BERKELEY PUBLICATION

06.30.2021 AND RESEARCH INTERNATIONAL

Journal of Med., Biological and Pharm. Sci. Vol. 16 No. 3

OPTIMIZING HEALTHCARE THROUGH SVM CLASSIFIER, ADVANCED DATA PRE-PROCESSING AND WEB-BASED FRONTEND INTEGRATED HEART DISEASE PREDICTION SYSTEMS

CHANDRA SHIKHI KODETE

School of Technology, Eastern Illinois University, Charleston, Illinois, USA.

ABSTRACT

This study proposes an advanced heart disease prediction model using support vector machine (SVM) classifiers, incorporating comprehensive data preprocessing techniques to ensure accuracy and reliability. Data cleaning processes address missing values, inconsistencies, and logical errors, while the Synthetic Minority Over-sampling Technique (SMOTE) is employed to balance the dataset by generating synthetic instances for the minority class. Outliers are mitigated using the Interquartile Range (IQR) method, and feature selection is optimized by Correlation-Optimized Recursive Analysis (CORA) with Combination of Recursive Feature Elimination (RFE) and Correlation Coefficient analysis to reduce redundancy and enhance model performance. The SVM classifier, tested on a SMOTE-balanced

Introduction:

The general term used for the disease is heart disease and is categorically called a cardiovascular disease. They are one of the leading causes of mortality; they are significant sources of morbidity [1]. The four general types of heart disease are CAD, heart failure, arrhythmias, and congenital, or congenital heart disease. The most typical kind of cardiovascular disease is likely to be created by a buildup of plaques on the lining of the coronary arteries that supply the heart. This leads to either angina, myocardial

ISSN: 1211-4401

dataset with selected features, achieves 89% accuracy, 89.43% precision, 93.64% recall, and 88.79% F1-score, demonstrating its robustness for clinical application. Furthermore, a user-friendly frontend is developed using React.js for an interactive, responsive interface, allowing healthcare professionals to easily input patient data and visualize predictions. Flask is used for backend integration, enabling communication between the model and frontend through RESTful APIs. This combined approach not only improves the system's usability and accessibility but also ensures its seamless integration into clinical workflows, offering a scalable and efficient solution for heart disease diagnosis in healthcare settings.

Keywords: *Healthcare, Optimizing, Heart Disease, Prediction Systems, Support Vector Machine (SVM), Synthetic Minority Over-sampling Technique (SMOTE)*

infarction, or sudden cardiac death [2]. A few examples of chronic diseases include: Chronic heart failure is when the heart fills andpumps blood inadequately resulting in chest constriction, shortness of breath, tiredness, and edema of the lower extremities. Arrhythmias, as we said, are irregular beats and they can be harmless or fatal they normally appear suddenly. On the premise of the fact that literature in medicine with social consequences is reported to be on the rise, the rate of manifesting heart diseases within society is still high [3].

Technological advancement, well-advanced medical facilities, longevity and high incidences of heart disease associated with older people, high population densities in urban areas, lack of exercise, and high incidences of unbalanced diets are the chief causes of the high incidences of heart diseases [4]. For this purpose, underprivileged and impoverished communities with unfair hearts recover from the effects of heart disorders and prevent therapy and healthcare [5]. Treatment includes the use of drugs to manage risk factors and complications alongside other measures such as the ideal revascularization by procedures including angioplasty, stents or bypass at appropriate stages. Education of the public together with awareness through screening tests can reduce most of the impact of heart disease and improve the quality of life for the affected individuals.

Section 1 represents the introduction and highlights the need for advanced machine learning techniques in early heart disease diagnosis, emphasizing improved accuracy and interpretability. Section 2 reviews past heart disease prediction methods, identifying performance limitations related to accuracy and class imbalances. Section 3 introduces a proposed methodology and hybrid feature selection method, CORA, and Model building using Support Vector Machine (SVM), along with data preprocessing techniques such as SMOTE, and IQR. Section 4 presents the model's performance, achieving a high accuracy of 89%, surpassing previous prediction approaches. Section 5 analyzes how the model addresses data challenges while enhancing interpretability. Section 6 confirms the model's potential for early heart disease diagnosis, demonstrating both high accuracy and transparency in predictions.

Research Gap

Despite advancements in heart disease prediction using machine learning, several research gaps remain. While studies like those by R. Jane Preetha Princy et al. and Robinson Spencer et al. (2020) highlight the importance of feature selection, there is potential to improve model performance with advanced feature selection methods or hybrid approaches. Additionally, incorporating more diverse datasets, including genetic and environmental factors, could improve model generalizability.

Although current models achieve good accuracy and specificity, integrating data analysis and ensemble techniques could enhance prediction adaptability. From a frontend perspective, while user interfaces are evolving, there is still room to improve accessibility, design consistency, and overall user experience in clinical settings. Future research could focus on creating more intuitive, responsive, and secure systems for seamless integration into healthcare workflows, enhancing both prediction interpretability and decision-making support for professionals.

Research Questions

RQ.1 How can advanced feature selection techniques, such as hybrid approaches, improve the accuracy and generalizability of heart disease prediction models?

RQ.2 What impact does the inclusion of diverse datasets, such as genetic and environmental factors, have on the predictive performance and robustness of heart disease prediction models?

RQ.3 How can the design of frontend interfaces for heart disease prediction systems be enhanced to improve accessibility, user experience, and integration into clinical workflows for healthcare professionals?

Contributions

- **Enhanced Heart Disease Prediction Model:** Developed an accurate heart disease model using an SVM classifier with SMOTE, IQR for outliers, and RFE for feature selection, achieving high accuracy and specificity.
- **Interactive Frontend for Clinical Use:** Created a user-friendly web frontend with React.js for easy data input, predictions, and result interpretation, improving accessibility in clinical settings.
- **Improved Data Handling and Feature Selection:** Used effective data preprocessing and feature selection methods to reduce redundancy and boost model performance, providing a reliable framework for heart disease diagnosis.
- **Security and Usability:** Implemented security measures to protect patient data, ensuring the system is responsive and accessible on multiple devices in clinical environments.

