JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 5
No. 3
Desember
2019

Prediksi Penyakit Jantung Menggunakan Metode-Metode *Machine Learning* Berbasis *Ensemble – Weighted Vote*

Apriyanto Alhamad[#1], Azminuddin I. S. Azis[#2], Budy Santoso[#3], Sunarto Taliki[#4]

[#]*Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Ichsan Gorontalo*
Jl. Raden Saleh No. 17, Telp. (0435) 829975, Fax (0435) 829976, Gorontalo

[1]apriyanto86@unisan.ac.id
[2]azminuddinazis@unisan.ac.id
[3]budysantoso@unisan.ac.id
[4]sunartotaliki@unisan.ac.id

Abstrak— **Kematian yang disebabkan penyakit jantung masih sangat tinggi, sehingga perlu peningkatan upaya-upaya pencegahannya, misalnya dengan meningkatkan capaian model prediksinya. Penerapan metode-metode *machine learning* pada *dataset* publik (*Cleveland*, *Hungary*, *Switzerland*, *VA Long Beach*, & *Statlog*) yang umumnya digunakan oleh para peneliti untuk prediksi penyakit jantung, termasuk pengembangan alat bantunya, masih belum menangani *missing value*, *noisy data*, *unbalanced class*, dan bahkan *data validation* secara efisien. Oleh karena itu, pendekatan imputasi *mean/mode* diusulkan untuk menangani *missing value replacement*, *Min-Max Normalization* untuk menangani *smoothing noisy data*, *K-Fold Cross Validation* untuk menangani *data validation*, dan pendekatan *ensemble* menggunakan metode *Weighted Vote* (WV) yang dapat menyatukan kinerja tiap-tiap metode *machine learning* untuk mengambil keputusan klasifikasi sekaligus untuk mereduksi *unbalanced class*. Hasil penelitian ini menunjukkan bahwa metode yang diusulkan tersebut memberikan akurasi sebesar 85,21%, sehingga mampu meningkatkan kinerja akurasi metode-metode *machine learning*, selisih 7,14% dengan *Artificial Neural Network*, 2,77% dengan *Support Vector Machine*, 0,34% dengan C4.5, 2,94% dengan *Naïve Bayes*, dan 3,95% dengan *k-Nearest Neighbor*.**

Kata kunci— ***machine learning*, *weighted vote*, *ensemble*, prediksi penyakit jantung, *unbalanced class***

I. PENDAHULUAN

Data *World Health Organization* (WHO) pada tahun 2012 menunjukkan 17.5 juta jiwa (31%) penduduk di dunia meninggal akibat penyakit jantung dan jumlah tersebut akan terus meningkat [1], [2]. Sementara data *The Institute for Health Metrics and Evaluation* (IHME) pada tahun 2016 menunjukkan 17.7 juta jiwa (32.26%) penduduk di dunia meninggal karena penyakit jantung [3]. Survei dari *Sample Registration System* pada tahun 2014 di Indonesia menunjukkan Penyakit Jantung Koroner berada pada posisi tertinggi setelah Stroke sebagai penyebab kematian, yaitu sebesar 12.9% [1].

Dengan demikian, perlu peningkatan upaya-upaya pencegahannya, seperti deteksi dini penyakit jantung yang dapat dilakukan melalui diagnosa oleh dokter. Namun masyarakat umumnya malas melakukan cek kesehatan secara berkala, minimnya pengetahuan mereka mengenai penyakit jantung, dan kendala biaya dapat mempersulit deteksi dini penyakit jantung [4]. Alternatif alat bantu untuk prediksi penyakit jantung dapat menjadi salah satu solusi dari kesulitan tersebut. Oleh karena itu, peningkatan capaian model prediksi penyakit jantung perlu terus ditingkatkan sehingga dapat digunakan untuk pengembangan alat bantunya.

Pendekatan *Machine Learning* (ML) memiliki potensi besar untuk mengatasi masalah-masalah dalam domain biomedis komputasi [5], seperti prediksi penyakit jantung. *Dataset* publik yang umumnya digunakan oleh para peneliti dalam menemukan model ML yang paling akurat untuk prediksi penyakit jantung tersedia di *UCI Machine Learning Repository* (*Cleveland*, *Hungary*, *Switzerland*, *VA Long Beach*, & *Statlog*) [6]. Kelima *dataset* tersebut memiliki karakteristik variabel yang sama (secara rinci ditunjukkan pada Tabel 1), sehingga dapat disatukan.

Kecuali *Statlog dataset*, keempat *dataset* lainnya memiliki masalah *Missing Value* (MV) dan *Unbalanced Class* (UC) yang memang sering terjadi pada data klinis harus dikendalikan dengan sangat hati-hati, karena sangat berefek pada hasil prediksi [7], [8]. Sementara adanya *outlier* atau anomali pada data dapat menyebabkan *Noisy Data* (ND) atau data yang tidak konsisten [7], [9], berdampak buruk terhadap kinerja model [10]. Begitupun dengan *Data Validation* (DV), tekniknya mesti tepat, partisi data latih dan data uji harus menjamin bahwa semua sampel digunakan untuk pelatihan dan pengujian model, setiap label *class* mestinya terwakili secara merata pada pelatihan model [9]. Secara rinci, masalah MV dan UC

pada lima *dataset* prediksi penyakit jantung ditunjukkan pada Tabel 2.

TABEL I
VARIABEL INPUT DATASET PENYAKIT JANTUNG

Variabel	Tipe	Keterangan
Umur	Integer	29 – 77 tahun
Jenis Kelamin	Binominal	0 (Wanita); 1 (Pria)
Jenis Sakit Dada	Ordinal	1 (Typical Angina); 2 (Atypical Angina); 3 (Non Anginal Pain); 4 (Asymptomatic)
Tekanan Darah	Integer	80 – 200
Kolestrol	Integer	85 – 603
Kadar Gula	Binominal	0 (<=120mg/dl); 1 (>120mg/dl)
Elektrokardiografi	Ordinal	0 (Normal); 1 (Kelainan ST-T); 2 (Hypertrofi Ventrikel)
Tekanan Jantung	Integer	60 – 202
Angina Induksi	Binominal	0 (No); 1 (Yes)
Oldpeak	Real	0 – 6,2
Slope	Ordinal	1 (Naik); 2 (Datar); 3 (Turun)
Denyut jantung	Ordinal	0 – 3
Thal	Ordinal	3 (Normal); 6 (Cacat Tetap); 7 (Cacat Reversibel)

