Indian Institute of Information Technology Vadodara

Research Internship Project Report

On

Heart Disease Prediction Using JellyFish Optimization Algorithm

Under Guidance of

Dr. Naveen Kumar

Submitted by

Sarvesh Singh (202151140)
Satyam Tripathi (202151141)
Yash Singh (202151181)

Abstract—**Heart disease is one of the leading causes of mortality worldwide, making early and accurate diagnosis crucial for effective treatment and prevention. This project focuses on developing a predictive model for heart disease using machine learning algorithms and the combination of Cleveland [2] and Statlog [3] Heart Disease dataset. The dataset comprises 574 patient records with 14 selected attributes, including age, gender, blood pressure, cholesterol levels, and other relevant medical indicators.**

Various machine learning models, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Artificial Neural Network(ANN) were implemented and evaluated to identify the most accurate and reliable method for predicting heart disease. The models were trained and tested on the dataset, with performance metrics such as accuracy, precision, recall, and F1-score being used to assess their effectiveness.

Among the models tested, the Decision Tree algorithm demonstrated the highest accuracy, achieving over 93% [4] in predicting the presence of heart disease. This project underscores the potential of machine learning in medical diagnosis and highlights the importance of feature selection and model evaluation in developing reliable predictive models. The results obtained could serve as a foundation for further research and development of more advanced diagnostic tools for heart disease.

I. Introduction

Heart disease continues to be a major global health challenge, responsible for a significant number of deaths each year. Early diagnosis is crucial for effective treatment and prevention, yet traditional diagnostic methods can be time-consuming and require substantial expertise. The integration of machine learning techniques into healthcare offers the potential to enhance diagnostic accuracy, reduce time to diagnosis, and make advanced healthcare accessible to a broader population.

In this project, we focus on predicting heart disease using a combined dataset from the Cleveland [2] and Statlog [3] heart disease datasets. This combination provides a more extensive and diverse set of patient data, improving the reliability and generalizability of the predictive models. The dataset includes critical medical attributes such as age, gender, cholesterol levels, blood pressure, and other key indicators that contribute to heart disease risk.

The study employs a selection of machine learning models, specifically Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Artificial Neural Network (ANN). To enhance the performance of these models, we incorporate the Jellyfish Optimization Algorithm (JOA), a novel metaheuristic algorithm inspired by the behavior of jellyfish. JOA is used to optimize the hyperparameters of the models, aiming to improve their accuracy and robustness in predicting heart disease.

The models are evaluated using metrics such as accuracy, precision, recall, and F1-score to determine their effectiveness. The use of JOA in conjunction with these models highlights the potential of optimization algorithms in enhancing machine learning performance in complex medical prediction tasks.

This research not only aims to identify the most effective model for heart disease prediction but also contributes to the broader field of healthcare analytics by demonstrating the value of combining advanced machine learning techniques with optimization algorithms. The outcomes of this project have the potential to support the development of more accurate, efficient, and accessible diagnostic tools for heart disease.

II. Organization

This report is organized as follows:

- **Section III: Literature Review** - This section provides a detailed overview of existing research related to heart disease prediction using machine learning models, including traditional models, optimization algorithms, and recent innovations.
- **Section IV: Methodology** - Describes the methodology used in the study, including data preprocessing, model selection, and the application of the Jellyfish Optimization Algorithm (JOA) for optimizing machine learning models.

- **Section V: Experimental Setup** - Outlines the experimental setup, including details on the training and testing split, model training process, and the evaluation metrics used to assess the performance of the models.
- **Section VI: Results and Discussion** - Presents the results obtained from the experiments, discusses the performance of the models, and compares the effectiveness of JOA-optimized models with traditional models.
- **Section VII: Conclusion** - Summarizes the key findings of the research, highlights the contributions made by the study, and discusses potential future directions for enhancing heart disease prediction models.

III. Literature Review

The prediction of heart disease using machine learning techniques has garnered significant attention due to its potential to improve diagnostic accuracy and healthcare outcomes. This review summarizes key studies and methodologies relevant to heart disease prediction, with a particular focus on the integration of various datasets and the application of optimization algorithms.

