Journal of Informatics Engineering and Software Applications

Classification of Heart Disease Risk using the Support Vector Machine

Suci Wulan Dari[1]*, Hendrawaty[2], Azhar[3]

[1,2,3] Jurusan Teknologi Informasi dan Komputer Politeknik Negeri Lhokseumawe, Jln. B.Aceh Medan Km.280 Buketrata 24301 INDONESIA
*Penulis Korespondensi : suciwd3@gmail.com

INFORMASI ARTIKEL

Riwayat artikel:
Diajukan pada 20 Mei 25
Direvisi pada 02 Juni 25
Publikasi pada 20 Juni 25

Kata kunci:
Penyakit Jantung
Klasifikasi
Support Vector Machine
SMOTE

Keywords:
Heart Disease
Classification
Support Vector Machine
SMOTE

ABSTRAK

Penyakit jantung merupakan salah satu penyebab kematian utama secara global, mempengaruhi jutaan orang setiap tahunnya. Pengembangan teknologi informasi telah memfasilitasi kemajuan signifikan dalam analisis data kesehatan melalui *machine learning* dan data *mining*. Salah satu teknik yang digunakan, *Support Vector Machine* (SVM), memungkinkan klasifikasi risiko penyakit jantung dengan akurat berdasarkan data klinis seperti usia, jenis kelamin, tekanan darah, dan faktor lainnya. Penelitian ini bertujuan untuk mengoptimalkan model SVM untuk klasifikasi risiko penyakit jantung, dengan hasil menunjukkan tingkat keberhasilan mencapai 90% dari 50 percobaan. Metode SVM bekerja dengan mencari *hyperplane* terbaik yang memaksimalkan margin antara kelas-kelas data, bahkan dalam ruang dimensi yang lebih tinggi melalui penggunaan *kernel*. Hasil penelitian menunjukkan bahwa model SVM dengan parameter optimal (C=10, γ=1) dan teknik SMOTE menghasilkan akurasi 92%, presisi 89%, *recall* 95%, dan f1-score 92%. Kesimpulan dari penelitian ini menegaskan bahwa SVM efektif dalam mengelompokkan data klinis untuk klasifikasi risiko penyakit jantung, meskipun tantangan tetap ada dalam mengenali beberapa kasus risiko yang lebih kompleks. Penelitian ini dapat memberikan landasan untuk pengembangan sistem klasifikasi yang lebih baik.

ABSTRACT

Heart disease is one of the leading causes of death globally, affecting millions of people every year. The development of information technology has facilitated significant advances in health data analysis through machine learning and data mining. One of the techniques used, the Support Vector Machine (SVM), allows for accurate classification of heart disease risk based on clinical data such as age, gender, blood pressure, and other factors. The study aimed to optimize the SVM model for heart disease risk classification, with results showing a success rate of 90% from 50 trials. The SVM method works by finding the best hyperplane that maximizes the margin between data classes, even in higher dimensional spaces through the use of the kernel. The results show that the SVM model with optimal parameters (C=10, γ=1) and SMOTE technique produces 92% accuracy, 89% precision, 95% recall, and 92% f1-score. The conclusions of this study confirm that SVM is effective in grouping clinical data for heart disease risk classification, although challenges remain in recognizing some of the more complex risk cases. This research can provide a foundation for the development of a better classification system.

1. Pendahuluan

Jantung merupakan organ vital dalam sistem tubuh manusia yang berperan sentral dalam menjaga peredaran darah yang penting untuk kelangsungan hidup. Fungsi utama jantung adalah memompa darah yang mengandung oksigen dan nutrisi ke seluruh organ dan jaringan dalam tubuh. Penyakit jantung, yang dikenal sebagai gangguan atau kegagalan pada jantung, dapat mengganggu sirkulasi darah normal dan mengakibatkan berbagai masalah kesehatan, bahkan bisa berakibat fatal[1-2].

