JUSIKOM PRIMA (Jurnal Sistem Informasi dan Ilmu Komputer Prima)
Vol. 5 No. 2, Februari 2022 E-ISSN : 2580-2879

PREDIKSI PENYAKIT GAGAL JANTUNG DENGAN MENGGUNAKAN RANDOM FOREST

Edric, Saut Parsaoran Tamba
Universitas Prima Indonesia
Jalan Sampul, Medan, Indonesia
E-mail : *sautparsaorantamba@unprimdn.ac.id

ABSTRAK- Lebih dari 4 dari 5 kematian atas *Cardiovascular disease* (CVD) pada jantung dan pembuluh darah yang termasuk diataranya: *coronary heart disease*, *cerebrovascular disease, rheumatic heart disease*, dan kondisi jantung lainnya. Faktor resiko penyakit ini seringnya disebabkan oleh diet yang tidak sehat, kurang berolahraga, serta penggunaan rokok dan alkohol yang berlebih. Pada penelitian ini, tim peneliti memutuskan untuk memprediksi probabilitas penyakit gagal jantung menggunakan *Random Forest*. Data yang dipakai untuk melatih algoritma *Random Forest* yang dipakai diambil dari kompilasi beberapa observasi yang mana total akhirnya berjumlah 918 observasi dengan 12 atribut. Adapun tujuan dari penelitian ini adalah untuk menunjukkan kemampuan *Random Forest* dalam memprediksi penyakit gagal jantung dengan hasil peforma yang baik serta membuat model klasifikasi yang sederhana namun beperforma baik. Manfaat dari penelitian ini adalah untuk menjembatani penelitian deteksi dan ekstraksi fitur EKG sebelumnya sehingga mampu dimanfaatkan dan dikembangkan untuk tahap selanjutnya hingga produk siap pakai. Proses implementasi algoritma Random Forest yang digunakan sukses dengan meraih tingkat akurasi sebesar 82,6087% yang kemudian dioptimasi dengan teknik *K-Fold* dan *GridSearchCV* menjadi 85,058%. Sistem klasifikasi dengan *Random Forest* yang dibuat dapat dimplementasikan kedalam rangkaian alat untuk menciptakan alat deteksi aritmia EKG otomatis *portable*. Hubungan antara informasi yang disajikan oleh alat juga dengan sukses dibuktikan kontribusi pentingnya terhadap diagnosa positif negatifnya seseorang mempunyai penyakit jantung terutama melalui Gelombang ST.

Kata kunci : Prediksi, penyakit gagal jantung, random forest.

1. PENDAHULUAN

Penyakit gagal jantung dapat disebabkan oleh beberapa kondisi penyakit jantung. Menurut WHO, penyakit jantung atau yang disingkat sebagai CVD (*Cardiovascular disease)* adalah kelompok gangguan pada jantung dan pembuluh darah yang termasuk diataranya: *coronary heart disease, cerebrovascular disease, rheumatic heart disease,* dan kondisi jantung lainnya. Lebih dari 4 dari 5 kematian atas CVD disebabkan oleh gagal jantung dan *strokes,* dan sepertiga dari angka kematian tersebut merupakan orang yang mati premature dibawah umur 70 [1,2]. Faktor resiko penyakit ini seringnya disebabkan oleh diet yang tidak sehat, kurang berolahraga, serta penggunaan rokok dan alkohol yang berlebih. Sifatnya yang menyerang mendadak menyebabkan dikembangkannya teknologi yang dapat memprevensi. EKG sebagai salah satu teknologi hasil yang dikembangkan berkontribusi memberikan tanda dan informasi akan aktifitas aliran listrik pada jantung.

Pada penelitian ini, tim peneliti memutuskan untuk memprediksi probabilitas penyakit gagal jantung menggunakan Random Forest. Data yang dipakai untuk melatih algoritma Random Forest yang dipakai diambil dari kompilasi beberapa observasi yang mana total akhirnya berjumlah 918 observasi dengan 12 atribut [3].

Pemilihan algoritma Random Forest dikarenakan algoritma tersebut adalah salah satu dari peringkat 10 terbesar algoritma *Machine Learning*. Kualitas algoritma yang dipilih diharapkan mampu menghasilkan hasil prediksi yang aktual, akurat dan dapat diandalkan. Penelitian ini juga diharapkan sebagai jembatan antar penelitian deteksi dan ekstraksi fitur EKG yang sebelumnya sering dilakukan yang diharapkan pada akhirnya dapat tercipta suatu model lengkap dari alat hingga klasifikasi dengan desain *user interface* yang menarik sehingga mampu dipakai dan diterapkan oleh para tenaga medis maupun untuk *personal use* [4,5].

Adapun tujuan dari penelitian ini adalah untuk menunjukkan kemampuan Random Forest dalam memprediksi penyakit gagal jantung dengan hasil peforma yang baik serta membuat model klasifikasi yang sederhana namun beperforma baik.