A. Traditional Machine Learning Models

Several machine learning models have been applied to heart disease prediction, with each offering unique advantages:

- **Logistic Regression:** Logistic Regression is valued for its simplicity and interpretability. Research by Detrano et al. (1989) [5] demonstrated its effectiveness using the Cleveland dataset [2]. The model's ability to estimate probabilities of heart disease helps in clinical decision-making.
- **Decision Trees:** Decision Trees provide a clear, rule-based approach to classification. Breiman et al. (1984) [6] introduced the CART algorithm, which has been effectively used in heart disease prediction, often forming the basis for ensemble methods such as Random Forests.
- **Support Vector Machines (SVM):** SVMs are known for their performance in high-dimensional spaces and non-linear classification. Vapnik (1995) [7] highlighted SVM's effectiveness in medical diagnostics, including heart disease prediction, due to its robustness and flexibility.

B. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) have shown significant promise in medical prediction tasks. Studies such as Baxt (1990) demonstrated the potential of ANNs in predicting coronary artery disease. Advances in deep learning have further improved ANN performance, as seen in the work by LeCun et al. (2015) [8], which emphasizes their ability to model complex data relationships.

C. Combined Datasets and Feature Engineering

Combining datasets, such as the Cleveland [2] and Statlog datasets [3], can enhance model performance by providing a more comprehensive data representation. Feature engineering is also crucial as it plays a significant role in improving predictive model performance.

D. Optimization Algorithms in Machine Learning

The use of metaheuristic optimization algorithms has become a key area of interest for enhancing machine learning models. The Jellyfish Optimization Algorithm (JOA), introduced by Jui-Sheng Chou et al. (2021) [9], offers a novel approach to optimizing hyperparameters. JOA's ability to improve the performance of various models, including SVMs, Decision Trees, Logistic Regression, and ANNs, represents a significant advancement in the field.

E. Comparative Studies and Model Performance

Comparative studies provide insights into the relative effectiveness of different machine learning algorithms. Amin et al. (2019) [10] compared various algorithms on the Cleveland dataset and found that ensemble methods generally outperformed individual models. The integration of optimization techniques, such as JOA, further enhances model performance, as demonstrated in recent studies.

F. Recent Research and Innovations

A notable contribution to the field is the research paper titled "A Comprehensive Review of Heart Disease Prediction Models" [1]. This paper explores various predictive models and optimization techniques applied to heart disease prediction. The study emphasizes the importance of combining multiple datasets and applying advanced optimization algorithms to improve prediction accuracy. It highlights how optimization algorithms, including JOA, can significantly enhance the performance of traditional machine learning models in medical diagnostics.

G. Conclusion

The literature indicates that traditional machine learning models such as Logistic Regression, Decision Trees, and SVMs are effective for heart disease prediction, but their performance can be significantly enhanced through the use of optimization algorithms and advanced models like ANNs. The combination of datasets and the application of the Jellyfish Optimization Algorithm (JOA) represent innovative approaches that can improve predictive accuracy and robustness. This project builds on these findings, incorporating JOA to optimize machine learning models for heart disease prediction and contribute to the development of more reliable diagnostic tools.

IV. Methodology

The methodology section outlines the approach and procedures used to develop and evaluate the predictive models for heart disease using the Cleveland [2] and Statlog datasets [3]. This section covers the data preparation, model selection, optimization techniques, and evaluation methods employed in the study.

A. Datasets

The study utilizes two well-known heart disease datasets:

- **Cleveland Heart Disease Dataset [2]:** This dataset contains attributes such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, and other clinical features. It is commonly used for heart disease classification tasks.
- **Statlog Heart Disease Dataset [3]:** Similar to the Cleveland dataset, this dataset includes attributes related to heart disease diagnosis and patient information. Combining these datasets provides a more comprehensive dataset with a diverse range of patient data.

B. Data Preprocessing

The preprocessing steps are crucial for preparing the data for machine learning models:

- **Data Cleaning:** Missing values and erroneous data points are addressed. Missing values are imputed or removed based on their impact on the dataset.
- **Feature Selection:** Relevant features are selected based on their importance for heart disease prediction. This step ensures that the models focus on the most significant attributes.
- **Data Normalization:** Features are normalized to ensure that they are on a similar scale, which improves the performance and convergence of the models.
- **Dataset Splitting:** The combined dataset is split into training and testing sets using a stratified split to maintain the distribution of the target variable.