Penyakit jantung menjadi beban kesehatan global yang signifikan dengan dampak yang sangat serius. Menurut Organisasi Kesehatan Dunia (WHO), setiap tahunnya penyakit jantung menyebabkan lebih dari 12 juta kematian di seluruh dunia. Hal ini menjadikan penyakit jantung sebagai salah satu penyebab utama kesakitan dan kematian di antara populasi global[3]. Di Indonesia, penyakit jantung menjadi penyebab kematian tertinggi, dengan penyakit jantung koroner dan serangan jantung sebagai jenis penyakit jantung yang paling umum terjadi[4]–[5]. Risiko penyakit jantung dapat dipengaruhi oleh berbagai faktor, seperti usia, jenis kelamin, riwayat medis, dan gaya hidup yang tidak seimbang.

Perkembangan teknologi informasi dan komputasi telah memberikan kontribusi signifikan dalam berbagai bidang, termasuk kesehatan. Salah satu inovasi penting adalah penggunaan *machine learning* dalam analisis data. *Machine learning* adalah cabang kecerdasan buatan yang berfokus pada pengembangan algoritma yang memungkinkan komputer untuk belajar dan membuat prediksi atau keputusan berdasarkan data.[6]–[7] Dalam konteks ini, data *mining* menjadi teknik penting yang digunakan untuk mengekstraksi informasi berharga dari sejumlah besar data. Salah satu metode dalam data *mining* adalah klasifikasi, yaitu proses pengelompokan data ke dalam kategori yang telah ditentukan berdasarkan fitur atau karakteristik tertentu[8].

Support Vector Machine (SVM) adalah salah satu metode *machine learning* yang digunakan untuk klasifikasi dan regresi. SVM bekerja dengan mencari *hyperplane* terbaik yang memaksimalkan margin antara kelas-kelas data dalam ruang fitur. Dengan menggunakan *kernel*, SVM dapat memetakan data non-linear ke dimensi yang lebih tinggi, sehingga memungkinkan pemisahan yang lebih baik antara kelas-kelas yang berbeda[9]–[10]. Metode SVM dapat membantu mengklasifikasi risiko penyakit jantung dengan lebih baik. Namun, masih banyak risiko penyakit jantung yang belum dapat dikenali dengan baik, sehingga perlu dilakukan penelitian ini untuk mengembangkan model klasifikasi risiko penyakit jantung menggunakan metode SVM berdasarkan data klinis pasien. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem klasifikasi dan pencegahan penyakit jantung, serta sebagai alat pendeteksi dini yang dapat membantu masyarakat mengurangi faktor risiko yang menyebabkan penyakit jantung.

2. Metode

2.1 Data dan Pengumpulan Data

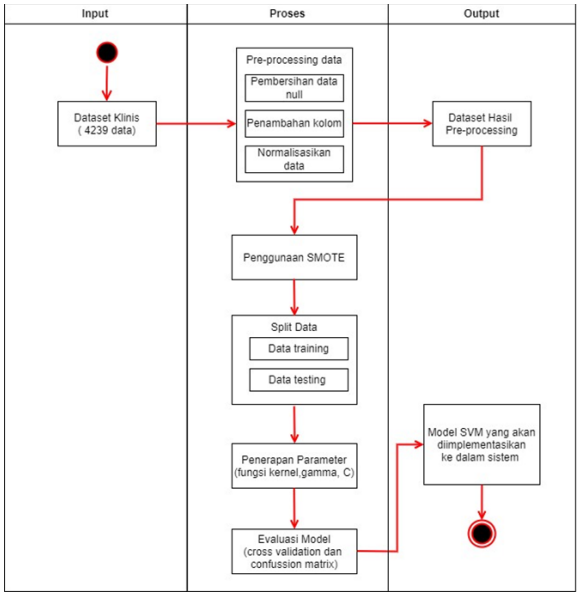
Dalam membangun sebuah sistem diperlukan data-data yang relevan dan dapat dipertanggungjawabkan kebenarannya tanpa adanya manipulasi. Selain data, hal lain yang dibutuhkan adalah pengumpulan informasi. Hal ini dilakukan untuk menjadi bahan referensi dan memperkuat penelitian yang akan dilakukan. Adapun data pada penelitian ini adalah data sekunder (*Secondary Data*).

Pada penelitian ini digunakan data sekunder yang berupa laporan klinis pasien pada *dataset* pasien penyakit jantung. *Dataset* yang digunakan dapat ditemukan pada *heart disease risk* dari kaggle.

