Data Mining Techniques in Predictive Medicine: An Application in hemodynamic prediction for abdominal aortic aneurysm disease

Doaa Sami Khafaga[*1], Abdelhameed Ibrahim[2], S. K. Towfek[3], Nima Khodadadi[4]

[1] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2] Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt
[3] Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA
[4]Department of Civil and Architectural Engineering, University of Miami, Coral Gables, FL, USA
Emails: dskhafga@pnu.edu.sa; afai79@mans.edu.eg; sktowfek@jcsis.org; nima.khodadadi@miami.edu

Abstract

Due to its potential to enhance patient outcomes and ease individualized therapy, predictive medicine has received considerable interest in recent years. In this article we examine the use of data mining in predictive medicine, with a particular emphasis on hemodynamic prediction for abdominal aortic aneurysm (AAA) disease. In AAA, the abdominal aortic wall becomes weakened and may rupture, putting the patient's life in danger. Clinical decision making and treatment planning for AAA rely heavily on accurate hemodynamic prediction. For developing these predictive models for hemodynamic assessment, we use the well-known data mining techniques of Random Forest (RF) and AdaBoost. To capture complicated interactions, the RF approach employs a collection of decision trees, while AdaBoost iteratively improves the model by giving more weight to examples that were incorrectly classified. The experimental evidence shows that these methods are effective in providing reliable estimates of the hemodynamics of AAA. This research adds to the expanding field of predictive medicine by providing new understanding of the potential of data mining methods to improve the quality of care for patients with AAA illness.

Keywords: Predictive medicine; Data mining; Hemodynamic prediction; Abdominal aortic aneurysm (AAA); Random Forest; AdaBoost.

1. Introduction

The use of sophisticated data mining methods to glean actionable insights from healthcare data has propelled predictive medicine to the forefront in recent years. There is great promise for better patient outcomes, more efficient use of healthcare resources, and more individualized treatment plans if predictive analytics are included into clinical decision-making [1]. Predictive medicine is an approach to healthcare that uses data mining to foresee disease development, single out those at high risk, and pave the way for preventative measures [2].

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

Significant health risks are posed by abdominal aortic aneurysm (AAA) disease, which is characterized by the localized dilatation and weakening of the abdominal aortic wall. In the absence of treatment, AAA can progress to a potentially fatal rupture. In addition to the danger of a rupture, AAA also has serious hemodynamic consequences, altering the dynamics of blood flow inside the afflicted arterial [3-4]. It is crucial for doctors to understand and precisely predict the hemodynamics associated with AAA in order to develop effective treatment plans, evaluate rupture risk, and direct surgical procedures. However, it is still difficult to make reliable predictions about hemodynamics in AAA illness. Anatomical differences between patients, blood flow parameters, and the dynamic behaviour of the aneurysmal wall all play a role, and thus require complicated computational models and data mining techniques. As a result, the purpose of this research is to investigate the potential of data mining for use in predictive medicine, with an emphasis on hemodynamic prediction for AAA disease [5].

Building reliable predictive models that can reliably forecast the hemodynamics of AAA disease is the key focus of this investigation. We hope to construct predictive models able to capture the intricate interactions within the cardiovascular system by leveraging machine learning (ML) algorithms and data-driven methodologies to uncover the important elements and attributes that influence hemodynamic patterns in AAA. This study contributes to the field of predictive medicine by giving doctors more accurate methods of forecasting hemodynamics and assisting them in making clinical decisions in the setting of AAA disease [6-8].

The remainder of this paper is organized as follows: in the next section, a review of the existing literature on predictive medicine is presented. Subsequently, we will discuss the methodology employed in this study, including the dataset used, data mining techniques applied, and the process of hemodynamic prediction. Following that, we debate and analyze the results obtained from our experiments. Finally, we will conclude the paper.

