



Evaluation Of Ensemble Based Hptrf ML Algorithm With Multiple Cardio Data Set

D.R. Krithika, Research Scholar, Department of Computer Science, Vels Institute of Science Technology and Science Technology and Advanced Studies
Chennai, India, krithikabanu@gmail.com.

Dr.K.Rohini Associate Professor, Department of Vels Institute of Advanced studies,
Chennai, India. rrohini16@gmail.com.

ABSTRACT:

Cardiovascular disease is leading cause of death. Athero means soft porridge like, and sclerosis is hardening so Atherosclerosis is fatty deposition of artery valves and stiffening of blood vessel valve is sclerosis. It affects our body large and medium arteries and chronic inflammation is caused and immune system activation of artery valve. lipids deposits in artery valves is the development of plaques. Three different problems of plaques are the first one is tightening of artery valve which leads to raise of BP , if BP will raised heart gets difficult to pump blood. Second problem is stenosis is narrowing because of the plaques of the space at blood can flow through in blood vessel and this which causes Blood flow in conditions like Angina and peripheral vascular disease. Another problem is rupture of plaque gets off thrombus leads to ischemia. It causes acute coronary syndrome. Proposed Algorithm (HPTRF) got better Accuracy Results.

Keywords: CVD - cardiovascular disease, Atherosclerosis, hypertension, blood pressure, ischemia, HPTRF- hyper parameter tuned random forest.

I. INTRODUCTION:

The process of huge volume of unstructured data into useful form of structured data is bigdata analytics. In this bigdata analytics five concepts are important thing, and the concepts are volume, variety, velocity, value, veracity.

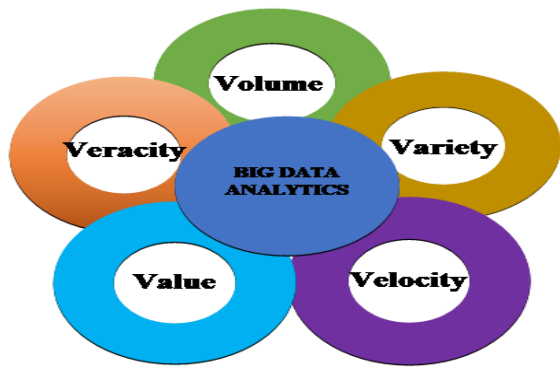


Fig 1 Five V's of Bigdata Analytics

Volume is huge amount of data set. Variety is multiple sources of file format like audio, video, jpg etc. The variety of file formats are structured, semi-structured, unstructured. Structured data is tabular format, semi-structured data is xml, csv, json format files. Unstructured is like audio, video, png, email and log files. Speed of all variety of data is velocity. The Value is extracted essential data. Veracity is Dumping more volume of data, some data packets lose in the process.

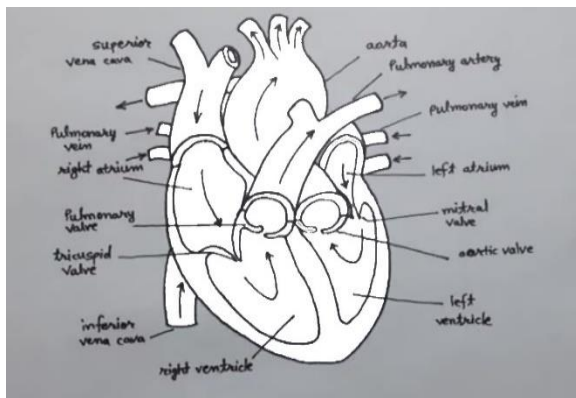


Fig 2 Human Heart

The Heart is very important organ, it has 4 chambers. The right and left atrium are called upper chambers. right ventricle and left ventricle are called lower chambers. de-oxygenated blood carry through Inferior and superior vena cava to the right-side atrium. The open tricuspid valve through flow of Blood from the right-side atrium into the right-side ventricle. The tricuspid valve shuts when the ventricles are full. The pulmonary valve through Blood leaves the right-side ventricle into the pulmonary artery for oxygenation. The oxygenated blood carry pulmonary veins to the left side atrium. The left atrium into the left ventricle Blood flows the open mitral valve. The mitral valve shuts when the ventricles are full. The Blood flows left ventricle through the aortic valve into the aorta, and this is how the heart functions continuously. Risk factors of developing Atherosclerosis is Modifiable and Non-

Modifiable. family history and older age are non-Modifiable. Alcohol and smoking are Modifiable. The certain comorbidities increased atherosclerosis risk and managed carefully to minimize the risk of type1 and type2 Diabetes, high BP, inflammation.