Pengumpulan data pada penelitian ini diperoleh melalui satu metode. Adapun metode yang dilakukan dalam pengumpulan data pada penelitian ini adalah Studi Literatur. Studi literatur dilakukan untuk memperoleh informasi, teori-teori, dan referensi penelitian yang berkaitan. Adapun studi literatur pada penelitian ini dilakukan dengan mengumpulkan jurnal-jurnal yang terkait tentang penyakit jantung. Selain jurnal, penelitian ini juga melibatkan penelusuran buku-buku, artikel ilmiah, dan publikasi lain yang relevan untuk mendapatkan gambaran yang komprehensif tentang topik yang diteliti. Dengan pendekatan ini, diharapkan dapat diperoleh dasar teori yang kuat dan mendalam, yang dapat mendukung analisis serta interpretasi hasil penelitian ini. Data yang digunakan sebagai *dataset* klasifikasi risiko penyakit jantung ini terdiri dari Usia, Jenis Kelamin, Kadar kolestrol, Tekanan darah sistolik, Tekanan darah diastolik, Detak Jantung, Diabetes, Merokok, Glukosa, Indeks massa tubuh, dan Risiko penyakit jantung.

2.2 Rancangan Metode *Support Vector Machine*

Proses perancangan metode SVM meliputi langkah-langkah seperti pemilihan parameter, termasuk jenis *kernel*, dan penyesuaian model untuk memastikan kinerja metode yang optimal untuk mengklasifikasi risiko penyakit jantung. Pemilihan parameter ini melibatkan pengaturan nilai-nilai seperti C (parameter regularisasi) dan gamma (parameter *kernel*), yang dapat mempengaruhi performa SVM. Langkah-langkah ini diperlukan untuk memastikan bahwa metode SVM dapat dengan tepat membedakan antara pasien atau individu yang berisiko terkena penyakit jantung dan yang tidak, berdasarkan fitur-fitur yang dianalisis dari data pasien yang tersedia. Selain itu, proses validasi silang (*cross-validation*) juga diterapkan untuk menguji model dan mencegah *overfitting*, sehingga menghasilkan model yang lebih *robust* dan dapat diandalkan. Adapun *activity* diagram dari perancangan metode SVM pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Diagram *Activity* Perancangan Implementasi Metode SVM

Input dataset klinis yaitu penginputan data merupakan tahap awal dalam proses pembangunan model. *Pre-processing* data adalah melakukan *preprocessing* data. *Dataset* yang telah diinputkan akan dilakukan pembersihan nilai kosong, pemisahan kolom dan normalisasi data. Selanjutnya *split* data dilakukan dengan

memisakan data *training* sebanyak 80% dari *dataset* dan data testing sebanyak 20%. Penerapan parameter digunakan untuk pemilihan parameter gamma dan C akan digunakan *grid search*. *Grid search* digunakan untuk menemukan kombinasi terbaik dari parameter-model (*hyperparameter*). Terakhir yaitu evaluasi model proses ini melibatkan penggunaan *K-Fold Cross Validation* dan *Confusion* Matrix, serta metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-score.

3. Hasil Dan Pembahasan

3.1 Implementasi Metode *Support Vector Machine*

Implementasi metode menjelaskan tahapan perancangan model dari metode yang digunakan. Adapun tahapan pembuatan model metode *Support Vector Machine* adalah *input dataset*, *dataset* awal yang digunakan untuk penelitian ini berjumlah 4237 data dengan memiliki 10 fitur dan 1 label target yaitu "*Heart Risk*" yang memiliki 2 kelas. Terdapat 3594 data yang termasuk dalam kelas "*No*", yang menunjukkan bahwa individu tersebut tidak memiliki risiko terkena penyakit jantung, dan 644 data termasuk dalam kelas "Yes", yang menunjukkan bahwa individu tersebut memiliki risiko terkena penyakit jantung. Bentuk *dataset* awal yang diinputkan dapat dilihat pada gambar 2.