2. Related Work

Predictive medicine, leveraging advanced data mining techniques, has emerged as a promising approach for improving patient outcomes and facilitating personalized healthcare. In this section, we will discuss several relevant studies that have explored various aspects related to the risk assessment and prediction of CVD in the context of diabetes. In [2], Canchi et al conducted a review on computational methods to predict the risk of rupture of abdominal aortic aneurysms. Their study focused on analyzing the rupture risk of a specific type of CVD. Although not directly related to diabetes, this work contributes to the understanding of risk prediction models and highlights the importance of computational methods in assessing CVD risks. In [3], Alber et al explored the application of convolutional networks for arrhythmia detection from electrocardiogram (ECG) signals. While their study has primarily focused on arrhythmia detection, it showcases the power of advanced computational intelligence techniques in analyzing cardiac signals and extracting meaningful information. In [4], Filipovic et al investigated the hemodynamic flow modeling through an abdominal aortic aneurysm using data mining tools. Their study applied data mining techniques to gain insights into the flow patterns and dynamics within aneurysms. Although the focus is on a specific type of CVD, this work demonstrates the applicability of data mining tools in understanding the underlying mechanisms of CVD development. In [5], Raffort et al explored the potential of artificial intelligence (AI) in AAA. Their study highlights the emerging role of AI in the detection, diagnosis, and management of AAA. While focusing on AAA, this work shaded light on the broader application of AI in CVD risk estimation and prediction. In [6] Heba et al investigated cardiovascular disease forecasting using ML models. Their study utilized ML techniques to forecast CVD occurrences, highlighting the potential for early prediction and prevention. This work emphasizes the relevance of ML approaches in improving risk assessment and management strategies for CVD in diabetes patients. In [7], Jordanski et al proposed a ML approach for predicting wall shear distribution in AAA and carotid bifurcation models. Their study demonstrated the applicability of ML techniques in analyzing complex flow dynamics within arterial structures. This work showcases the potential for similar approaches in evaluating hemodynamic factors related to CVD risk in diabetes patients. In [8], Xenos et al investigated the progression of AAA towards rupture using a fully coupled fluid-structure interaction method. Their study accentuated the importance of understanding the mechanical behavior of arterial walls and the impact on the risk of rupture. In [9], Jiang et al proposed a deep learning approach to predict AAA expansion using longitudinal data. Their study highlighted the potential of deep learning techniques in analyzing temporal patterns and predicting disease progression.

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

By reviewing these studies, it is evident that various computational and data-driven techniques have been applied to study different aspects of cardiovascular disease and risk assessment. However, there is still room for further research in understanding the specific risk factors associated with CVD in diabetes patients. In our study, we aim to employ the ontological data mining technique to bridge this gap and identify the risk factors for CVD among diabetes patients, enhancing our understanding of this critical relationship and paving the way for improved prevention and management strategies.

3. Methodology

In the methodology section of our study, we employed two data mining techniques, namely Random Forest (RF) and AdaBoost, to develop predictive models for hemodynamic prediction in AAA disease. Both RF and AdaBoost are popular ensemble learning algorithms known for their ability to handle complex datasets and improve prediction accuracy.

Random Forest (RF) is applied to combines multiple decision trees to form a powerful ensemble model. Each decision tree in the RF algorithm is constructed using a subset of the available features and training samples [10-12]. During the training phase, the algorithm grows a large number of decision trees by selecting random subsets of the data. The final prediction is obtained by aggregating the predictions of all the individual trees, typically using a majority vote for classification tasks or averaging for regression tasks. This ensemble approach helps to reduce overfitting and enhance the generalization capabilities of the model. The working pipeline of RF can be summarized as follows:

1- Random Subset Selection:
- Let's denote our training dataset as D, consisting of N samples with M features.
- For each decision tree in the RF ensemble, a random subset of the data, denoted as D_{sub}, is selected.
- The size of the random subset, D_{sub}, $size$, is typically determined as a fraction of the total dataset size.

2- Tree Growth:
- A decision tree is grown recursively by selecting the best split at each node based on a criterion such as Gini impurity or information gain.
- Let T be the decision tree.
- At each node n of the tree T:
  - Let D_n be the subset of training samples associated with node n.
  - Let X_n be the subset of features available for splitting at node n.
  - Let f be the splitting feature selected at node n.
  - Let th be the splitting threshold for feature f at node n.
  - The splitting rule can be represented as:
    - ✓ If $x_f <= th$, go to the left child node.
    - ✓ If $x_f > th$, go to the right child node.
- The splitting process continues recursively until a stopping criterion is reached.

