forest, and ensembles, so that a reader gets a strategic understanding of the various techniques that exist in the area.
- Dimensionality reduction emphasis: Noting the importance of PCA and other dimensionality reduction techniques through feature selection is beneficial by establishing that advancements are linked to sound management of high-dimensional data.
- Concrete performance references: By citing the ranges of accuracies and F-measures from various studies, it provides a real sense of what the performance could be like using various approaches on heart disease datasets.

Weakness

- Absence of systematic review methodology: The survey does not provide a clear description of the methodology used for systematic search or for selecting/excluding studies for review.
- Heterogeneous datasets and metrics: The results shown are from various datasets, experiments, or metrics, such that direct numerical comparison of various studies is difficult; the survey is more into citing numbers without focusing too much on their comparison.
- Limited depth on evaluation protocols: Information such as cross-validation techniques, train-test ratios, approaches for dealing with class imbalance, or hyperparameter adjustment techniques in other research works cited is not explored in-depth.
- Minimal discussion of clinical context: Although the introduction provides some context about the clinical significance of cardiovascular disease, there is a minimal discussion about how these models work with the clinical context or how they compare with existing models.

Validity and Reliability of Results

As it is a secondary study, the “results” here are actually collective results of other research papers. How valid the conclusion is would entirely depend on the caliber of the primary research sources, which is not formally appraised by the authors. The qualitative trends, such as “Random Forest and ensembles tend to work well” or “overfitting is a problem for individual decision trees,” actually make sense and are not unheard of in the realm of ML research, and that is what adds face validity. But it would be difficult to vouch for the validity of performance differences, which could be dependent on various experimental factors.

Contribution and Relevance

Importance to the field

This study makes its own contribution by pointing to the need for a more organized presentation of the existing research on machine learning for heart disease prediction methods. This allows researchers who are concerned with medical data mining to gain a starting point of understanding regarding the algorithms that have been used along with the techniques of dimensionality reduction that work well.

Practical or theoretical value

In practice, it can help practitioners or students selecting candidate algorithms and preprocessing techniques in designing their own heart disease prediction systems by indicating that SVM, Random Forest, and ensembling could be good initial options, especially with effective feature selection. On the theoretical front, it does not present any models, but it reiterates the importance of understanding that performance for classified models in biomedicine is not only dependent on the classifier but also on the features used, and the risk of overfitting.

Relevance to artificial intelligence, research, or real-world problems

This paper can be used in conjunction with the empirical paper by Jindal et al. for a course on machine learning or data mining:

- The paper by Jindal et al. describes a concrete system implementation;
- Ramalingam et al. illustrate the larger context of algorithms and design decisions that exist in many systems.

A research application of this survey could be in understanding the often utilized approaches as well as the shortcomings (e.g., small data sizes, overfitting) that could act as a driving force for more robust research in the future. In the real-world health care setting, it is emphasized that various approaches of ML have already been utilized for predicting heart diseases with encouraging outcomes.

Conclusion

Total evaluation of the paper

On the whole, “Heart disease prediction using machine learning techniques: a survey” is a compact and informative introduction to the application of classical machine learning approaches for heart disease prediction. It is primarily good for its pertinent topic, discussion of important algorithms, emphasis on dimensionality reduction, and can be used as supplementary or related work material for student projects or researches. It is weakened by its lack of systematic review procedure, insufficient critical appraisal of primary literature, and inability for direct comparison of diverse findings, among other things.

Recommendations (areas for improvement or future research)

- Systematic review practices: Future research could hone search approaches, inclusion/exclusion criteria, and quality measurement systems to provide a more systematic evidence base.
- Standardized reporting: A standardized set of reporting practices, incorporating standardized benchmarking indicators (accuracy, AUC ROC, sens, spec), would make it easier to compare the results of various experiments.
- Clinical integration: Research should investigate comparisons of ML models with current risk scores, integrating the models into practice, or the hurdles that must be overcome before implementation.
- Meta analysis/synthesis: If possible, quantitative meta analysis or at least some form of comparison of performance in similar conditions would improve the conclusion of which algorithms actually work best.
- New approaches: As deep learning techniques and Explainable ML algorithms become more mature, these should be explored in more recent surveys, especially for high dimensional data like ECG, images, or multimodal data.