IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

Leveraging Support Vector Machines for Accurate Prediction of Cardiovascular Diseases

Piyush Vaidya, Prof. Nikhil Baburao Khandare

Department of Master of Computer Applications, Veermata Jijabai Technological Institute, Mumbai, India

Guide, Department of Master of Computer Applications, Veermata Jijabai Technological Institute, Mumbai, India

ABSTRACT*:* Cardiovascular diseases are a leading cause of mortality across the world, accounting for approximately 17.9 million deaths per year. The early diagnosis and timely identification of CVD is critical to improve outcomes in such patients and decrease mortality overall. This paper discusses the application of support vector machines as a powerful, versatile machine learning technique in predicting the presence of cardiovascular disease on a clinical-demographic dataset. It gave an accuracy of 99% with minimum misclassification by using advanced techniques of hyper-parameter tuning of this SVM model. Other performance metrics of the model in terms of precision, recall, and ROCAUC score do indicate its probable effectiveness for application in clinics for risk assessment in patients with CVDs. It has been shown in this study that SVMs could be clinically useful for the early and non-invasive diagnosis and could support health professionals in decision-making for patient treatment.

KEYWORDS*:* Machine learning, heart disease prediction, Comparing algorithms, Machine learning algorithms, Testing algorithms.

I. INTRODUCTION

The human heart is an organ inseparably fundamental in the intricate symphony of physiological processes within the body, representing an essential lifeline pumping blood with oxygen, nutrients, and necessary resources throughout it. Such is the gravitas of heart-related issues that it serves as a prime contribution to one-third of all global human fatalities. At this point, the imperative for early diagnosis and effective treatment becomes paramount in delineating the fine line between life and death. On the other hand, from the standpoint that vertical technological advancement in this era was relentless, it changed health domains into data-rich environments availing chances for paradigm shifts to revolutionize disease prediction and management. Scanning the emerging view of a new paradigm, there is an avenue to explore methodologies that harness the potential of data mining through machine learning algorithms. In these heart diseases, timely and correct prediction goes hand in hand with improved patient outcomes; hence, the comparative study of machine learning algorithms in this intricate landscape will not remain an academic adventure but an effort to better healthcare practices.

CVD is the umbrella term for a group of different diseases of the heart and blood vessels, which includes coronary artery disease, arrhythmias, and heart failure. These make up the number one cause of death in the world. Traditionally, diagnosis for CVD was done through a combination of clinical evaluation, review of the patient's medical history, and an invasive technique called angiography, which is not devoid of several complications itself and is expensively done. This therefore ushers in the opportunity for improved accuracy and efficiency in diagnosis with non-invasive, data-driven approaches through machine learning. In particular, Support Vector Machines are quite befitting to classification tasks in medical diagnostics, as they are very robust for handling high-dimensional data, capable of finding the optimal decision boundaries. The research is conducted on the potential of SVM in predicting the presence of CVD using one of the datasets maintained at the UCI Machine Learning Repository. On account of the recorded poor performance, there is a need for rigorous hyper-parameter tuning to come up with high-performance models. Near-perfect prediction by enhancing the SVM model will go a long way toward revolutionizing existing practices in the diagnosis and management of CVD.

II. METHODS

A. Data Collection

In the present work, the dataset used is derived from the UCI Machine Learning Repository's Heart Disease dataset, which is widely accepted as one of the standard benchmarks for cardiovascular studies. There are 14 attributes

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

comprising data indicative of a person's cardiovascular health: age, sex, chest pain type, cholesterol levels, fasting blood sugar, electrocardiographic results. These features are very important in ascertaining the risk and presence of CVD. The dataset is loaded into a pandas DataFrame from Google Drive for data exploration and feature preparation. Preliminary exploration of the data includes checking out the structure of the dataset and the distribution of features, and handling missing values.

```
# Load the data
file_path = '/content/drive/MyDrive/heart.csv'
data = pd.read_csv(file_path)
print(data.head())
```