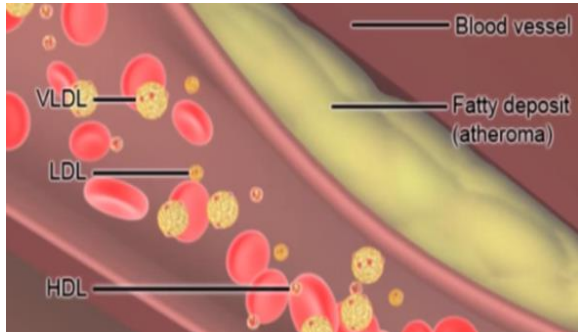


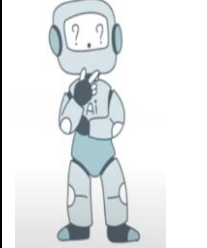


Fig 3 Blood Vessel

CVD Prevention is one of two main categories firstly primary prevention that prevent CVD in patients that have never heart disease and secondary prevention this is when after Angina, Heart attack and TIA trying to prevent further CVD. So we optimize Modifiable risk factors try to improve our health. Primary prevention is checking lipids every 3 month and aim for 40% reduction in Non-HDL Cholesterol.

 <p>Supervised Learning</p>	 <p>Unsupervised Learning</p>	 <p>Reinforcement Learning</p>
<p>Supervised Learning is Machine Learn under Guidance like Teacher guiding students. labeled data and explicitly</p>	<p>Unsupervised Learning is Machine Learn without Guidance. Here the data is not labeled and hidden part of the prediction of</p>	<p>Reinforcement Learning is experience learn from experience. It's a learning method agent interact environment while</p>

telling input, and output must look. The data in this case training data.	output. So no need of guidance to help activity.	producing some actions it discovers errors or rewards.
---	--	--

Table 1 ML Algorithm

II. LITERATURE REVIEW:

The paper discussed about blood viscosity-based heart disease prediction [1]. The paper discusses about Takayasu arteries and its significant challenges [2]. In this paper discussed about cardiovascular model for research hotspots [3]. In this paper proposed about automated diagnosis of cardiovascular disease inpatient and outpatient using KNN [4]. In this paper SVM and Gaussian kernel together got high accuracy of processing ECG signal [5]. The paper discussed about I MoT – smart phone-based CVD management of Blood lipid data acquisition [6]. In this paper proposed about hybrid machine learning techniques to predict CVD [7]. The paper discussed about LSTM (long short-term memory) based CVD prediction [8]. In this paper discussed with CVD understanding mechanism and methodologies focused [9]. In this paper discusses about inner wall thickness of carotid artery [10]. The paper proposed about cardio prediction using artificial neural network [11]. In this paper collected world covid data and future prediction of cardio death rate and covid confirmed cases using logistic, polynomial and SVM algorithms [12]. This paper proposed tuned CNN model pascal dataset [13]. In this paper using supervised combined with forward reasoning algorithm to diagnosis of cardio disease [14].