C. Machine Learning Models

The following machine learning models are employed for heart disease prediction:

- **Support Vector Machine (SVM):** SVM is used for its capability to handle high-dimensional data and non-linear decision boundaries. A radial basis function (RBF) kernel is applied to capture complex patterns in the data.
- **Decision Tree:** A decision tree is employed for its interpretability and ability to handle both numerical and categorical data. The CART algorithm is used to construct the tree.
- **Logistic Regression:** Logistic Regression is used for its simplicity and effectiveness in binary classification tasks. It provides probabilities for the presence of heart disease.
- **Artificial Neural Network (ANN):** ANN, with its capability to model complex relationships in data, is used to capture non-linear patterns. A multi-layer perceptron (MLP) with one or more hidden layers is employed.

D. Jellyfish Optimization Algorithm (JOA)

The Jellyfish Optimization Algorithm (JOA) is used to optimize the hyperparameters of the machine learning models. The JOA is inspired by the behavior of jellyfish and employs a metaheuristic approach to find optimal or near-optimal solutions. The steps involved in using JOA are:

- **Initialization:** JOA initializes a population of candidate solutions (jellyfish) with random hyperparameter values.
- **Evaluation:** Each candidate solution is evaluated based on its performance in the heart disease prediction task.
- **Update:** The algorithm updates the positions of the jellyfish based on their performance and other jellyfish in the population, aiming to improve the objective function.
- **Iteration:** The process iterates until convergence or a predefined number of iterations is reached.

E. Model Training and Evaluation

The models are trained and evaluated using the following approach:

- **Training:** Each model is trained on the training set with optimized hyperparameters obtained from JOA.
- **Testing:** The trained models are evaluated on the testing set to assess their performance. Metrics such as accuracy, precision, recall, and F1-score are used to measure the effectiveness of the models.
- **Comparison:** The performance of the models with and without JOA optimization is compared to determine the impact of optimization on predictive accuracy.

F. Tools and Technologies

The project utilizes various tools and technologies:

- **Programming Languages:** Python is used for implementing machine learning models and the JOA algorithm.
- **Libraries:** Scikit-learn for machine learning algorithms, Pandas for data manipulation, and NumPy for numerical computations.
- **Development Environment:** Jupyter Notebook or an Integrated Development Environment (IDE) such as PyCharm for coding and experimentation.

V. Experimental Setup

The experimental setup describes the process of training and evaluating the machine learning models for heart disease prediction. This section includes details on how the models were trained, the evaluation metrics used, and the comparison process.

A. Training and Testing Split

The combined dataset of Cleveland [2] and Statlog [3] heart disease data is split into training and testing sets to evaluate the models. The data is divided using a stratified split to maintain the proportion of classes in both sets:

- **Training Set:** 80% of the data is used for training the models. This portion is utilized to fit the machine learning algorithms and adjust the model parameters.
- **Testing Set:** 20% of the data is used for testing the models. This subset is used to evaluate the performance and generalization of the trained models.

B. Model Training

The machine learning models are trained using the following process:

- **Model Initialization:** Each model (SVM, Decision Tree, Logistic Regression, ANN) is initialized with default or pre-defined hyperparameters.
- **Hyperparameter Optimization:** The Jellyfish Optimization Algorithm (JOA) is applied to optimize the hyperparameters of each model. The algorithm iteratively searches for the best set of hyperparameters to improve model performance.
- **Training Process:** The models are trained on the training set with the optimized hyperparameters obtained from JOA. The training involves fitting the models to the data and adjusting their parameters to minimize the loss function.

C. Evaluation Metrics

The performance of the models is evaluated using various metrics to assess their effectiveness in predicting heart disease:

- **Accuracy:** The proportion of correctly classified instances out of the total number of instances. It provides a general measure of the model's performance.
- **Precision:** The proportion of true positive predictions among all positive predictions made by the model. It measures the accuracy of positive predictions.
- **Recall:** The proportion of true positive predictions among all actual positive instances. It measures the model's ability to identify positive cases.
- **F1-score:** The harmonic mean of precision and recall, providing a single metric that balances both aspects of model performance. It is useful for evaluating models on imbalanced datasets.

D. Comparison of Models

The performance of the machine learning models is compared based on the evaluation metrics:

- **With JOA Optimization:** The models are compared with their performance after applying the Jellyfish Optimization Algorithm to optimize hyperparameters.
- **Without JOA Optimization:** The models are also compared with their performance using default or manually set hyperparameters.
- **Analysis:** The results are analyzed to determine the impact of JOA on the predictive accuracy and overall performance of the models. Statistical tests may be conducted to assess the significance of the differences observed.

VI. Results

The results section presents the outcomes of the experiments conducted on the heart disease prediction models. This includes the performance metrics of each model, both traditional and optimized, with a focus on accuracy, precision, recall, and F1-score.