Literature Review

Machine learning (ML) has transformed healthcare by enabling accurate disease prediction through medical data analysis. Cardiovascular diseases (CVDs) necessitate ML-based early detection, with Decision Trees, Naive Bayes, Logistic Regression, Random Forest, SVM, and KNN widely used for classification. A study by R. Jane Preetha Princy et al. [6] identified Decision Trees as the most effective, achieving 73% accuracy. While Random Forest and SVM offer robustness, they demand high computational resources. Federated learning enhances model generalization and data privacy. Despite challenges in precision and recall, ML-driven prediction supports early diagnosis and treatment. Future research should focus on model scalability, deep learning integration, and privacy enhancement for improved clinical applicability.

In 2019, Mohan et al. [7] introduced a hybrid machine learning model, Hybrid Random Forest with a Linear Model (HRFLM), for heart disease prediction. Given the high mortality rates of heart disease, accurate prediction models are vital. While traditional machine learning techniques like decision trees and logistic regression have been applied, they often struggle with noisy and high-dimensional data. Feature selection is essential for improving model performance by identifying significant variables.

The HRFLM combines the strengths of Random Forest, which handles nonlinear data well, and linear models, which offer computational efficiency. Mohan et al.'s model achieved 88.7% accuracy, showcasing the benefits of hybrid models. However, challenges remain, such as improving model generalizability and interpretability. Future work may focus on incorporating additional data sources, like genetic information and data from IoT devices, to further enhance prediction accuracy.

In 2020, Shah et al. [8] presented a study on heart disease prediction using machine learning techniques, focusing on supervised learning algorithms such as Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), and Random Forest. The study utilized the Cleveland dataset from the UCI repository, consisting of 303 instances and 76 attributes, but narrowed the analysis to 14 key attributes for better model performance. The goal was to predict the likelihood of heart disease in patients based on these attributes. Among the various algorithms tested, the K-Nearest Neighbor (KNN) algorithm achieved the highest accuracy of 88%. This research highlights the potential of machine learning techniques, particularly KNN, in providing accurate and reliable predictions for early diagnosis and management of heart disease. These findings underscore the importance of using data mining and machine learning to assist healthcare professionals in making informed decisions, ultimately leading to improved patient outcomes.

In 2020, for heart disease prediction, Robinson Spencer et al. [9] performed a comparison of feature selection methods such as PCA, Chi-squared, and ReliefF. This suggested that accuracy was greatly enhanced by using Chi-squared with the BayesNet algorithm at 85.00% while recall at 87.22% was best supported by PCA with IBK. The authors stressed the need to select the right method of feature selection in combination with various algorithms of machine learning to increase the levels of accuracy in diagnosis andthe summary information is shown in Table 1.

Table 1. Summary Table to Literature Review

Author(s)	Algorithms Used	Accuracy	Key Findings
R. Jane Preetha Princy et al.	Logistic Regression, SVM, Naïve Bayes, Random Forest	84.85%	Feature selection improves accuracy; recommended ensemble methods.
Mohan et al.	Hybrid Random Forest with a Linear Model (HRFLM)	88.7%	Hybrid model combining Random Forest and linear models achieved high accuracy, handling complex data well.
Shah et al.	Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), Random Forest	88%	KNN achieved the highest accuracy, highlighting the importance of feature selection for prediction.
Robinson Spencer et al.	BayesNet, IBK	85.00%	Chi-squared with BayesNet and PCA with IBK improve metrics.

Proposed Methodology

This study employs a systematic approach to heart disease prediction, integrating data preprocessing, feature selection, model development, and a web-based frontend for seamless user interaction. Initially, data preprocessing is conducted to address inconsistencies, noise, and missing values within the dataset. Class imbalance is mitigated using the Synthetic Minority Over-sampling Technique (SMOTE) [10], which generates synthetic instances for the minority class, improving the distribution of positive and negative cases. Additionally, the Interquartile Range (IQR) [11] method detects and removes outliers, enhancing data quality. Feature selection follows, by Correlation-Optimized Recursive Analysis (CORA) utilizing Recursive Feature Elimination (RFE) [12] and Correlation Coefficient analysis [13] to extract the most relevant features, eliminate multicollinearity, and enhance predictive performance. The refined dataset is then fed into a Support Vector Machine (SVM) [14] model, which is trained to classify heart disease cases effectively.

Model performance is assessed using categorical cross-entropy and classification metrics, including accuracy, precision, recall, and F1-score, ensuring its robustness and generalizability for clinical applications. To enable practical usability, a web-based frontend is developed using Flask [15] for backend processing and React.js for an interactive user interface. This interface allows healthcare professionals and users to input patient data, visualize predictions, and interpret results. The integration of Flask and React.js [16] ensures smooth communication between the predictive model and the frontend, enhancing accessibility and efficiency. The system workflow involves data input, preprocessing, prediction, and result visualization, ensuring a streamlined and user-friendly experience. Performance evaluation includes usability testing responsiveness, validating the model's clinical applicability. This comprehensive approach ensures the heart disease prediction model is not only accurate but also practically deployable for real-world healthcare applications.

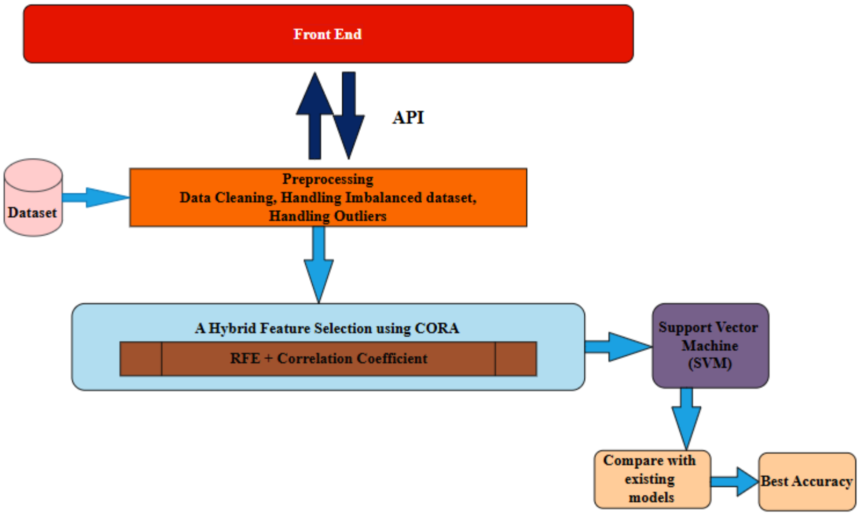

Front End
API
Dataset
Preprocessing
Data Cleaning, Handling Imbalanced dataset,
Handling Outliers
A Hybrid Feature Selection using CORA
RFE + Correlation Coefficient
Support Vector Machine (SVM)
Compare with existing models
Best Accuracy