III. RESULTS AND DISCUSSION:

Cardio online and real 96655 data collected and applied few Algorithms. My proposed algorithm is HPTRF. So, I applied these data to HPTRF. Attributes of 96655 data's are age, gender, height, weight, systolic, diastolic, cholesterol, glucose, smoke, alcohol, cardio, bmi, bmihigh, pulse pressure.

age	gender	height	weight	systolic	diastolic	cholesterol	gluc	smoke	alco	cardio	bmi	bmi_high	pulse pressure
50	1	168	62.0	110	80	1	1	0	0	0	21.97	0	30
55	0	156	85.0	140	90	3	1	0	0	1	34.93	1	50
52	0	165	64.0	130	70	3	1	0	0	1	23.51	0	60
48	1	160	82.0	150	100	1	1	0	0	1	28.71	0	50
48	0	156	56.0	100	60	1	1	0	0	0	23.01	0	40

Fig 4 96655 cardio data

age	gender	height	weight	systolic	diastolic	cholesterol	gluc	smoke	alco	cardio	bmi	bmi_high	pulse pressure
Less than 18	1	188	103.0	120	90	3	3	1	1	1	38.77	1	58
19-24	1	182	105.0	160	112	3	3	1	1	1	44.00	1	70
25-34	1	186	129.0	190	135	3	3	1	1	1	47.86	1	80
35-44	1	198	200.0	14020	8500	3	3	1	1	1	278.12	1	13840
45-54	1	1576	200.0	16020	10000	3	6	1	1	1	298.67	1	15940
55-64	1	1576	732.0	14020	11000	3	6	1	1	1	2397.96	1	13940
65-74	1	180	114.0	220	130	3	3	1	1	1	51.84	1	150
75-84	1	180	103.0	231	108	3	3	1	1	1	43.26	1	170
85+	1	164	87.0	170	90	3	3	1	0	1	35.56	1	92

Fig 5 Max Value details

age	gender	height	weight	systolic	diastolic	cholesterol	gluc	smoke	alco	cardio	bmi	bmi_high	pulse pressure
Less than 18	0	70	8.0	88	48	1	1	0	0	0	13.88	0	20
19-24	0	140	35.0	80	50	1	1	0	0	0	16.20	0	20
25-34	0	139	36.0	80	50	1	1	0	0	0	14.98	0	10
35-44	0	67	28.0	-120	0	1	1	0	0	0	13.51	0	-9300
45-54	0	59	10.0	-140	0	1	1	0	0	0	0.37	0	-9850
55-64	0	14	11.0	-150	-70	1	1	0	0	0	0.26	0	-10800
65-74	0	134	31.0	14	40	1	1	0	0	0	15.43	0	-66
75-84	0	114	31.0	90	40	1	1	0	0	0	13.96	0	20
85+	0	138	43.0	90	50	1	1	0	0	0	16.73	0	30

Fig 6 min Value Details

age	gender	height	weight	systolic	diastolic	cholesterol	gluc	smoke	alco	cardio	bmi	bmi_high	pulse pressure
Less than 18	0.454545	142.090909	45.909091	104.909091	69.545455	1.454545	1.545455	0.181818	0.090909	0.181818	20.473636	0.090909	35.363636
19-24	0.431034	161.344028	64.396552	111.224138	74.756921	2.000000	1.413793	0.137931	0.172414	0.224138	24.685862	0.126960	36.465517
25-34	0.547278	162.408877	73.446991	118.541547	76.730659	1.845272	1.421203	0.063095	0.120344	0.217785	27.788080	0.320917	41.810888
35-44	0.390369	165.162756	72.624705	122.869697	89.984187	1.243610	1.161395	0.110069	0.067714	0.300128	26.685642	0.210404	32.905510
45-54	0.353304	164.477617	73.894255	128.964703	94.798981	1.336249	1.215832	0.090255	0.058106	0.449493	27.440558	0.251129	34.164822
55-64	0.351945	163.532114	74.700950	131.648152	98.695328	1.489125	1.299651	0.078272	0.051106	0.598852	28.100779	0.296881	32.952823
65-74	0.628423	157.700186	66.377095	132.862197	77.047486	1.764432	1.702048	0.119161	0.121974	0.288156	26.711593	0.214153	55.614711
75-84	0.721713	156.669725	64.403670	136.422018	74.614679	1.758410	1.669725	0.067278	0.082569	0.284404	26.317492	0.180428	61.807339
85+	0.541667	151.075000	58.833333	127.416667	70.625000	1.791667	1.625000	0.041667	0.000000	0.333333	25.509593	0.166667	56.791667