	Gender	age	Smoker	diabetes	Chol	sysBP	diaBP	BMI	heartRate	glucose	Heart_risk
0	Male	39	0	0	195.0	106	70	26.97	80.0	77.0	No
1	Female	46	0	0	250.0	121	81	28.73	95.0	76.0	No
2	Male	48	1	0	245.0	128	80	25.34	75.0	70.0	No
3	Female	61	1	0	225.0	150	95	28.58	65.0	103.0	yes
4	Female	46	1	0	285.0	130	84	23.10	85.0	85.0	No
...	...	...	...	...	...	...	...	...	...	...	...
4233	Male	50	1	0	313.0	179	92	25.97	66.0	86.0	yes
4234	Male	51	1	0	207.0	127	80	19.71	65.0	68.0	No
4235	Female	48	1	0	248.0	131	72	22.00	84.0	86.0	No
4236	Female	44	1	0	210.0	127	87	19.16	86.0	NaN	No
4237	Female	52	0	0	269.0	134	83	21.47	80.0	107.0	No

4238 rows × 11 columns

Gambar 2. Dataset Awal

Dalam pre-processing data meliputi pembersihan nilai kosong, normalisasi data untuk menghilangkan perbedaan skala antar fitur, serta transformasi fitur agar sesuai dengan kebutuhan model SVM.

	Gender	age	Smoker	diabetes	Chol	sysBP	diaBP	BMI	heartRate	glucose	Heart_risk	Obesity
0	1.120405	-1.239365	-0.981607	-0.167181	-0.939466	-1.202193	-1.089537	0.285610	0.357599	-0.206805	0	-0.385997
1	-0.892535	-0.422830	-0.981607	-0.167181	0.289273	-0.522695	-0.169896	0.718402	1.614169	-0.248899	0	-0.385997
2	1.120405	-0.189534	1.018738	-0.167181	0.177569	-0.205596	-0.253500	-0.115215	-0.061257	-0.501462	0	-0.385997
3	-0.892535	1.326889	1.018738	-0.167181	-0.269245	0.791000	1.000556	0.681516	-0.898970	0.887636	1	-0.385997
4	-0.892535	-0.422830	1.018738	-0.167181	1.071197	-0.114996	0.080915	-0.666042	0.776456	0.129946	0	-0.385997
...	...	...	...	...	...	...	...	...	...	...	...	...
3820	1.120405	2.143425	-0.981607	-0.167181	-1.363939	1.606397	1.167764	-0.656206	-1.317827	-0.122617	1	-0.385997
3821	1.120405	0.043762	1.018738	-0.167181	1.696737	2.104696	0.749745	0.039705	-0.815199	0.172040	1	-0.385997
3822	1.120405	0.160410	1.018738	-0.167181	-0.671378	-0.250896	-0.253500	-1.499659	-0.898970	-0.585650	0	-0.385997
3823	-0.892535	-0.189534	1.018738	-0.167181	0.244591	-0.069697	-0.922329	-0.936537	0.692685	0.172040	0	-0.385997
3824	-0.892535	0.277058	-0.981607	-0.167181	0.713746	0.066203	-0.002688	-1.066867	0.357599	1.056012	0	-0.385997

3825 rows × 12 columns

Gambar 3. Hasil Normalisasi Data

Teknik SMOTE digunakan untuk menyeimbangkan kelas yang minoritas agar memiliki rentang data yang tidak jauh berbeda diantara kedua kelas tersebut. Diketahui dari 3825 *dataset*, data yang memiliki label "*yes*" hanya berjumlah "585". Artinya, data dengan label "*yes*" merupakan kelas minoritas yang akan dilakukan penyeimbangan oleh SMOTE.

```
Jumlah data untuk setiap kelas setelah SMOTE:
Heart_risk
0    3240
1    3240
Name: count, dtype: int64
Jumlah data training: 5184
Jumlah data testing: 1296
```

Gambar 4. Penggunaan SMOTE

Split Data dilakukan untuk membagi *dataset* yang digunaka menjadi data *training* dan data testing. Dari jumlah seluruh data yaitu 3825 sebelum SMOTE dibagi 80% data *training* dan 20% data testing. Oleh karena itu, didapatkan data *training* sejumlah 3060 dan data testing sejumlah 765 data. Sedangkan, jumlah seluruh data yang telah dilakukan SMOTE adalah 6480. Setelah dilakukan *split* data, data *training* berjumlah 5184 dan data testing berjumlah 1296 data.