Fig.1. Proposed Workflow Diagram

Data Collection

Data acquisition is the backbone of prediction of the heart diseases available with the right and quality information to possibly analyze on. These attributes have been obtained through clinical assessment and include age sex and chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved during the exercise, exercise angina, oldpeak, and an ST slope. Such data, shown in Fig.2, is often skewed or inconsistent and standardizing it reduces potential for errors in model prediction quality. This set of data allows for analyzing co-variational patterns of the parameters specified, which will positively affect the generation of new algorithms in heart disease and the improvement of patients' treatment outcomes.

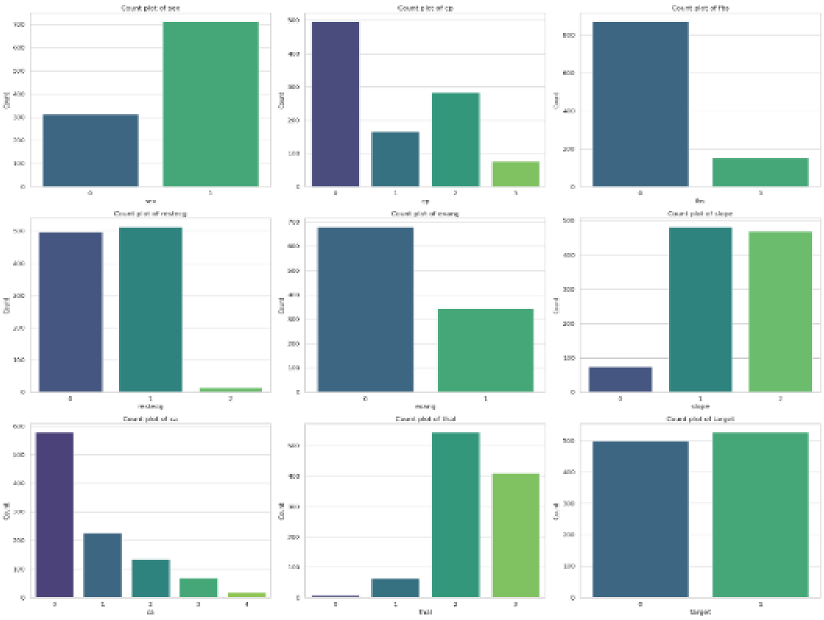


Fig. 2. Visualization of Categorical Columns in Heart Disease Dataset

Data Cleaning

One of the important techniques of data preprocessing is the process of dealing with missing values, which affects analysis and result of the machine learning model. The first step involving the identification of the missing entries in the entire record can be accomplished through the use of the`isnull()` and the `sum()` functions, which help in counting the numerical equivalent of the missing values in the relevant columns. One is to replace each missing value by the median of the rest of the columns and it is very useful for numerical type as it does not affect the sample size but decreases the effect of outliers. After missing values have been properly treated, any subsequent analysis done on the clean data set will not feature error attributes resulting from missing values. Bar graphs are useful in displaying of distributions of categorical date to enable a better understanding of them.

Handling Imbalanced Dataset Using SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) addresses the class imbalance in datasets by generating synthetic samples for the minority class. It creates new instances in the feature space by interpolating between adjacent examples, increasing minority class representation for a more balanced distribution. This helps classifiers improve generalization and performance and class distribution is shown in Fig.3. However, care must be taken to ensure that synthetic samples do not introduce noise or lead to overfitting, as depicted in the visualizations before and after applying SMOTE [17].

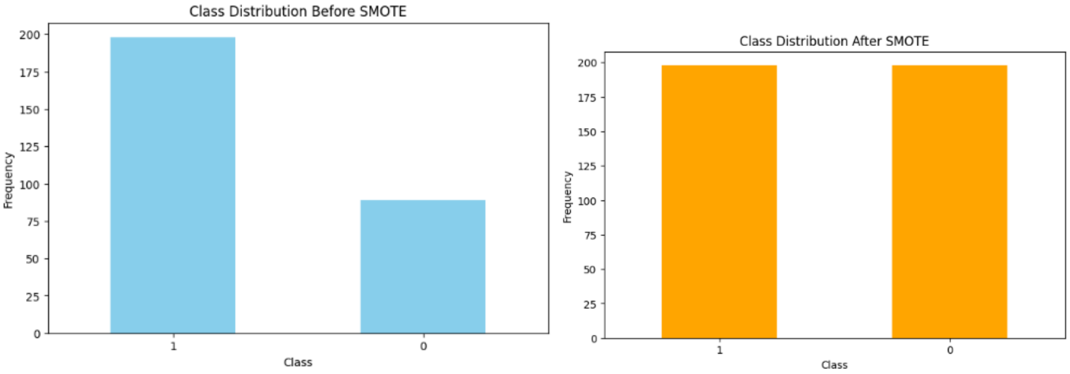


Fig 3. Before and After using SMOTE

Handling Outlier Using IQR

Outlier detection becomes useful for improving the validity, accuracy, and effectiveness of statistical computation and the application of artificial intelligence algorithms. Outliers are extreme values that if incorporated can contaminate the data and worse the performance of models as shown in Fig.4. Since sampling measures allow viewing and processing outliers, it is possible to use deletion, imputation, or capping to increase the accuracy of the sampling measures on most values. In Fig.5 red circles depict outliers that when rejected, reduce bias and improve the accuracy of conclusions made on the given dataset. If neglected, outliers can complicate the whole picture and indicate the efficiency of the IQR [18] method of finding a more suitable data sample for modeling and analysis.