Fig 7 Mean Value

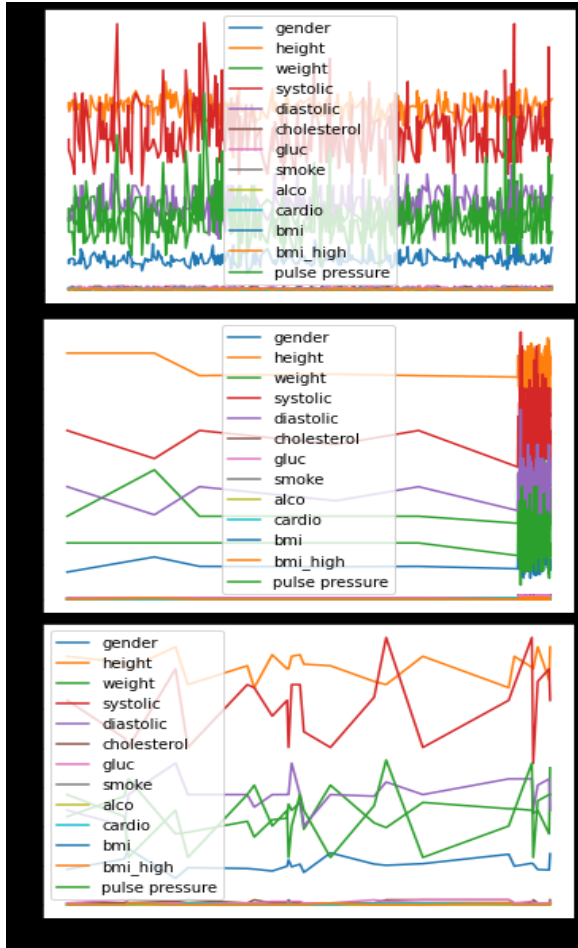


Fig 8 Graphical representation of Data

```
confusion matrix
[[8204 1746]
 [2037 7344]]
```

Train Accuracy of HPTRF: 96.54

Test Accuracy of HPTRF: 80.43039677202421

	precision	recall	f1-score	support
0	0.80	0.82	0.81	9950
1	0.81	0.78	0.80	9381
accuracy			0.80	19331
macro avg	0.80	0.80	0.80	19331
weighted avg	0.80	0.80	0.80	19331

Fig 9 HPTRF- 96655 Data Result

SNO	MODEL	TRAIN ACCURACY	TEST ACCURACY
0	Logistic Regression	70.46	70.011898
1	Naive Bayes	58.43	58.672598
2	Random Forest	73.38	72.143190
3	Extreme Gradient Boost	73.59	71.781077
4	K-Nearest Neighbor	72.99	68.118566
5	Decision Tree	72.07	71.465522
6	Support Vector Machine	72.58	71.460349
7	ANN	71.66	70.930000
8	StackingCVClassifier	74.09	72.308727
9	HPTRF	96.54	80.430397

Fig 10. 96655 data Comparison of Algorithm Results

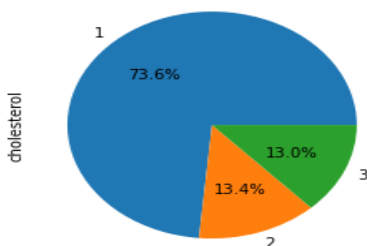


Fig 11 Cholesterol Level

The blood contains large number of substances and various specific test were developed to assess its composition one such test is the hematocrit which measures the percentage of erythrocytes commonly known as red blood cells. Hematocrit normal ranging for females 36% to 48 % and for males hematocrit ranging 41% to 50%. If the hematocrit value is very low, it indicates anemia and the high hematocrit values including dehydration low oxygen availability as a result very high hematocrit value can indicate the presence of health condition such as pulmonary fibrosis congenital heart disease or polycythemia vera. Very

high hematocrit levels affect the blood viscosity significantly thickening and slowing its rate of blood flow around the body this impaired flow can increase the risk of blood clotting.

