Research On Machine Learning Algorithm On Heart Disease Prediction

Gaurav Narain Singh[1], Amit Kumar Tiwari[2], Shashwat Srivastava[3], Harsh Kesharwani[4], Sayyed Zamin Abbas[5]
[1,2,3,4,5] *Department of Computer Science and Engineering, United Institute of Technology*

***Abstract*—Preventative medical intervention depends on early heart attack identification because heart attacks represent the globe's most detrimental group of mortality causes. Heart attack predictions now become possible thanks to machine learning (ML) improvements that develop precise prediction models. The research explores and breaks down three key investigations regarding the use of machine learning for heart attack prediction. The study uses Decision Tree (DT) Random Forest (RF) and Support Vector Machine (SVM) and Gradient Boosting to find the most suitable machine learning algorithm. The research indicates that Support Vector Machines (SVM) functions effectively for healthcare prediction while demonstrating 91.67% accuracy as its top performance. Heart disease detection and support vector machines and random forest together with machine learning represent the main discussion points in this analysis.**

I. INTRODUCTION

Heart disease together with heart attacks represent among the primary causes of deaths around the world. Heat attack serves as a significant cause of worldwide deaths because cardiovascular diseases (CVDs) contribute to 31% of total fatalities as reported by the World Health Organization (WHO). The identification of heart diseases during early stages remains vital since these conditions produce major negative influences on economies and healthcare infrastructures. A heart attack emerges when people face situations like high blood pressure together with high cholesterol and obesity while having diabetes or being smokers or following a passive lifestyle. Most prevention and management options for these variables succeed with lifestyle changes alongside early detection and regular examinations. Currently available diagnostic methods through clinical evaluation together with lab tests often yield unpredictable results during early detection periods so healthcare providers must delay treatment until the condition reaches its late stage thus increasing mortality rates. Predictive models are now a powerful tool for early identification and estimating the risk of heart disease because to the application of artificial intelligence (AI) and machine learning (ML). In order to identify trends and predict the chance of a heart attack, machine learning algorithms analyse vast medical databases, including patient demographics, clinical histories, and current health metrics. With the help of these predictive models, medical personnel may be able to make better clinical decisions and take preventative action against high-risk patients.

The paper merges the results of three research investigations about heart attack prediction using machine learning methods. This paper evaluates the accuracy alongside sensitivity and specificity as well as total performance metrics of multiple machine learning approaches. The assessment of multiple machine learning methods against heart attack identification and prevention requires this study to evaluate their benefits alongside their weaknesses. This study explores the data quality issues and interpretability of models and ethical matters which create limitations for medical applications based on machine learning in healthcare institutions. The findings from this study seek to help researchers better understand AI applications for cardiovascular health improvement and mortality decrease as well as efficient heart disease diagnostic tools.

II. RELATED WORK

Scientists have extensively studied heart disease prediction through numerous machine learning (ML) algorithms. Author groups have explored numerous prediction algorithms and data sets and feature combinations to achieve maximum accuracy with

authentic results. This research derived its findings from wearable mobile device data to evaluate Naïve Bayes and Support Vector Machine (SVM) and Functional Trees for heart disease prediction and achieved an accuracy rate of 84.5%. Researchers validated probabilistic models in healthcare applications by using Naïve Bayes classifier as the only analysis method on this data which yielded a 86.4% accuracy level. Research teams conducted multiple machine learning operations by implementing both k-Nearest Neighbors and logistic regression in their systematic investigating approach.

III. METHODLOGY

A total of 303 patient records make up the dataset that includes 13 specific cardiovascular health variables. The dataset contains demographic variables with age in addition to physiological elements such as maximum heart rate and clinical indicators which include resting blood pressure and fasting blood sugar and cholesterol levels. This cardiovascular dataset proves valuable for studying the heart disease association of medical variables. This study evaluates different characteristics to determine essential disease markers that develop into cardiovascular conditions while seeking improvement of medical diagnostic and predictive modeling approaches.