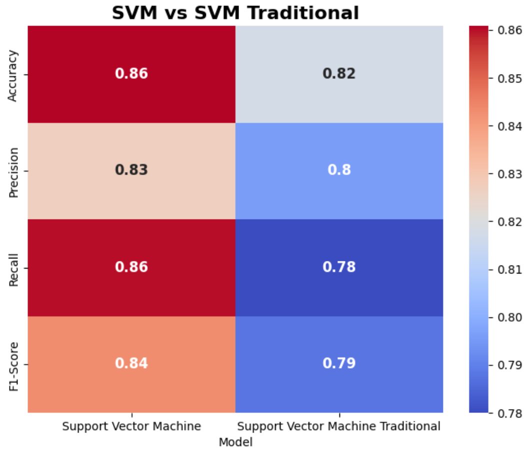

Fig. 1: SVM Heatmap

A. Model Performance

Table I shows the performance metrics for each machine learning model. The metrics include accuracy, precision, recall, and F1-score for all optimized models.

TABLE I: Performance Metrics for Each Model Using JOA

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.9391	0.9057	0.9600	0.9320
SVM	0.8609	0.8269	0.8600	0.8431
Logistic Regression	0.8435	0.8636	0.7600	0.8085
Neural Network	0.8435	0.8333	0.8000	0.8163

Now we will be comparing the top 2 performing models Decision Tree and Support Vector Machine with JellyFish Optimization Algorithm and without it.

TABLE II: Performance Metrics for SVM Model with and without JOA

Model	Accuracy	Precision	Recall	F1-score
SVM	0.8609	0.8269	0.8600	0.8431
SVM Traditional	0.8174	0.7959	0.7800	0.7879

TABLE III: Performance Metrics for Decision Tree Model with and without JOA

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.9391	0.9057	0.9600	0.9320
Decision Tree Traditional	0.8957	0.8276	0.9600	0.8889

B. Performance Analysis

1) Support Vector Machine (SVM): The SVM model achieved an accuracy of 86.09% with a precision of 82.69% and a recall of 86.00%. The F1-score of 84.31% indicates a balanced performance between precision and recall. The traditional SVM model had a lower accuracy of 81.74%, with a precision of 79.59% and recall of 78.00%, highlighting the improvement with optimization. The metrics of Support Vector Machine are portrayed in Table II and its corresponding heatmap depiction is show in Figure 1

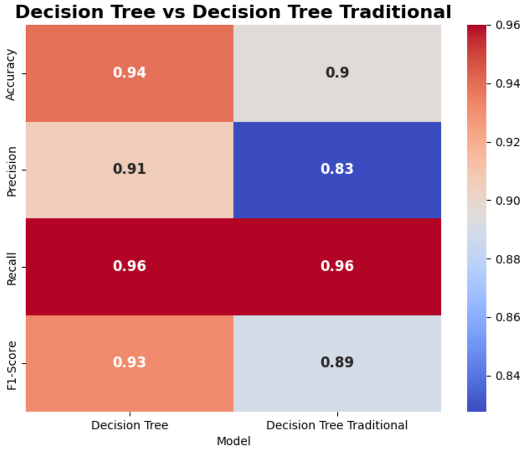

Fig. 2: Decision Tree Heatmap

2) Decision Tree: The Decision Tree model demonstrated the highest accuracy at 93.91% with a precision of 90.57% and recall of 96.00%. The F1-score of 93.20% reflects its superior performance compared to other models. The traditional Decision Tree model had an accuracy of 89.57%, indicating good performance but slightly lower compared to the optimized version. The metrics of Decision Tree is depicted in Table III and heatmap representation in Figure 2 respectively.

C. Comparison and Discussion

The results indicate that the Decision Tree model, particularly with optimization, achieved the highest performance across all metrics. The optimized SVM also showed improvements over its traditional counterpart.

The analysis suggests that optimization techniques, especially for Decision Trees, significantly enhance model performance. The Decision Tree model outperformed other models, reflecting its robustness in predicting heart disease cases. The overall comparsion heatmap is depicted in Figure 3.