Penerapan Parameter dilakukan untuk meningkatkan model dan akurasi hasil klasifikasi. Fungsi *kernel* yang dipakai adalah RBF dikarenakan RBF memiliki kemampuan untuk menangani data non linear. Dalam mencari parameter C dan gamma yang memiliki kinerja terbaik digunakan teknik *grid search*.

```
SVC(C=1, gamma=0.01)
              precision    recall  f1-score   support

           0       0.86      1.00      0.92       654
           1       0.25      0.01      0.02       111

    accuracy                           0.85       765
   macro avg       0.55      0.50      0.47       765
weighted avg       0.77      0.85      0.79       765
```

Gambar 5. Hasil *Grid Search* tanpa SMOTE

Parameter dengan kinerja terbaik didapatkan pada parameter C=1 dan gamma=0,01 akurasi mencapai 85%. Namun, jika menggunakan SMOTE, parameter terbaik yang didapatkan berbeda dengan akurasi yang berbeda. Parameter dengan kinerja terbaik didapatkan pada parameter C= 10 dan gamma =1 dengan akurasi mencapai 92%.

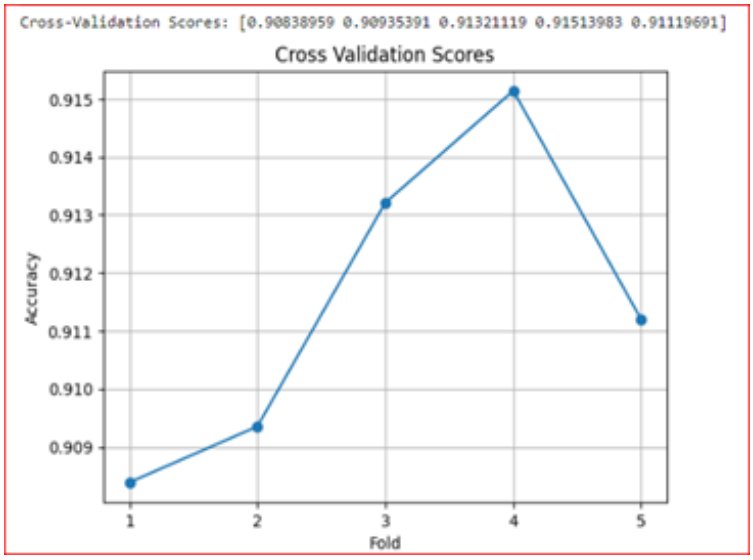
```
SVC(C=10, gamma=1)
              precision    recall  f1-score   support

           0       0.95      0.89      0.92       657
           1       0.89      0.95      0.92       639

    accuracy                           0.92      1296
   macro avg       0.92      0.92      0.92      1296
weighted avg       0.92      0.92      0.92      1296
```

Gambar 6. Hasil *Grid Search* dengan SMOTE

3.2 Evaluasi Model SVM



Gambar 7. Hasil Cross Validation C=10 dan gamma=1 dengan SMOTE

Proses mengevaluasi kinerja model menggunakan teknik *k-fold cross validation*, di mana nilai k yang digunakan adalah 5. Dari gambar 7 dijelaskan bahwa hasil dari parameter C=10 dan gamma=1 dengan SMOTE memperoleh rata-rata akurasinya 91%. Selanjutnya, evaluasi model dilakukan menggunakan Confusion Matrix dan menghitung akurasi, presisi, recall, serta f1-score. Berikut hasil Confusion Matrix dengan parameter C=10 dan gamma=1 dengan SMOTE.

