Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

Sistem Klasifikasi Penyakit Jantung Menggunakan Teknik Pendekatan SMOTE Pada Algoritma Modified K-Nearest Neighbor

Fitria Novitasari, Elin Haerani*, Alwis Nazir, Jasril, Fitri Insani

Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim, Pekanbaru, Indonesia
Email: [1]11850125022@students.uin-suska.ac.id , [2,*]elinhaerani@uin-suska.ac.id, [3]alwis.nazir@uin-suska.ac.id,
[4]jasril@uin-suska.ac.id, [5]fitri.insani@uin-suska.ac.id
Email Penulis Korespondensi: elinhaerani@uin-suska.ac.id
Submitted: **10/06/2023**; Accepted: **29/06/2023**; Published**: 29/06/2023**

Abstrak−Jantung adalah organ vital yang berperan penting dalam mengalirkan darah yang terdapat oksigen dan nutrisi ke seluruh tubuh. Penyakit jantung merujuk pada kerusakan pada jantung yang dapat terjadi dalam berbagai bentuk, baik disebabkan oleh infeksi maupun kelainan bawaan. Organisasi Kesehatan Dunia (WHO) memaparkan tercatat hampir 17,9 juta kematian setiap tahun yang disebabkan akibat penyakit jantung, di Indonesia prevalensi penyakit jantung sekitar 1,5%, dengan arti pada tahun 2018 sekitar 15 dari 1.000 jiwa atau hampir 2.784.06 jiwa akan menderita penyakit ini berdasarkan data Riset Kesehatan Dasar (Riskesdas) 2018. Banyak masyarakat kurang paham tentang kesehatan jantung, sehingga mereka sering tidak menyadari penyakit jantung yang mereka derita. Hal ini disebabkan kurangnya kesadaran akan pentingnya pemeriksaan medis terkait kesehatan jantung. MKNN termasuk salah satu metode data mining digunakan dalam klasifikasi penyakit jantung. Data yang digunakan pada penelitian didapat dari repositori UCI dataset yang mana terdapat 918 record dengan 12 atribut Untuk menyeimbangkan dataset yang memiliki kelas minoritas, digunakan pendekatan Synthetic Minority Over-sampling Technique (SMOTE), yang menghasilkan data sampel baru dari kelas minoritas. Tujuan dari pengembangan sistem berbasis web untuk klasifikasi penyakit jantung adalah membantu masyarakat dalam melakukan pemeriksaan risiko penyakit jantung sedini mungkin. Dengan demikian, mereka dapat mengambil tindakan preventif lebih awal. Hasil akurasi dari algoritma MKNN dengan rasio 90:10 adalah 88,05%, sedangkan dengan pendekatan MKNN+SMOTE, akurasinya meningkat menjadi 90,20%. Penggunaan pendekatan SMOTE mampu meningkatkan akurasi data dengan kinerja rendah.

Kata Kunci: Penyakit jantung; MKNN; Klasifikasi; SMOTE; Data mining*;*

Abstract−The heart is a vital organ that plays a crucial role in pumping oxygenated blood and nutrients throughout the body. Heart disease refers to damage to the heart that can occur in various forms, caused by infections or congenital abnormalities. The World Health Organization (WHO) reports nearly 17.9 million deaths each year due to heart disease. In Indonesia, the prevalence of heart disease is around 1.5%, meaning that in 2018, approximately 15 out of 1,000 people, or nearly 2,784,060 individuals, were affected by this disease, according to the Basic Health Research data (Riskesdas) 2018. Many people have limited knowledge about heart health, leading to a lack of awareness of their heart conditions. This can be attributed to a lack of understanding regarding the importance of medical checkups related to heart health. Modified K-Nearest Neighbors (MKNN) is one of the data mining methods applied for classifying the risk of heart disease. The research utilized data obtained from the UCI dataset repository, which consists of 918 records with 12 attributes. To balance the imbalanced dataset with minority classes, the Synthetic Minority Over-sampling Technique (SMOTE) approach was used to generate new synthetic samples from the minority class. The objective of developing a web-based system for heart disease classification is to assist the public in assessing their risk of heart disease as early as possible, enabling them to take preventive actions sooner. The accuracy results of the MKNN algorithm with a 90:10 ratio are 88,05%, while with the MKNN+SMOTE approach, the accuracy increased to 90,20%. The use of the SMOTE approach improved the accuracy of low-performing data.

Keywords: Heart disease; MKNN; Classification; SMOTE; Data mining;

1. PENDAHULUAN

Jantung ialah salah satu organ vital yang mempunyai peranan penting bagi kehidupan individu termasuk manusia, karena berfungsi mengalirkan darah yang terdapat oksigen dan nutrisi ke seluruh bagian tubuh [1]. Ketika terjadi disfungsi atau kerusakan pada jantung maka akan mempengaruhi fungsi organ lain dalam tubuh manusia [2]. Penyakit jantung mengacu pada kondisi di mana jantung rusak dan dapat terjadi dalam berbagai bentuk, seperti penyakit arteri koroner, gangguan katup jantung, atau gangguan pada otot jantung. Penyakit jantung juga dapat terjadi karena penyebaran infeksi atau kelainan bawaan [3]. Pencegahan dan pengelolaan penyakit jantung melibatkan perubahan gaya hidup sehat yang melibatkan melakukan perubahan seperti menjaga pola makan yang seimbang, rutin berolahraga, tidak merokok, dan mengelola stres. Dalam beberapa kasus, pengobatan medis dan prosedur bedah mungkin diperlukan untuk mengatasi masalah jantung yang lebih serius [4].

World Health Organization (WHO) memaparkan bahwa penyakit jantung ialah penyebab utama dari kematian di seluruh dunia, tercatat hampir 17,9 juta kematian setiap tahun [5]. Jumlah ini mewakili sekitar 32% dari seluruh penyebab kematian global pada tahun 2019 dan diperkirakan mengalami peningkatan hingga mencapai 23,6 juta kematian dikarenakan penyakit jantung pada kiraan tahun 2030 [5]. Di Indonesia prevalensi penyakit jantung sekitar 1,5%, dengan arti pada tahun 2018 sekitar 15 dari 1.000 jiwa atau hampir 2.784.064 jiwa akan menderita penyakit ini [6]. Untuk tahun 2021, angka kematian oleh penyakit jantung diperkiraan bertambah sebesar 22-23% akibat situasi pandemi COVID-19 [7]. Prevalensi penyakit jantung diperkirakan akan terus meningkat di Indonesia dan akan menjadi penyebab kematian yang paling umum, terutama pada kelompok usia yang masih produktif [6].