TABEL II
KARAKTERISTIK DATASET PENYAKIT JANTUNG

Code	Dataset	Instances	MV	UC
D1	Cleveland	303	6	No
D2	Hungary	294	782	Yes
D3	Switzerland	123	273	Yes
D4	VA Long Beach	200	698	Yes
D5	Statlog	270	0	No
	Total	1,190	1,759	Yes

Namun penerapan metode-metode ML dan varian-variannya pada *dataset* penyakit jantung masih kurang memperhatikan masalah MV, ND, UC, dan bahkan DV. Sehingga walaupun akurasi yang tinggi telah dicapai, namun belum efisien secara keseluruhan. Begitupun dari sisi pengembangan aplikasinya, tiga aplikasi berbasis *web* yang telah dikembangkan menggunakan metode ML (*Naïve Bayes*) yang justru akurasinya lebih rendah dan bahkan sangat tidak memperhatikan masalah MV, ND, UC, dan DV pula. Secara rinci, kinerja akurasi dari penerapan metode-metode ML yang diusulkan untuk prediksi penyakit jantung maupun pengembangan aplikasinya ditunjukkan pada Tabel 3. Sedangkan studi-studi survei pada topik prediksi penyakit jantung ditunjukkan pada Tabel 4.

Bahkan, akurasi 100% ditunjukkan *Hidden Naïve Bayes – Inter Quartile Range* pada *dataset Statlog* yang tidak memiliki masalah MV dan UC [11]. Namun mentransformasi data yang sudah terstruktur menjadi data *image* justru menyebabkan adanya MV yang mereka ganti dengan nilai *mean/mode*. Mentransformasi data yang sudah terstruktur menjadi tidak terstruktur itu justru malah bisa meningkatkan ND. Penelitian tersebut mengejar akurasi tapi mengorbankan efisiensi secara keseluruhan, karena justru menyebabkan *bias* pada data.

TABEL III
LITERATUR STUDI (TECHNICAL PAPERS)

Year, Ref.	Dataset	Methods	MV	UC	DV	Acc
2008, [12]	D1	NB	?	☑	⊗	95%/web
2010, [13]	D1	DT–GA	?	☑	☑	99,2%
2011, [14]	D1	DT–GA	?	☑	☑	99,2%
2012, [15]	D1&D5	ANN	☑	?	?	99,25%
2012, [16]	?	NB	?	?	?	web
2013, [4]	D1	NB	?	☑	⊗	95%/web
2015, [17]	D5	ANN	☑	☑	?	85%
2016, [11]	D5	HNB–IQR	⊗	☑	☑	100%
2016, [18]	D1	J48	☑	?	☑	56,76%
2016, [19]	D1	k-NN–Apriori	⊗	?	☑	99,19%
2016, [20]	D1-D4	Ensemble–ML	⊗	?	☑	92%
2016, [21]	D1	Rule methods	☑	?	☑	86,7%
2017, [22]	D2	DT	?	?	☑	67,7%
2018, [23]	D1	RF	☑	?	☑	83,15%
2018, [24]	D1	ICO	⊗	?	?	84,17%
2018, [25]	D1	Ensemble–ML	⊗	?	?	90%

Keterangan:
☑ Tepat/baik atau tidak perlu ditangani
? Tidak diketahui atau tidak teridentifikasi
⊗ Kurang/tidak tepat, tidak ditangani, diabaikan, atau dibuang

TABEL IV
LITERATUR STUDI (SURVEY PAPERS)

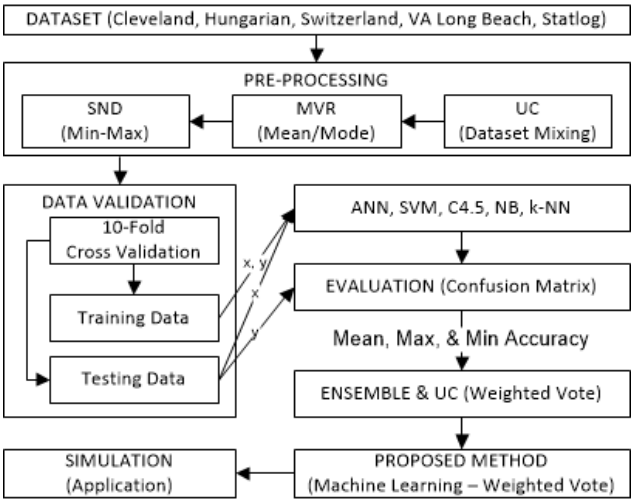
Year, Ref.	Authors	Proposed Method
2012, [26]	Shouman, Turner and Stocker	DT – GA = 99,2%
2012, [27]	Vijayarani	SVM
2016, [28]	Banu and Swamy	DT = 99,62 acc
2017, [29]	Gnaneswar and Jebarani	ANN
2017, [30]	Sowmiya and Sumitra	ML – Apriori

Membuang MV bisa menghilangkan informasi yang mungkin penting, mengakibatkan *bias* [31]. Pendekatan imputasi merupakan strategi yang efisien untuk menangani MV [9], [31]. Selanjutnya, cara klasik dan umum digunakan untuk mereduksi ND, yaitu pendekatan normalisasi [10]. Sementara itu, selain dapat memberikan keputusan prediksi yang lebih baik daripada yang diperoleh metode ML masing-masing, pendekatan *ensemble* juga dapat menangani masalah UC dengan baik [7], [32]. Penerapan pendekatan *ensemble* menggunakan metode *Weighted Vote* (WV) [7], *Mayotiry Vote* [25], maupun *Bagging*, *Boosting*, dan *Stacking* [20], terbukti memberikan akurasi yang lebih baik daripada metode ML masing-masing.

Dengan demikian, penelitian ini bertujuan untuk meningkatkan efisiensi dan kinerja metode-metode ML untuk prediksi penyakit jantung melalui *Missing Value Replacement* (MVR) menggunakan pendekatan inputasi *mean/mode*, *Smoothing Noisy Data* (SND) menggunakan pendekatan *Min-Max Normalization*, *Unbalanced Class Handling* (UCH) menggunakan pendekatan *ensemble* dengan metode *Weighted Vote*, dan *Data Validation* menggunakan metode *K-Fold Cross Validation*. Diharapkan metode yang diusulkan ini dapat digunakan untuk pengembangan sistem cerdas deteksi dini penyakit jantung, sehingga berdampak dalam mereduksi tingkat kematian yang disebabkan penyakit jantung.