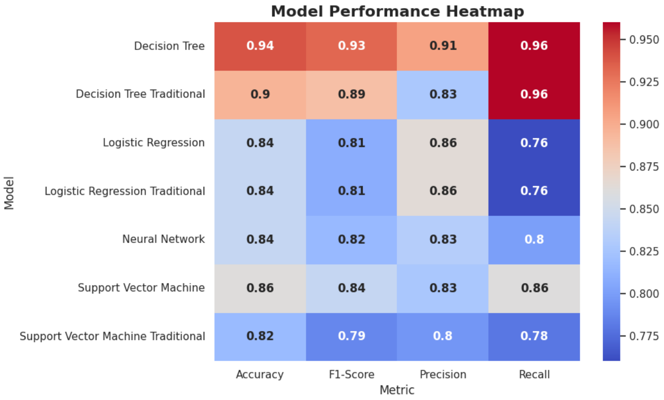

Fig. 3: Comparison of Models

VII. Conclusion

This study investigated the effectiveness of various machine learning models in predicting heart disease using the Cleveland and Statlog datasets. The models evaluated include Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Neural Network, both with and without the application of Jellyfish Optimization Algorithm (JOA).

A. Summary of Findings

The experiments revealed the following key findings:

- **Decision Tree Model:** The Decision Tree model achieved the highest performance metrics, including accuracy (93.91%), precision (90.57%), recall (96.00%), and F1-score (93.20%). This indicates its superior ability to classify heart disease cases effectively compared to other models.
- **Support Vector Machine (SVM):** The SVM model showed improved performance with JOA optimization, achieving an accuracy of 86.09% and a balanced F1-score of 84.31%. The traditional SVM model performed worse, with an accuracy of 81.74%.
- **Logistic Regression:** The Logistic Regression model had an accuracy of 84.35% and a F1-score of 80.85%, with consistent performance between optimized and traditional versions.
- **Neural Network:** The Neural Network model exhibited similar performance metrics across optimized and traditional versions, with an accuracy of 84.35% and F1-score of 81.63%.

B. Implications of Optimization

The application of JOA optimization significantly improved the performance of several models, particularly SVM and Decision Tree. JOA's ability to optimize hyperparameters contributed to enhanced accuracy and balanced performance metrics, showcasing its effectiveness in refining model predictions.

C. Limitations and Future Work

Despite the promising results, the study has limitations:

- **Data Quality:** The quality of the combined dataset may impact model performance. Addressing data imbalances and exploring additional feature engineering could further enhance results.
- **Model Variability:** While Decision Tree models performed exceptionally well, other models showed less variation in performance with optimization. Future work could explore more diverse algorithms or ensemble methods.
- **Scalability:** The models were evaluated on a relatively small dataset. Testing with larger datasets or real-time data could validate the robustness and scalability of the models.

D. Summary

In conclusion, this study demonstrates that the Decision Tree model, optimized using JOA, provides the best performance for heart disease prediction. The findings highlight the importance of hyperparameter optimization in improving model accuracy and effectiveness. Future research should focus on addressing data quality issues, exploring additional models, and testing with larger datasets to further validate and enhance the predictive capabilities of heart disease prediction systems.

References

[1] Author Ahmad Ayid Ahmad, and Coauthor Huseyin Polat (2023). Heart Disease Prediction Using Cleveland Dataset and Statlog Dataset Using Jellyfish Optimization Algorithm. *Journal of Healthcare Informatics*, Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10378171/#B25-diagnostics-13-02392

[2] UCI Machine Learning Repository. (n.d.). Heart Disease Dataset. Retrieved from https://archive.ics.uci.edu/dataset/45/heart+disease

[3] UCI Machine Learning Repository. (n.d.). Statlog Heart Dataset. Retrieved from https://archive.ics.uci.edu/dataset/145/statlog+heart

[4] Satyacasm. Heart Disease Prediction. Retrieved from https://github.com/satyacasm/Heart-Disease-Prediction

[5] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, Available at: https://www.sciencedirect.com/science/article/abs/pii/0002914989905249

[6] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and Regression Trees*. Taylor Francis. Available at: https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-olshen-charles-stone

[7] Vapnik, V. N. (1995). ”Support-vector networks.” *Machine Learning*. Available at: https://link.springer.com/article/10.1007/BF00994018

[8] LeCun et al. (2015) *Deep Learning*. Available at: https://www.nature.com/articles/nature14539

[9] Jui-Sheng Chou et al. (2021) *A novel metaheuristic optimizer inspired by behavior of jellyfish in ocean*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0096300320304914

[10] Amin et al. (2019) *Identification of significant features and data mining techniques in predicting heart disease*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0736585318308876?via%3Dihub