

Evaluation of feature scaling for improving the performance of supervised learning methods

Tsehay Admassu Assegie¹, Vadivel Elanangai², Josephin Shermila Paulraj³, Mani Velmurugan⁴,
Daya Florance Devesan⁵

¹Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia

²Department of of Electrical and Electronics Engineering, St. Peter's Institute of Higher Education and Research, Avadi, India

³Department of Artificial Intelligence and Data Science, R.M.K College of Engineering and Technology, Chennai, India

⁴Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India

⁵Department of Computer Science and Engineering, Velammal Engineering College, Chennai, India

Article Info

Article history:

Received Nov 11, 2022

Revised Nov 27, 2022

Accepted Dec 16, 2022

Keywords:

Decision tree

Feature selection

Heart failure

Random forest

Support vector machine

ABSTRACT

This article evaluates the performance of the support vector machine (SVM), decision tree (DT), and random forest (RF) on the dataset that contains the medical records of 299 patients with heart failure (HF) collected at the Faisalabad Institute of Cardiology and the Allied hospital in Pakistan. The dataset contains 13 descriptive features of physical, clinical, and lifestyle information. The study compared the performance of three classification algorithms employing pre-processing techniques such as min-max scaling, and principal component analysis (PCA). The simulation result shows that the performance of the DT, and RF decreased with dimensionality reduction while the SVM improved with dimensionality reduction. The SVM achieved 84.44%. Thus, feature scaling improves the performance of the SVM. The RF performs at 82.22%, the DT at 81.11%, and the SVM shows an improvement of 1.64% with scaled features, compared to the original dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tsehay Admassu Assegie

Department of Computer Science, College of Engineering and Technology, Injibara University

Injibara, Ethiopia

Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

In recent years, machine learning has been widely applied in the healthcare industry for medical decision-making. Various machine-learning models such as supervised learning methods have been widely researched for the automation of medical decision-making aiding the early diagnosis of heart disease [1]. One of the most widely employed pre-processing methods is feature selection, which improves the effectiveness of supervised learning for heart disease diagnosis [2].

Research by Assegie *et al.* [3] developed a cox-based model for heart failure (HF) survival prediction. The simulation of the study revealed that the cox-based model performs with a C-index receiver operating characteristic (ROC) of 0.74 and a log-likelihood ratio of 81.95 on 11 degrees of freedom on the validation dataset. While the study has shown the encouraging result of the use of the cox-based model for HF survival prediction, the study did not focus on the pre-processing of the original HF dataset.

Research by Zhang *et al.* [4] highlighted the importance of class balancing for improving the performance of light gradient boosting machine (light-GBM) for coronary artery diagnosis. The simulation of the study has shown that the performance of the proposed model improved on the balanced dataset. In the

study, synthetic minority oversampling (SMOTE) is applied for class balancing. Despite the use of the class balancing method for model performance improvement, the study did not suggest the significance of other pre-processing methods such as feature scaling and feature reduction with the principal component analysis (PCA).

Habib and Tasnim [5] developed an ensemble-voting model for cardiovascular disease diagnosis. The study suggested the early diagnosis of HF with a supervised learning model is significant to save the patient life. The model is implemented using gradient boosting, Gaussian naïve Bayes (GNB), random forest (RF), and multi-layer perceptron (MLP) as base classifiers. The simulation of the performance of the developed model shows that the GNB outperforms the other model with an accuracy of 74%.

Several studies [6], [7] have proposed a neural network (NN) based heart disease prediction model. The study [6] compared RF and logistic regression (LR) for HF prediction and tested the performance of the RF, and LR on an electronic HF dataset. The result reveals that the RF model performs better than the LR for the HF dataset. In addition, the study highlighted that feature selection improves the performance of NN for HF prediction [7].

In addition, another study applied support vector machine (SVM) for HF prediction [8]. The researcher tested the developed SVM model on the HF dataset and the result highlights that the SVM model has shown 92.22% accuracy on the HF simulation dataset. The result obtained shows that the SVM model appears effective in HF prediction although the model has scope for improvement with pre-processing and feature selection.