II. METODE

Beberapa metode ML yang populer digunakan oleh para peneliti, yaitu *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), C4.5, *Naive Bayes* (NB), *dan k-Nearest Neighbor* (k-NN). Jika kinerja kelima metode tersebut disatukan dengan pendekatan *ensemble–Weighted Vote* yang sekaligus menangani UC, maka akan diperoleh suatu metode yang lebih handal daripada menggunakan salah satu dari metode ML tersebut. Terlebih lagi apabila data yang diolah telah melalui *pre-processing* yang efisien (MVR, SND, dan DV). Metode yang diusulkan ini ditunjukkan pada Gambar 1.



Gambar 1. Metode yang diusulkan

A. Mean/Mode untuk MVR

MV yang jumlahnya sedikit dapat diatasi dengan cara membuangnya. Namun jika MV jumlahnya banyak, seperti pada beberapa *dataset* prediksi penyakit jantung, maka membuangnya dapat menghilangkan informasi yang mungkin penting, mengakibatkan *bias* [31]. Pendekatan imputasi merupakan salah satu strategi yang dapat digunakan untuk mengatasi MV [9], [31]. Imputasi secara sederhana dapat dilakukan dengan mengganti MV menjadi nilai *mean* (1) pada variabel numerik dan nilai *mode* pada variabel kategorikal [7], [9], [31], [33].

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

B. Min-Max Normalization untuk SND

Adanya *outlier* atau anomali pada data dapat menyebabkan ND [7], [9], berdampak buruk terhadap kinerja model [10]. Intinya, *outlier* mestinya dibuang dengan cara mendeteksinya lebih dahulu, misalnya dengan model prediksi. Namun cara lain yang klasik dan dapat diterapkan untuk mereduksi ND yaitu dengan melakukan normalisasi data, salah satunya menggunakan pendekatan *Min-Max Normalization* (2) [10].

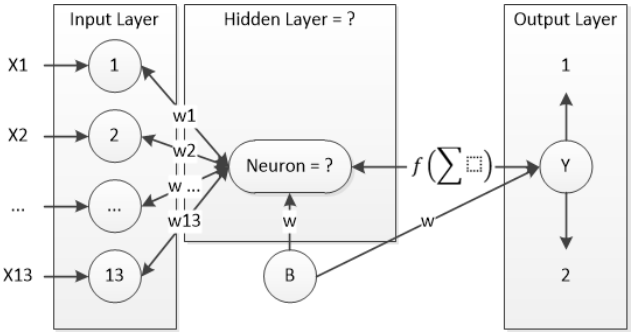
$$X_{ij}^{*} = \frac{(X_{ij} - X_{\min j})}{(X_{\max j} - X_{\min j})}\left((New_{\max j} - New_{\min j}) + New_{\min j}\right) \quad (2)$$

C. K-Fold Cross Validation untuk DV

Selanjutnya, strategi DV bukan hanya sekedar untuk memperoleh akurasi yang tinggi. Metode yang dapat menjamin bahwa semua sampel digunakan untuk pelatihan dan pengujian model, begitupun setiap label *class* terwakili secara merata pada pelatihan model, yaitu *K-Fold Cross Validation* dan *Leave One Out* [9]. Ketika ukuran sampel cukup besar, *K-Fold Cross Validation* adalah pilihan terbaik [9], [34].

D. Artificial Neural Network (ANN)

Metode ANN dapat digunakan untuk estimasi/regresi dan klasifikasi. Algoritma ANN yang digunakan adalah *Backpropagation* dengan arsitektur jaringan ditunjukkan pada Gambar 2.



Gambar 2. Jaringan ANN untuk prediksi penyakit jantung

Arsitektur jaringan tersebut menunjukkan bahwa suatu jaringan ANN memiliki tiga *layer*, yaitu *input layer*, *hidden layer*, dan *output layer*. Suatu jaringan ANN memiliki pula tiga komponen, yaitu *synapse* ($w_1, w_2, \ldots, w_n$), *adder*, dan fungsi aktifasi (*f*). Model matematika ketika komponen tersebut diformulasikan dengan Persamaan (3) yang diperkenalkan oleh Mc. Culloch & Pitts.

$$y = f(\sum_{i=1}^{n} x_i w_i) \quad (3)$$

Dimana signal *x* berupa vektor berdimensi *n* ($x_1, x_2, \ldots, x_n$) akan mengalami penguatan oleh *synapse w* ($w_1, w_2, \ldots, w_n$). Selanjutnya, akumulasi dari penguatan tersebut akan mengalami transformasi oleh fungsi aktifasi *f*. Fungsi *f* ini akan memonitor, bila akumulasi penguatan signal itu telah melebihi batas tertentu.

E. Support Vector Machine (SVM)

Metode SVM yang digunakan adalah algoritma Binary SVM, ditunjukkan pada Persamaan (4), dimana *n* merupakan jumlah data yang menjadi *support vector*, dibawah Kendala (5). Apabila data bersifat non-linier, maka pendekatan *kernel* dapat diterapkan.

$$\min_{w,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (4)$$

$$y_i(w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\dots,n \quad (5)$$

F. C4.5

Metode *Decision Tree* yang digunakan adalah algoritma C4.5. *Entropy* diperoleh melalui Persamaan (6), dimana *m* merupakan jumlah nilai yang berbeda pada *class* dan $P(\omega_i|s)$ merupakan proporsi *class* atau nilai fitur ke-*i* dalam data yang diproses di *node s*.

$$Entropy(s) = -\sum_{i=1}^{m} P(\omega_i|s) Log_2 P(\omega_i|s) \quad (6)$$

Sementara *gain* diperoleh melalui Persamaan (7), *split info* diperoleh melalui Persamaan (8), dan *gain ratio* diperoleh melalui Persamaan (9).

$$Gain(s,j) = E(s) - \sum_{i=1}^{n} P(v_i|s)\, E(s_i) \quad (7)$$

$$Split\ Info(s,j) = -\sum_{i=1}^{k} P(v_i|s) Log_2 P(v_i|s) \quad (8)$$

$$Ratio\ Gain(s,j) = \frac{G(s,j)}{SP(s,j)} \quad (9)$$

$P(v_i|s)$ merupakan proporsi nilai *v* muncul pada *class* dalam *node*. $E(s_i)$ merupakan *Entropy* komposisi nilai *v* dari *class* ke-*j* dalam data ke-*i* suatu *node*. Selanjutnya *n* merupakan jumlah nilai yang berbeda dalam *node*. Sementara *k* merupakan jumlah pemecahan.