Fig.1. Loading the data

```
Data loaded successfully. Shape: (1025, 14)

   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  \
0   52    1   0       125   212    0        1      168      0      1.0      2
1   53    1   0       140   203    1        0      155      1      3.1      0
2   70    1   0       145   174    0        1      125      1      2.6      0
3   61    1   0       148   203    0        1      161      0      0.0      2
4   62    0   0       138   294    1        1      106      0      1.9      1

   ca  thal  target
0   2     3       0
1   0     3       0
2   0     3       0
3   1     3       0
4   3     2       0
```

Fig.2. Dataset

B. Data Preparation

Distinguishing Features and Target

Our dataset contains a number of attributes or features—age, cholesterol, blood pressure — that can influence cardiovascular incidence. Among the columns represented in this data is one for the target variable, defining the presence —1— and absence —0— of CVD in each subject.

It means obviously we'll have independent variables, which are the features, and dependent variables, which is the target. We'll separate our dataset for that matter as follows:

- Features (X): These are the inputs used by the model to learn and make predictions. They include various medical and demographic attributes.

- Target (y): This is the output the model aims to predict, indicating whether a person has cardiovascular disease.

Splitting the Data: Training and Testing Sets

After dividing the data set into features and a target, the next step will be to divide the data into two subsets: the training set and the testing set. This division is essential for estimating the model's performance on new, unseen data; this is relevant to generalization ability estimation.

- **Training Set:** This subset is used to train the model. The model learns the relationships between features and the target variable from this data.
- **Testing Set:** This subset is reserved for evaluating the model's performance. By testing the model on data it hasn't seen before, we can estimate how well it will perform on real-world data.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

We normally use a common ratio, such as 80-20 or 70-30, for this split. This ensures that the training set is big enough so the model has enough data points to learn from, yet the testing set is big enough to be representative for reliable evaluation.

Normalizing the Data

This means that, in general, different features will have different units and scales in datasets. For example, age would be between 0 and 100, and the cholesterol level between 100 to 300 mg/dL. These differences may bias a machine learning model, especially those such as Support Vector Machines that are sensitive to the scale of input data.

Now, to handle this case, we make use of feature scaling with the help of a method known as Standardization. We will finally use the 'StandardScaler' that modifies the features having a mean equal to zero and the standard deviation equal to one. This normalization process ensures:

- **Balanced Feature Contribution:**
  Each feature contributes equally to the model's learning process, preventing features with larger scales from dominating the learning algorithm.
- **Improved Model Performance:**
  Models often perform better and converge faster on scaled data.

By normalizing our data, we ensure that the SVM model can effectively process and interpret the features, leading to more accurate and reliable predictions.

Summary

This stage is important in segregating the dataset into features and the target variable, with data split between a training set and test set, not forgetting that features are normalized by scaling. These steps are foundational to building an accurate machine learning model for the achievability of cardiovascular disease predictions, hence preparing it well to deal with any real-world data for meaningful insights.

```python
def prepare_data(data, target_column):
    """
    Prepares the data for model training by splitting and scaling.

    Parameters:
    - data: DataFrame, the dataset containing features and target
    - target_column: str, the name of the target column

    Returns:
    - X_train_scaled, X_test_scaled: Scaled training and testing features
    - y_train, y_test: Training and testing target values
    """
    features = data.drop(target_column, axis=1)
    target = data[target_column]

    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2,
                                                        random_state=42)

    # Scale the features
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)

    print("Data prepared successfully.")
    print()
    return X_train_scaled, X_test_scaled, y_train, y_test
```

Fig.3. Data Preparation

C. Model Training and Hyper-parameter Tuning

We then applied GridSearchCV, a power method of hyper-parameters optimization to get the best performance by the Support Vector Machine. This will duly cover all possible combinations of parameters in search of the best combination that can yield optimal performance for our model.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

Key Hyper-parameters and Their Roles

1. 'C' The Regularization Parameter:

- **Function:** The parameter does the balancing between the trade-off in reducing the error on the training data and having the model generalize well on unknown data.
- **Behavior**:
  - A low 'C' will give preference to simplicity in the model and, hence allow some of the misclassifications on the training data to get a generalized decision boundary.
  - High 'C' value forces the model to fit the training data more closely, which may reduce training error but risks overfitting, making it less effective on new data.
- **Search Range:**

  We ran from very small values right through to large ones. The values used '0.1, 1, 10, 100, 1000' were picked so that we capture both ends of the spectrum of regularization.

