Improving Heart Failure Prediction Accuracy with PCA and Information Gain in Machine Learning Models

Literature Review

The chapter analyzes the different research methods used to increase the accuracy of heart failure prediction. This section studies dimension reduction techniques and introduces PCA and Information Gain methods for feature selection. The chapter highlights research gaps and displays how to utilize PCA and Information Gain techniques effectively in heart failure prediction systems.

Role of PCA in dimensionality reduction

PCA converts large datasets into primary components through dimensionality reduction to preserve important details in the input data. PCA makes heart failure predictions more accurate because it eliminates unnecessary information which lets the system detect patterns better. PCA helps systems operate faster while going against overfitting patterns to create forecasts that work equally with other data inputs (Reddy et al, 2022). When working with datasets having multiple interrelated attributes this technique produces outstanding results by finding meaningful connections and recognizing key changes that shape output. Heart failure prediction models perform better when PCA joins the process to focus on important data information (Panyamit et al, 2022).

Significance of Feature Selection and the Impact of Information Gain

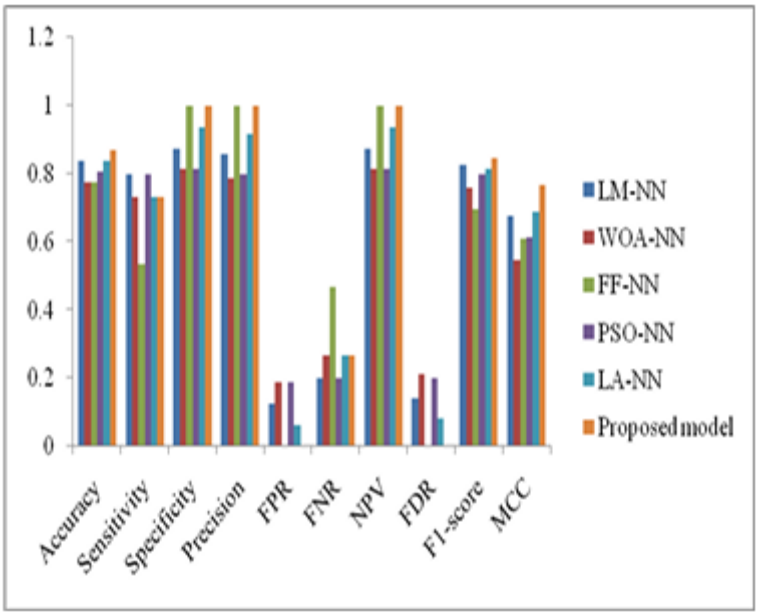
The success of machine learning algorithm performance enhancement requires identification of the most important features. The feature selection process operates on multiple objectives such as decreasing dataset dimensions and eliminating noise elements to achieve more precise and efficient and interpretable models. Information gain represents a standard method to choose features by measuring their ability to reduce uncertainties for target value prediction. The method determines data uncertainty reduction through the examination of partitioned data entropy (Rani *et al*, 2021). High information gain features provide better predictive results to models because the maximum predictive power allows for rapid training and reduced prospect of overfitting. Digital systems achieve better performance and enhanced generalization abilities through this process which serves as a vital operation to develop modern machine learning systems (Win and Kham, 2019).

Recent analyses of PCA and Information Gain in predicting heart failure

The paper by Gárate-Escamila, El Hassani, and Andrès (2020) investigated heart disease prediction through the utilization of feature selection and dimensionality reduction methods to address resources needed for processing 74 features. The researchers obtained their information from the UCI Machine Learning Repository through the combination of Cleveland datasets and Hungarian and Cleveland-Hungarian (CH) datasets. A combined examination of cholesterol and chest pain features took place through Chi-square (CHI) and PCA while performing feature selection. CHI-PCA when combined with Random Forest (RF) obtained the best accuracy levels

of 98.7% (Cleveland), 99.0% (Hungarian) and 99.4%(CH) datasets respectively. Raw data had poor performance after direct application of PCA techniques.