G. Naïve Bayes (NB)

Metode NB (10) bekerja berdasarkan teori probabilitas, dimana $P(C_k)$ adalah probabilitas dari *class* ke-*k = 1, 2, …, l*. $P(X_{ij}|C_k)$ adalah probabilitas atribut ke-*j = 1, 2, …, m*. $C(X_i)$ adalah *class* hasil klasifikasi *instance ke-i = 1, 2, … n*.

$$C(X_i) = \underset{C_k \in Y}{\operatorname{argmax}} P(C_k) \prod_{j=1}^{m} P(X_{ij}|C_k) \quad (10)$$

Jika variabel bertipe numerik, maka distribusi *Gaussian* (11) dapat digunakan.

$$P(x_k|C_i) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (11)$$

H. k-Nearest Neighbor (k-NN)

Algoritma k-NN (12) bekerja berdasarkan perhitungan jarak atau kemiripan antar data.

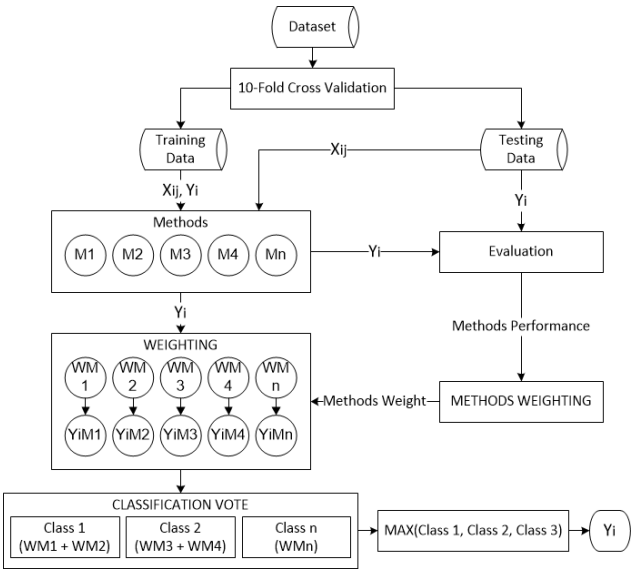
$$y' = \underset{v}{argmax} \sum_{i=1}^{n} x_i y_i \in D_z I(v = y_i) \quad (12)$$

Standarnya, perhitungan jarak mengunakan metode *Euclidean* (13). Simpan hasilnya di dalam *D*, dan pilih D_z ϵ *D* yang merupakan *K* tetangga terdekat dari *z*. Sementara *v* merupakan jumlah data yang masuk dalam *class* y_i.

$$d(X_{ij}) = \sqrt{\sum_{k=1}^{n}(X_{ik} - X_{jk})^2} \quad (13)$$

I. Ensemble - Weighted Vote (WV) untuk UC

Pendekatan *ensemble* dapat digunakan untuk menangani UC [7], [32]. Metode *ensemble* yang digunakan yaitu WV. Pada kasus klasifikasi, algoritma WV memberikan bobot untuk output dari tiap-tiap metode ML yang digunakan berdasarkan kinerja dari metode ML masing-masing, selanjutnya menjumlahkan hasil dari seluruh metode ML pada tiap-tiap label output, akhirnya memilih label output yang memiliki nilai terbesar. Cara kerja WV ditunjukkan pada Gambar 3.



Gambar 3. Cara Kerja Ensemble - WV

Algoritma WV adalah sebagai berikut:

1. Hitung akurasi rata-rata (1), maksimum, dan minimum dari *k-Fold Cross Validation* (k=10) untuk tiap-tiap metode *machine learning* yang digunakan.
2. Normalisasikan nilai akurasi rata-rata, maksimum, dan minimum tersebut menggunakan *Min-Max Normalization* (2) dengan *range* [0.1, 1] untuk tiap-tiap metode *machine learning* yang digunakan.
3. Hitung selisih antara m_i dengan 1 (14), dimana m_i adalah hasil normalisasi metode *machine learning* ke-*i = 1, 2, …, j*.

$$m_i = 1 - m_i \quad (14)$$

4. Hitung *weight* (15) untuk tiap-tiap metode *machine learning* yang digunakan.

$$w_i = \frac{m_i}{\sum_{i=1}^{j} m_i} \quad (15)$$

5. Dapatkan keputusan klasifikasi (16), dimana $y(x_d)$ adalah hasil keputusan klasifikasi (label *class*) *instance*

ke-*d* dari metode *Weighted Vote*, y_c adalah label *class* ke-*c*, y_i adalah label *class* prediksi m_i (metode ke-i = *1, 2, …, j*).

$$y(x_d) = \underset{y_c \epsilon Y}{\operatorname{argmax}} \sum_{i=1}^{j} w_i(if\ y_i = y_c) \quad (16)$$

J. Evaluasi Model Klasifikasi

Pengukuran kinerja suatu model klasifikasi dapat menggunakan *Confusion Matrix* untuk memperoleh nilai *accuracy*, *recall* (*specificity* dan *sensitivity*), *precision*, dan *F-Measure* yang ditunjukkan pada Tabel 5.

TABEL V
CONFUSION MATRIX

	Actual +	Actual -	Precision
Predicted +	TP	FP	TP/(TP+FP) *
Predicted -	FN	TN	TN/(TN+FN)
Recall	TP/(TP+FN)	TN/(TN+FP)	
F-Measure	(2*Precision*Sensitivity)/(Precision+Sensitivity)		
Accuracy	(TP+TN) / (TP+TN+FN+FP)		

Keterangan: T (True); F (False); P (Positive); N (Negative)

III. HASIL PENELITIAN

Tools yang digunakan untuk ekperimen dan pengembangan aplikasi (simulasi) yaitu Matlab. Prosedur yang dilakukan sesuai dengan langkah-langkah dalam metode yang diusulkan, yaitu: (1) Pra pengolahan yang terdiri dari *dataset mixing*, MVR, SND, dan DV; (2) Pemodelan dan Evaluasi; dan (3) Simulasi.

A. Hasil Pra Pengolahan

Langkah pertama yang dilakukan setelah mengumpulkan data yaitu *dataset mixing* untuk menangani masalah UC (ditunjukkan pada Tabel 6 dan Tabel 7). Selain dengan pendekatan *dataset mixing*, masalah UC ditangani pula dengan pendekatan *ensemble* menggunakan metode WV.