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

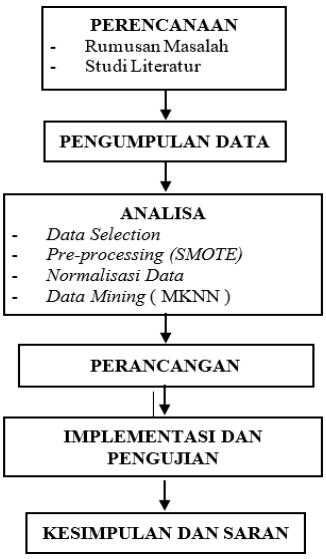
Banyak orang di masyarakat yang kurang memiliki pengetahuan tentang kondisi kesehatan jantung mereka, sehingga mereka sering tidak menyadari bahwasannya mereka sedang mengalami penyakit jantung. Hal ini dapat disebabkan oleh kurangnya kesadaran akan pentingnya pemeriksaan medis terkait kesehatan jantung. Padahal, penyakit jantung dapat terjadi pada individu manapun termasuk orang yang sebelumnya tidak menunjukkan riwayat penyakit jantung. Jumlah kematian yang signifikan dan prevalensi yang terus meningkat, penelitian klasifikasi penyakit jantung menjadi penting untuk mengidentifikasi faktor risiko, pola geografis, dan perubahan tren epidemiologi. Ini membantu dalam merencanakan intervensi yang tepat, meningkatkan pengobatan, dan mengurangi beban penyakit jantung secara global. Karena itu, memiliki sistem klasifikasi yang dapat memberikan informasi rinci tentang penyakit jantung menjadi sangat penting. Dengan adanya sistem ini, kita dapat melakukan pemeriksaan dini terhadap serangan jantung pada manusia, sehingga memungkinkan tindakan pencegahan yang lebih efektif [1].

Klasifikasi penyakit jantung telah banyak diteliti dengan berbagai metode yang ada pada data mining seperti penelitian yang dilakukan oleh [1], [2], [8]–[15], [16]. Penelitian yang dilakukan oleh Riski Annisa dalam penelitiannya mendapatkan hasil dimana algoritma random forest (RF) dan decision menunjukkan kinerja terbaik dalam klasifikasi pada dataset yang digunakan. Algoritma C4.5 juga menunjukkan hasil yang baik. Namun, algoritma k-NN kurang baik dengan akurasi 60,77% [14]. Penelitian lain dari Yovi Pratama menemukan hasil klasifikasi penyakit jantung dengan metode k-nearest neighbor didapatkan hasil akurasi sebesar 70,65%, nilai presesinya adalah 75 %, sedangkan recall menghasilkan 70,73% [12]. Metode K-Nearest Neighbor (KNN) ialah salah satu metode klasifikasi dasar dan paling sederhana selain KNN, terdapat juga algoritma pengembangan KNN yang dikenal sebagai Modified K-Nearest Neighbor (MKNN). MKNN meningkatkan perhitungan validitas dan bobot *vote* dimana nantinya validitas menggambarkan informasi tambahan tentang hubungan antara sampel data pelatihan dalam ruang fitur, hal ini mempertimbangkan stabilitas dan ketahanan nilai dari setiap sampel terhadap tetangganya dengan tujuan memberikan klasifikasi yang lebih baik dibanding KNN [17].

Klasifikasi penelitian ini menggunakan variabel target data yaitu kategori penyakit jantung. Namun, terdapat ketidakseimbangan data dalam kumpulan data ini, yang membuat perbandingan setiap kategori penyakit jantung menjadi sedikit tidak seimbang dengan perbandingan kelas 0 410 data dan kelas 1 508 data. Ketidakseimbangan kelas ini dapat menyebabkan kinerja klasifikasi yang buruk dan kurang optimal [18] . Masalah ini bisa disebabkan apabila jumlah data pada kelas mayoritas dan minoritas pada data sangat berbeda [19]. Adapun untuk solusi yang digunakan untuk menyelasaikan permasalahan ini dapat menggunakan dua pendekatan yang biasa digunakan, salah satu pendekatan yang umum adalah *oversampling* atau *undersampling*. Pada pendekatan oversampling, penyeimbangan data pada kelas minoritas dilakukan dengan menggunakan teknik acak. Namun, replikasi pengamatan dapat menyebabkan overfitting, di mana model terlalu sesuai dengan data pelatihan sehingga dapat mengurangi akurasi ketika diuji dengan data yang berbeda. Untuk mengatasi masalah ini, digunakanlah teknik yang populer yang dikenal sebagai *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE mengatasi overfitting dengan mempertimbangkan tetangga terdekat dalam menghasilkan data sintetis, sehingga batas keputusan yang lebih baik dapat menangkap contoh kelas minoritas yang saling berdekatan. Berdasarkan penjelasan di atas, penelitian tugas akhir ini akan mempelajari Sistem klasifikasi penyakit jantung menggunakan teknik pendekatan SMOTE pada algoritma Modified Knearest Neighbor.

2. METODOLOGI PENELITIAN

Penelitian memerlukan langkah-langkah yang terstruktur agar tercapainya tujuan penelitian. Adapun langkah demi langkah pada Gambar 1.



Gambar 1. Bagan Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

2.1 Perencanaan

2.1.1 Rumusan Masalah

Tahap awal dimulai dengan mengenali dan memahami permasalahan yang ada. Tujuannya adalah mendapatkan solusi untuk memecahkan masalah tersebut. Pada penelitian, rumusan masalahnya adalah bagaimana penerapan Modified K-Nearest Neighbor (MKNN) dengan pendekatan SMOTE dalam proses klasifikasi penyakit jantung.