Age	Gender	Familyhist	Smoke	Alco	Height	Weight	BMI	SpO2	Systolic	Diastolic	HeartRate	ECG	ECHO	TMT	HBA1C	HCT	TCHOL	LDL	cardio
56	1	0	0	0	167	62.5	22.4	99	149	86	116	0	0	0	6.9	48.5	274	176	0
65	1	0	0	0	156	89.0	36.6	96	133	66	84	0	0	0	7.5	44.8	127	82	0
46	1	0	0	0	167	67.9	24.3	99	126	80	104	0	0	0	5.8	42.9	197	126	0
45	1	0	0	0	159	73.3	29.0	99	119	83	74	1	0	1	10.6	52.0	200	180	1
63	1	0	0	0	168	71.5	25.3	99	110	60	80	0	0	1	6.9	50.3	115	63	1

Fig 12. 100 master Health checkup sample data

```

Train Accuracy of HPTRF- Hyper Parameter Tunned Random Forest Classifier: 100.0
Test Accuracy of HPTRF- Hyper Parameter Tunned Random Forest Classifier: 100.0

      precision    recall  f1-score   support

0         1.00      1.00      1.00         20

 accuracy          1.00      1.00      1.00         20
 macro avg         1.00      1.00      1.00         20
weighted avg         1.00      1.00      1.00         20

Number of people predicted with and without heart-disease: 0  95
                                                            1   5
    
```

Fig 13. 100 Data Algorithm Result

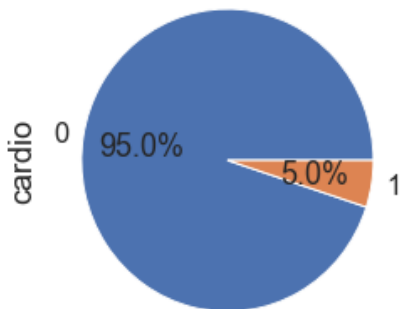


Fig 14. cardio patient pie chart

```

blood viscosity above limit cardio Patient False 97
blood viscosity above limit cardio Patient True  3
    
```

Fig 15 HCT above limit cardio patient Result

Age, Sex, Resting BP, Cholesterol, Fasting Blood sugar, Exercise, Heart Disease are Attributes of 919 data.

	Age	Sex	RestingBP	Cholesterol	FastingBS	ExerciseA0gi0a	HeartDisease
0	40	1	140	289	0	0	0
1	49	0	160	180	0	0	1
2	37	1	130	283	0	0	0
3	48	0	138	214	0	1	1
4	54	1	150	195	0	0	0