TABEL VI
VARIABEL OUTPUT (Y)

Dataset	Class Label	UC
D1	0 (Sehat)	54,13%
	1 (Resiko Rendah)	18,15%
	2 (Resiko Sedang)	11,88%
	3 (Resiko Tinggi)	11,55%
	4 (Resiko Sangat Tinggi)	4,29%
D2	0 (Sehat)	63,95%
	1 (Sakit)	36,05%
D3	0 (Sehat)	6,50%
	1 (Resiko Rendah)	39,02%
	2 (Resiko Sedang)	26,02%
	3 (Resiko Tinggi)	24,39%
	4 (Resiko Sangat Tinggi)	4,07%
D4	0 (Sehat)	25,50%
	1 (Resiko Rendah)	28,00%
	2 (Resiko Sedang)	20,50%
	3 (Resiko Tinggi)	21,00%
	4 (Resiko Sangat Tinggi)	5,00%
D5	1 (Negatif)	55,56%
	2 (Positif)	44,44%

TABEL VII
DATASET MIXING

Dataset	Class Label	UC	
D1	0 ← 0; 1 ← 1, 2, 3, 4	54,13%; 45,87%	No
D2	0 ← 0; 1 ← 1	63,95%; 36,05%	Yes
D3	0 ← 0; 1 ← 1, 2, 3, 4	6,50%; 93,50%	Yes
D4	0 ← 0; 1 ← 1, 2 ,3 ,4	25,50%; 74,50%	Yes
D5	0 ← 1; 1 ← 2	55,56%; 44,44%	No
All	**0; 1**	**41,13%; 58,87%**	**No**

Label output (*class*) tiap-tiap *dataset* berbeda-beda namun memiliki karakteristik yang sama. Dengan demikian, masalah UC yang terjadi pada semua *dataset* kecuali *Statlog dataset* diatasi dengan menggabungkan data yang termasuk *class* 1 (ada indikasi positif penyakit jantung), 2 (>1), 3 (>2), dan 4 (positif penyakit jantung) menjadi *class* 1, sementara *class* 0 tetap *class* 0, sehingga akhirnya antara *class* 0 dan 1 menjadi seimbang. Hal ini juga membuat *Statlog dataset* yang hanya memiliki *class* 1 (negatif penyakit jantung) dan 2 (positif penyakit jantung) yang tidak memiliki masalah UC dapat digabungkan dengan *dataset* lainnya dengan cara yang sama.

Sementara masalah ND diatasi dengan melakukan normalisasi (*Min-Max Normalization*) data dalam jangkauan 0-1. Misalnya nilai x_{ij} = 8, nilai maksimum dari atribut j = 10, dan nilai minimum dari atribut j = 1, maka normalisasi data dengan jangkauan [0 1] adalah:

$$\frac{(8-1)}{(10-1)}\big((1-0)+0\big) = 0{,}78$$

Tabel 8 menunjukkan hasil normalisasi data yang telah dilakukan menggunakan metode *Min-Max Normalization*.

TABEL VIII
HASIL NORMALISASI DATA

No	X1	X2	X3 - X 12	X13	Y
1	0,71	1	…	0,75	0
2	0,80	1	…	0,00	1
…	…	…	…	…	…
1190	0,80	1	…	0,00	1

DV dilakukan menggunakan metode *K-Fold Cross Validation* dengan parameter *K=10*. Metode ini dapat menjamin bahwa setiap *class* dan data terbagi merata untuk data latih dan data uji di setiap iterasi *K*. Metode lainnya yang juga mampu melakukan hal tersebut yaitu *Leave One Out*. Namun metode tersebut memberikan kompleksitas komputasi yang sangat besar, terlebih pada *dataset* yang berukuran besar. Oleh karena itu, metode ini menjadi pilihan yang lebih efisien. Penerapan *K-Fold Cross Validation* untuk DV juga mendukung penerapan *ensemble* – WV.

B. Hasil Pemodelan, Evaluasi, dan Pembahasan

Opsi percobaan yang dilakukan ditunjukkan pada Tabel 9. Kami tidak melakukan komparasi kinerja terhadap opsi dibuangnya MV, karena ukuran *dataset* akan berbeda. Selain itu, membuang MV bukan solusi yang benar jika *dataset* memiliki terlalu banyak MV. Tabel 2 menunjukkan

begitu banyak jumlah MV pada *dataset*, sehingga membuangnya merupakan suatu kesalahan. Membuang MV yang banyak tentu saja dapat membuang beberapa informasi yang mungkin penting, sehingga dapat menimbulkan *bias*. Oleh karena itu kami melakukan MVR.

TABEL IX
OPSI EKSPERIMEN

A	B (proposed)	C (not compare)
Dataset Mixing	Dataset Mixing	Dataset Mixing
MVR	MVR	No MVR (Removed)
No SND	SND	No SND
ML-WV	ML-WV	ML-WV

Parameter terbaik yang diberikan tiap-tiap metode ML untuk prediksi penyakit jantung ditunjukkan pada Tabel 10.

TABEL X
PARAMETER TERBAIK METODE-METODE ML

Method	Parameters
ANN	hd=1; hd neurons=50; error target=1e-6; epoch=100
SVM	kernel=rbf
C4.5	discretization=Entropy
NB	numeric probability=Gaussian
k-NN	k=5; distance measure=Euclidean; vote rule=Nearest

Kinerja tiap-tiap metode diukur/dievaluasi berdasarkan tingkat *accuracy*, *precision*, *recall* (*specificity* dan *sensitivity*) menggunakan teknik *Confusion Matrix*. Hasil evaluasi metode-metode ML dan metode yang diusulkan (ML–WV) pada opsi percobaan A dan B ditunjukkan pada Tabel XI.

TABEL XI
HASIL EVALUASI

Method	Precision		Specificity		Sensitivity		Accuracy	
	A	B	A	B	A	B	A	B
ANN	75,77	70,74	69,15	80,77	76,34	76,70	70,40	78,07
SVM	**93,59**	**92,86**	74,58	77,98	**92,99**	**92,70**	80,25	82,44
C4.5	80,94	83,24	79,75	**84,62**	82,82	85,34	81,26	84,87
NB	81,82	82,35	80,04	80,55	83,55	83,99	81,76	82,27
k-NN	67,22	80,55	70,28	80,05	71,81	82,65	70,93	81,26
ML-WV	85,93	85,01	**82,73**	84,07	87,26	86,60	**84,79**	**85,21**