2.1.2 Studi Literatur

Pada tahap literatur dilakukan pencarian dalam memperoleh informasi teoritis dari beberapa sumber seperti buku, *paper*, artikel, sumber internet, dan penelitian terkait yang berhubungan dengan topik pada penelitian.

2.2 Pengumpulan data

Pada tahap pengumpulan data terdiri dari data pasien yang menderita penyakit jantun. Data tersebut selanjutnya digunakan dianalisis. Informasi pasien didapatkan dari survei yang tersedia di repositori UCI *machine learning. Dataset* yang digunakan terdiri dari 918 sampel dengan 12 atribut. Atribut ke-12 merupakan pemilihan atribut yang digunakan sebagai prediksi serta sebagai kelas dalam proses analisis.

2.3 Tahap Analisa

2.3.1 Data Selection

Dataset penyakit jantung yang digunakan terdiri dari 918 sampel dengan 12 atribut yang dibagi menjadi dua kelas yakni, terdapat kelas 0 410 data yang berisiko rendah terkena penyakit jantung dan kelas 1 508 data yang beresiko terkena penyakit jantung.

2.3.2 *Synthetic Minority Oversampling Technique*

Ketidakseimbangan kelas dapat menghasilkan hasil klasifikasi yang tidak baik dan tidak selalu optimal. Teknik oversampling minoritas sintetik (SMOTE) adalah bagian dari teknik *oversampling* yang dapat menjadi solusi untuk mengatasi masalah. Nithes V. Chawla merupakan orang yang pertama kali memperkenalkan teknik ini dan juga melibatkan penciptaan duplikat data sintetik dari beberapa data yang ada. Ketidakseimbangan kelas dapat menyebabkan hasil klasifikasi yang kurang baik dan optimal [19].

Metode SMOTE melakukan proses dengan mencari K-Nearest Neighbor (KNN) dari semua data dalam kelas yang sedikit. Kemudian, metode ini menghasilkan data sintetis dari persentase duplikasi data minoritas yang diinginkan (N%). Proses pemilihan k tetangga terdekat dilakukan secara random. Dapat diliat pada persamaan (1):

$$N\% = \frac{Total\ data\ kelas\ mayor}{Total\ data\ kelas\ minor} \times 100\% \qquad (1)$$

2.3.3 Normalisasi

Normalisasi data ialah proses mengubah skala atau rentang data menjadi rentang yang lebih terstandarisasi atau normal. Hal ini dilakukan untuk memastikan bahwa semua fitur atau variabel memiliki pengaruh yang seimbang dalam analisis data. Min-Max Normalization adalah suatu teknik normalisasi yang menggunakan transformasi linier pada data asli untuk mencapai nilai yang seimbang dalam perbandingan antara data sebelum dan setelah proses normalisasi. Rentang yang digunakan dalam metode ini adalah 0-1 [20]. Berikut rumus Min-Max Normalization dapat diliat pada persamaan (2):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} (new_{max} - new_{min}) + new_{min} \qquad (2)$$

2.3.4 Data Mining

Data mining adalah tahapan pengelolaan data dalam menemukan, menganalisis pola serta pengetahuan yang terdapat dalam data menggunakan metode statistic matematika dan pembelajaran mesin. *Data mining* melibatkan kegiatan analisis berulang dalam *database* besar untuk mendapatkan informasi dan pengetahuan yang tepat pada target untuk bisa digunakan pada pengambilan keputusan dalam pemecahan masalah [21]. Metode klasifikasi yang akan digunakan ialah metode *Modified K-Nearest Neighbors.*

2.3.4.1 Modified K-Nearest Neighbors

Modified K-Nearest Neighbors (MKNN) ialah sebuah algoritma peningkatan lebih lanjut dari metode Knearest Neighbor yang menambahkan dua tahap komputasi yakni pada perhitungan nilai validitas dan bobot.

Penjelasan singkat mengenai proses klasifikasi algoritma MKNN sebagai berikut:

a. Menentukan data latih dan data *testing* serta nilai K. Dimana nilai K ganjil.
b. Mencari nilai dari semua jarak pada data pelatihan dengan menggunakan rumus Euclidean Distance pada persamaan (3).

$$d(x,\ y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^{\cdot 2}} \qquad (3)$$

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

Dimana :
D(xi,yi) : Jarak antar data latih
xi : Sampel data
yi : Data yang diuji

c. Mencari nilai validitas antar data latih. Dapat dilihat pada persamaan (4).

$$Validitas_{(x)} = \frac{1}{k}\sum_{i=0}^{n} S\left(lbl(x), lbl\left(Ni(x)\right)\right) \quad (4)$$

Dimana:
i : Total banyak data latih
K : Banyak titik terdekatt
Lbl(x) : Kelas-x pada label x
Lbl*(Ni(x)) : Label kelas titik terdekat pada x

Untuk mempermudah, digunakan fungsi "S" pada titik "x" dan data tetangga terdekat ke-i. Adapun untuk mendapatkan fungsi S dapat dicari dengan menggunakan Persamaan (5) di bawah ini.

$$S(a,b) = \begin{cases} 1_{Jika\ a=b} \\ 0_{Jika\ a\neq b} \end{cases} \quad (5)$$

Dimana:
S : kemiripan pada tiap data latih
a : kelas a pada data latih.
b : kelas lainnya selain kelas a pada data latih

d. Mencari nilai jarak antar data latih dan data uji sesuai persamaan (3) perhitungan Euclidien Distance.
e. Menghitung Weight Voting (WV). Pada persamaan (6).

$$W_{(i)} = Validitas_{(x)}\ x \frac{1}{d_e + a} \quad (6)$$

Dimana:
W(i) : Weight Voting
Validity(i): Nilai Validitas
de : Rentang Euclidien
a : 0,5; untuk menghindari nilai 0

f. Mengurutkan nilai WV dengan pengurutan *descanding*. Ambil sesuai nilai K.
g. Mendapatkan hasil klasifikasi dari wv.