The research targeted heart disease prediction by testing several machine learning methods. Different optimization and deep learning systems were tested by researchers to boost prediction precision. Most research used the Cleveland heart disease dataset available on the UCI Machine Learning Repository with its 74 features. The study aimed to improve predictive results by taking advantage of PCA to shrink data dimensions and develop the best possible classification method. The study worked to build advanced computer systems that process medical data and clinical work better automatically. The AOA-Neural Network approach based on Archimedes Optimization delivered better heart disease prediction than Levenberg-Marquardt-NN and other techniques like PSO-NN, WOA-NN, FF-NN, and LA-NN with improved accuracy levels ranging from 16% to 23%.



Performance analysis of AOA-NN model (Anand, 2021)

The study (Kakoly, Hoque, and Hasan, 2023) aimed to predict diabetes risk factors using ML algorithms, employing two feature selection approaches: PCA and IG. A total of 738 records were obtained for this paper while operating under the Helsinki Declaration guidelines. Random Forest (RF), Logistic Regression (LR), and Decision Tree alongside SVM and K-Nearest Neighbors formed a set of five ML algorithms assessed during testing. The feature selection process using PCA and IG identified glucose and eating meat together with eating fruits and area size with age being the critical characteristics. With Information Gain feature selection and five features Logistic Regression produced 82.2% accuracy with an AUC of 87.2%, as shown below.

Group 4: Features after IG (Top 5 Features)

Model	AUC	CA	F1	Precision	Recall	Speci
LR	88.5%	82.2%	70.6%	82.6%	61.6%	92.5%
RF	79.4%	80.2%	68.7%	75.8%	62.7%	85.5%
DT	72.9%	77.5%	65.8%	69.3%	62.7%	91.1%
KNN	75.4%	77.4%	63.3%	72.0%	56.5%	82.2%
SVM	80.0%	71.7%	63.4%	57.3%	71.0%	98.6%

Outcomes of various machine learning models (Kakoly, Hoque, and Hasan, 2023)

The research by Bhatt et al (2023) designs a ML predictive model for cardiovascular diseases which works to reduce mortality rates through improved diagnostic precision. The dataset used is from Kaggle, containing 70,000 instances and the models were trained using an 80:20 data split. The analysis used RF, DT, MLP, and XGBoost (XGB) algorithms parallel to GridSearchCV implementation for hyperparameter optimization. Cross-validated Multilayer Perceptron (MP) delivered an accuracy rate of 87.28% which proved highest among all trained models. The decision tree achieved 86.37% accuracy followed by XGBoost with 86.87% accuracy then Random Forest with 87.05% accuracy. Decision Tree model achieved an AUC value of 0.94 while XGBoost, RF, and MP attained 0.95.

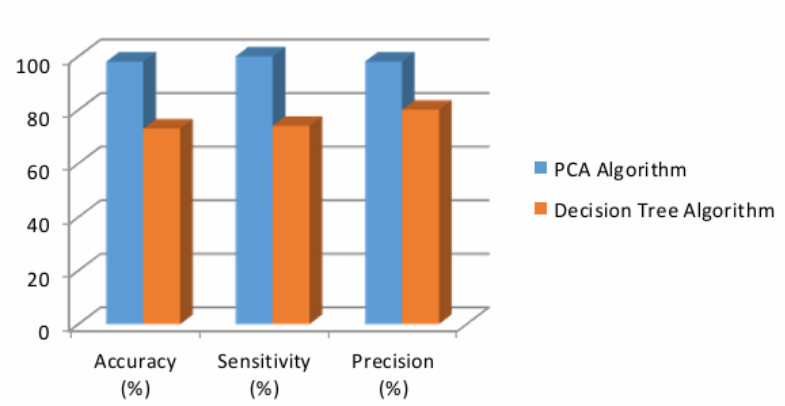
The researchers at Aviny, Ghasemi and Fazlazad (2023) developed new methods for detecting cardiovascular diseases early by refining diagnostic precision through data mining methodologies. A total of 270 patient samples were analyzed through the study using 14 distinct features. The analysis utilized PCA to lower the original 14 features down to 8 features thus demonstrating more effective investigation techniques. The Weka tool run the regression tree algorithm to enable diagnosis while incorporating essential process sequences for feature selection and tree generation as well as pruning. When PCA integrated with the regression tree algorithm the diagnostic accuracy rose to 81.48%. Advanced algorithm applications point toward improved precision which will further strengthen cardiovascular disease diagnostic accuracy.