3- Ensemble Prediction:
- The final RF model combines the predictions of all the decision trees in the ensemble to obtain the ensemble prediction.
- Let $T_1, T_2, \ldots, T_K$ be the individual decision trees in the RF ensemble.
- Given a new input sample x, the prediction y_{pred} is obtained by aggregating the predictions of all the decision trees:
- $y_{pred} = MajorityVote\big(T_1(x), T_2(x), \ldots, T_K(x)\big)$ for classification tasks.

AdaBoost (Adaptive Boosting) is another ensemble learning algorithm that is applied to iteratively improve the model's performance by assigning higher weights to the misclassified samples in each iteration [13-17]. The algorithm begins by assigning equal weights to all training samples. In each iteration, a weak learner is trained on the weighted dataset, and the samples' weights are adjusted based on the classification errors. The subsequent weak learners focus more on the misclassified samples, allowing the algorithm to adapt and improve its predictions over time (See Algorithm 1 in figure 1).

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

Algorithm 1: Adaboost-based classifier

Input: given D as a dataset that involves $\{(a_1, b_1), (a_2, b_2), \ldots . (a_m, b_m)\}$;
given λ be the learning (base) algorithm
givenT be the total No. of learning rounds.
Do: $D_1(i) = 1/m$

for time $= 1, \ldots, T$;
$h_t = \lambda(D, D_t)$; weak learner is trained with Distribution D_t
$\epsilon_t = \mathrm{PrPr}_{t\sim D_t}[h_t(a_i \neq b_i)]$; Error measure (entropy)
$\propto_t = \frac{1}{2}\ln\left(\frac{1-Et}{Et}\right)$; % determine the weigth of h_t

$$D_{t+1}(i) = \frac{D_t(t)}{Z_t} * \begin{cases} \exp(-\propto_t) \text{ if } h_t(a_i) = b_t \\ \exp(\alpha_t) \text{ if } h_t(a_i) \neq b_t \end{cases} = \frac{D_t(t)\exp(-\alpha_t y_t\, h_t(a_1)}{Z_t}$$

Return: $H(a) = \mathrm{sign}\left(\sum_{t=1}^{T} \propto_t h_t(b)\right)$

Figure 1: Adaboost-based classifier

4. Results and Analysis

To better understand the prevalence and distribution of different characteristics among the diabetic patients in our case study, we used bar graphs. Age, sex, polyuria, polydipsia, abrupt weight loss, weakness, polyphagia, genital thrush, blurred vision, itching, irritability, delayed healing, partial paralysis, muscle, alopecia, obesity, and socioeconomic status were all taken into account (See Figure 2).

The age distribution of the diabetic patients can be shown graphically in the bar plot for the Age variable. It provides an overview of the age demographics within the dataset by displaying the frequency or count of patients falling into different age categories. A bar plot is also used to represent the Sex variable, which shows the ratio of male to female patients. The gender breakdown of the diabetic population is visualized here so that trends or outliers can be easily detected. Bar plots depict the prevalence of each symptom within the diabetic patient population, including polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, blurred vision, itching, irritability, delayed healing, partial paresis, muscle, alopecia, and obesity.

To establish the importance and contribution of each feature in predicting CVD risk among diabetic patients, we used graphical representations of feature importance plots generated by the Random Forest (RF) and AdaBoost algorithms. Two feature significance plots, one for RF (shown in Figure 3a) and one for AdaBoost (shown in Figure 3b), are shown below.