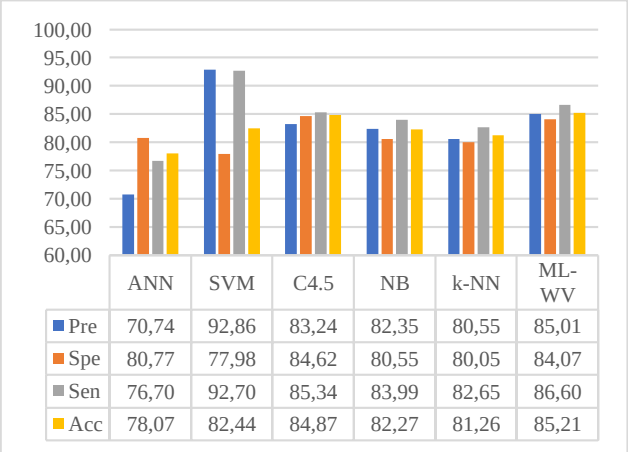
Hasil evaluasi menunjukkan bahwa kinerja akurasi terbaik diberikan oleh metode ML–WV sebagai metode yang diusulkan, dengan akurasi = 85,21% pada opsi percobaan B, lebih baik dari pada opsi percobaan A. Sedangkan kinerja *precision* dan *sensitivity* terbaik diberikan oleh SVM, dengan *precision* = 93,59% pada opsi percobaan A yang lebih baik dari pada opsi percobaan B dan *sensitivity* = 92,99% pada opsi percobaan A yang lebih baik pula dari pada opsi percobaan B. Sementara itu, kinerja *specificity* terbaik diberikan oleh C4.5, dengan *specificity* = 84,62% pada opsi percobaan B yang lebih baik dari pada ML–WV pada opsi percobaan A.

Kinerja *precision*, *specificity*, dan *sensitivity* ML–WV tidak lebih baik dari pada metode-metode ML karena cara kerja WV hanya menggunakan nilai akurasi (nilai rata-rata akurasi), nilai akurasi maksimum, dan nilai akurasi minimum metode-metode ML yang digunakan dari setiap iterasi *k-Fold Cross Validation*. Selain itu, tujuan WV dalam hal ini untuk meningkatkan akurasi metode-metode ML, sekaligus mereduksi UC. Berikut ini adalah contoh perhitungan ML-WV (Tabel 12). Akurasi metode-metode ML yang digunakan adalah akurasi dari opsi percobaan B sebagai opsi percobaan yang memberikan akurasi terbaik terhadap metode-metode ML.

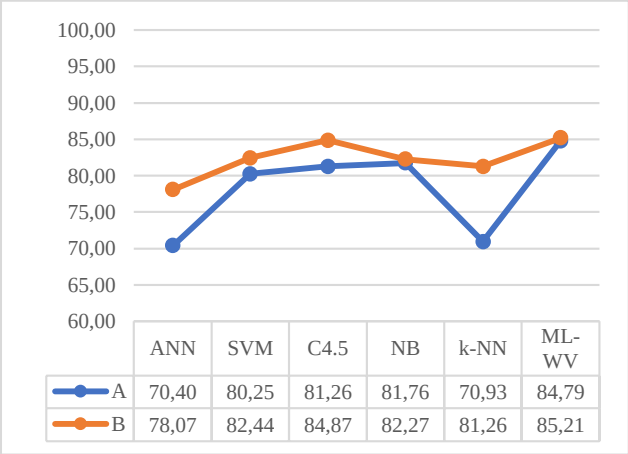
TABEL XII
PERHITUNGAN MANUAL ML-WV

		ANN	SVM	C4.5	NB	k-NN	Jumlah
K-Fold Cross Validation	K=10						
	1	76,01	85,20	99,29	72,90	81,10	
	2	73,01	86,10	78,47	91,09	89,70	
	3	91,01	89,30	88,25	86,80	78,20	
	4	67,02	79,50	79,64	77,10	79,70	
	5	67,00	72,90	87,63	72,70	79,30	
	6	82,02	85,70	78,72	74,20	79,90	
	7	86,00	78,50	87,87	88,71	91,40	
	8	93,60	80,10	79,88	92,30	77,20	
	9	67,00	91,30	89,76	76,50	77,50	
	10	78,00	75,80	79,20	90,40	78,60	
Mean		78,07	82,44	84,87	82,27	81,26	
Max		93,60	91,30	99,29	92,30	91,40	
Min		67,00	72,90	78,47	72,70	77,20	
Normalization		0,42	0,52	0,31	0,49	0,29	
		0,58	0,48	0,69	0,51	0,71	2,98
Weight		0,20	0,16	0,23	0,17	0,24	1,00
Pred	Class 0	0	0	1	1	1	
	Class 1	1	1	0	0	0	
Vote	Class 0	0,00	0,00	0,23	0,17	0,24	0,64
	Class 1	0,20	0,16	0,00	0,00	0,00	0,36
Maka keputusan klasifikasi adalah Class 0 = 0,64							

Kinerja tiap-tiap metode pada opsi percobaan B sebagai opsi terbaik ditunjukkan pada Gambar 4. Sedangkan selisih akurasi antara opsi percobaan B dan A dari tiap-tiap metode ditunjukkan pada Gambar 5. Hasil evaluasi menunjukkan bahwa adanya peningkatan kinerja akurasi setiap metode setelah dilakukan MVR dan SND. Hasil evaluasi juga menunjukkan bahwa metode yang diusulkan (ML–WV) memberikan kinerja akurasi yang lebih baik dari pada metode ML masing-masing, baik dilakukan MVR dan SND ataupun tidak.



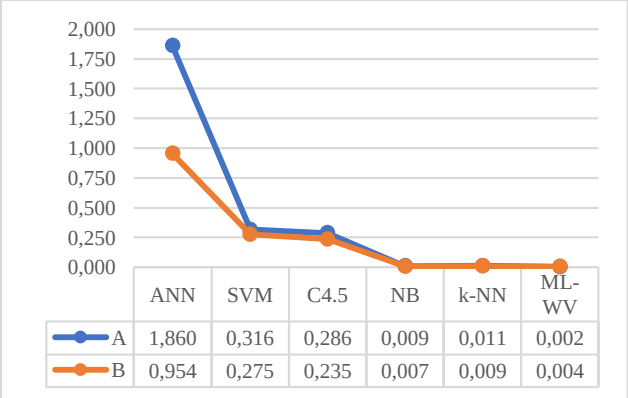
Gambar 4. Kinerja tiap-tiap metode pada opsi percobaan B



	ANN	SVM	C4.5	NB	k-NN	ML-WV
A	70,40	80,25	81,26	81,76	70,93	84,79
B	78,07	82,44	84,87	82,27	81,26	85,21

Gambar 5. Akurasi tiap-tiap metode pada opsi percobaan A dan B

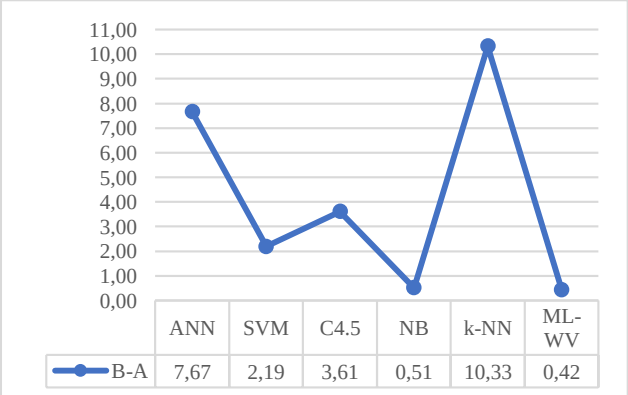
Sementara itu, waktu proses (detik) yang dibutuhkan tiap-tiap metode ditunjukkan pada Gambar 6. Metode ANN membutuhkan waktu proses yang paling lama. Waktu proses juga menunjukkan bahwa adanya penurunan waktu proses setelah dilakukan MVR dan SND.