Tabel 1. Hasil *Confussion Matrix*

Data Aktual	Data Prediksi	
	Yes (Risk)	**No (No Risk)**
Yes (Risk)	607	32
No (No Risk)	74	583

Adapun performa dari seluruh parameter yang digunakan dalam grid search dapat dilihat pada tabel 2. Tabel 2 menunjukkan bahwa kombinasi Cost 10 dengan Gamma 1 memberikan kinerja terbaik secara keseluruhan, dengan akurasi, presisi, recall, dan F1-Score masing-masing mencapai 92%, 95%, 89%, dan 92%. Di sisi lain, kombinasi dengan Cost 100 menunjukkan hasil yang mirip, tetapi dengan presisi yang sedikit lebih rendah (95% dengan recall 88%), menandakan adanya trade-off yang harus dipertimbangkan saat menentukan parameter. Hal ini menunjukkan bahwa penyesuaian parameter Cost dan Gamma memiliki dampak signifikan terhadap performa model. Cost yang lebih tinggi tidak selalu menjamin hasil yang lebih baik, dan Gamma juga harus dipilih dengan cermat untuk mencapai keseimbangan yang optimal. Hasil ini menekankan pentingnya eksperimen sistematis dalam memilih parameter model untuk meningkatkan kinerja dalam aplikasi nyata. Penentuan parameter yang optimal dapat mendorong ketepatan model di lingkungan yang lebih kompleks.

Hasil ini mencerminkan bahwa pemilihan parameter yang tepat sangat penting untuk mencapai kinerja model yang optimal. Karena pemilihan parameter tersebut bukan hanya meningkatkan kinerja tetapi juga meningkatkan daya prediksi model, memberikan informasi yang lebih akurat dan relevan untuk pengambilan keputusan yang lebih baik.

Tabel 2. Performa model SVM kernel RBF

SVM kernel RBF dengan SMOTE					
Cost	**γ**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
0.1	1	80	79	82	81
	0.1	66	69	59	63
	0.01	65	67	60	63
	0.001	63	63	66	65
1	1	91	92	90	91
	0.1	71	73	67	70
	0.01	66	68	61	64
	0.001	65	66	61	64
10	1	92	95	89	92
	0.1	78	81	73	77
	0.01	67	70	62	66
	0.001	65	67	61	64
100	1	92	95	88	91
	0.1	85	90	80	84
	0.01	69	71	64	67
	0.001	66	68	61	64

4. Kesimpulan

Berdasarkan pembahasan dari bab-bab sebelumnya mengenai penelitian "Klasifikasi Risiko Penyakit Jantung dengan Metode *Support Vector Machine*" dapat diambil simpulan yaitu metode *Support Vector Machine* dapat diterapkan pada sistem klasifikasi risiko penyakit jantung berdasarkan data klinis. Hasil pengujian menunjukkan bahwa kinerja model SVM dengan parameter C=10 dan γ = 1 dengan SMOTE mencapai akurasi sebesar 92%, presisi 89%, recall 95%, dan f1-score 92%. Sehingga kinerja metode SVM terbaik untuk melakukan klasifikasi data klinis pasien adalah parameter C=10 dan γ = 1 dengan SMOTE.

Referensi

[1] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," Mob. Inf. Syst., vol. 2020, 2020.

[2] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 302–305, 2020.

[3] V. harsha vardhan:Uppala rajesh kumar;Yanumu vardhini;Sabbi leela varalakshmi;A. sura. Kumar, "Heart Disease Prediction Using Machine Learning," vol. 14, no. 04, 2023.

[4] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, pp. 1–16, 2021.

[5] A. A. Permana et al., Machine Learning, vol. 45, no. 13. 2023. [Online].

[6] I. Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 10–19, 2021.

[7] I. Rashad, Sistem Prediksi Untuk Risiko Penyakit Jantung Menggunakan Algoritma Klasifikasi Linear Discriminant Analysis. 2023.

[8] N. Fitriyani, D. R. Amalia, H. H. Handayani, A. Fitri, and N. Masruriyah, "Aplikasi Berbasis Web Berdasarkan Model Klasifikasi Algoritma SVM dan Logistic Regression Terhadap Data Diabetes," vol. 7, pp. 1762–1771, 2023.

[9] B. Akalin, Ü. Veranyurt, and O. Veranyurt, "Classification of Individuals at Risk of Heart Disease Using Machine Learning," Cumhur. Med. J., no. September, pp. 283–289, 2020.

[10] S. T. P. Prasanna and T. Veeramani, "Supervised study of Novel Random Forest Algorithm for prediction of heart disease in Comparison With The Decision Tree Algorithm," Cardiometry, no. 25, pp. 1483–1490, 2023.