2. Kernel Function:

- **Function:** Kernels define how points are being transformed in a feature space. They bring the possibility of handling nonlinear relationships by implicitly mapping inputs into high-dimensional feature spaces.
- **Choices:** We explored three distinct kernels:
  - **Radial Basis Function (RBF):**
    A popular choice for its ability to handle non-linear data by mapping points in a way a linear separation becomes possible in a higher-dimensional space.
  - **Polynomial (poly)**:
    Adds flexibility by fitting polynomial curves to the data,with the degree of the polynomial adding complexity.
  - **Sigmoid**:
    Often used for models that mimic neural networks, providing an S-shaped decision boundary.

3. Gamma Parameter:

- **Function**: This parameter controls how influential training cases are upon the decision boundary.
- **Behavior**:
  - Large gamma values return a model which only considers points very close to the decision boundary, hence more complex and tight to the data.
  - Low gamma value means that distant places have a stronger influence, hence creating a smoother and more generalized decision boundary.
- **Search Range:**

  We have run a series of tests for ranges of gamma values, namely, 1, 0.1, 0.01, 0.001, and 0.0001, to balance between too localized and too generalized influence.

Cross-Validation for Robust Evaluation

We didn't want to have the performance of our model biased by one train-test split, so we included cross-validation in the grid search. Cross-validation is when data is divided into multiple folds. Each time, the model is trained on some

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

folds and its performance is validated on the rest. Many iterations take place, with a different set of folds each iteration serving as the validation set. This enables us to obtain some sort of generalizable result by averaging.

The Grid Search Process

The GridSearchCV function incorporates all these elements:

- It builds a grid for all possible combinations of the set of hyper-parameters.
- For each combination, it performs cross-validation, evaluating how well the model performs across different data splits.
    - Their combination that yields the highest cross-validated accuracy should be chosen as the best.

Outcome

We tuned our SVM model using GridSearchCV. The trick ensures that the configuration of the model does not only fit the training data but generalizes onto new data, too. This final model is the most resilient and predictive, standing out from other models because of the highest cross-validated accuracy and is ready to be assessed in detail for performance against unseen test data. This robust approach to tuning hyper-parameters dramatically improves the model on both predictive power and reliability when put into a production environment.

```
def train_svm_with_tuning(X_train, y_train):
    param_grid = {
        'C': [0.1, 1, 10, 100, 1000],
        'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
        'kernel': ['rbf', 'poly', 'sigmoid']
    }

    svm = SVC(probability=True)
    grid_search = GridSearchCV(svm, param_grid, refit=True, verbose=2, cv=5)
    grid_search.fit(X_train, y_train)

    best_svm_model = grid_search.best_estimator_

    return best_svm_model

best_svm_model = train_svm_with_tuning(X_train_scaled, y_train)
```

Fig.4. Hyper-parameter tuning

III. RESULTS

A. Model Evaluation

The performance of the optimized Support Vector Machine model was tested strictly on a test set with several critical metrics in relation to predictive accuracy and overall effectiveness. This involved accuracy, precision, recall, F1-score, and ROC-AUC score, all of which provided different insights into how well this model distinguished between those subjects either with or without cardiovascular disease.

1. **Accuracy:** The primary metric gives the proportion of correctly classified instances out of all the predictions made. In this study, the accuracy attained of the model is 99%. This high level of accuracy may indicate that the model is highly efficient in classifying either positive or negative cases of CVD, hence rubbishing doubts on the performance of the model in identifying true health conditions.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