2.4 Perancangan Sistem

Melanjutkan dari proses Analisa slabgkah selanjutnya ialah melakukan perancangan pada system yang akan dibuat. Tahapan tersebut yaitu:

a. Perancangan *database*
Tahapan ini berisi table *database* serta atribut yang digunakan untuk pembuatan sistem.
b. Perancangan struktur menu
Tahapan ini mendeskripsikan menu-menu yang akan digunakan pada system.
c. Perancangan Antarmuka
Tahapan terakhir ialah tampilan yang akan digunakan pada system. Tampilan tersebut harus bmudah dipahami bagi penggunanya.

2.5 Implementasi Sistem dan Pengujian

Pembangunan system akan dilakukan pada spesifikasi sebagai berikut :

a. Perangkat Keras
CPU : 11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz 2.42 GHz
Memory : 16 GB
SSD : 512 GB
b. Perangkat Lunak
System operasi : Windows 11 Home Single Language
Web Browser dan server : Microsoft Edge dan Apache
Bahasa pemprograman pengembangan : PHP (Framework Laravel)
Tools dan DBMS : Visual Studio Code dan MySQL

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

2.5.1 Confusions Matrix

Tabel yang digunakan untuk menyimpan hasil dari klasikasi dikenal juga dengan tabel *confusion matrix* [22]. Tabel berikut adalah contoh matriks konfusi pada model klasifikasi biner (dua kelas), seperti 0 dan 1. Pada tiap sel data mewakili kumpulan data kelas-i dengan hasil *prediction* terdapat di kelas-j. Tersaji pada Tabel 1.

Tabel 1. Klasifikasi biner dengan *Confusions Matrix*

Data		**Hasil prediksi (*j*)**	
		1	**0**
Kelas-asli(*i*)	**1**	TP	FP
	0	FN	TN

Dari Tabel 1 kita dapat menghitung akurasi, laju eror, presisi, dan recall dari hasil klasifikasi dan tingkat kesalahan dari klasifikasi yang dilakukan menggunakan persamaan (7), (8), (9), (10) :

$$\text{Akurasi} = \frac{TP + TN}{semua\ data} \quad (7)$$

$$\text{Laju } Error = \frac{FP + FN}{semua\ data} \quad (8)$$

$$Presisi = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

3. HASIL DAN PEMBAHASAN

Data jantung dianalisis pada beberapa tahap dalam *Knowledge Discovery in Database* (KDD). Fase-fase ini meliputi pemilihan data, preprocessing, normalisasi, dan data mining.

3.1 Pemilihan Data

Penggunaan data pada penelitian berasal dari UCI dataset. Keseluruhan data yang digunakan berjumlah 918 *record* dengan 12 atribut. Parameter data yang digunakan tersaji dalam Gambar 2 dibawah ini. Sedangkan untuk data yang digunakan tersaji pada Gambar 3.

Atribut	Deskripsi	Kode Pengisian
Age	Umur	
Sex	Jenis	Pria
		Wanita
Cp	Jenis nyeri dada	*Typical angina*
		Atypical angina
		Asymptomatic
		Non angina
RestingBP	Hipertensi(Tekanan darah istirahat)	(dalam mm Hg)
Chol	Kolesterol	(dalam mg/dl)
FastingBS	Kadar gula	*True* = gula darah > 120 mg/dl
		False = gula darah < 120 mg/dl
Restecg	hasil tes elektrokardiografi	Normal
		Stt *abnormality*
		lv *hypertrophy*
MaxHR	Detak jantung maksimum	
ExerciseANogiNoa	Tes latihan	*TRUE*
		FALSE
Oldpeak	Depresi ST yang diinduksi oleh olahraga relatif terhadap istirahat	
ST_Slope	Kemiringan segmen ST latihan puncak	*Upsloping* (menanjak)
		Flat (Datar)
		Downsloping (Menurun
Class	Kelas klasifikasi	0 = Beresiko rendah
		1 = Beresiko tinggi

Gambar 2. Atribut Data

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseANogiNoa	Oldpeak	ST_Slope	kelas
1	40	Male	ATA	140	289	FALSE	Normal	172	No	0	Up	0
2	49	Female	NAP	160	180	FALSE	Normal	156	No	1	Flat	1
3	37	Male	ATA	130	283	FALSE	ST	98	No	0	Up	0
4	48	Female	ASY	138	214	FALSE	Normal	108	Yes	1,5	Flat	1
5	54	Male	NAP	150	195	FALSE	Normal	122	No	0	Up	0
6	39	Male	NAP	120	339	FALSE	Normal	170	No	0	Up	0
7	45	Female	ATA	130	237	FALSE	Normal	170	No	0	Up	0
8	54	Male	ATA	110	208	FALSE	Normal	142	No	0	Up	0
	...		....	....	...		...	...	....	....	....	....
909	48	Female	ATA	120	284	FALSE	Normal	120	No	0	Up	0
910	37	Female	NAP	130	211	FALSE	Normal	142	No	0	Up	0
911	58	Male	ATA	136	164	FALSE	ST	99	Yes	2	Flat	1
912	39	Male	ATA	120	204	FALSE	Normal	145	No	0	Up	0
913	49	Male	ASY	140	234	FALSE	Normal	140	Yes	1	Flat	1
914	42	Female	NAP	115	211	FALSE	ST	137	No	0	Up	0
915	54	Female	ATA	120	273	FALSE	Normal	150	No	1,5	Flat	0
916	38	Male	ASY	110	196	FALSE	Normal	166	No	0	Flat	1
917	43	Female	ATA	120	201	FALSE	Normal	165	No	0	Up	0
918	60	Male	ASY	100	248	FALSE	Normal	125	No	1	Flat	1

Gambar 3. Data Penelitian Penyakit Jantung yang digunakan

3.2 *Synthetic Minority Oversampling Techniquesmote (SMOTE)*

Pada titik ini, pendekatan dengan metode SMOTE diterapkan. Metode ini membangun data replikasi dari kelas data minoritas menggunakan k-nearest neighbor untuk semua data di kelas minoritas. Data sintetik kemudian dibuat sesuai dengan proporsi duplikat data minoritas.