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

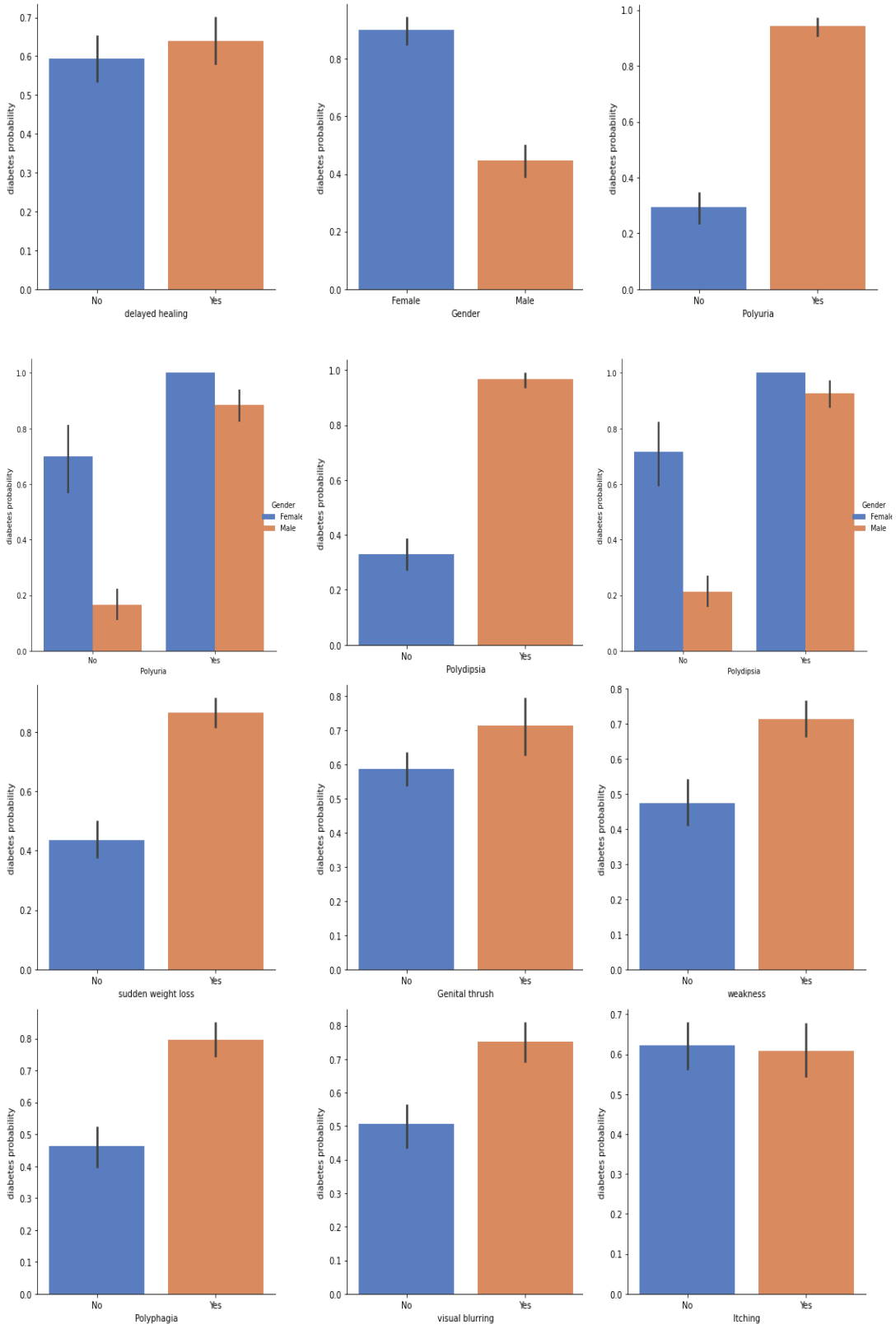
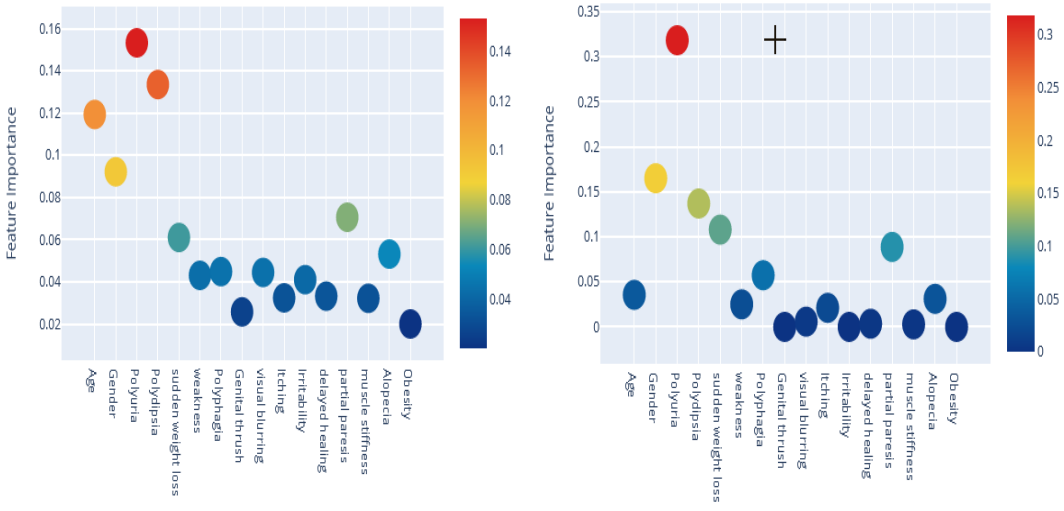