Kazangirler and Özkaynak (2024) created an automated cardiac condition detection tool by processing medical records from Cleveland Eye Hospital, the Hungarian Institute of Cardiology, University Hospitals Switzerland, and Zurich VA Medical Center. They used PCA to reduce data dimensions and added artificial variables to improve their classification method. They explored nine distinct ML algorithms: SVM, DT, Naive Bayes, KNN, Bagging, RF, Gradient Boosting, LR, AdaBoost, and LDA along with ANN. They reached better performance by testing Grid Search Cross-Validation and Bagging/Boosting methods. Gaussian Naive Bayes outperformed other methods by achieving 91.0% accuracy which represents a 3.0% gain above the baseline method. The result is shown below.

Algorithm	Accuracy	ROC	AUC
DTC	85.0%	85.0%	85.0%
AdaBoostC	87.0%	87.0%	87.0%
KNN	89.0%	89.0%	89.0%
SVC	89.0%	89.0%	89.0%
LDA	89.0%	89.0%	89.0%
LR	89.0%	89.0%	89.0%
LinearSVC	89.0%	89.0%	89.0%
RFC	90.0%	90.0%	90.0%
GaussianNBC	**91.0%**	**91.0%**	**91.0%**

Algorithm performance comparison (Kazangirler and Özkaynak ,2024)

Hambali Gbolagade and Olasupo (2023) studied how data mining tools help find cardiovascular disease cases through health records from UCI repository. The research team used PCA to find essential data elements before achieving better results with classification tasks. PCA yielded greater accuracy than Decision Trees because it found 98.4% of cases correctly with 100% sensitivity and 98% precision. The system successfully found patients at high risk which led junior cardiologists to recommend them to specialists for extra evaluation. Data mining tools help find early signs of disease which improves doctors' diagnosis skills and helps patient well-being.



Results (Hambali, Gbolagade, and Olasupo, 2023)

Pasha and Mohamed (2022) enriched disease risk forecasting by fixing the inadequate selection process of features in uneven datasets. Their Advanced Hybrid Ensemble Gain Ratio Feature Selection (AHEG-FS) model combined multiple algorithms to produce an enhanced feature selection system. The researchers experimented with nine ML algorithms including Boosted Regression Trees (BRT), RF, Stochastic Gradient Boosting (SGB), LR, SVM, KNN, NB, Classification via Clustering (CVC), and AdaBoost for heart disease prediction using four datasets from UCI (Cleveland, Statlog, Hungarian,

Switzerland). The RF model produced the highest results of 99.00% AUC and 95.47% accuracy while cutting the number of input variables by 46.15% to outperform past studies by 6.18%.

Comparative Analysis

Study	Dataset(s)	Aim	Feature Selection Methods	Models Evaluated	Best Performing Model(s)	Performance Metrics
Gárate-Escamila, El Hassani, and Andrès (2020)	UCI Heart Disease Dataset.	To predict heart disease using efficient dimensionality reduction and selection techniques	CHI, PCA, CHI-PCA	LOG, DT,GBT, RF,MPC, NB,	CHI-PCA with RF	99.4% -CH, 99.0% - Hungarian and 98.7% - Cleveland
Kakoly, Hoque, and Hasan (2023)	Bangladesh Diabetes Survey	To predict diabetes risk factors using ML and feature selection techniques	PCA and IG	LR, RF, DT, SVM, KNN	LR	Accuracy of 82.2% and an AUC of 87.2%
Aviny, Ghasemi and Fazlazad (2023)	Cardiovascular Patient Dataset	Enhance cardiovascular disease diagnosis accuracy using data mining and feature selection techniques	PCA	Regression tree	Regression tree with PCA	81.48%
Kazangirler and Özkaynak (2024)	Hungarian Institute of Cardiology, Cleveland , Zurich VA Medical Center and		PCA	SVM, Naive Bayes, DT, KNN, Bagging, RF,	GNB	91%