Total data penyakit jantung terdiri dari 918 record, dengan kelas 0 berisi hingga 410 record (kelas minor) dan kelas 1 berisi hingga 508 record (kelas mayor). Membandingkan kelas 0 dan kelas 1 dari 1:1.24. Metode SMOTE diterapkan dengan mensintesis data kelas 0. Ini memberikan persentase oversampling :

$$N\% = \frac{508}{410} \times 100\% = 123{,}9\ \%$$

Persentase duplikasi data minor dan banyak data yang akan dilakukan replikasi sebagai berikut :

Tabel 2. Persentase data minor

Kelas	Total Data	Data yang akan di Sintetis	Percentage oversampling	Jumlah sintetis tiap data
0	410	98	123,9%	2

Berikut hasil data sintetis yang berjumlah 98 data tersaji pada Gambar 4. Jadi total data yang akan di gunakan dan di tambah dengan data baru berjumlah 1116 data.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseANogiNoa	Oldpeak	ST_Slope	kelas
1	53	Male	ATA	120	181	FALSE	Normal	132	No	0	Up	0
3	37	Male	ATA	130	283	FALSE	ST	98	No	0	Up	0
4	35	Female	ASY	138	183	FALSE	Normal	182	No	1,4	Up	0
5	48	Female	ATA	120	284	FALSE	Normal	120	No	0	Up	0
6	50	Male	ATA	120	168	FALSE	Normal	160	No	0	Up	0
7	55	Male	ASY	140	229	FALSE	Normal	110	Yes	0,5	Flat	0
8	49	Female	NAP	130	207	FALSE	ST	135	No	0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
90	63	Male	ASY	126	0	FALSE	ST	120	No	1,5	Down	0
91	29	Male	ATA	130	204	FALSE	LVH	202	No	0	Up	0
92	74	Female	ATA	120	269	FALSE	LVH	121	Yes	0,2	Up	0
93	37	Male	NAP	130	194	FALSE	Normal	150	No	0	Up	0
94	51	Male	ASY	140	261	FALSE	LVH	186	Yes	0	Up	0
95	39	Male	ASY	130	307	FALSE	Normal	140	No	0	Up	0
96	29	Male	ATA	120	243	FALSE	Normal	160	No	0	Up	0
97	48	Female	NAP	130	275	FALSE	Normal	139	No	0,2	Up	0
98	54	Female	ATA	132	288	TRUE	LVH	159	Yes	0	Up	0

Gambar 4. Hasil Data Sintetis

3.3 Normalisasi

Setelah data mengalami penambahan karena proses SMOTE selanjutnya data akan dinormalisasikan. Tujuan nya agar data memiliki keseimbangan nilai dengan rentang nilai 0-1 menggunakan min-max normalization. Normalisasi data hanya di lakukan dengan beberapa atribut data yakni *RestingBP*, *Cholesterol*, *MaxHR*, dan *Oldpeak*. Data tersaji pada Gambar 5:

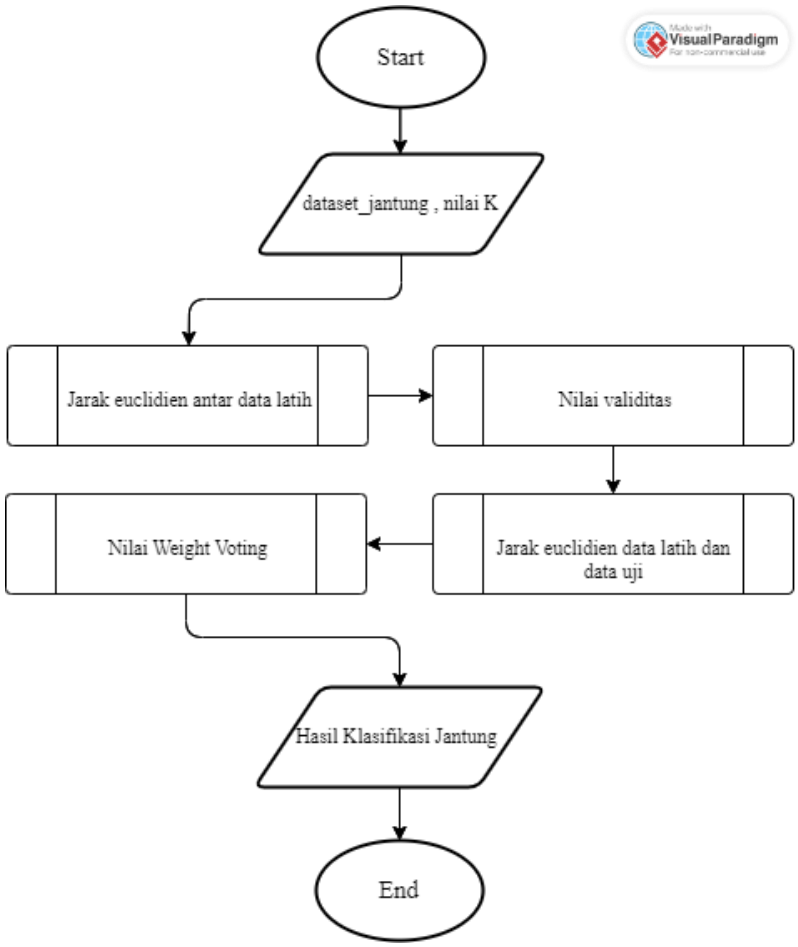
Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

ın1	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseANogiNoa	Oldpeak	kelas
1	40	Male	ATA	0,70	0,479	FALSE	Normal	0,789	No	0,333	0
2	49	Female	NAP	0,80	0,299	FALSE	Normal	0,676	No	0,444	1
3	37	Male	ATA	0,65	0,469	FALSE	ST	0,268	No	0,333	0
4	48	Female	ASY	0,69	0,355	FALSE	Normal	0,338	Yes	0,556	1
5	54	Male	NAP	0,75	0,323	FALSE	Normal	0,437	No	0,333	0
6	39	Male	NAP	0,60	0,562	FALSE	Normal	0,775	No	0,333	0
...	...	...	...	...	...	...	...	...	...	...	...
1009	54	Male	ATA	0,55	0,345	FALSE	Normal	0,577	No	0,333	0
1010	37	Male	ASY	0,70	0,343	FALSE	Normal	0,493	Yes	0,556	1
1011	48	Female	ATA	0,60	0,471	FALSE	Normal	0,423	No	0,333	0
1012	37	Female	NAP	0,65	0,350	FALSE	Normal	0,577	No	0,333	0
1013	58	Male	ATA	0,68	0,272	FALSE	ST	0,275	Yes	0,556	1
1014	39	Male	ATA	0,60	0,338	FALSE	Normal	0,599	No	0,333	0
1015	49	Male	ASY	0,70	0,388	FALSE	Normal	0,563	Yes	0,444	1
1016	42	Female	NAP	0,58	0,350	FALSE	ST	0,542	No	0,333	0