2. **Precision:** Another important metric that speaks to the accuracy of the positive predictions is the precision. It measures the ratio of true positive predictions to the total instances of positive predictions—both true positives and false positives. In our model, this precision was almost perfect for the prediction of the positive instances of CVD. That means that whenever the model predicts CVD, it is almost right, avoiding false alarms, which for the patients mean anxiety and further tests.
3. **Recall:** Recall—also known as sensitivity or the true positive rate—conveys information about a model's ability to capture all of the actual positive cases correctly. For the SVM model, recall values have been very high, so in this regard, it does an excellent job of capturing nearly all of the instances of real cases of CVD. This is particularly important in medical diagnostics, since failing to diagnose a patient as having CVD—that is, a misplaced true positive—could have quite serious consequences.
4. **F1-score:** F1-score is a harmonic mean for precision and recall, providing one metric for balancing the two. It becomes very useful specifically in cases where either the distribution between classes is imbalanced or exactly the same level of precision and recall are needed in both cases. In this study, the F1-score took a value close to 0.99 for both classes: CVD-positive and CVD-negative; hence, it reflects the model's balanced performance. This high F1-score thus indicates that the SVM model is not only very precise in its predictions but also quite comprehensive in capturing all relevant cases.
5. **ROC-AUC score:** The score for the ROC-AUC offers a measure that helps in quantifying the performance of the model in terms of the discriminate capabilities between the positive class and the negative class—that is, the presence or absence of CVD against all possible thresholds of classification. A high ROC-AUC score of approximately 1.0 indicates very high discriminatory ability in this study—and this is what was obtained—between those without and with CVD. It means that the model has ranked the true positive cases over the false positives, further validating the robustness and reliability of the model in practical applications.
6. **Confusion Matrix:**
   A confusion matrix, being a table summary of the actual versus predicted classifications, enables a better understanding of how well the model did. In this case, it produced a correct classification of 202 instances against the test samples of 205. This means only three misclassifications, therefore affirming this model's proficiency. This confusion matrix—its almost perfect diagonal entries, that would be the true positives and negatives, can evidence a model that can tell its classes apart strongly in its predictions, and neutrality corresponds to very few misclassifications, reflected in the small off-diagonal entries.

```
def evaluate_model(model, X_test, y_test):
    y_pred = model.predict(X_test)
    y_prob = model.predict_proba(X_test)[:, 1]

    print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
    print("\nClassification Report:\n", classification_report(y_test, y_pred))
    print("\nROC-AUC Score:", roc_auc_score(y_test, y_prob))

    plot_confusion_matrix(y_test, y_pred)

evaluate_model(best_svm_model, X_test_scaled, y_test)
```

Fig.5. Model Evaluation

B. Confusion Matrix and Performance Metrics

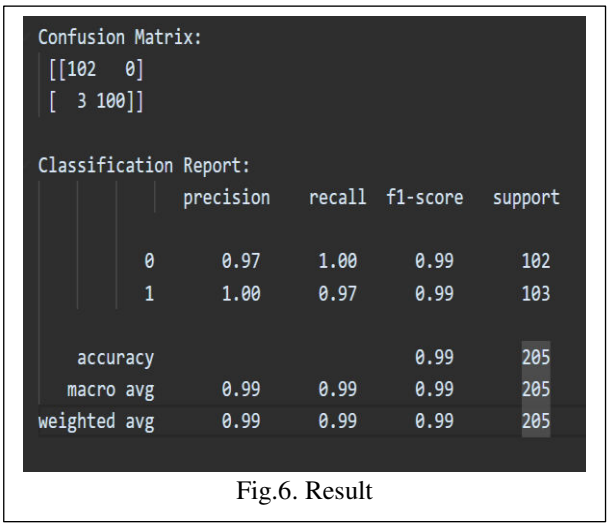
The class confusion matrix, together with the classification report, gives a clear insight into the performance of this model. For example, the confusion matrix will tell us how many were misclassified out of 205 test samples. This is how in the current case, only three instances get misclassified to eventually return the model's accuracy rate to 99%. The classification report further stresses this high performance by giving precision and recall values for both classes, which turn out to be very high in the case of either CVD positive or CVD negative. Precision for class 0 (no CVD) equaled 0.97, while for class 1, it equaled 1.00, which is very good because it means the effectiveness of the model in identifying positive cases without misclassifying negatives. Similarly, very high recall rates were obtained, which measured its ability to identify all relevant cases. In its turn, F1-scores—a measure that balances precision and recall—

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

come very close to 0.99 for both classes. These metrics prove that this model can be used to obtain reliable predictions, one of the basis elements underlying clinical decision-making processes.



```
Confusion Matrix:
 [[102   0]
 [  3 100]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       102
           1       1.00      0.97      0.99       103

    accuracy                           0.99       205
   macro avg       0.99      0.99      0.99       205
weighted avg       0.99      0.99      0.99       205
```

