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

Chapter-3

Methodology

This section depicts the research methodology to improve the heart failure prediction accuracy using Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE). The feature selection methods are integrated with various machine learning models in order to improve the predictive accuracy and reliability, reduce redundancy, as well as improve efficiency.

Project Design

There is a structured pipeline in the project, which includes collecting data, preprocessing the data, feature selection, model training, evaluation, and validation for prediction of heart failure. As the main source of data, the ECG Arrhythmia Classification Dataset from Kaggle is used. Preprocessing is done by handling missing values, normalizing features and handling class imbalance using SMOTE. PCA is used as a dimensionality reduction technique, Information Gain is utilized to rank feature importance, and finally the research uses the Recursive Feature Elimination technique to perform the feature selection step by iteratively discarding irrelevant features. The data is divided into train, validate and test sets. Logistic Regression, Naïve Bayes, Random Forest, LSTM, RNN and ANN are trained with and without feature selection. Hyperparameter tuning is done using GridSearchCV and for robustness 2-fold cross validation is done. Accuracy, precision, recall, F1-score, confusion matrix and AUC-ROC are used for assessment of performance. Feature selection techniques are validated with optimal combination of feature selection method and machine learning for predicting heart failure.

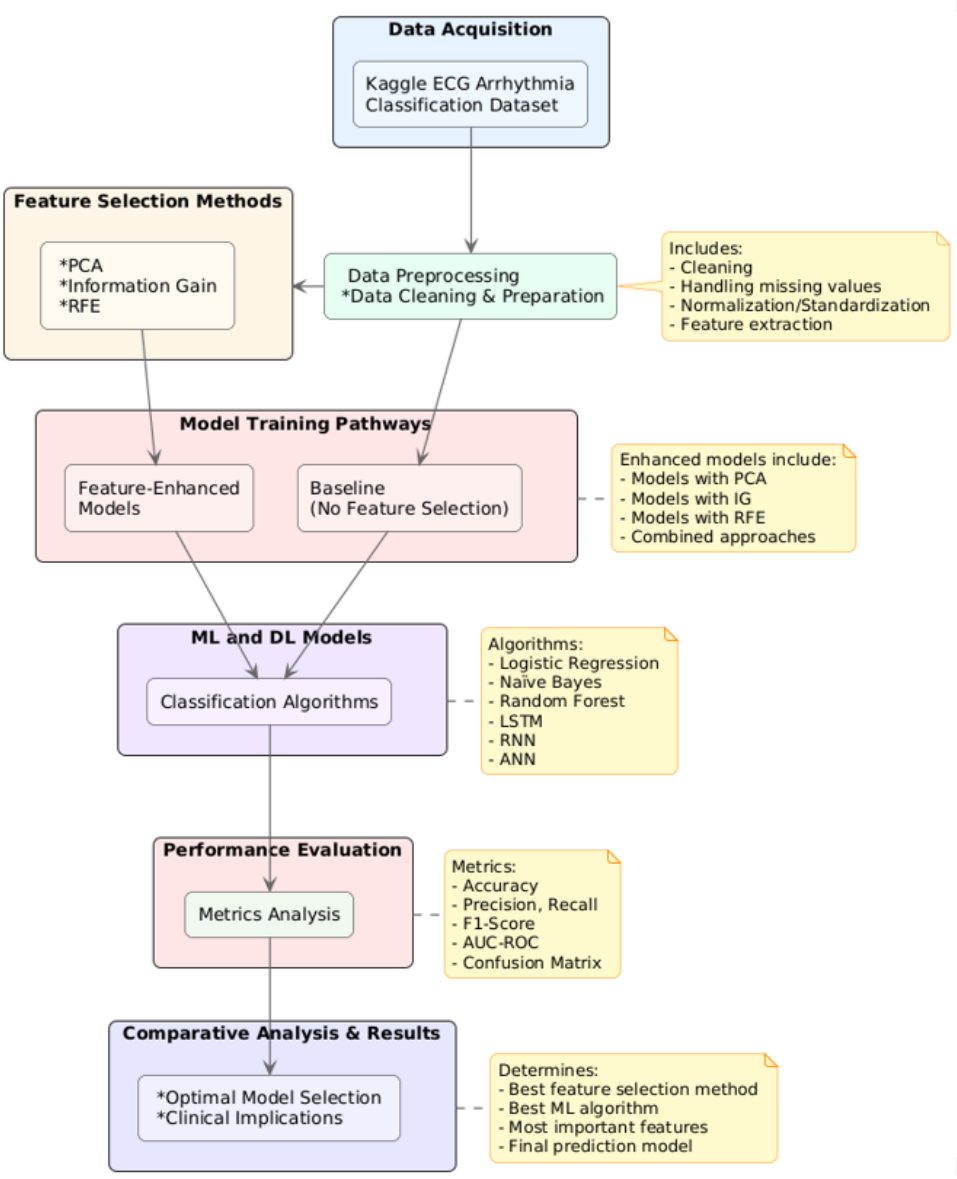

Figure Project Architecture

Dataset Description

In this research, the used dataset is ECG Arrhythmia Classification Dataset by Kaggle for heart failure prediction. It comprises of several ECG signal features that will aid in arrhythmia classification, including arrhythmia that are linked with heart failure. The dataset has numerical and categorical attributes which need normalizing, handling missing values and selecting features before training models. To have a uniformity, features are normalized and to deal with class imbalance, Synthetic Minority Over-Sampling Technique (SMOTE) is used. The training, validation and testing datasets are split into 80%, 20% and 20% respectively for the purpose of an

effective model evaluation. Handling data related to patient information, data privacy and security standards are adhered to maintain ethical considerations.

Choice of Methods