Gambar 5. Data Normalisasi

3.4 Analisa dan Perancangan

Tahap analisis menggunakan algoritma MKNN untuk mempores data yang telah ada. Atribut pada data berjumlah 11 atribut terlihat pada Tabel 2. Berikut proses algoritma MKNN dalam bentuk bagan alir tersaji pada Gambar 6:



Gambar 6. Flowchard MKNN

Langkah awal pada proses MKNN menentukan nilai K, nilai K disini menggunakan K=3 lalu mencari jarak antar semua data latih dengan cara euclidien distance seperti persamaan (3) berikut perhitungan antar data 1 dan data 2.

$$= \sqrt{((40-49)^2 + (1-2)^2} + (2-4)^2 + (0.70-0.80)^2 + (0.479-0.299)^2 + (2-2)^2 + (1-1)^2 + (0.789-0.676)^2 + (2-2)^2 + (0.333-0.444)^2 + (1-2)^2 + (0-1)^2)$$

$$= \sqrt{81} + 1 + 4 + 0.11 + 0.0324 + 0 + 0 + 0.0127 + 0 + 0.0123 + 1 + 1$$

$$= 9.384444$$

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

Selanjutnya dilakukan validitas data antar jarak data latih yang telah di lakukan sesuai persamaan (4) dengan similaritas pada persamaan (5)

Jika nilai perbandingan jarak terdekat pertama data latih 1 adalah 1 dan jarak terdekat kedua data latih 1 adalah 0 maka :

$$Validitas = \frac{1}{3} \times (1 + 0 + 1) = 0.6666$$

Selanjutnya dilakukan perhitungan jarak antar data latih 1 dan uji 1 dengan persamaan (3)

$$= \sqrt{((40-39)^2 + (1-1)^2 + (2-4)^2 + (0.70-0.60)^2 + (0.479-0.562)^2 + (2-2)^2 + (1-1)^2 + (0.789-0.775)^2 + (2-2)^2 + (0.333-0.333)^2 + (1-1)^2 + (0-0)^2)}$$

$$= \sqrt{2.22398}$$

Selanjutnya mencari nilai weight voting menggunakan nilai validitas pada data latih 1 dan uji 1 dengan persamaan (6)

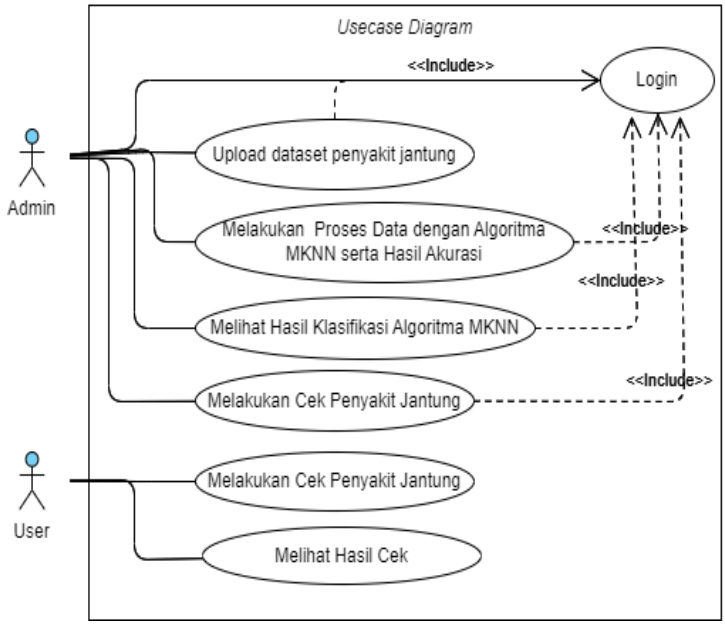
$$W_{(i)} = 0.6666\, x \frac{1}{2.2240 + 0.5}$$

$$W_{(i)} = 0.2447$$

Selanjutnya mengurutkan nilai weight voting terbesar ke terkecil sesuai nilai K yang di tentukan. Dan melakukan perbandingan antar nillai output klasifikasi. Dimana jika nilai terbesar memiliki output 1 0 0 maka di ambil hasil klasifikasi yaitu 0 dengan arti 0 ialah beresiko rendah.

3.4.1 Perancangan Sistem

Perancangan sistem ialah tahap penting dalam pengembangan suatu sistem. Tujuannya adalah untuk merancang sistem yang dapat memenuhi kebutuhan pengguna dan mencapai tujuan yang telah ditetapkan. Berikut usecase pada system klasifikasi penyakit jantung yang tersaji pada Gambar 7:



Gambar 7. Usecase Diagram Sistem

3.5 Implementasi Sistem dan Pengujian

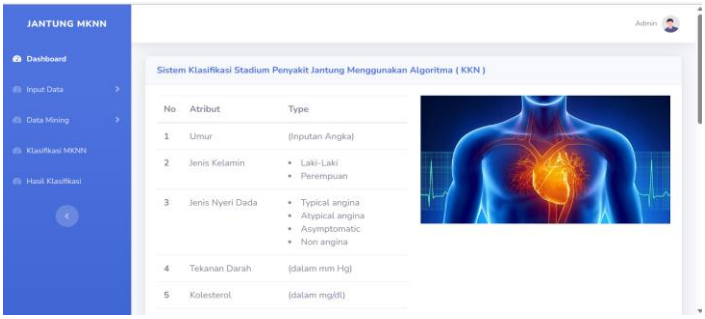
Implementasi sistem adalah tahap di mana desain sistem yang telah dirancang diubah menjadi sistem yang berfungsi secara nyata. Berikut akan di jabarkan implementasi system serta pengujian pada system.