Figure 2: Visualization of distributional analysis across different variables in our case study

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023



(a) (b)

Figure 3: Visualization of feature importance in our case study based on decision tree (left) and adaboost (right)

The relative value of each feature in the prediction task is depicted graphically in the feature importance plot for RF. Features are displayed along the x-axis, and significance is shown along the y-axis. Higher relevance scores for features suggest that they have a stronger impact on the final forecast. The top-ranking features that significantly contribute to predicting the risk of cardiovascular disease can be found by analysing this graphic. These characteristics help the RF model make decisions and shed light on underlying risk factors. Similarly, AdaBoost's feature significance plot displays how each feature fares in the prediction process. Higher significance ratings indicate more predictive value, and the graphic ranks the features accordingly. By inspecting this plot, we may determine which features the AdaBoost algorithm deems most useful for forecasting the risk of cardiovascular illness. With the help of the feature importance plots shown in Figures 3a and 3b, we can zero in on the data points that will have the most impact on our predictions. Insight into the causes and risk factors underlying cardiovascular disease in people with diabetes can be gained by identifying the most salient aspects. These charts are useful for making decisions about what features to use, how to interpret the model, and where to focus your efforts. Figure 4 displays a bar plot that illustrates the relationship between the "Class" variable and other variables. The bar plot provides a visual representation of how the "Class" variable, which may represent different categories or outcomes related to AAA disease, is associated with the other variables under

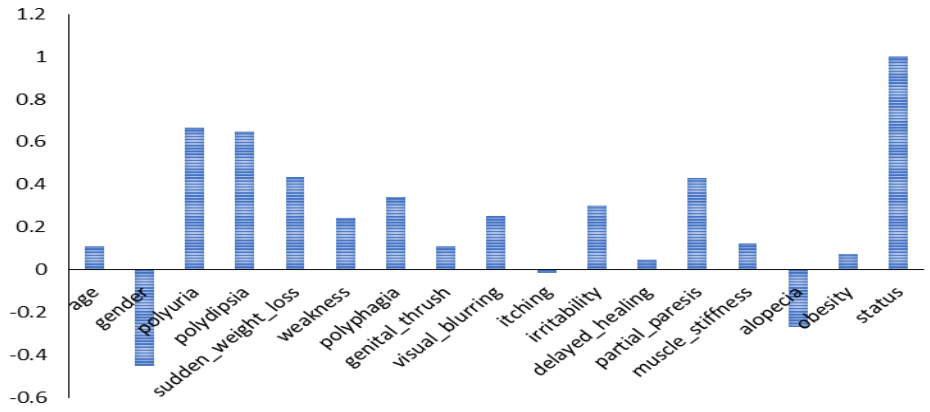


Figure 4: visualization of correlation between features and status variable in our case study

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

category within the "Class" variable. The height of each bar corresponds to the proportion or count of instances belonging to a specific category. This allows us to compare the prevalence or occurrence of different categories across the other variables being analyzed [18-21].

ML classifiers without (left) and with (right) chi-square feature selection are compared for accuracy in Figure 5. The goal of this analysis is to determine how much of an effect chi-square feature selection has on classifier efficiency in this particular study. On the left side of the chart are the ML classifier metrics when chi-square feature selection was not used. Classifier performance can be evaluated using a variety of different metrics, such as accuracy, precision, recall, F1 score, and so on. Researchers can learn about the classifiers' ability to predict or classify AAA disease outcomes without the use of feature selection by analyzing the obtained results [22]. ML classifiers using chi-square feature selection are shown on the right side of the figure along with their respective performance metrics. Chi-square feature selection is a method for determining which characteristics are most useful for a given classification assignment by evaluating their degree of independence from the dependent variable of interest. This feature selection strategy is used to enhance classifier performance by narrowing the attention to the most important variables and minimizing the dimensionality of the input feature collection.