This study utilizes the data science approach on the combination of machine learning algorithms and feature selection methods in order to improve the accuracy in heart failure prediction. Data preprocessing, feature selection by using PCA, Information Gain (IG) and Recursive Feature Elimination (RFE), model training and performance evaluation are included in its methodology. Gárate-Escamila et al, (2020) utilized PCA to reduce dimensionality and increase the classification efficiency using Chi-Square for predicting heart disease. In Kakoly et al, (2023), IG method has ranked features based on their contribution to the target variable for diabetes risk prediction. Kazangirler & Özkaynak (2024) apply RFE, where they iteratively remove less important features to achieve better model interpretability and accuracy. According to Bhatt et al, (2023), five machine learning models, namely Logistic Regression, Naïve Bayes and Random Forest are chosen due to their high effectiveness in cardiovascular disease classification. The combination of both ensures a solid comparison of feature selection techniques for heart failure prediction. For capturing complex pattern and dependency information in the medical data, especially for sequential features, Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) are used (Tariq and Ismail, 2024). These methods are selected based on the literature and help improve the reliability of the research.

Justification and Support of Choices

PCA, IG, and RFE are shown to be some of the feature selection methods that can improve the prediction of cardiovascular disease. As per prior studies, PCA has proven to be effective as Gárate Escamila et al, (2020) used PCA with Chi-Square to improve heart disease prediction accuracy and Aviny et al. (2023) diminished features from 14 to 8 using PCA resulting in better diagnosis. Kakoly et al, (2023) demonstrated that IG was effective at increasing classification performance in diabetes risk prediction and is therefore applicable to heart failure datasets. The article (Bhatt et al, 2023) states that RF and LR have been commonly used in heart disease studies and RF is the most accurate model for cardiovascular disease prediction. In particular, deep learning models like ANN, RNN and LSTM are included because they are capable of learning complex patterns in large datasets (Ladeira and Silva, 2025) which is important for predicting complex medical condition.

Tools and Techniques utilised

Tool/Technology	**Purpose**
Python	Programming language for model implementation
Scikit-learn	Machine learning library for model training and feature selection
Pandas	Data manipulation and preprocessing
NumPy	Numerical computation
Matplotlib & Seaborn	Data visualization
GridSearchCV	Hyperparameter tuning

SMOTE (Imbalanced-learn)	Balancing dataset classes
Jupyter Notebook	Interactive coding environment
K-Fold Cross-Validation (K=2)	Model validation

Feature Selection Techniques Used

Feature selection techniques select the most relevant features and make a dimensionality reduction on data to enhance the model performance. To improve heart failure prediction accuracy, PCA, RFE and Information Gain are employed in this project to eliminate redundant features.