3.5.1 Implementasi Sistem

Berikut beberapa tampilan pada system yang telah di buat:

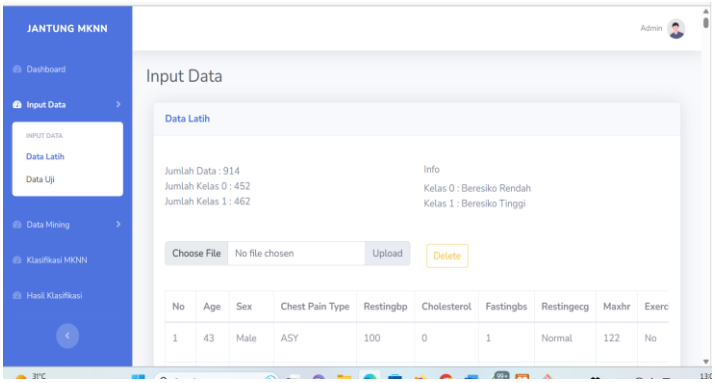
a. Halaman Tampilan *Dashboard*
   Halaman ini mencantumkan informasi atribut dan tipe dari atribut dataset penyakit jantung yang akan digunakan. Berikut tampilan halaman *dashboard* tersaji pada Gambar 8:

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610



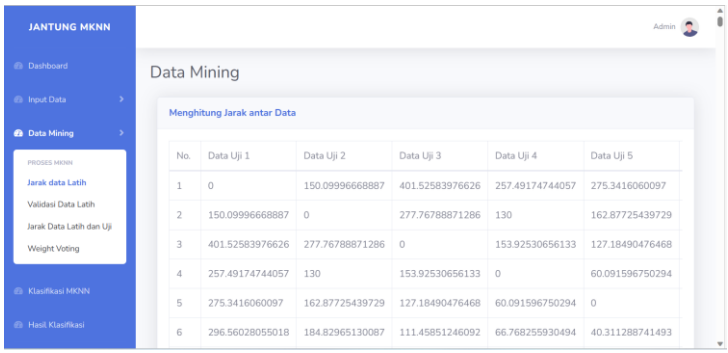
Gambar 8. Tampilan *Dashboard*

b. Halaman Input Data
Halaman ini berisikan informasi tentang data latih serta data uji yang akan dikelola. Berikut tampilan halaman Data latih tersaji pada Gambar 9:



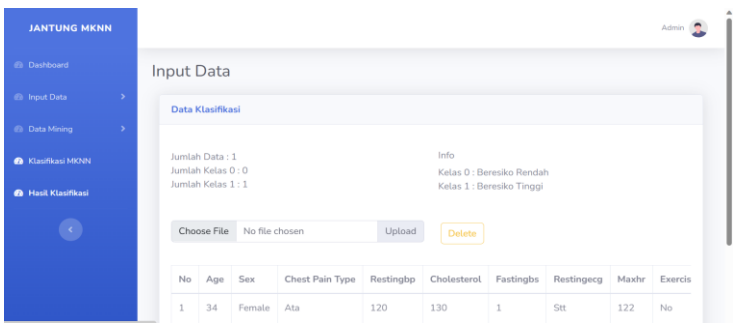
Gambar 9. Halaman Input Data

c. Halaman Data Mining Hitung Jarak
Halaman ini berfungsi untuk salah satu proses metode yang digunakan. Berikut beberapa halaman untuk proses mencari jarak data tersaji pada Gambar 10:



Gambar 10. Halaman Data Mining

d. Halaman Hasil Klasifikasi
Halaman ini berfungsi untuk menampilkan data hasil klasifikasi cek risiko penyakit jantung yang diinput oleh *user* dan admin. Berikut tampilan halaman klasifikasi tersaji pada Gambar 11:



Gambar 11. Hasil Klasifikasi

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

3.5.2 Pengujian

Confusion Matrix dilakukan untuk mendapatkan hasil akurasi dari metode yang digunakan dengan hasil akurasi dari data sebenarnya. Pengujian dilakukan sebanyak tiga kali. Pada pengujian pertama data latih yang digunakan sebanyak 90% dan data uji sebanyak 10%. Pada pengujian yang kedua digunakan data latih sebanyak 80% dan data uji sebanyak 20%. Sedangkan pengujian ketiga menggunakan 70% data latih dan data uji sebanyak 30%.

Perbandingan data menggunakan rasio 80:10 data latih dan uji. Adapun jumlah penggunaan data yang diujikan berjumlah 184 data uji asli dan 202 data uji SMOTE yang terbagi atas 101 data kelas 0 dan 101 data kelas 1 yang dipilih secara random dari *dataset*.

Tabel 3. Tabel perbandingan 80:10

Aktual	Kelas Hasil Klasifikasi MKNN		Kelas Hasil Klasifikasi MKNN+SMOTE	
	Kelas1	Kelas0	Kelas1	Kelas0
Kelas 1	97	1	102	20
Kelas 0	27	59	6	74

Adapun nilai akurasi berdasarkan hasil prediksi serta laju error prediksi dapat ditentukan dengan menggunakan *confusion matrix*, sebagai berikut:

a. confusion matrix kelas hasil klasifikasi MKNN

$$\text{Akurasi} = \frac{97 + 59}{97 + 59 + 27 + 1} = 84{,}78\%$$

b. confusion matrix kelas hasil klasifikasi MKNN+SMOTE

$$\text{Akurasi} = \frac{102 + 74}{102 + 74 + 6 + 20} = 87{,}13\%$$

Dari hasil perhitungan didapat akurasi sebesar 84,78% dan jika menggunakan pendekatan SMOTE akurasi MKNN+SMOTE mengalami peningkatan sebesar 2,35% yaitu menjadi 87,13%.

Berikut pengujian dengan beberapa perbandingan :

Tabel 4. Tabel hasil pengujian perbandingan akurasi

Rasio	Akurasi		Peningkatan
	MKNN	MKNN+SMOTE	
90:10	88,05%	90,20%	2,15%
80:20	84,78%	87,13%	2,35%
70:30	79,64%	83,22%	3.58%

Dari hasil pengujian tersebut, dapat ditarik kesimpulan bahwa algoritma MKNN mencapai akurasi tertinggi sebesar 88,05% dengan perbandingan data latih dan uji 90:10. Sedangkan, algoritma MKNN+SMOTE mencapai akurasi tertinggi sebesar 90,20% dengan perbandingan yang sama.