Fig 16 online 919 cardio data set

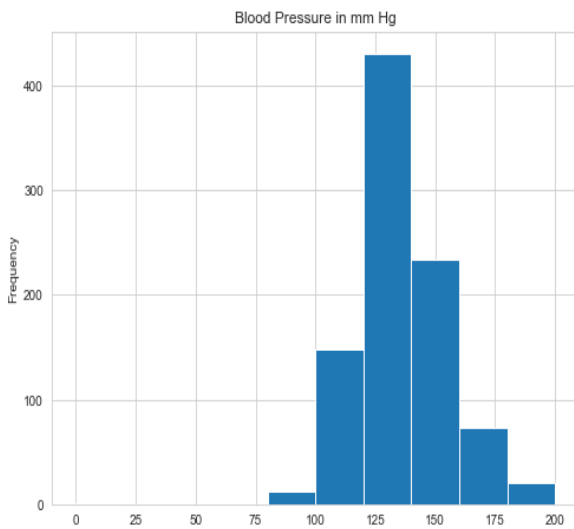


Fig 17 Blood Pressure

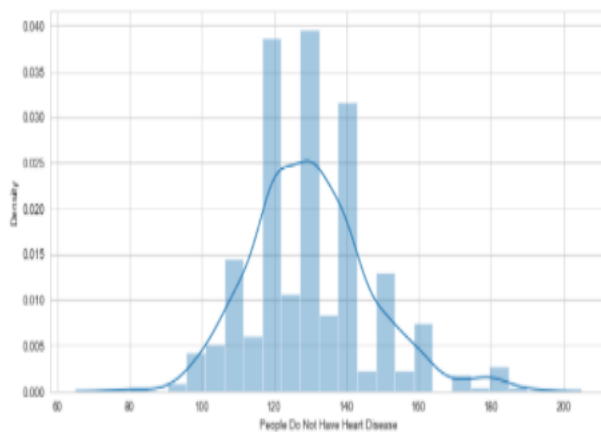


Fig 18 People do not have heart disease

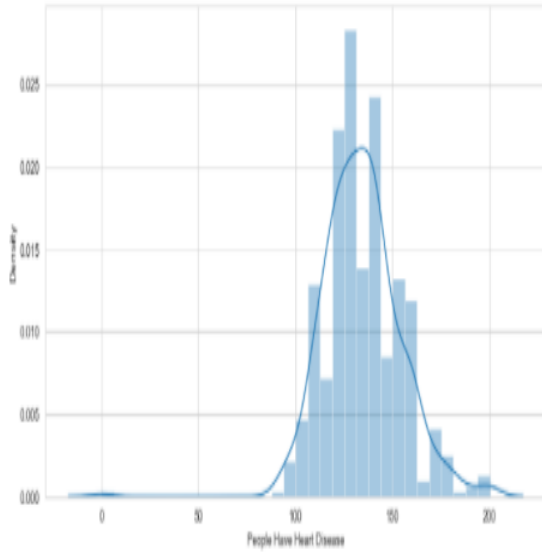


Fig 19 People have heart disease

```
confusion matrix
[[61 16]
 [17 90]]
```

Train Accuracy of HPTRF- Hyper Parameter Tunned Random Forest Classifier: 99.86

Test Accuracy of HPTRF- Hyper Parameter Tunned Random Forest Classifier: 82.06521739130434

	precision	recall	f1-score	support
0	0.78	0.79	0.79	77
1	0.85	0.84	0.85	107
accuracy			0.82	184
macro avg	0.82	0.82	0.82	184
weighted avg	0.82	0.82	0.82	184

Fig 20 online 919 cardio data set result

	Model	Train Accuracy	Test Accuracy
0	Logistic Regression	78.20	78.260870
1	Naive Bayes	78.34	78.260870
2	K-Nearest Neighbour	80.38	80.434783
3	Decision Tree	82.29	79.347826
4	Support Vector Machine	81.20	81.521739
5	HPTRF	99.86	82.065217

Fig 21 Algorithm comparison of 919 data Result

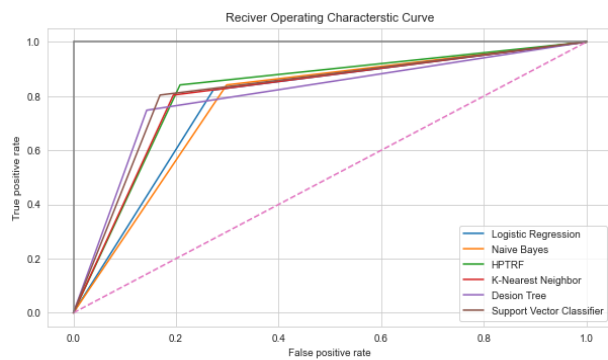


Fig 22 ROC Curve

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig 23 UCI data 303 rows

	Model	Train Accuracy	Test Accuracy
0	Logistic Regression	82.23	86.585246
1	Naive Bayes	83.47	78.688525
2	K-Nearest Neighbour	84.30	81.967213
3	Decision Tree	92.15	80.327869
4	Support Vector Machine	94.21	90.163934
5	HPTRF	100.00	86.885246

Fig 24 Applied Algorithm Results of 303 Data

V. CONCLUSION

Bigdata analytics I implemented huge volume of data more than one lakh data. After preprocessing I got 96655 only. In this paper 96655 cardio data applied to my proposed algorithm HPTRF (hyper parameter tuned random forest) and few machine learning Algorithm. I got better accuracy result in HPTRF comparing to other algorithms. I checked HPTRF using less amount of data like 100 master checkup data and 919 online cardio dataset and base paper dataset applied to HPTRF Algorithm, and I got better accuracy result. So my proposed algorithm is very useful to clinical people. Hematocrit is important role plays to find heart disease. In future prediction based upon some other diseases.