	ANN	SVM	C4.5	NB	k-NN	ML-WV
A	1,860	0,316	0,286	0,009	0,011	0,002
B	0,954	0,275	0,235	0,007	0,009	0,004

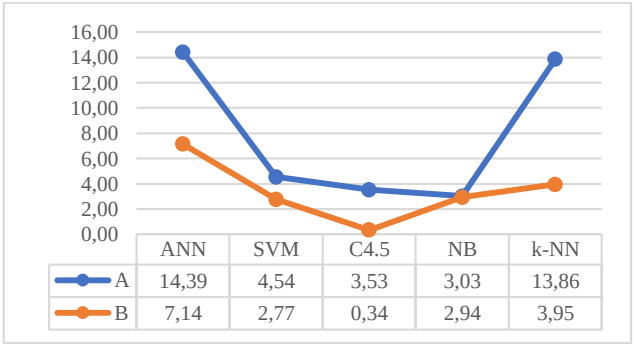
Gambar 6. Waktu proses antara opsi percobaan A dan B

Hasil evaluasi menunjukkan peningkatan kinerja akurasi setelah dilakukan MVR dan SND (opsi percobaan B) pada setiap metode yang selisihnya ditunjukkan pada Gambar 7. Sedangkan metode yang diusulkan (ML-WV) juga menunjukkan peningkatan kinerja akurasi dari tiap-tiap metode ML yang selisihnya ditunjukkan pada Gambar 8.



	ANN	SVM	C4.5	NB	k-NN	ML-WV
B-A	7,67	2,19	3,61	0,51	10,33	0,42

Gambar 7. Selisih akurasi metode antara opsi percobaan B – A



	ANN	SVM	C4.5	NB	k-NN
A	14,39	4,54	3,53	3,03	13,86
B	7,14	2,77	0,34	2,94	3,95

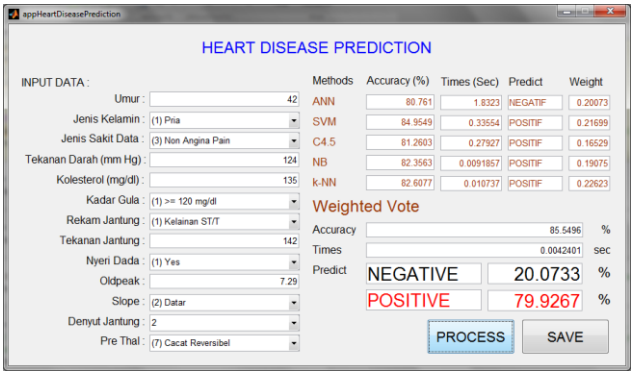
Gambar 8. Selisih akurasi ML-WV dengan metode-metode ML

C. *Simulasi*

Setelah model terbaik untuk prediksi penyakit jantung diperoleh, tahap terakhir yang dilakukan adalah melakukan simulasi terhadap model tersebut yang ditunjukkan pada Gambar 9. Adapun prosedur menggunakan aplikasi simulasi ini adalah sebagai berikut:

1. *User* menginput data deteksi penyakit jantung.
2. *User* menekan tombol *process* yang akan menampilkan informasi:
   - Keputusan atau output deteksi dini penyakit jantung, apakah negatif atau positif kemungkinan terkena penyakit jantung.
   - Output tersebut diberikan oleh tiap-tiap metode ML dengan tingkat akurasinya masing-masing.
   - Output dengan tingkat kepercayaan yang paling tinggi diberikan oleh metode ML-WV dengan tingkat bobot dari tiap-tiap label output (negatif atau positif), label output yang memiliki tingkat bobot tertinggi adalah keputusan deteksi (berwarna merah).
3. *User* menekan tombol *save* untuk menyimpan data tersebut.

Selanjutnya, pengembangan aplikasi untuk deteksi dini penyakit jantung dapat dikembangkan bagi masyarakat umum, sehingga masyarakat tidak perlu mengeluarkan biaya yang besar untuk itu. Selain itu, dapat pula memudahkan pekerjaan dokter jantung yang tidak mudah dijumpai di daerah-daerah. Bahkan bisa jadi dapat mereduksi tingkat kematian yang disebabkan penyakit jantung, dimana penyakit ini termasuk jenis penyakit *silent killer*, sehingga deteksi dininya sangatlah dibutuhkan.



Gambar 9. Simulasi deteksi dini penyakit jantung berbasis ML–WV

IV. KESIMPULAN

Berdasarkan hasil penelitian yang diperoleh, maka dapat disimpulkan bahwa:

1. Penerapan pendekatan *mean/mode* untuk MVR, dan penerapan pendekatan *Min-Max Normalization* untuk SND mampu meningkatkan kinerja akurasi metode-metode ML dengan selisih 7.67% pada ANN, 2.19% pada SVM, 3.61 pada C4.5, 0.51 pada NB, dan 10.33 pada k-NN.
2. Selanjutnya penerapan pendekatan *ensemble* menggunakan metode WV untuk menyatukan kinerja tiap-tiap metode ML dalam mengambil keputusan klasifikasi melalui pembobotan dan untuk mereduksi UC memberikan akurasi sebesar 85.21%, lebih tinggi dari setiap metode-metode ML, yaitu selisih 7.14% dengan ANN, 2.77% dengan SVM, 0.34% dengan C4.5, 2.94% dengan NB, dan 3.95% dengan k-NN.
3. Dengan demikian, penerapan pendekatan *ensemble* – WV, MVR, dan SND mampu meningkatkan kinerja metode-metode ML.