Despite the wider application of various supervised learning methods such as SVM [9], [10], deep learning [11], RF, decision tree (DT) [12], MLP [13], and LR [14] for HF prediction, the effectiveness of these HF prediction methods have scope for improvement and requires much research effort. The literature review shows that most of the supervised methods are developed on the original dataset and the significance of pre-processing such as feature scaling, and dimensionality reduction for the linear model such as SVM is widely ignored [15]–[20]. To address the research gap, this study investigates feature scaling, and PCA as a method for improving the performance of DT, RF, and SVM models for HF prediction. The objective of this study are: i) to provide a literature review of the HF prediction model; ii) to apply feature scaling and PCA and evaluate the performance of DT, RF, and SVM; and iii) to compare the performance of DT, SVM, and RF on original and pre-processed data. The rest of the article is organized as follows: section 2 presents the method, section 3 presents the result and discusses the result, and the section 4 concluded the work.

2. METHOD

This study employed an HF dataset obtained from the Institute of Cardiology and Allied hospital by Ahmad *et al.* [21] previously studied by [22]–[25] for HF survival prediction. The dataset contains 299 samples of patients aged above 40 years. The HF dataset contains 105 womens and 194 mens. The researchers followed the following steps to conduct this study. The first step involved a dataset collected from the public Kaggle data repository. The second step involves the manual assessment of the dataset checking the difference in the magnitude of different continuous features. The third step pre-processed the dataset by feature scaling before training DT, RF, SVM, and model to avoid the impact of the magnitude on model performance. The final step applied PCA to the dataset and compared the performance of the model on reduced and original features. The model is trained using k-1 of the folds as training data. The remaining part uses testing data to measure the accuracy of each fold. Table 1 describes the HF features used in the study.

Table 1. The feature description of the HF dataset

Feature	Description	Statistics
Age	Age of patient	Minimum age 40 and maximum age 95
Anemia	Whether the patient has anemia or not	0=absence (170 patients) and 1=presence (129)
Diabetes	Whether the patient has diabetes or not	0=absence (174 patients) and 1=presence (125 patients)
Sex	Sex of patient	0=female (105 patients) and 1=male (194)
Smoking	whether the patient has a smoked or not	0=not smoked (203 patients) and 1=smoked (96 patients)
Platelets	Platelets in the blood in kiloplatelets/mL	Range: 25.01–850.00 and mean=263.358
Blood pressure	Absence/presence of hypertension	0=absence (194 patients) and 1=presence (105 patients)
CPK	CPK enzyme in blood in mcg/L	Range: 23–7861 and mean=581.839
Serum creatinine	creatinine in blood in mg/ dL	Range: 0.50–9.40 and mean=1.394
Serum sodium	sodium in blood in mEq/L	Range: 114–148 and mean=136.625
DEATH_EVENT	label	0=survived (203 patients) and 1=deceased (96 patients)
Ejection fraction	Blood leaving the heart at each contraction	Range: 14–80 and mean=38.084

3. RESULTS AND DISCUSSION

In this section, the simulation results of the research are presented. The performance of the DT, SVM, and RF models is evaluated on the original 299 of which 203 are negative or not suffering from HF, and 96 belong to the positive class or suffer from HF. In the simulation, 70% of the dataset, or 142 samples belonging to the negative class and 67 belonging to the positive class were used to train the model. Then the model is tested or evaluated on 30% or 61 samples belonging to the negative class and 29 samples belonging to the positive class. The details of the simulation results are presented in subsections 3.1 and 3.2.

3.1. Performance of decision tree, and random forest on the original dataset

Figures 1 and 2 illustrate the effect of depth on the performance of the DT and RF model respectively. As demonstrated in Figure 1, the DT model performance is better when the depth of the tree is lower. The highest cross-validated accuracy is obtained at depth of 1 as illustrated in Figure 1. However, the accuracy increases with an increased depth value as the DT model overfits. Thus, a better cross-validation value is achieved at depth of 1 for the DT model.