Berdasarkan Tabel 7, dapat dilihat bahwa pendekatan SMOTE berhasil meningkatkan akurasi pada data uji pada setiap rasio. Peningkatan akurasi terbesar terjadi pada rasio data 70:30, di mana tingkat akurasi meningkat dari 79,64% menjadi 83,22%, atau meningkat sebesar 3,58%. Pengaruh pendekatan SMOTE terhadap peningkatan akurasi lebih signifikan pada data uji dengan akurasi awal yang rendah. Dapat disimpulkan bahwa pendekatan SMOTE sangat efektif digunakan pada data dengan tingkat akurasi yang rendah.

4. KESIMPULAN

Berdasarkan dari penelitian yang telah dilakukan, dapat ditarik kesimpulan bahwa aplikasi berbasis web untuk klasifikasi penyakit jantung dapat mengklasifikasikan risiko penyakit jantung yang diinginkan dengan demikian penggunaan metode MKNN dapat digunakan pada data klasifikasi penyakit jantung. Sistem klasifikasi ini juga dapat mneghasilkan ouput penyakit jantung dari data baru yang di inputkan user dengan mengambil 3 jarak terdekat weight voting. Pengujian akurasi menggunakan confusion matrix dengan dua perbandingan antara algoritma MKNN dan MKNN-SMOTE. Dari hasil pengujian yang dilakukan dengan 918 data dan 1016 data dapat disimpulkan akurasi tertinggi pada rasio 90:10 dimana pada algoritma MKNN ialah sebesar 88,05% sedangkan algoritma MKNN+SMOTE sebesar 90,20%. Peningkatan SMOTE tertinggi di peroleh dari akurasi dengan rasio 70:30 dimana terjadi kenaikan sebesar 3,58%. Dari table pengujian akurasi dapat diliat bahwa semakin rendah akurasi yang dihasilkan semakin tinggi peningkatan yang didapatkan dengan teknik SMOTE.

REFERENCES

[1] M. A. Bianto, K. Kusrini, and S. Sudarmawan, "Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 75, 2020, doi: 10.24076/citec.2019v6i1.231.

[2] A. B. Wibisono and A. Fahrurozi, "Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 274−284
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3610

Koroner," *J. Ilm. Teknol. dan Rekayasa*, vol. 24, no. 3, pp. 161–170, 2019, doi: 10.35760/tr.2019.v24i3.2393.
[3] Alodokter, "Penyakit Jantung," 2019,
[4] A. Nurmasani and Y. Pristyanto, "Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class," *Pseudocode*, vol. 8, no. 1, pp. 21–26, 2021, doi: 10.33369/pseudocode.8.1.21-26.
[5] World Health Organizatotion, "Cardiovascular diseases," 2021.
[6] Tim Riskesdas, "Laporan Kesehatan Nasional RISKESDES 2018," Indonesia, 2019.
[7] Kementerian Kesehatan Republik Indonesia, "Penyakit Jantung Koroner Didominasi Masyarakat Kota," 2021
[8] P. D. Putra, S. Sukemi, and D. P. Rini, "Peningkatan Akurasi Klasifikasi Backpropagation Menggunakan Artificial Bee Colony dan K-NN Pada Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 208, 2021, doi: 10.30865/mib.v5i1.2634.
[9] A. Ariani and S. Samsuryadi, "Classification of Kidney Disease Using Genetic Modified Knn and Artificial Bee Colony Algorithm," *Sinergi*, vol. 25, no. 2, p. 177, 2021, doi: 10.22441/sinergi.2021.2.009.
[10] J. Patel, A. A. Khaked, J. Patel, and J. Patel, "Heart Disease Prediction Using Machine Learning," *Lect. Notes Networks Syst.*, vol. 203 LNNS, no. 3, pp. 653–665, 2021, doi: 10.1007/978-981-16-0733-2_46.
[11] E. Prasetyo and B. Prasetiyo, "Increased Classification Accuracy C4 . 5 Algorithm Using Bagging Techniques in Diagnosing Heart Disease," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 5, pp. 1035–1040, 2020, doi: 10.25126/jtiik.202072379.
[12] Y. Pratama, A. Prayitno, D. Nazrian, N. Aini, Y. R. R, and E. Rasywir, "BULLETIN OF COMPUTER SCIENCE RESEARCH Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma K-Nearest Neighbor," vol. 3, no. 1, pp. 52–56, 2022, doi: 10.47065/bulletincsr.v3i1.203.
[13] A. N. Sari and S. Alfionita, "Klasifikasi Penyakit Jantung Menggunakan Metode Naïve Bayes," *AMRI (Analisa Metod. Rekayasa Inform.*, vol. 1, no. 1, pp. 22–26, 2022, doi: 10.12487/AMRI.v1i1.xxxxx.
[14] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019.
[15] D. A. Ryfai, N. Hidayat, and E. Santoso, "Klasifikasi Tingkat Resiko Serangan Penyakit Jantung menggunakan Metode K-Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 10, pp. 4701–4707, 2022.
[16] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, "Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods," *Comput. Electr. Eng.*, vol. 84, p. 106628, 2020, doi: 10.1016/j.compeleceng.2020.106628.
[17] B. Ismanto and N. Amalia, "Peningkatan Akurasi Pada Modified K-NN Untuk Klasifikasi Pengajuan Kredit Koperasi Dengan Menggunakan Algoritma Genetika," *IC-Tech*, vol. 3, no. 2, pp. 66–70, 2018.
[18] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, no. October 2017, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.
[19] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. PP, pp. 1–15, 2022, doi: 10.1109/TNNLS.2021.3136503.
[20] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.
[21] Y. Yin, L. Long, and X. Deng, "Dynamic Data Mining of Sensor Data," *IEEE Access*, vol. 8, pp. 41637–41648, 2020, doi: 10.1109/ACCESS.2020.2976699.
[22] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. because*, vol. 2018-Janua, pp. 294–298, 2018, doi: 10.1109/ICITISEE.2017.8285514.