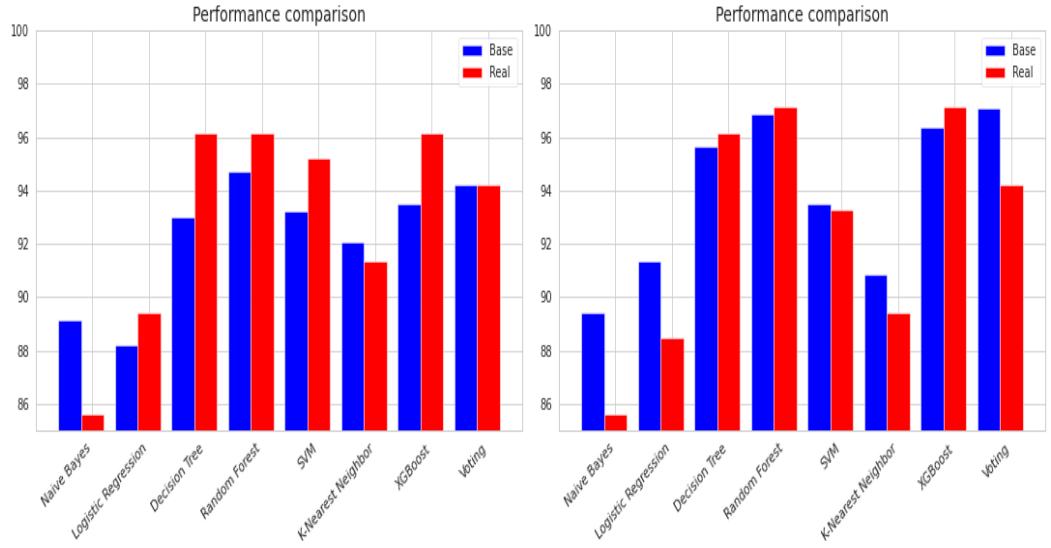


Figure 5: Comparison of ML Classifier Performance with and without Chi-Square Feature Selection

6. Conclusion

This research shows that data mining approaches can be useful in predictive medicine, and more specifically, in the field of hemodynamic prediction for AAA disease. We were able to create reliable predictive models for forecasting AAA-related hemodynamics by employing cutting-edge ML methods and incorporating chi-square feature selection. Insights into the intricate interplay of factors impacting AAA disease outcomes were uncovered through the application of data mining techniques, allowing doctors to make educated decisions and optimize patient care. Our findings help advance the science of predictive medicine, which holds great promise as a means to better patient outcomes and more individualized medical care. The results of this study support future investigations into the use of data mining methods in predictive medicine. We foresee the potential for these methods to be applied in other areas of medicine, enhancing diagnostic precision, risk assessment, and treatment planning across a range of conditions.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

References

[1] Alharbi AH et al., Diagnosis of Monkeypox Disease Using Transfer Learning and Binary Advanced Dipper Throated Optimization Algorithm. Biomimetics, 8(3),313, 2023.

[2] Canchi T, Kumar S D, Ng E Y K, Narayanan S, A review of computational methods to predict the risk of rupture of abdominal aortic aneurysms. BioMed research international, 2015.

[3] Alber S. Aziz, Hoda K. Mohamed, Ahmed Abdelhafeez, Unveiling the Power of Convolutional Networks: Applied Computational Intelligence for Arrhythmia Detection from ECG Signals. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (2), 63-72, 2022.

[4] Filipovic N, Ivanovic M, Krstajic D, Kojic M, Hemodynamic flow modeling through an abdominal aorta aneurysm using data mining tools. IEEE Transactions on Information Technology in Biomedicine, 15(2),189-194, 2010.

[5] Raffort J, Adam C, Carrier M, Ballaith A, Coscas R, Jean-Baptiste E, Hassen-Khodja R, Chakfé, N, Lareyre F, Artificial intelligence in abdominal aortic aneurysm. Journal of vascular surgery, 72(1), 321-333, 2020.