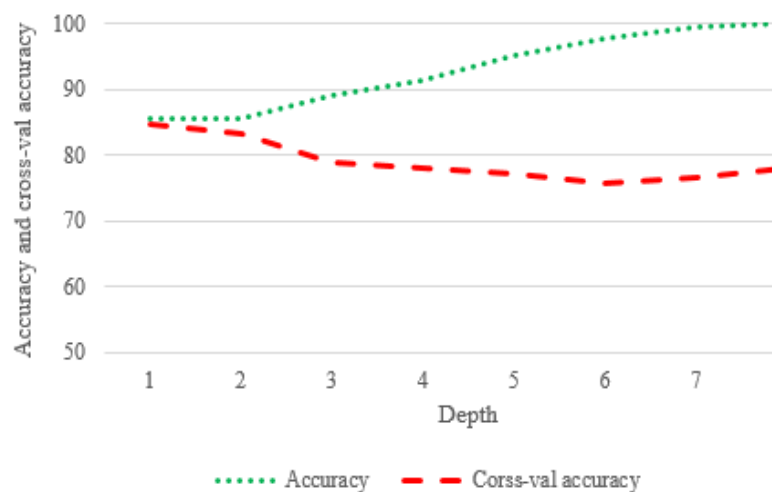


Figure 1. Depth vs the performance of the DT model

Figure 2 demonstrates the performance of the RF model on the original HF dataset. The cross-validated accuracy varies with variation in the depth of the RF model as illustrated in Figure 2. The highest cross-validated accuracy is achieved at depth of 5. Thus, the RF model achieves the highest accuracy for HF prediction when trained with a depth value of 5.

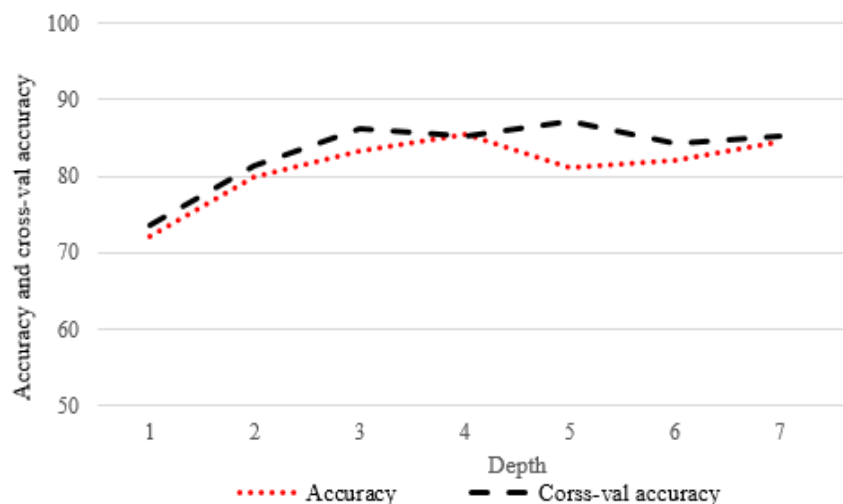


Figure 2. Depth vs the performance of the RF model

3.2. The effect of principal component analysis on the performance of the model

This section discusses the results obtained from the simulation of the DT, RF, and SVM models with the HF test set. Figure 3 illustrates the effect of the PCA on the performance of the DT, RF, and SVM models. The DT and RF models performed well on the original HF dataset compared to the PCA-reduced feature set. However, the performance of the SVM model improves with the PCA illustrated in Figure 3.

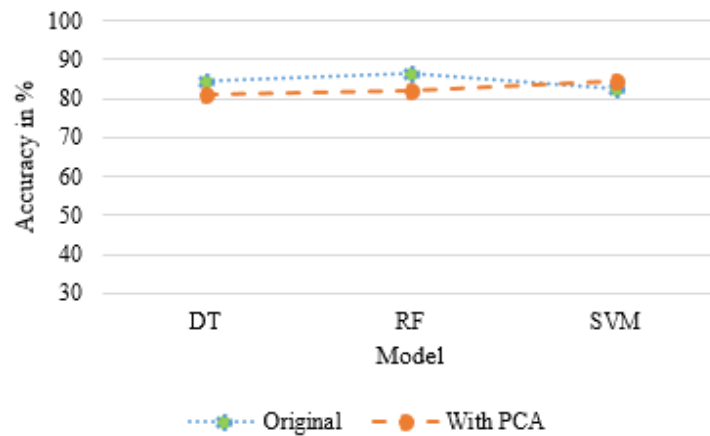


Figure 3. The effect of PCA on the performance of DT, RF, and SVM

4. CONCLUSION

This study presented a comparative analysis of the performance of three supervised learning methods such as SVM, DT, and RF. The study compared the effectiveness of these methods on the original and scaled 299 HF dataset with a min-max scaler. Furthermore, the study compared the effectiveness of the models on pre-processed and PCA-component-reduced HF datasets. The simulation result shows that the performance of the RF classifier technique performed scoring the highest accuracy without PCA and feature scaling having an accuracy of 86.62%. Thus, it is been shown that the RF model can assist the decision-making process for identifying the HF.