Dalam penelitian ini, terdapat batasan masalah dari permasalahan yang ada, yaitu :

1. Dataset terdiri dari 918 observasi dengan 12 atribut yang merupakan integrasi antar 11 fitur umum dari beberapa observasi, diantaranya: 303 observasi Cleveland, 294 observasi Hungarian, 123 observasi Switzerland, 200 observasi Long Beach VA, dan 270 observasi Stalog (Heart) Data Set. Total keseluruhan terdapat 1190 observasi namun diantaranya terdapat 272 data observasi duplikat / sama.
2. Algoritma yang digunakan adalah Random Forest.

3. Bahasa pemrograman yang digunakan adalah bahasa Python dengan memanfaatkan beberapa *library* yang telah ada, yaitu: *pandas, numpy, seaborn, matplotlib, sklearn, yellowbrick, colorama,* dan *termcolor.*
4. Hasil penelitian difokuskan pada penerapan algoritma Random Forest yang kemudian akan dilakukan validasi dengan teknik *cross validation* dan GridCV. Data independen terdiri dari 11 kolom hasil integrasi fitur penyakit jantung yang umum dan sama. Dataset yang dipakai tidak memiliki data hilang dan data dependen yang akan dijadikan sebagai target adalah kolom HeartDisease yang sudah bertipe data Boolean. Dari keseluruhan dataset terdapat 5 kolom bertipe data kategori dan sisanya bertipe integer. Data kategori tersebut nantinya akan diberi data *dummy* (*dummy variables*) terlebih dahulu.

Penelitian yang berkaitan sebelumnya dimulai dengan penelitian dengan judul "*Abnormalities State Detection from P-Wave, QRS Complex, and T-Wave in Noisy ECG*" [6]. Tujuan penelitian tersebut adalah membangun sistemasi algoritma yang dapat diterapkan kedalam alat portabel. Penelitian tersebut secara spesifik meneliti tentang proses filterasi EKG hingga ekstraksi fitur yang dilakukan pada subjek dengan tiga kondisi (duduk-istirahat, berjalan dan jalan cepat). Hasil penelitian tersebut yang menggunakan FIR dan Butterworth filter menunjukan data sinyal mentah yang direkam berhasil di filter dengan baik dan desain model ekstraksi fitur QRS complex dapat dengan akurat menandai puncak QRS. Kesimpulan penelitian yang didapat menunjukkan bahwa data EKG dan ekstraksi fitur yang dilakukan belum diterapkan algoritma klasifikasi dan belum diimplementasi kedalam alat portable siap pakai.

Penelitian berikutnya dari tahapan yang sama mengubah penerapan dari subjek manusia remaja-dewasa menjadi fetus dalam kandungan (ibu hamil) dengan judul "*An Application of Modified Filter Algorithm Fetal Electrocardiogram Signals with Various Subjects*". Penelitian tersebut berfokuskan pada proses deteksi sinyal fetus dalam kandungan untuk mendapat hasil EKG dini. Hasil penelitian menunjukkan proses filterasi noise terhadap FEKG yang direkam namun tidak meneruskannya kedalam tahap ekstraksi fitur maupun klasifikasi EKG [7].

Penelitian yang ditulis peneliti ini diharapkan dapat dijadikan sebagai penyelesaian tahapan proses sinyal EKG yaitu proses klasifikasi EKG berdasarkan data aktual dan data medis dengan diagnosa yang telah dilakukan oleh professionalnya. Hasil penelitian ini dapat juga digunakan sebagai bahan acuan untuk melanjutkan penelitian sebelumnya ke tahap selanjutnya. Proses klasifikasi pada proses sinyal EKG adalah yang terpenting untuk dapat memberikan informasi yang dapat digunakan baik pengguna maupun tenaga medis.

Penelitian sebelumnya mengenai klasifikasi EKG masih banyak yang menggunakan dataset aritmia dari MIT-BIH. Salah satu dari penelitian tersebut yang dijadikan sebagai referensi penelitian ini berjudul "Deteksi Arritmia pada Sinyal EKG dengan Deep Neural Network". Penelitian tersebut menyajikan hasil validasi data akhir sebesar 99.62% dan sensitivitas 98.92%. Algoritma yang digunakan adalah *Dense Neural Network* dengan beberapa nilai parameter yang dicoba. Sumber referensi penelitian tersebut menunjukkan bahwa tidak sedikit penelitian sebelumnya dengan database yang sama telah memperoleh hasil signifikan diatas 95% tersebut. Confusion Matriks hasil penelitian tersebut namun menunjukkan kelemahan dataset yang dipakai dimana lebih banyak data normal dari abnormal pada dataset MIT-BIH **[**8**].**

2. ISI PENELITIAN

2.1. Metode Penelitian

Adapun jenis penelitian yang dilakukan dalam penelitian ini yaitu penelitian eksperimental kuantitatif. Penelitian eksperimental adalah penelitian yang dilakukan untuk mengetahui akibat yang ditimbulkan dari suatu perlakuan yang diberikan secara sengaja oleh peneliti. Penelitian kuantitatif adalah suatu proses menemukan pengetahuan yang menggunakan data berupa angka sebagai alat menganalisis keterangan mengenai apa yang ingin diketahui.