Untuk lebih menguji kinerja dari metode yang diusulkan ini, maka penerapannya pada berbagai objek yang lain dengan karakteristik masalah yang berbeda-beda dapat diuji coba pada penelitian-penelitian selanjutnya. Selain itu, teknik pembobotan pada WV dengan pendekatan yang berbeda dapat diuji coba pula. Model prediksi penyakit jantung yang diusulkan ini dapat pula digunakan untuk pengembangan alat bantu deteksi dini penyakit jantung untuk mereduksi kematian yang disebabkan penyakit jantung.

UCAPAN TERIMA KASIH

Penelitian ini didukung dan didanai oleh: (1) Direktorat Riset dan Pengabdian Masyarakat; (2) Kementrian Riset dan Pendidikan Tinggi Republik Indonesia.

REFERENSI

[1] Kementerian Kesehatan Republik Indonesia, "Penyakit Jantung Penyebab Kematian Tertinggi, Kemenkes Ingatkan CERDIK," 2017. [Online]. Available: http://www.depkes.go.id/article/view/17073100005/penyakit-jantung-penyebab-kematian-tertinggi-kemenkes-ingatkan-cerdik-.html. [Accessed: 19-Aug-2018].

[2] PT. Jawa Pos Grup Multimedia Redaksi, "Sepertiga Kematian di Dunia Dipicu Penyakit Jantung, Angkanya Segini," 2018. [Online]. Available: https://www.jawapos.com/kesehatan/29/09/2017/sepertiga-kematian-di-dunia-dipicu-penyakit-jantung-angkanya-segini. [Accessed: 19-Aug-2018].

[3] katadata, "Penyakit Kardiovaskular, Penyebab Kematian Terbanyak di Dunia," 2018. [Online]. Available: https://databoks.katadata.co.id/datapublish/2018/03/13/penyakit-kardiovaskular-penyebab-kematian-terbesar-di-dunia. [Accessed: 30-Aug-2018].

[4] S. H. Ishtake and S. A. Sanap, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," *Int. J. Healthc. Biomed. Res.*, vol. 1, no. 3, pp. 94–101, 2013.

[5] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: big data for personalized healthcare," in *IEEE Journal of Biomedical and Health Informatics*, 2015, vol. 19, no. 4, pp. 1209–1215.

[6] University of California Irvine Machine Learning Repository, "Heart Disease Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease.

[7] S. Bashir, U. Qamar, and F. H. Khan, "Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble," *Qual. Quant.*, vol. 49, no. 5, pp. 2061–2076, 2015.

[8] I. Fakhruzi, "An Artificial Neural Network with Bagging to Address Imbalance Datasets on Clinical Prediction," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, no. 1, pp. 895–898.

[9] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting Breast Cancer Recurrence Using Machine Learning Tehniques: A Systematic Review," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 52:1-52:40, 2016.

[10] M. M. Suarez-Alvarez, D.-T. Pham, M. Y. Prostov, and Y. I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," in *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2012, vol. 468, no. 2145, pp. 2630–2651.

[11] M. A. Jabbar and S. Shirina, "Heart disease prediction system based on hidden naïve bayes classifier," in *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, 2016.

[12] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 8, pp. 343–350, 2008.

[13] M. Anbarasi, E. Anupriya, and N. C. S. N. Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.

[14] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

[15] C. S. Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, 2012.

[16] S. A. Pattekari and A. Parveen, "Prediction System for Heart Disease using Naive Bayes," *Int. J. Adv. Comput. Math. Sci.*, vol. 3, no. 3, pp. 290–294, 2012.

[17] E. O. Olaniyi, O. K. Oyedotun, and A. Helwan, "Neural Network Diagnosis of Heart Disease," in *International Conference on Advances in Biomedical Engineering (ICABME)*, 2015, pp. 21–24.

[18] J. Patel, T. Upadhyay, and S. Patel, "Heart Disease Prediction Using Machine Learning and Data Mining Technique," *Comput. Sci. Electron. Journals*, vol. 7, pp. 129–137, 2016.

[19] J. Singh, A. Kamra, and H. Singh, "Prediction of Heart Diseases Using Associative Classification," in *International Conference on Wireless Networks and Embedded Systems (WECON)*, 2016.

[20] R. El Bialy, M. A. Salama, and O. Karam, "An ensemble model for Heart disease data sets: a generalized model," in *Proceedings of the 10th International Conference on Informatics and Systems - INFOS '16*, 2016, pp. 191–196.

[21] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," in *Procedia Computer Science*, 2016, vol. 85, pp. 962–969.

[22] S. Ekiz and P. Erdogmus, "Comparative Study of Heart Disease Classification," in *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2017.

[23] D. Kinge and S. K. Gaikwad, "Survey on Data Mining Techniques for Disease Prediction," *Int. Res. J. Eng. Technol.*, vol. 5, no. 1, pp. 630–636, 2018.

[24] G. Manikandan, A. Vasudev, and A. Balasubramanian, "A Survey to Identify an Efficient Classification Algorithm for Heart Disease Prediction," *Int. J. Pure Appl. Math.*, vol. 119, no. 12, pp. 13337–13345, 2018.

[25] R. A. Kurian and K. S. Lakshmi, "An Ensemble Classifier for the Prediction of Heart Disease," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 3, no. 6, pp. 25–31, 2018.

[26] M. Shouman, T. Turner, and R. Stocker, "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment," in *Japan-Egypt Conference on Electronics, Communications and Computers*, 2012, pp. 189–193.

[27] S. Vijayarani, "A Study of Heart Disease Prediction in Data Mining," *Int. J. Comput. Sci. Inf. Technol. Secur.*, vol. 2, no. 5, pp. 1041–1045, 2012.

[28] N. K. S. Banu and S. Swamy, "Prediction of Heart Disease at early

stage using Data Mining and Big Data Analytics: A Survey," in *nternational Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, 2016, pp. 256–261.

[29] B. Gnaneswar and M. R. E. Jebarani, "A Review on Prediction and Diagnosis of Heart Failure," in *International Conference on Innovations in Information, Embedded and Communication System (ICIIECS)*, 2017.

[30] C. Sowmiya and P. Sumitra, "Analytical Study of Heart Disease Diagnosis Using Classification Techniques," in *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2017.

[31] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, 2003.

[32] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, 2010.

[33] S. Zhang, Z. Jin, and X. Zhu, "Missing data imputation by utilizing information within incomplete instances," *J. Syst. Softw.*, vol. 84, no. 3, pp. 452–459, 2011.

[34] M. S. Santos, P. H. Abreu, P. J. Garcia-Laencina, A. Simao, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *J. Biomed. Inform.*, vol. 58, pp. 49–59, 2015.