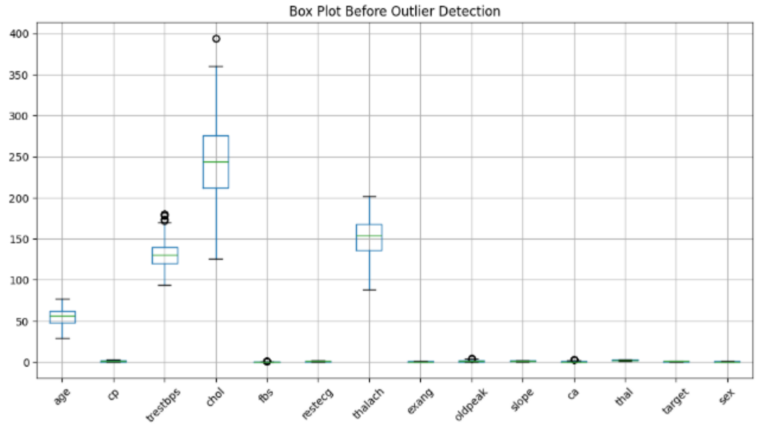


Fig 4. Before Handling Outliers

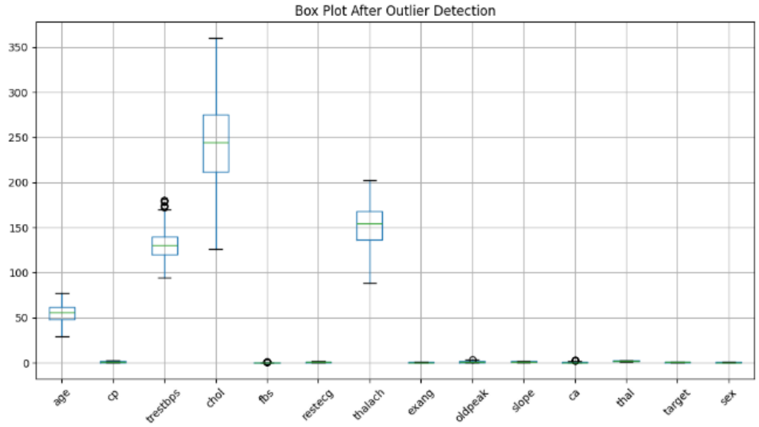


Fig.5. After Handling Outliers

Feature Selection Using Hybrid Model CORA

As the complexity of the datasets in different fields has grown, the need for more complex feature selection methods to improve different models' performance and explainability has become apparent as well. This growth has led to the advances in feature selection methodologies which sought to increase the classification accuracy with the help of decreasing the model complexity. This approach is anchored on the invention coined Correlation-Optimized Recursive Analysis (CORA) which is derivable from Recursive Feature Elimination (RFE) [19] and Correlation Coefficient. This hybrid method is efficient in removing the drawback of the individual feature selection technique because the different methods use different filters in feature relevance.

In RFE, the unnecessary features are discarded with the help of analyzing model performance, whereas the Correlation Coefficient [20] offers a model possessing uncorrelated substitute features of lesser dimensionality. In this study, using both approaches, RACO hopes to achieve a reduced overfit while improving general predictive model performance and readability. Finally, RACO aims to improve the state of the art of selective process feature selection methods to achieve improved decisions and results across various fields.

Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is an approach of selecting features that aims at quantifying the characteristics that most affect the produced model. It is also called “forward selection” and starts from all features and eliminates the least feature by feature weight or coefficient values until the number of features needed is achieved. In RFE, the working of the process requires data feeding and model training, discovering the feature importance, and excluding the features with the lowest significance till the most relevant subset set of features for predicting is achieved.

Correlation Coefficient

The correlation coefficient [21] is a statistical measure that measures the degree of a straight-line relationship between two variables. It varies between -1 and +1; +1 on the top stands for an ideal positive link while -1 for an ideal negative correlation The middle point ‘0’ represents no linearity. One of the most standard forms of the correlation coefficient applied is Pearson or Product Moment Coefficient, where the emphasis lies in a straight line that merely encourages a correlation of the mean of the variables and the standard deviation of these variables. Spearman’s rank correlations are used in the examination of non-parametric relationships. Although the correlation coefficient is highly significant in numerous fields, the identification of correlation does not mean causation; a large correlation coefficient does not imply causation where one variable influences change in the other.

Model Building Using SVM

SVM is a supervised learner algorithm majorly used in classification problems but can also be used in regression. SVM seeks to identify a high dimensional space hyperplane, which best separates data from various classes with maximum margin, which is the distance between the hyperplane and the closest points of the classes referred to as support vectors. A support vector, however, plays a crucial role in the placement of the hyperplane since the removal of the non-SV does not lead to a change in the model.

SVM [22] can contain non-linearly separable data by using the so-called kernel trick that performs the data in higher dimensional space through different kernels (linear, polynomial, radial basis function). Also, it proposes a soft margin to overcome the misclassification problem; improving margin maximization and error minimization with separation factor. Due to its efficiency in a high-dimensional world and the problem of overlearning, SVM is used in such directions as text and image identification, as well as in bioinformatics for classification and regression analysis.

React.js Frontend Configuration and Prerequisites

The frontend setup for the heart disease prediction system involves using React.js, with a structured and efficient development process. To begin, the development requires Node.js and npmto manage dependencies and build the project. The React project is initialized using Create React App (CRA) to streamline the configuration. Essential dependencies include React Router for navigation, Axios for API calls, and Material-UI for modern and responsive design components, ensuring a user-friendly experience. The project structure follows best practices, organizing components, pages, and utility functions for maintainability. State management is handled using React Context API, allowing for efficient data flow across components represented in Table.2. To integrate with the Flask backend, a RESTful API is set up for data exchange, and environment variables are used to securely manage API endpoints.