Principal Component Analysis (PCA)

PCA reduces dimensionality by transforming correlated features into uncorrelated principal components while retaining maximum variance.

Information Gain (IG)

IG measures the reduction in entropy when a feature is used for classification, helping rank the most informative features.

Recursive Feature Elimination (RFE)

RFE iteratively removes the least important features based on model performance until the optimal subset is selected. It uses a feature ranking function

Algorithms Used in the Research:

The implementation of five machine learning models, namely, the Logistic Regression, Naïve Bayes, Random Forest, LSTM, RNN and RNN is done. To evaluate the effect that PCA, IG, and RFE have on the prediction accuracy, these models are trained and evaluated with and without feature selection.

Logistic Regression (LR)

LR models the probability of heart failure occurrence using the sigmoid function

Naïve Bayes (NB)

NB is a probabilistic classifier based on Bayes' Theorem

Random Forest (RF)

RF is an ensemble method that builds multiple decision trees and aggregates their predictions

Long Short-Term Memory (LSTM)

RNN is a type of neural network that can handle the long time dependencies in the sequential data, which makes it suitable for analyzing the time series patient health records in heart failure prediction tasks.

Recurrent Neural Network (RNN)

RNN is a type of neural network that operates on data in sequence and maintains memory of the previous inputs to capture the patterns from past medical records to predict heart failure risk.

Artificial Neural Network (ANN)

ANN is made of connected layers which simulate human brain neurons enabling the machine to comprehend elaborate relationships in patient data to improve the precision of heart failure diagnosis. Performance Evaluation

Performance Evaluation

Accuracy, precision, recall, F1-score, confusion matrix and AUC-ROC are used to assess the performance of heart failure prediction models. Accuracy measures the overall correctness of model predictions while precision measures the proportion of correctly predicted positive cases out of all the total predicted positives. (sensitivity) Recall determines how well the model can identify actual heart failure cases. The harmonic mean of precision and recall is called F1-score, which balances both the metrics. Confusion matrix gives us idea of false positives and false negatives and also helps us in analyzing misclassification rates. Finally, the model discrimination ability is evaluated using AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and a high value implies better classification performance..

	Predicted positive	Predicted negative
Actual positive	TP	FN
Actual negative	FP	TN

Abbreviations: FN, false negative; FP, false positives; TN, true negative; TP, true positive.

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Precision (P)} = \frac{TP}{TP + FP}$$

$$\text{F1 score (F1)} = \frac{2 \times P \times R}{P + R} \qquad \text{Recall (TPR) (R)} = \frac{TP}{TP + FN}$$

Figure Confusion Matrix and Evaluation metrics formula (Nizam-Ozogur and Orman, 2024)

Test Strategy

To incur overfitting and guarantee the robustness of the model, a stratified K-fold cross-validation (K=2) is used. The training and testing is done separately for each fold and the average performance is recorded. The effect of ML models is finally quantitatively analyzed by testing the models with and without the feature selection techniques (PCA, IG, and RFE). Evaluation is conducted using performance metrics like accuracy, precision, recall, F1 score, Confusion Matrix and AUC ROC.

Testing and Results

The model is run with various feature selection techniques and performance is compared. The results indicate which is the best combination of machine learning algorithm and feature selection method to use for accurate heart failure prediction. It also evaluates the classification performance by visualizing confusion matrix and ROC curves. The results are compared with each other to ensure that accuracy has improved.

Validation

Validation process tests models based on the independent validation set and also with the help of the cross validation technique to make sure the results are accurate and reliable. Feature selection methods are investigated and validated as effective methods to improve the accuracy of heart failure prediction thus identifying the most efficient combination of feature selection methods and machine learning models. Evaluation metrics are used to select the best performing model that also enhances predictive accuracy and generalizability of the proposed approach.