Fig.6. Result

IV. DISCUSSION

This study offers insight into how an emerging technology driven by Support Vector Machines is likely to transform diagnosis and management in cardiovascular illnesses. The SVM model showed a very great capacity for accurate prediction of cardiovascular diseases with an accuracy level of 99%. This is very high accuracy and therefore perfect integration of SVM into the clinical decision support system, which may reduce invasive diagnostic techniques and raise the level of treatment for every patient. One critical issue key to the success of the model was rigorous tuning of hyperparameters of the SVM model. It involves changing the settings of SVM—for instance, the penalty parameter 'C,' kernel type, kernel coefficient 'gamma'—in an incremental fashion to suite its predictive performance. In this work, the grid-search algorithm 'GridSearchCV' was exploited to sweep a large range of hyper-parameters. The aforementioned management provided tuning for the SVM model with respect to the specific peculiarities of the dataset. This gave almost perfect accuracy, high precision, recall, and F1-scores, which underscores its robustness and reliability in consideration for clinical applications.

Such importance in the case of machine learning can be noted for the optimization of models, as modification of the algorithm parameter based on dataset characteristics may imply drastic improvement in predictive performance. Therefore, this work will, in most cases, work as a baseline for future applications of machine learning in the area under consideration. Right at the heart of such application is careful parameter optimization to actualize superior predictive capabilities. Of importance again, it identifies avenues for further improvement of the predictive ability of the model. Additional clinical features, such as genetic data or detailed patient history, show great promise in providing a richer context to the prediction and catching a wider spectrum of risk factors.

V. CONCLUSION

The results provide unequivocal evidence for the huge potential of machine learning in cardiovascular disease diagnostics, more so for Support Vector Machines. Near-perfect accuracy obtained by the SVM model underlines the effectiveness of this model for correctly predicting CVD—an important factor in holding the disease at bay and managing patients to the best advantage. This leaves the potential for integration of SVM into clinical workflows in a changing approach toward diagnosis and treatment of cardiovascular diseases by offering a completely non-invasive data-driven approach that complements conventional techniques for diagnosis. Further work on bettering these models to expand their application into a wide array of clinical scenarios will be a critical part of this process as machine learning evolves further. Further elaboration of prediction tools with continuous validation and inclusion of more data sources will enhance their reliability and strength in general. Conclusively, the potentials for using support vector machines and other innovative machine learning algorithms are huge and hold great promise for significantly improving patient outcomes and revolutionizing cardiovascular healthcare.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206092 |

REFERENCES

[1] UCI Machine Learning Repository. (n.d.). Heart Disease Data Set. Retrieved from UCI Repository
[2] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
[3] Arghandabi, Hidayatullah & Shams, Parvaneh. (2020). A Comparative Study of Machine Learning Algorithms for the Prediction of Heart Disease. International Journal for Research in Applied Science and Engineering Technology. 8. 677-683. 10.22214/ijraset.2020.32591.
[4] Isreal Ufumaka (2021); Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction; International Journal of Scientific and Research Publications (IJSRP) 11(1) (ISSN: 2250-3153)
[5] Patidar, Sanjay & Kumar, Deepak & Rukwal, Dheeraj. (2022). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. 10.3233/ATDE220723.
[6] Comparative Analysis of Machine Learning Algorithms for Heart Disease PredictionRuby Hasan ITM Web Conf. 40 03007 (2021) DOI: 10.1051/itmconf/20214003007 5.
[7] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
[8] Journal, IJERT. "IJERT-Heart Disease Prediction Using Machine Learning."International Journal of Engineering Research & Technology (IJERT), 2020.
[9] Amanpreet Singh and Narina Thakur" A review of supervised machine learning algorithms" in 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016.
[10] Halima EL HAMDAOUI, Saïd BOUJRAF1, Nour El Houda CHAOUI, Mustapha MAAROUFI, "A Clinical support System for Prediction of Heart Disease using Machine Learning Techniques", 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2-5 Sept. 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details