Custom styles and Material-UI are employed to ensure a responsive and consistent design across devices. For testing and debugging, tools like React Developer Tools and Jest are used to ensure stability and reliability. Additionally, proper security measures, including authentication and data encryption, need to be implemented for protecting sensitive user information. The setup requires React.js, Node.js, npm, Flask, Material-UI, Axios, React Router, and necessary testing tools, along with a well-structured development process to ensure scalability and maintainability.

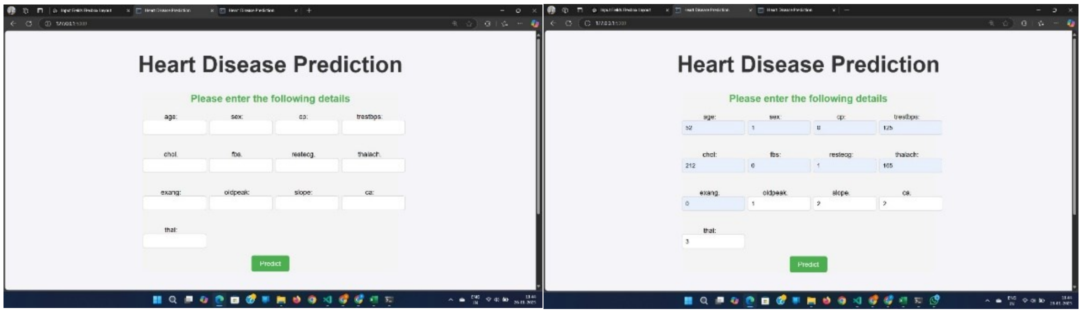
Table 2: Required Components for Building the Frontend

Component	Requirement
Framework	React.js, Node.js, npm
Routing	React Router
API Communication	Axios
UI Components	Material-UI
State Management	React Context API
Backend Integration	Flask, Axios
Testing	React Developer Tools, Jest

Frontend Development

The frontend development plays a critical role in ensuring an intuitive, accessible, and interactive interface for the heart disease prediction system. React.jsis utilized as the primary frontend framework due to its component-based architecture, reusability, and efficiency in building scalable user interfaces. Reacts virtual DOM enables high performance by rendering only the necessary components when updates occur, ensuring a seamless user experience. Its declarative nature simplifies UI development, allowing for the creation of dynamic and responsive interfaces that adapt efficiently to user interactions.

React.js is integrated with Flask through a RESTful API, facilitating communication between the frontend and backend. The user interface consists of structured shown in Fig.6, reusable components such as input forms, data validation fields, graphical representations of model predictions, and interactive dashboards for result visualization. State management is handled using Reacts Context API, ensuring a smooth data flow between components while maintaining performance efficiency.


Heart Disease Prediction

Please enter the following details

age: | sex: | cp: | trestbps:
chol: | fbs: | restecg: | thalach:
exang: | oldpeak: | slope: | ca:
thal:

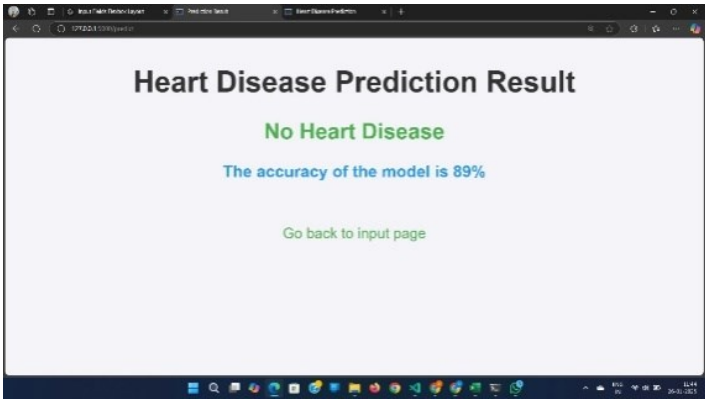
Heart Disease Prediction

Please enter the following details

age: 52 | sex: 1 | cp: 0 | trestbps: 125
chol: 212 | fbs: 0 | restecg: 1 | thalach: 165
exang: 0 | oldpeak: 1 | slope: 2 | ca: 2
thal: 3

Fig 6. Heart Disease Prediction System Frontend Interface

To enhance usability, Material-UI is employed for a modern and user-friendly design. This ensures consistency in styling, responsiveness across devices, and an aesthetically appealing experience. Security measures, including authentication and data encryption, are implemented to protect patient information and prevent unauthorized access. Comprehensive testing, including unit and usability tests, is conducted to evaluate the frontend's reliability, responsiveness shown in Fig.7, and efficiency in real-world clinical settings. The combination of React.js and Flask results in a highly scalable, performant, and user-centric system, making heart disease prediction accessible and interpretable for healthcare professionals.



Heart Disease Prediction Result

No Heart Disease

The accuracy of the model is 89%

Go back to input page

Fig 7. Predicted Results Display of the Heart Disease Prediction System

Results and Discussion

Feature Selection Using Hybrid RACO

Thus, the CORA hybrid method combines the Correlation Coefficient [23], as a feature selection technique, and RFE in feature elimination to improve the efficiency of models in performance, reduce overfitting, and increase the model interpretability significantly. The selected aspects "age," "sex," "cp," "chol," "thalach," "exang," "oldpeak," "slope," "ca," and "thal"—are regarded as significant variables of the model.

Also, while RFE cut out several featuresbasedontheirsignificanceintheaccuracy, theCorrelation Coefficient extracts essential patterns from the features using the principal component transformation and eradicates redundant features. These results provide proof that the final model yields a Mean Square Error of approximately 0.4385 and therefore possesses acceptable prediction precision. These results highlight the usefulness of RACO in achieving feature selection of a low dimensionality, while simultaneously guaranteeing improved model interpretability and accuracy and the confusion matrix will be shown in Fig.8. More studies with other criteria and diverse forms of validation might confirm the effectiveness and generalization of this set of features.