	University Hospitals of Switzerland			Gradient Boosting, LR, AdaBoost, LDA,ANN		
Hambali, Gbolagade, and Olasupo (2023)	UCI repository dataset	Predict cardiovascular diseases using data mining to improve diagnostic accuracy and outcomes	PCA	PCA, DT	PCA	98.4%
Pasha and Mohamed (2022)	UCI Heart Disease dataset	Enhance disease risk prediction by improving feature selection and MLmodels	AHEG-FS model	RF,BRT, SGB, LR, KNN, NB, CVC, SVM, and AdaBoost	AHEG-FS with RF	AUC - 99.00% Accuracy- 95.47%

Research findings demonstrate PCA and Information Gain (IG) effectively enhance the feature selection process for machine learning models while focusing on health risk prediction through studies by Kakoly et al. (2023) and Pasha and Mohamed (2022). According to Aviny et al. (2023) the benefits of dimension reduction and improved computational speed from PCA come at the cost of reduced feature interpretability. The ranking system of IG features according to their importance encounters problems when dealing with datasets that contain redundant or correlated data as Pasha and Mohamed (2022) point out. The direct implementation of PCA on raw data by Hambali et al. (2023) frequently results in unsatisfactory model performance. The shortcomings of current predictive modeling methods emphasize the requirement to develop and implement sophisticated hybrid methodologies.

Identification of gaps

Machine learning (ML) systems have successfully predicted diseases like diabetes and cardiovascular problems yet research into the use of feature selection approaches PCA and IG for heart failure prediction remains limited. While PCA and IG methods show success across diabetes risk prediction research they have rarely been applied to heart failure prediction

modeling. Research demonstrates substantial support for the usage of ML algorithms while neglecting the performance improvement presented by customized feature selection strategies. The research gap blocks enhanced effectiveness of prediction models for heart failure which requires careful selection of impactful features for precise prediction accuracy. Research initiatives must explore how PCA and IG impact heart failure prediction models as a way to enhance both early diagnosis protocols and patient health outcomes.

REFERENCES

Gárate-Escamila, A.K., El Hassani, A.H. and Andrès, E., 2020. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, p.100330.

Anand, S., 2021. Archimedes optimization algorithm: Heart disease prediction: archimedes optimization algorithm: heart disease prediction. *Multimedia research*, *4*(3).

Kakoly, I.J., Hoque, M.R. and Hasan, N., 2023. Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique. *Sustainability*, *15*(6), p.4930.

Bhatt, C.M., Patel, P., Ghetia, T. and Mazzeo, P.L., 2023. Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), p.88.

Aviny, H.R., Ghasemi, M. and Fazlazad, M., 2023. Cardiovascular Disease Diagnosis Using the Combination of Principal Component Analysis Algorithm and Regression Tree. *Transactions on Machine Intelligence*, *6*(2), pp.114-125.

Kazangirler, B.Y. and Özkaynak, E., 2024. Conventional Machine Learning and Ensemble Learning Techniques in Cardiovascular Disease Prediction and Analysis. *Journal of Intelligent Systems: Theory and Applications*, *7*(2), pp.81-94.

Hambali, M.A., Gbolagade, M.D. and Olasupo, Y.A., 2023. Heart disease prediction using principal component analysis and decision tree algorithm. *Journal of Computer Science and Engineering (JCSE)*, *4*(1), pp.1-14.

Pasha, S.J. and Mohamed, E.S., 2022. Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction. *Informatics in Medicine Unlocked*, *32*, p.101064.

Reddy, K.V.V., Elamvazuthi, I., Abd Aziz, A., Paramasivam, S. and Chua, H.N., 2021, July. Heart disease risk prediction using machine learning with principal component analysis. In *2020 8th international conference on intelligent and advanced systems (ICIAS)* (pp. 1-6). IEEE.

Panyamit, T., Sukvivatn, P., Chanma, P., Kim, Y., Premratanachai, P. and Pechprasarn, S., 2022. Identification of factors in the survival rate of heart failure patients using machine learning

models and principal component analysis. *Journal of Current Science and Technology*, *12*(2), pp.336-348.

Rani, P., Kumar, R., Ahmed, N.M.S. and Jain, A., 2021. A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), pp.263-275.

Win, T.Z. and Kham, N.S.M., 2019. *Information gain measured feature selection to reduce high dimensional data* (Doctoral dissertation, MERAL Portal).

Chandrasekhar, N. and Peddakrishna, S., 2023. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, *11*(4), p.1210.

Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), p.345.

Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, pp.81542-81554.