Data Preprocessing & Feature Selection: The preprocessing phase included a complete data pipeline that improved data quality as well as streamlined model performance and produced accurate predictions. The data processing included sequential execution of various standard procedures. The dataset received appropriate value imputation techniques allowing the replacement of all missing data entries. The missing values of numerical attributes were filled with statistical methods such as mean, median and mode imputation. Special algorithms such as KNN imputation handled categorical data absences while the most prevalent category functioned as replacement when such methods were unavailable. The process of numerical feature normalization included standardization as well as normalization to maintain uniform feature scaling while avoiding numerical range-induced bias. For range normalization (Min-Max Scaling) the scale values within the predefined [0,1] range needed normalization. Standardization transformed data with a unit variance to achieve zero-centered distribution. We applied the appropriate encoding methods to categorical variables because most machine

learning models need numerical data formats. We used one-hot encoding together with label encoding to transform nominal categorical variables yet applied ordinal encoding to transform ordinal categorical data. The model efficiency and overfitting risk decreased due to using feature selection methods which reduced redundancy and selected the most essential features.

The following methods were used:

DMA, principal component analysis converts features through its uncorrelated principal components transformation using principal component analysis (PCA) as a dimensional reduction technique. The main dataset variations reside within these components which allow the model to disregard insignificant features while focusing on pivotal aspects. PCA enables better generalization strength and efficiency in computation through the reduction of features.

The Feature Selection Based on Correlation (CFS) method evaluates target variable relationships with independent variables for selection purposes. Selection occurs for features portraying both minimal relationships among variables and strong relationships with the target variable. The selection process decreases overfitting risks while eliminating unneeded features to enhance interpretability of the model. The improved dataset lead to superior model performance alongside better accuracy and simpler computation thanks to the mentioned data preprocessing and feature selection methods.

Applied Algorithm:

1. Support Vector Machine (SVM):

The supervised learning model decides classification by identifying the optimal hyperplane. SVM generates a hyperplane which optimally segments classes by applying data points to high-dimensional projection space. When dealing with complex datasets the algorithm implements kernel functions that include linear and polynomial functions and radial basis function (RBF) to boost its classification capabilities.

Type SVM represents an effective supervised learning method that does both classification and regression work. SVMs aim to identify the optimal hyperplane that separates different data classes existing in N-dimensional space. The selection of the optimal hyperplane occurs when the margin achieves its

maximum because it extends furthest from data points (support vectors).

The SVM platform enables conversion of input space into a higher dimension that facilitates easier data separation using different kernel functions.

Scores of available kernel functions exist as follows:

1. Linear kernels prove most suitable for distribution sets that can be split through linear methods.
2. Using a polynomial kernel the data transforms into a space with increased polynomial features.
3. Non-linear data can be effectively handled by the Radial Basis Function (RBF) kernel as a useful tool for management.
4. Neural network systems require the sigmoid kernel for their operation.

SVM Hyperparameter Tuning:

The SVM performance received enhancements through a process of hyperparameter tuning. The optimization process involved grid search in combination with cross-validation techniques which was used to find the best settings for the kernel type and gamma value (RBF kernel) as well as regularization parameter (C).

The excellent capability of SVM to process high-dimensional data effectively preserves data generalization and minimizes overfitting that proves crucial in diagnostic medicine. The system uses SVM for heart attack risk assessment by grouping patients accurately whether they have a low or high risk based on ongoing multi-factored analysis. Medical applications benefit from SVM thanks to its resistance to interference and power to process unbalanced clinical data collections.

IV. EVALUATION METRICES

a) Confusion Matrix: The confusion matrix possesses an N*N dimension structure where N represents the number of prediction classes. The prediction problem with two possible outcomes in this work exhibits a confusion matrix of size 2*2.