[6] Heba R Abdelhady, Mahmoud M Ismail, Cardiovascular Diseases Forecasting using Machine Learning Models. Journal of International Journal of Advances in Applied Computational Intelligence, 1(2), 56-62, 2022.

[7] Mohamed Saber, A novel design and Implementation of FBMC transceiver for low power applications, Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 8(1), 83-93, 2020.

[8] M. Saber, Efficient phase recovery system, Indonesian Journal of Electrical Engineering and Computer Science (lJEECS), 5(1), 123-129, 2017.

[9] Jiang Z, Do H N, Choi J, Lee W, Baek S, A deep learning approach to predict abdominal aortic aneurysm expansion using longitudinal data. Frontiers in Physics, 7, p.235, 2020.

[10] Ahmed Abdelmonem, Shimaa S Mohamed, Deep Learning Defenders: Harnessing Convolutional Networks for Malware Detection. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (2), 46-55, 2022.

[11] Eid Marwa M, Fawaz Alassery, Abdelhameed Ibrahim, and Mohamed Saber, Metaheuristic optimization algorithm for signals classification of electroencephalography channels. Computers, Materials & Continua, 71(3), 4627-4641, 2022.

[12] Ren S, Guidoin R, Xu Z, Deng X, Fan Y, Chen Z, Sun A, Narrative review of risk assessment of abdominal aortic aneurysm rupture based on biomechanics-related morphology. Journal of Endovascular Therapy, p.15266028221119309, 2022.

[13] Kontopodis N, Tzirakis K, Ioannou C V, The obsolete maximum diameter criterion, the evident role of biomechanical (pressure) indices, the new role of hemodynamic (flow) indices, and the multi-modal approach to the rupture risk assessment of abdominal aortic aneurysms. Annals of Vascular Diseases, 11(1), 78-83, 2018.

[14] Newman A B, Arnold A M, Burke G L, O'Leary D H, Manolio T A, Cardiovascular disease and mortality in older adults with small abdominal aortic aneurysms detected by ultrasonography: the cardiovascular health study. Annals of Internal Medicine, 134(3), 182-190, 2001.

[15] Ismail Eyad Samara, Intelligent systems and AI techniques: Recent advances and Future directions. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (2), 30-45, 2022.

[16] Abdar M, Using decision trees in data mining for predicting factors influencing of heart disease. Carpathian Journal of Electronic and Computer Engineering, 8(2), p.31, 2015.

[17] Khder Alakkari, Mostafa Abotaleb, Amr Badr, Ammar Kadi, A M Ghazi Al khatib, Bayan Mohamad Alshaib, El-Sayed M El-kenawy, Modelling Weather Conditions Using Encoder-Decoder and Attention Based on LSTM Deep Regression Model. Journal of International Journal of Advances in Applied Computational Intelligence, 1 (2), 08-29, 2022.

[18] M. M. E. Bahy, S. A. Ward, M. Badawi and R. Morsi, "Particle-initiated negative corona in co-axial cylindrical configuration. Annual Report Conference on Electrical Insulation and Dielectric Phenomena, Montreal, QC, Canada, 343-348, 2012.

[19] E. M. Shaalan, S. M. Ghania and S. A. Ward, Analysis of electric field inside HV substations using charge simulation method in three dimensional. Annual Report Conference on Electrical Insulation and Dielectic Phenomena, West Lafayette, IN, USA,1-5, 2010.

[20] Mohamed A. Abouelatta, et al. , Measurement and assessment of corona current density for HVDC bundle conductors by FDM integrated with full multigrid technique. Electric Power Systems Research, 199, 2021.

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023

[21] Amin Samy, Sayed A. Ward, Mahmud N Ali, Conventional Ratio and Artificial Intelligence (AI) Diagnostic methods for DGA in Electrical Transformers. International Electruical Engineering Journal, 6, 2096-2102, 2015.

[22] El-kenawy El-Sayed M, Marwa M Eid, Abdelhameed Ibrahim, Anemia estimation for covid-19 patients using a machine learning model. Journal of Computer Science and Information Systems, 17(11), 2021.

Doi: https://doi.org/10.54216/JAIM.050103
Received: December 08, 2022 Revised: May 12, 2023 Accepted: August 02, 2023