REFERENCES:

1. R. Latha and P. Vetrivelan, "Blood Viscosity based Heart Disease Risk Prediction Model in Edge/Fog Computing," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), 2019, pp. 833-837, doi: 10.1109/COMSNETS.2019.8711358.
2. C. Zhao et al., "Role of Contrast-Enhanced Ultrasound Sonography in the Medical Diagnostics of the Disease Activity in Patients With Takayasu Arteritis," in IEEE Access, vol. 7, pp. 23240-23248, 2019, doi: 10.1109/ACCESS.2019.2896386.
3. J. Hou et al., "Research Hotspots Analysis of Cardiovascular Model by PubMed," 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 931-934, doi: 10.1109/ITME.2018.00207.
4. K. Karboub, M. Tabaa, F. Monteiro, S. Dellagi, F. Moutaouakkil and A. Dandache, "Automated Diagnosis System for Outpatients and Inpatients With Cardiovascular Diseases," in IEEE Sensors Journal, vol. 21, no. 2, pp. 1935-1946, 15 Jan.15, 2021, doi: 10.1109/JSEN.2020.3019668.
5. N. V. Khandait and A. A. Shirolkar, "ECG signal processing using classifier to analyses cardiovascular disease," 2019 3rd International Conference on Computing

- Methodologies and Communication (ICCMC), 2019, pp. 855-859, doi: 10.1109/ICCMC.2019.8819777.
6. X. Huang et al., "Smartphone-Based Blood Lipid Data Acquisition for Cardiovascular Disease Management in Internet of Medical Things," in *IEEE Access*, vol. 7, pp. 75276-75283, 2019, doi: 10.1109/ACCESS.2019.2922059.
 7. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
 8. M. S. Islam, H. Muhamed Umran, S. M. Umran and M. Karim, "Intelligent Healthcare Platform: Cardiovascular Disease Risk Factors Prediction Using Attention Module Based LSTM," 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2019, pp. 167-175, doi: 10.1109/ICAIBD.2019.8836998.
 9. L. Athanasiou, F. R. Nezami and E. R. Edelman, "Computational Cardiology," in *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 4-11, Jan. 2019, doi: 10.1109/JBHI.2018.2877044.
 10. R. Gupta, R. Pachauri and A. K. Singh, "Despeckle and Segmentation of Carotid Artery for Measurement of Intima-media Thickness," 2019 International Conference on Signal Processing and Communication (ICSC), 2019, pp. 345-348, doi: 10.1109/ICSC45622.2019.8938322.
 11. G. Suseendran, N. Zaman, M. Thyagaraj and R. K. Bathla, "Heart Disease Prediction and Analysis using PCO, LBP and Neural Networks," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019, pp. 457-460, doi: 10.1109/ICCIKE47802.2019.9004357.
 12. D. R. Krithika and K. Rohini, "Comparative Interpretation Of Machine Learning Algorithms In Predicting The Cardiovascular Death Rate For Covid-19 Data," 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2021, pp. 394-400, doi: 10.1109/ICCIKE51210.2021.9410777.
 13. M. Boulares, T. Alafif and A. Barnawi, "Transfer Learning Benchmark for Cardiovascular Disease Recognition,"
 14. in *IEEE Access*, vol. 8, pp. 109475-109491, 2020, doi: 10.1109/ACCESS.2020.3002151. *Computer Science (NICS)*, 2019, pp. 246-250, doi: 10.1109/NICS48868.2019.9023903.
 15. H. Le Ngoc, H. T. Cong, S. D. Phuoc and K. B. Dan, "Developing the New Proposal of Intelligence System in Heart and Cardiovascular Diagnosis," 2019 6th NAFOSTED Conference on Information and