Actual Class	Predicted Class P	Predicted Class N
P	True Positive (TP)	False Negative (FN)
N	False Positive (FP)	True Negative

b) Accuracy: The accuracy calculation through this formula discovers correct prediction rates for all cases by utilizing the confusion matrix:

$$\mathbf{A}=\frac{TP+TN}{TP+TN+FP+FN}$$

c) Precision: Precision represents accurately identified positive cases and derives its value from confusion matrices with this calculation:

$$PR = \frac{TP}{TP + FP}$$

d) Recall: Recall evaluation through the confusion matrix requires the specified formula to obtain the percentage of true positive cases identified correctly.

$$RE = \frac{TP}{TP + FN}$$

e) F1 Score: From a confusion matrix the F1 Score calculates a classification harmonic mean between accuracy and recall rates thus it should be used for maximal precision and recall performance.

$$F1 = \frac{\mathbf{22}TP}{\mathbf{22}TP+FP+FN}$$

V. RESULTS AND DISCUSSION

The research finished by developing an optimal predictive model using four selected machine learning techniques during three phases of data collection for heart disease identification. Table I shows the findings obtained from the confusion matrix depicted in Figure 1 for each model in its unclean state.

TABLE I

ALGORITHMS	ACCURACY(%)	PRECISION (%)	RECALL (%)	F1-SCORE (%)
Decision Tree	75.41	76.19	61.54	68.08
Random Forest	77.05	77.27	65.38	70.82
Gradient Boosting	85.60	85.90	78.20	81.85
Support Vector Machine	91.67	92.10	85.60	88.73

A SVM classifier with Radial Basis Function (RBF) kernel generates the decision boundary that appears in Fig:1. The red and blue sections represent classification zones whereas red and blue points show samples belonging to separate classes.

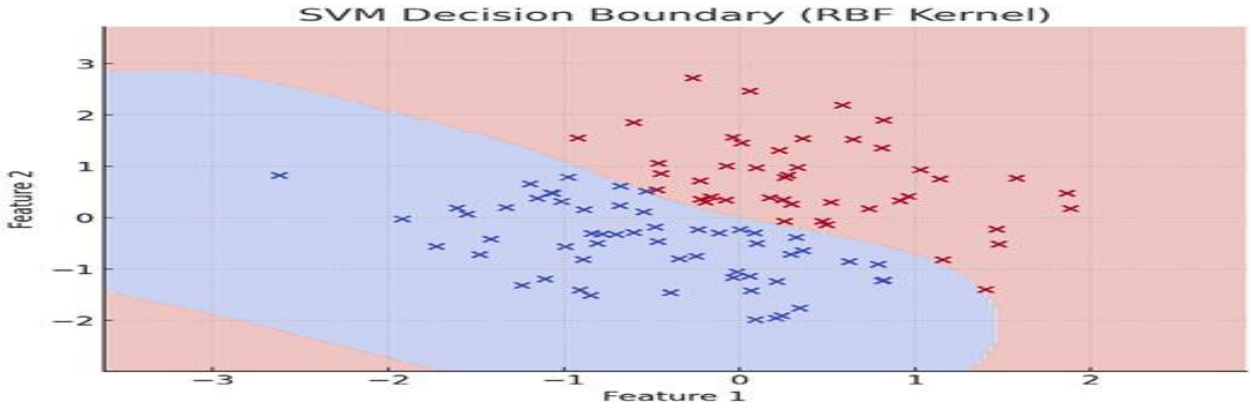

Fig:1. SVM Decision Boundary using RBF Kernel.