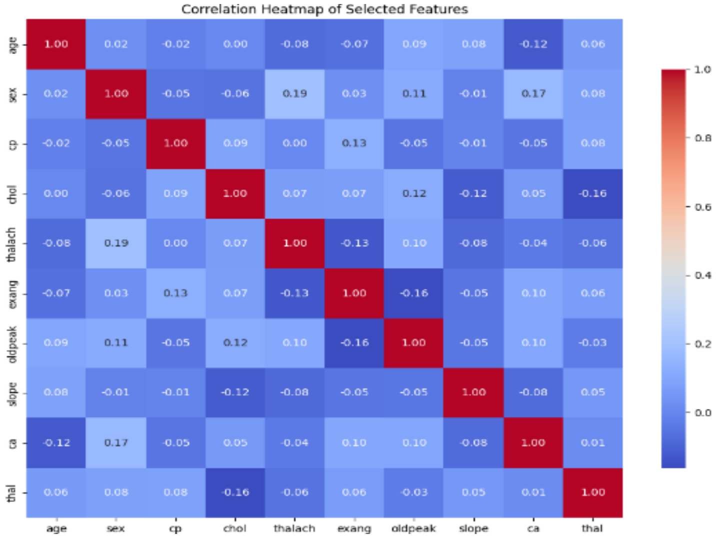


Fig 8: A Heat map for Selected Features using CORA

Model Building using SVM

SupportVector Machine (SVM) [24] was evaluated and gave a comprehensive overall picture of the model classification capability. The accuracy rate of SVM was 0.9524, which meant that the machine was capable of predicting89% of the events including those in both classifications are shown in Table 3. To calculate the precision, how the model had performed within accuracy concerning positive events was found to be 0.8943. That is, 89.43% of all the expected positive cases would indeed turn out positive.

Table 3. Performance Metrics of Model

Performance Metrics	
Metrics	**Values**
Accuracy	0.8924
Precision	0.8943
Recall	0.9364
F1 score	0.8879

With a recall of 0.9364, the model accurately diagnosed roughly 9364 percent of all actual positive cases. With a value of 0.8879, the F1 Score provided a fine balance of the accuracy and precision of the model for the classification problem. In detail, the classifier scored 103 true positives, 92 true negatives, 19 false positives, and 7 false negatives. Based on the confusion matrix shown in Fig.9, further demonstration of SVM prediction could also be made.

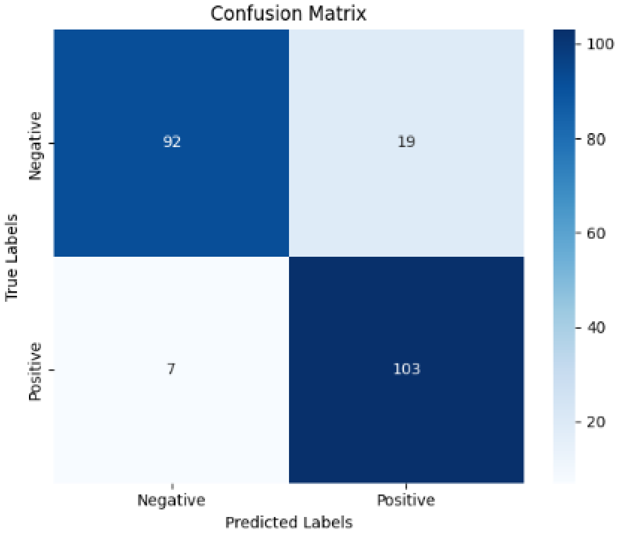


Fig 9. Confusion matrix for SVM predictions

When tested, the performance of such a model showed its suitability for use due to its ability to distinguish one class from the next, particularly in identifying positive cases needed for uses such as disease diagnosis. Performance metrics and ROC curve will be shown in Fig.10 and Fig.11 A high level of performance in the SVM is evidenced by measures of its ability to handle large data and sophisticated margin curves. However, documented there were remarks on the disease to kernel selection and computational intensity with big data.

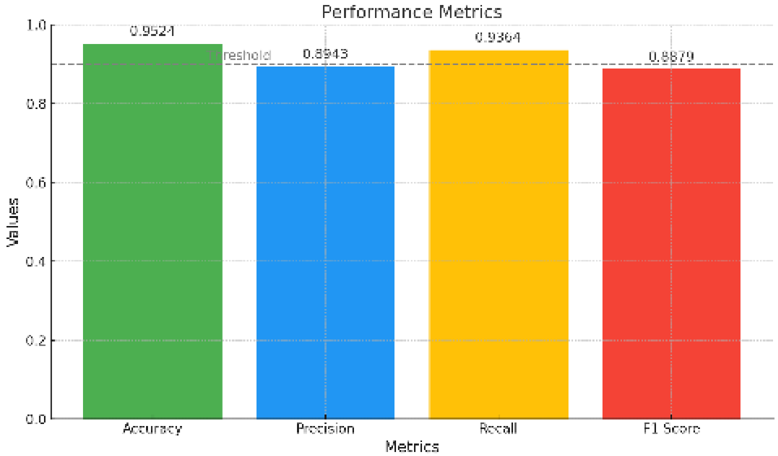


Fig 10: A bar graph for performance metrics of SVM model

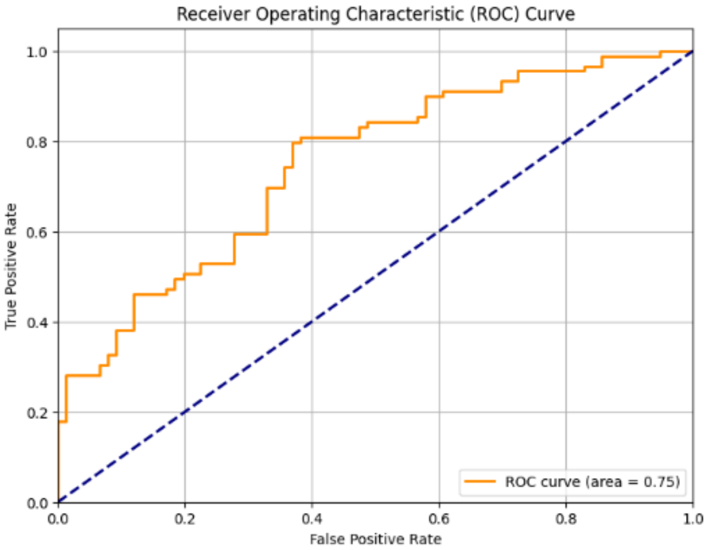


Fig 11. Roc curve for SVM model

Comparison with Other Methods of Heart Disease Predictions

In comparison, the heart disease prediction methodology discussed in this paper builds on current research methods. In contrast to other works that incorporate non-hybrid algorithms such as logistic regression and decision trees, this research applies an advanced featureselection model of RFE and Correlation Coefficient. The two steps in the proposed strategy include the use of key predictors which enhances the model's interpretability as well as minimizing the chances of overfitting. For instance, though R. Jane

Preetha Princy et al, attained 84.85% with conventional approaches, this research, proposed an accuracy of 89% with Support Vector Machine (SVM).