Ethical Issues

With regards to ethical considerations, this project ensures that the data is public and no authorization is required as it is open source. Anonymization and secure storage of the information provided protects it from misuse. All these methods abide by ethical guidelines by ensuring responsible model deployment without compromising on predictive accuracy and consciously avoiding unintended discrimination. The bias mitigation techniques include trained on balanced datasets and diverse training data. The second issue is to retain interpretability and to prevent automation bias, addressing ethical concerns in decision making. Not only does this project carefully evaluating its impact and avoiding unintended harm, but it also does it to minimise unintended harm through the implementation of responsible machine learning.

Legal Issues

It is important to note that the project strictly complies to data protection laws such as GDPR and CCPA by securely following handling of the data and using it within the applicable boundary. None of the information is processed without consent, and all the data sources are legitimately obtained for research purposes. Intellectual property rights are honored and no copyrighted material is inappropriately used. Encryptions and secure access protocols are included within the project to safeguard against data breaches. Data handling and model development takes into account the legal side, such as complying with licensing agreements. Data leakage or misuse is controlled through access and proper use of the personal information is ensured through these ethical practices.

Professional Issues

The project is structured in accordance with a systematic method with clearly defined objectives executed systematically within possible timeframe. Since these processes will be performed by computer, this output file will be the best model available where all best practices in machine

learning and data science are applied to ensure accuracy, reliability, and reproducibility. Data handling, preprocessing and model evaluation is properly documented to ensure accountability and standardization. Rigorous model validation, performance analysis, error handling techniques are applied at the model development stages to prioritize quality assurance. The workflow integrates ethical considerations so that the predictions are unbiased and the AI is responsibly deployed. This project maintains industry standards, risky development practices, and sound ROI by prioritizing transparency in methodology, long term and scalability.

Social Issues

The positive contribution of the project is in the improvement of the decision making process with respect to societal impact. It treads on order by fixing data bias and making sure its predictions are fair in a sense. Robust machine learning models are applied to supply solutions to sectors without causing disproportionate disadvantages. Automation bias safeguards are implemented to prevent the use of automation in faulty ways. Measures of security ensure the privacy and the trust in the user data. Through careful validation potential risks such as a misinterpretation of predicted values from the model are minimized. However, the project is designed to encourage innovation, but with a conscience, and to be deployed in the real world in an ethical manner.

Practicality

The integration of PCA, Information Gain (IG), and Recursive Feature Elimination (RFE) in this project allows the heart failure prediction accuracy to improve and solves the real world constraints. Dimensionality reduction using PCA optimizes computational efficiency while IG ranks relevant features so as to improve model performance. It eliminates nonessential attributes systematically until the prediction accuracy is refined. It also uses Synthetic Minority Over Sampling Technique (SMOTE) to handle class imbalance so as to achieve robust predictions. Hyperparameter tuning is used for model optimization using GridSearchCV. The mitigation of the computational complexity, dataset limitation and feature selection trade off problems are handled through cross validation and benchmark comparisons, to obtain a reliable and a generalizable outcome in predictive healthcare analytics.

Conclusion Summary

The methodology is composed of PCA, Information Gain, and Recursive Feature Elimination that is integrated to optimize model performance over the ECG Arrhythmia Classification Dataset. Logistic Regression, Naïve Bayes, Random Forest, LSTM, RNN and ANN are trained and validated. Data integrity and fairness are carefully maintained in consideration of ethical, legal and social aspects. In the next chapter, it attempts to perform experimentation, model training, performance evaluation and comparative analysis of feature selection techniques.

References

Ladeira, E.F. and Silva, B.M., 2025. A Machine Learning-based Platform for Monitoring and Prediction of Hazardous Gases in Rural and Remote Areas. *IEEE Access*.

Tariq, M.U. and Ismail, S.B., 2024. Deep learning in public health: Comparative predictive models for COVID-19 case forecasting. *Plos one*, *19*(3), p.e0294289.

Nizam-Ozogur, H. and Orman, Z., 2024. A heuristic-based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data. *Expert Systems*, p.e13596.

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.

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.