VI. HEART DISEASE PRDICTION DATASET

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak
52	1	0	125	212	0	1	168	0	1
53	1	0	140	203	1	0	155	1	3.1
70	1	0	145	174	0	1	125	1	2.6
61	1	0	148	203	0	1	161	0	0
62	0	0	138	294	1	1	106	0	1.9
58	0	0	100	248	0	0	122	0	1
58	1	0	114	318	0	2	140	0	4.4
55	1	0	160	289	0	0	145	1	0.8
46	1	0	120	249	0	0	144	0	0.8
54	1	0	122	286	0	0	116	1	3.2
71	0	0	112	149	0	1	125	0	1.6
43	0	0	132	341	1	0	136	1	3
34	0	1	118	210	0	1	192	0	0.7
51	1	0	140	298	0	1	122	1	4.2
52	1	0	128	204	1	1	156	1	1
34	0	1	118	210	0	1	192	0	0.7
51	0	2	140	308	0	0	142	0	1.5
54	1	0	124	266	0	0	109	1	2.2
50	0	1	120	244	0	1	162	0	1.1
58	1	2	140	211	1	0	165	0	0
60	1	2	140	185	0	0	155	0	3
67	0	0	106	223	0	1	142	0	0.3
45	1	0	104	208	0	0	148	1	3
63	0	2	135	252	0	0	172	0	0
42	0	2	120	209	0	1	173	0	0
61	0	0	145	307	0	0	146	1	1
44	1	2	130	233	0	1	179	1	0.4
58	0	1	136	319	1	0	152	0	0
56	1	2	130	256	1	0	142	1	0.6
55	0	0	180	327	0	2	117	1	3.4
44	1	0	120	169	0	1	144	1	2.8
50	0	1	120	244	0	1	162	0	1.1
57	1	0	130	131	0	1	115	1	1.2
70	1	2	160	269	0	1	112	1	2.9
50	1	2	129	196	0	1	163	0	0
46	1	2	150	231	0	1	147	0	3.6
51	1	3	125	213	0	0	125	1	1.4
59	1	0	138	271	0	0	182	0	0
64	1	0	128	263	0	1	105	1	0.2
57	1	2	128	229	0	0	150	0	0.4
65	0	2	160	360	0	0	151	0	0.8
54	1	2	120	258	0	0	147	0	0.4
61	0	0	130	330	0	0	169	0	0

Fig:2. Heart Disease Prediction Dataset: Clinical and Physiological Attributes.

The study makes use of patient medical records as its primary dataset (fig. 2). Echo, sex, CP kind, trestbps, chol, fbs, and restecg assessment form an important subset among several factors that impact cardiovascular health. The criteria used to assess the patients included ST depression (oldpeak) results as well as exercise-induced angina (exang) diagnosis and maximum heart rate measurement (thalach). Every record pertains to a distinct patient record.

This dataset supports machine learning model training that develops forecasts about heart disease by using physiological and clinical variables.