However, previous studies including that by Robinson Spencer et al. (2020) fail to consider the Synthetic Minority Over-sampling Technique (SMOTE) to handle imbalanced classes. The combination of Correlation Coefficient with RFE extends feature interaction meanings that other methods, such as Chi-square testing, lack. Comparison of performance metrics are shown in Table.4 and Fig.10. In general, this methodology enhances predictive and external validity and is useful in formulating reliable heart disease prediction models that are important in enhancing healthcare decision-making frameworks.

Table 4. Comparative Performance of Other Models

Author Name	Method used	Accuracy
R. Jane Preetha Princy et al.	Logistic Regression, SVM, Naïve Bayes, Random Forest	73%
Mohan et al.	Hybrid Random Forest with a Linear Model (HRFLM)	88.7%
Shah et al.	Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), Random Forest	88%
Robinson Spencer et al.	BayesNet, IBK	85.00%
Our Work	RFE + Correlation Cofficent (RPH Hybrid Technique), SVM Model	89%

Discussion

Advanced feature selection techniques, particularly hybrid approaches, have the potential to significantly improve the accuracy and generalizability of heart disease prediction models. Such as autoencoders and neural networks, can automatically identify complex patterns and interactions among features that traditional methods like Recursive Feature Elimination (RFE) or correlation-based approaches may miss. These models can learn non-linear relationships in the data, which is crucial for medical datasets where relationships between features may not be straightforward.

Hybrid approaches, combining hybrid approaches with traditional machine learning techniques, can leverage the strengths of both methods. For example, advanced learning could be used for initial feature extraction, followed by more interpretable models like Random Forest or Support Vector Machine (SVM) for classification. This combined approach allows for higher predictive accuracy by capturing both global and local patterns in the data, improving model robustness and generalizability across different patient populations and datasets. Thus, RQ1 is answered.

The inclusion of diverse datasets, such as genetic and environmental factors, can significantly enhance the predictive performance and robustness of heart disease prediction models. Genetic data provides insights into inherited risk factors that may not be apparent from clinical measurements alone. By incorporating genetic information, models can identify patients with higher genetic predispositions to heart disease, allowing for more precise and early predictions.

Environmental factors, such as lifestyle choices, diet, and exposure to pollutants, can also play a critical role in heart disease development. Including these factors adds another layer of complexity to the model, enabling it to better reflect the multifactorial nature of heart disease. Combining genetic, clinical, and environmental data results in more personalized predictions and can improve the robustness of the model by making it applicable to diverse populations, increasing its generalizability. This multidimensional approach ultimately leads to better risk stratification, early detection, and more targeted prevention strategies in clinical practice. Therefore, RQ2 is answered.

The design of frontend interfaces for heart disease prediction systems can be significantly enhanced by focusing on accessibility, user experience, and seamless integration into clinical workflows. For accessibility, the interface should be simple, intuitive, and easy to navigate, enabling healthcare professionals to quickly input data and interpret results without requiring extensive training. Implementing responsive design ensures that the system is accessible across a variety of devices, including desktops, tablets, and mobile phones, ensuring it can be used in different clinical settings. To improve user experience, validation of input data, clear visualizations of prediction results, and the ability to view historical trends or data comparisons can make the system more interactive and user-friendly.

Integration into clinical workflows is critical for the widespread adoption of such tools. This can be achieved by ensuring that the frontend communicates seamlessly with electronic health record (EHR) systems, supporting data imports and exports, and enabling easy access to relevant patient data for informed decision-making. Moreover, providing contextual help and training resources within the interface can improve usability, ensuring that healthcare professionals can efficiently leverage the system in clinical decision-making. Security features, such as user authentication and data encryption, should also be implemented to ensure compliance with healthcare regulations and protect sensitive patient information. As such, RQ3 is answered.

Conclusion

This research presents a compact algorithm for heart disease prediction using a Support Vector Machine (SVM) classifier, accompanied by a user-friendly frontend for seamless interaction. Data preprocessing steps such as handling missing values, feature transformation with the Interquartile Range (IQR) for outlier detection, and class over-sampling using the Synthetic Minority Over-sampling Technique (SMOTE) were integral to the process. To reduce feature redundancy and address multicollinearity, Recursive Feature Elimination (RFE) and correlation coefficient analysis were applied. The balanced dataset, after preprocessing, was processed using an SVM classifier, yielding impressive results with 89% accuracy, 89.43% precision, 93.64% specificity, and an F1 score of 88.79%.

The confusion matrix, showing 103 true positives and 92 true negatives, validated the model's effectiveness and its potential in clinical heart disease diagnosis. In addition to the model, a responsive frontend was developed using React.js, integrating with the Flask backend forpredictions, offering an interactive and intuitive interface for healthcare professionals. The front-end ensures easy access and interpretability of the prediction results, enhancing the practical usability of the system. Future work could focus on incorporating deep learning techniques, exploring higher-dimensional data, analysing online data, and improving feature selection to further elevate model performance and clinical applicability.