REFERENCES

- [1] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of machine learning tools in healthcare decision making," *Journal of Healthcare Engineering*, pp. 1–20, Jan. 2021, doi: 10.1155/2021/6679512.
- [2] R. Porto, J. M. Molina, A. Berlanga, and M. A. Patricio, "Minimum relevant features to obtain explainable systems for predicting cardiovascular disease using the statlog data set," *Applied Sciences*, vol. 11, no. 3, pp. 1–18, Jan. 2021, doi: 10.3390/app11031285.
- [3] T. A. Assegie, T. Karpagam, S. Subramanian, S. M. Janakiraman, J. Arumugam, and D. O. Ahmed, "Prediction of patient survival from heart failure using a cox-based model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 3, pp. 1550–1556, Sep. 2022, doi: 10.11591/ijeecs.v27.i3.pp1550-1556.
- [4] S. Zhang, Y. Yuan, Z. Yao, J. Yang, X. Wang, and J. Tian, "Coronary artery disease detection model based on class balancing methods and LightGBM algorithm," *Electronics*, vol. 11, no. 9, pp. 1–44, May 2022, doi: 10.3390/electronics11091495.
- [5] A.-Z. S. bin Habib and T. Tasnim, "An ensemble hard voting model for cardiovascular disease prediction," in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dec. 2020, pp. 1–6, doi: 10.1109/STI50764.2020.9350514.
- [6] B. Wang *et al.*, "A multi-task neural network architecture for renal dysfunction prediction in heart failure patients with electronic health records," *IEEE Access*, vol. 7, pp. 178392–178400, 2019, doi: 10.1109/ACCESS.2019.2956859.
- [7] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," *Mobile Information Systems*, pp. 1–11, Aug. 2020, doi: 10.1155/2020/8843115.
- [8] L. Ali *et al.*, "An expert system based on optimized stacked support vector machines for effective diagnosis of heart disease," *IEEE Access*, vol. 7, pp. 54007–54014, 2019, doi: 10.1109/ACCESS.2019.2909969.
- [9] J. Vijayashree and H. P. Sultana, "A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier," *Programming and Computer Software*, vol. 44, no. 6, pp. 388–397, Nov. 2018, doi: 10.1134/S0361768818060129.
- [10] E. Owusu, P. Boakye-Sekyerehene, J. K. Appati, and J. Y. Ludu, "Computer-aided diagnostics of heart disease risk prediction using boosting support vector machine," *Computational Intelligence and Neuroscience*, pp. 1–12, Dec. 2021, doi: 10.1155/2021/3152618.
- [11] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational Intelligence and Neuroscience*, pp. 1–11, Jul. 2021, doi:




- 10.1155/2021/8387680.
- [12] L. Chandrika and K. Madhavi, "A hybrid framework for heart disease prediction using machine learning algorithms," in *E3S Web of Conferences*, Oct. 2021, pp. 1–7, doi: 10.1051/e3sconf/202130901043.
- [13] P. Gupta and D. Seth, "Comparative analysis and feature importance of machine learning and deep learning for heart disease prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 451–459, Jan. 2022, doi: 10.11591/ijeecs.v29.i1.pp451-459.
- [14] M. O. Rahaman *et al.*, "Internet of things based electrocardiogram monitoring system using machine learning algorithm," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3739–3751, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3739-3751.
- [15] A. A. Almazroi, "Survival prediction among heart patients using machine learning techniques," *Mathematical Biosciences and Engineering*, vol. 19, no. 1, pp. 134–145, 2022, doi: 10.3934/mbe.2022007.
- [16] K. Wang *et al.*, "Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP," *Computers in Biology and Medicine*, vol. 137, pp. 1–9, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104813.
- [17] S. Y. Hwang, K. A. Kim, and O. J. Choi, "Predictive factors on the incidence of HF in patients with ischemic heart disease: Using a 10-year population-based Korea national health insurance cohort data," *International Journal of Environmental Research and Public Health*, vol. 17, no. 22, pp. 1–9, Nov. 2020, doi: 10.3390/ijerph17228670.
- [18] A. Cai, Y. Zhu, S. A. Clarkson, and Y. Feng, "The use of machine learning for the care of hypertension and heart failure," *JACC: Asia*, vol. 1, no. 2, pp. 162–172, Sep. 2021, doi: 10.1016/j.jacasi.2021.07.005.
- [19] J. Chu, W. Dong, and Z. Huang, "Endpoint prediction of heart failure using electronic health records," *Journal of Biomedical Informatics*, vol. 109, pp. 1–12, Sep. 2020, doi: 10.1016/j.jbi.2020.103518.
- [20] Z. Wang, B. Wang, Y. Zhou, D. Li, and Y. Yin, "Weight-based multiple empirical kernel learning with neighbor discriminant constraint for heart failure mortality prediction," *Journal of Biomedical Informatics*, vol. 101, pp. 1–10, Jan. 2020, doi: 10.1016/j.jbi.2019.103340.
- [21] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLOS ONE*, vol. 12, no. 7, pp. 1–8, Jul. 2017, doi: 10.1371/journal.pone.0181001.
- [22] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–16, Dec. 2020, doi: 10.1186/s12911-020-1023-5.
- [23] A. Roy, C. Bruce, P. Schulte, L. Olson, and M. Pola, "Failure prediction using personalized models and an application to heart failure prediction," *Big Data Analytics*, vol. 5, no. 1, pp. 1–19, Dec. 2020, doi: 10.1186/s41044-020-00044-2.
- [24] A. Newaz, N. Ahmed, and F. S. Haq, "Survival prediction of heart failure patients using machine learning techniques," *Informatics in Medicine Unlocked*, vol. 26, pp. 1–10, 2021, doi: 10.1016/j.imu.2021.100772.
- [25] M. M. Nishat *et al.*, "A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyper parameter optimization for imbalanced HF dataset," *Scientific Programming*, pp. 1–17, Mar. 2022, doi: 10.1155/2022/3649406.

BIOGRAPHIES OF AUTHORS






Mr. Tsehay Admassu Assegie    holds a Master of Science degree in Computer Science from Andhra University, India 2016. He received his B.Sc. in Computer Science from Dilla University, Ethiopia in 2013. His research includes machine learning, data mining, health informatics, network security, and software-defined network. He has published over 42 papers in reputed international journals and international conferences. Tsehay is an active member of the International Association of Engineers (IAENG), with membership number: 254711. He can be contacted at email: tsehayadmassu2006@gmail.com.



Ms. Vadivel Elanagai    is currently working as an Assistant Professor in the Department of Electrical and Electronics Engineering at St. Peter's Institute of Higher Education and Research, AVADI, Chennai. She has 11 years of Teaching Experience. She is currently doing her research in image processing. Her current research interest includes image processing, VLSI design, fuzzy logic, and artificial neural network. She has also published research papers in reputed journals and conference proceedings. She can be contacted at email: elanagai123@gmail.com.






Dr. Josephin Shermila Paulraj    is Associate Professor at Department of Artificial Intelligence and Data Science R.M.K. College of Engineering and Technology, Chennai, India. She was born in Kanyakumari District, Tamilnadu, India in 1983. She received her B.E. degree in Electronics and Communication Engineering from Noorul Islam College of Engineering, Kumaracoil, Anna University, India in 2005, and obtained M.E. degree in Computer Communication Engineering from National Engineering College, Kovilpatti, Anna University, India in 2007. She has completed her Ph.D degree in Information and Communication Engineering, Anna University, India. She worked as a Programmer Analyst in Cognizant Technology Solutions from October 2007 to October 2010. She is in teaching profession since November 2010. Currently she is working as an Associate Professor in the Department of Artificial Intelligence and Data Science, R. M. K. College of Engineering and Technology, Chennai. She is a member of Academia, IAENG, and IFERP. She works in the field of Image Processing, Machine Learning, and Deep Learning. Her research area of interest is nutrition estimation from food images. She can be contacted at email: blossomshermi@gmail.com.



Mr. Mani Velmurugan    is graduate from Shanmuganathan Engineering College, Pudukkottai under Trichy Anna University in 2011 and received Master Degree Programme in SNS College of Technology, Coimbatore under Anna University in 2014. He has four years' experience in PHP developer in reputed companies. Currently he was working as a Assistant Professor in Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India. He can be contacted at email: velbesec@gmail.com.



Ms. Daya Florance Devesan    is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Velammal Engineering College, Velammal Nagar, Surapet, Chennai. Her research interests include computer networks and machine learning. She can be contacted at email: dayaflorance@gmail.com.