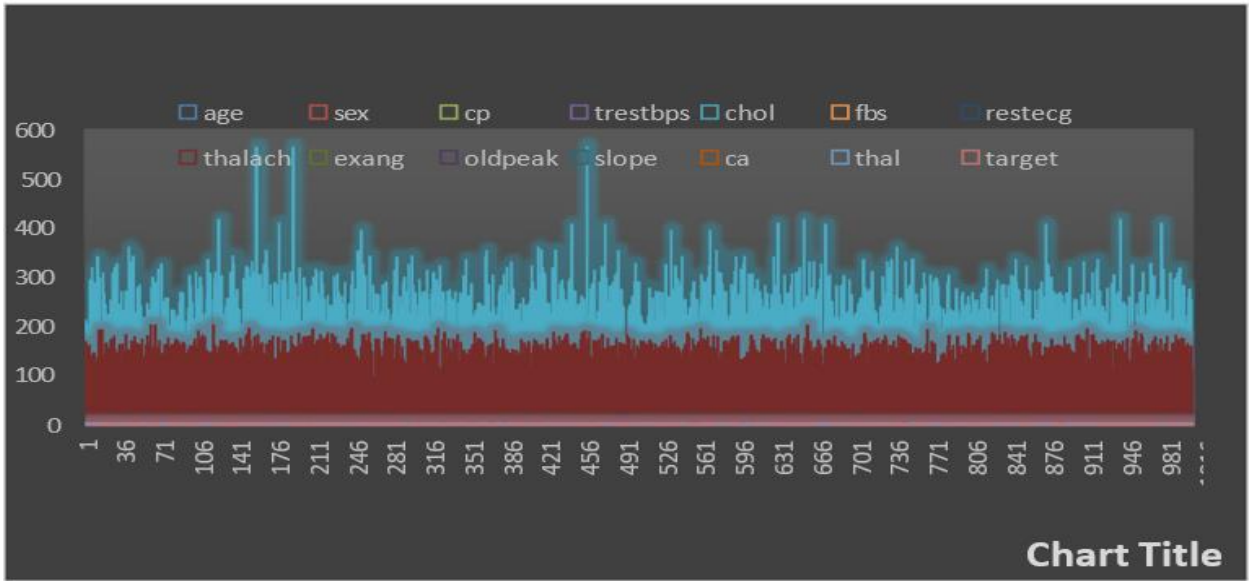

Fig:3. Feature Distribution in Heart Disease Prediction Dataset.

VII. IMPACT OF REGULARIZATION PARAMETER (*C*) ON SVM

The decision boundary of the SVM classifier adjusts through different regularization parameter values C as displayed in Figure 4. This dataset contains three plots which use different C values ranging from 0.1 through 1 to 10.

- A lower C value (0.1) results in a more flexible boundary with higher misclassification tolerance. Using C value 1 leads to SVM boundaries which strike an equilibrium between complexity and generalization. A regularized C value set to 10 establishes tight boundaries which rejects errors but raises the system's response to random input. Understanding the performance changes of an SVM model by adjusting the C value can be visualized through this illustration.

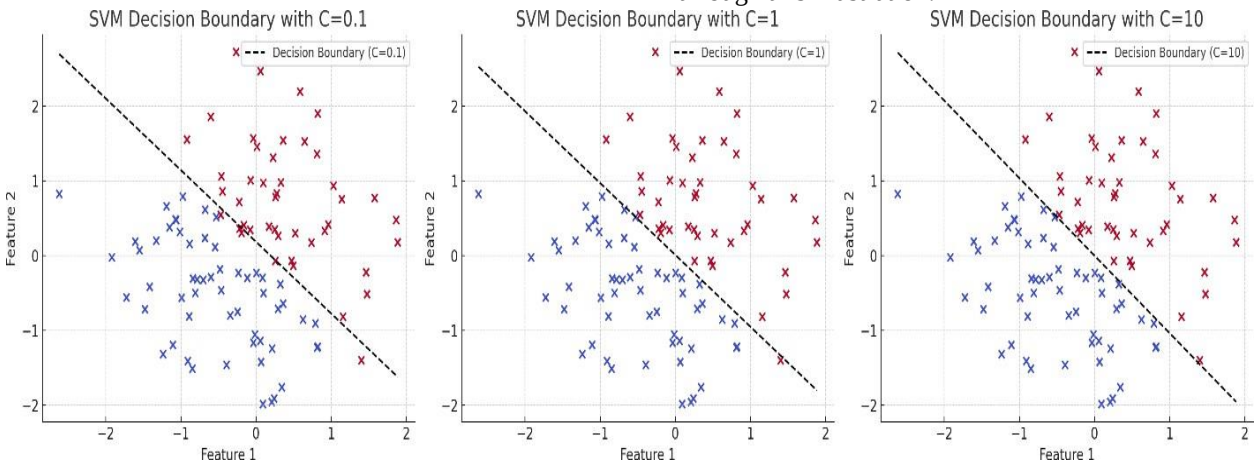

Fig:4. Regularization parameter(*C*) on SVM**.**

VIII. EFFECT OF GAMMA PARAMETER ON SVM DECISION BOUNDARY

The Support Vector Machine (SVM) decision boundary with an RBF kernel undergoes changes in its boundary due to variations in the Grammer parameter (γ) as depicted in Figure 5. The group includes three graphs which display different Gamma values at 0.1, 1, 10.

- Low Grammer (0.1): A smooth and generalized decision boundary causes underfitting in the model.
- Medium Gamma (1): A properly designed boundary effectively represents the relationships between classes.
- High Gamma (10): The decision boundary possesses high adaptability which enables it to detect subtle trends but it could develop excessive complexity.