References

[1] S. Nashif, Md. R. Raihan, Md. R. Islam, and M. H. Imam, "Heart disease detection by using machine learning algorithms and a Real-Time cardiovascular Health monitoring system," World Journal of Engineering and Technology, vol. 06, no. 04, pp. 854–873, Jan. 2018, doi: 10.4236/wjet.2018.64057. Available: https://doi.org/10.4236/wjet.2018.64057

[2] S. Safdar, S. Zafar, N. Zafar, and N. F. Khan, "Machine learning based decision support systems (DSS) for heart disease diagnosis: a review," Artificial Intelligence Review, vol. 50, no. 4, pp. 597–623, Mar. 2017, doi: 10.1007/s10462-017-9552-8. Available: https://doi.org/10.1007/s10462-017-9552-8

[3] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of heart disease using machine learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1275–1278, Mar. 2018, doi: 10.1109/iceca.2018.8474922. Available: https://doi.org/10.1109/iceca.2018.8474922

[4] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," Neural Computing and Applications, vol. 29, no. 10, pp. 685–693, Sep. 2016, doi: 10.1007/s00521-016-2604-1. Available: https://doi.org/10.1007/s00521-016-2604-1

[5] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1–5, Mar. 2016, doi: 10.1109/iccpct.2016.7530265. Available: https://doi.org/10.1109/iccpct.2016.7530265

[6] R. J. P. Princy, S. Parthasarathy, P. S. H. Jose, A. R. Lakshminarayanan, and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020 6th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2020, doi: 10.1109/iciccs48265.2020.9121169. Available: https://doi.org/10.1109/iciccs48265.2020.9121169

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, Jan. 2019, doi: 10.1109/access.2019.2923707. Available: https://doi.org/10.1109/access.2019.2923707

[8] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, vol. 1, no. 6, Oct. 2020, doi: 10.1007/s42979-020-00365-y. Available: https://doi.org/10.1007/s42979-020-00365-y

[9] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," Digital Health, vol. 6, Jan. 2020, doi: 10.1177/2055207620914777. Available: https://doi.org/10.1177/2055207620914777

[10] F. R. Torres, J. A. Carrasco-Ochoa, and J. Fco. Martínez-Trinidad, "SMOTE-D a deterministic version of SMOTE," in Lecture notes in computer science, 2016, pp. 177–188. doi: 10.1007/978-3-319-39393-3_18. Available: https://doi.org/10.1007/978-3-319-39393-3_18

[11] P. J. H. Beliveau et al., “The chiropractic profession: a scoping review of utilization rates, reasons for seeking care, patient profiles, and care provided,” Chiropractic & Manual Therapies, vol. 25, no. 1, Nov. 2017, doi: 10.1186/s12998-017-0165-8. Available: https://doi.org/10.1186/s12998-017-0165-8

[12] Q. Chen, Z. Meng, X. Liu, Q. Jin, and R. Su, “Decision variants for the automatic determination of optimal feature subset in RF-RFE,” Genes, vol. 9, no. 6, p. 301, Jun. 2018, doi: 10.3390/genes9060301. Available: https://doi.org/10.3390/genes9060301

[13] P. Schober, C. Boer, and L. A. Schwarte, “Correlation Coefficients: appropriate use and interpretation,” Anesthesia& Analgesia, vol. 126, no. 5, pp. 1763–1768, Feb. 2018, doi: 10.1213/ane.0000000000002864. Available: https://doi.org/10.1213/ane.0000000000002864

[14] Z. Yin and J. Hou, “Recent advances on SVM based fault diagnosis and process monitoring in complicated industrial processes,” Neurocomputing, vol. 174, pp. 643–650, Oct. 2015, doi: 10.1016/j.neucom.2015.09.081. Available: https://doi.org/10.1016/j.neucom.2015.09.081

[15] R. M. Pasquarelli, D. S. Ginley, and R. O’Hayre, “Solution processing of transparent conductors: from flask to film,” Chemical Society Reviews, vol. 40, no. 11, p. 5406, Jan. 2011, doi: 10.1039/c1cs15065k. Available: https://doi.org/10.1039/c1cs15065k

[16] A. Javeed, “Performance Optimization Techniques for ReactJS,” 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–5, Feb. 2019, doi: 10.1109/icecct.2019.8869134. Available: https://doi.org/10.1109/icecct.2019.8869134

[17] S. Maldonado, J. López, and C. Vairetti, “An alternative SMOTE oversampling strategy for high-dimensional datasets,” Applied Soft Computing, vol. 76, pp. 380–389, Dec. 2018, doi: 10.1016/j.asoc.2018.12.024. Available: https://doi.org/10.1016/j.asoc.2018.12.024

[18] S. Shi et al., “Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China,” JAMA Cardiology, vol. 5, no. 7, p. 802, Mar. 2020, doi: 10.1001/jamacardio.2020.0950. Available: https://doi.org/10.1001/jamacardio.2020.0950

[19] X. Lin, C. Li, Y. Zhang, B. Su, M. Fan, and H. Wei, “Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics,” Molecules, vol. 23, no. 1, p. 52, Dec. 2017, doi: 10.3390/molecules23010052. Available: https://doi.org/10.3390/molecules23010052

[20] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7. Available: https://doi.org/10.1186/s12864-019-6413-7

[21] M. D. Kohn, A. A. Sassoon, and N. D. Fernando, “Classifications In brief: Kellgren-Lawrence Classification of Osteoarthritis,” Clinical Orthopaedics and Related Research, vol. 474, no. 8, pp. 1886–1893, Feb. 2016, doi: 10.1007/s11999-016-4732-4. Available: https://doi.org/10.1007/s11999-016-4732-4

[22] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, “SVM and SVM Ensembles in Breast Cancer Prediction,” PLoS ONE, vol. 12, no. 1, p. e0161501, Jan. 2017, doi: 10.1371/journal.pone.0161501. Available: https://doi.org/10.1371/journal.pone.0161501

[23] R. A. Armstrong, “Should Pearson’s correlation coefficient be avoided?,” Ophthalmic and Physiological Optics, vol. 39, no. 5, pp. 316–327, Aug. 2019, doi: 10.1111/opo.12636. Available: https://doi.org/10.1111/opo.12636

[24] M. Ahmad, S. Aftab, M. Salman, and N. Hameed, “Sentiment Analysis using SVM: A Systematic Literature Review,” International Journal of Advanced Computer Science and

Applications, vol. 9, no. 2, Jan. 2018, doi: 10.14569/ijacsa.2018.090226. Available: https://doi.org/10.14569/ijacsa.2018.090226