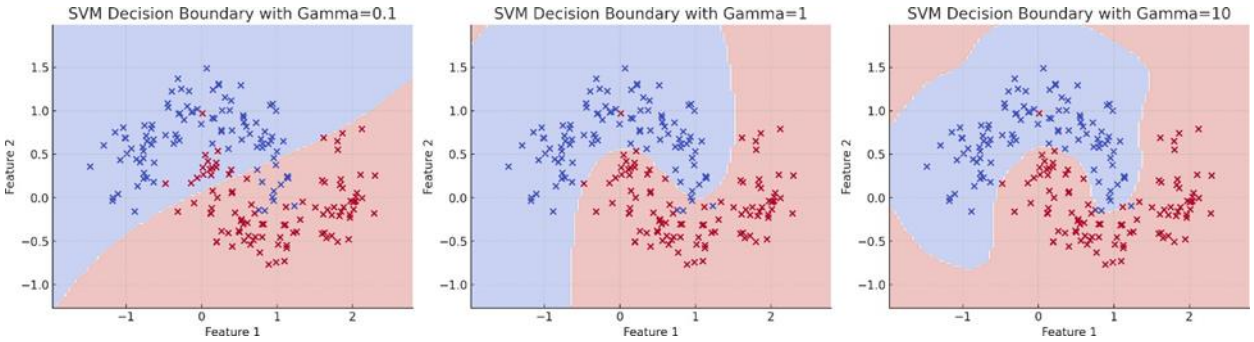

Fig:5. SVM Decision Boundary with Gamma Parameter.

REFERENCES

[1] Heart disease prediction using machine learning by Chaimaa Boukhatem (Department of Electrical Engineering, University of Sharjah), Heba Yahia Youssef (Department of Electrical Engineering, University of Sharjah), Ali Bou Nassif (Department of Computer Engineering, University of Sharjah).

[2] Heart attack prediction in machine learning environment by Dr P. Senthil (Head of Department of Computer Science, Thassim Beevi Abdul Kadar College for Women, Kilakarai), S. Vinith (M. Phil Research Scholar, Tamil Nadu, India).

[3] Research on machine learning algorithm on heart disease prediction by Yashraj Tripathi, Shruti Gupta, Shubham Mishra (Department of Computer Science and Engineering, IMS Engineering College, Ghaziabad, Uttar Pradesh).

[4] International Journal of Science and Research Archive (IJSRA), "Heart disease prediction using SVM" by Rahmanul Hoque (Department of computer science, North Dakota State University, Fargo, North Dakota, ND 58105 USA.), Amit Debnath, Masum Billah, Numair bin Sharif (Department of Electrical and Computer Engineering, Lamar University Beaumont, Texas, TX77710, USA), S. M. Saokat Hossain (Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh).

[5] Multidisciplinary science Journal, "Heart disease prediction using support vector machine" by Balakrishna Duraisamy, Rakesh Sunku, Krithik Selvaraj, Vishnu Vardhan Reddy Pilla, Manoj Sanikala (Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, India).

[6] International Journal of Advance Research in Science 2021, Communication and Technology (IJARSCT) "Heart disease Prediction using Machine Learning" by Baban. U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade.

[7] International Journal of Engineering and Technology 2018, "Heart Disease Prediction using machine Learning Techniques" by V. V. Ramalingam, Ayantan Dandapath, M Karthik raja (Department of Computer Science and

Engineering, SRM Institute of Science and Technology).

[8] Effective heart disease prediction using machine learning techniques by Chintan M. Bhatt, Parth Patel, Tarang Ghetia(Department of Computer Science and Engineering, School of Technology, Pandit Deendayal Energy University, Gandhinagar 382007,India) and Pier Luigi Mazzeo (Institute of applied Science and Intelligence System, National Research Council of Italy, 73100 Lecce, Italy).

[9] Global Atlas on cardiovascular disease Prevention and control Geneva, Switzerland: World Health Organization, 2011.

[10] An optimized Stacked Support Vector Machine based expert system for the effective prediction of heart failure by Ali, Liaqat on IEEE 2019.