2.2 Prosedur Penelitian

Metode yang digunakan untuk menyusun prosedur kerja laporan akhir ini adalah metode *waterfall* ditambah dengan visualisasi diagram alur penelitian. Tahapan-tahapan yang diperlukan adalah:

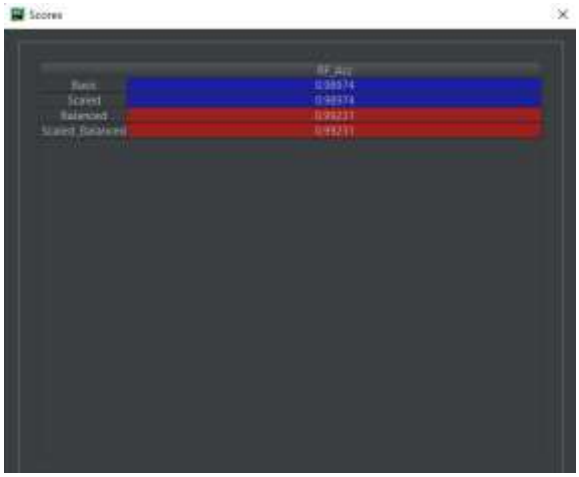
1. Pengumpulan Data
   Pada tahapan ini, metode pengumpulan data yang digunakan mencakup:
   a. Metode Analisa Data Kuantitatif, dimana data hasil survey yang akan terlebih dahulu dilakukan analisa untuk menentukan target klasifikasi hingga menyaring fitur penting dari dataset yang kemudian dilakukan visualisasi terhadapnya demi keperluan penyajian data hasil analisa.
   b. Metode Studi Literatur, dimana akan dikumpulkan informasi dan bahan yang diperlukan dalam penelitian ini, yang diperoleh dari berbagai sumber baik dari jurnal ilmiah maupun artikel.
2. *Preprocessing*
   Dataset terdapat 12 kolom dengan keseluruhan 918 entri data dimana diantaranya terdapat 5 kolom kategori yaitu: Sex, ChestPainType, RestingECG, ExerciseAngina, dan ST_Slope yang mana semua kolom tersebut akan dibuat kolom data *dummy*. Pada dataset tidak terdapat *missing* values sehingga tidak diperlukan

imputasi data. Data kemudian dibagi menjadi data dependen dan independen, HeartDisease sebagai target klasifikasi untuk memprediksi ada tidaknya penyakit jantung pada seseorang. Algoritma Random Forest merupakan bagian kelompok *ensemble* yang tidak memerlukan *scaler*, namun data tetap diskala untuk mendukung proses klasifikasi dengan memanfaatkan algoritma MinMax.

3. Analisa
   Setelah *preprocessing* data, analisa dapat dimulai dengan melakukan penjabaran terhadap fitur data pada database. Fitur pada database terdiri dari:

- Age: umur pasien [tahun]
- Sex: jenis kelamin [M: Laki-laki, F: Perempuan]
- ChestPainType: Jenis sakit pada dada [*TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic*]
- RestingBP: tekanan darah dalam kondisi istirahat [mm Hg]
- Cholesterol: serum kolesterol [mm/dl]
- FastingBS: gula darah setelah puasa [*1: if FastingBS > 120 mg/dl, 0: otherwise*]
- RestingECG: hasil rekam EKG dalam kondisi istirahat [*Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria*]
- MaxHR: detak jantung tertinggi [Numerik antara 60 dan 202]
- ExerciseAngina: exercise-induced angina [Y: Ya, N: Tidak]
- Oldpeak: oldpeak = ST [Numerik diukur pada *depression*]
- ST_Slope: lembah dari puncak segmen *exercise ST* [*Up: upsloping, Flat: flat, Down: downsloping*]
- HeartDisease: *output class/target* [1: penyakit jantung, 0: Normal]

Beberapa data lainnya juga yang layak dianalisa mengenai bagaimana *feature importance* tiap fitur agar dapat terlihat fitur mana yang paling berdampak pada klasifikasi dan mana yang paling sedikit berkontribusi. Implementasi data dependen dan independent yang telah melewati proses *preprocessing* kemudian dilakukan split data untuk data latih dan data uji dengan perbandingan 0.15. Uji akurasi data training dengan algoritma Random Forest kemudian dilakukan dalam empat tahapan: "Basic", "Scaled tanpa penyeimbangan", "Basic dengan penyeimbangan", dan "Scaled dengan penyeimbangan" sebagai perbandingan.



Gambar 1. Tabel Nilai Akurasi

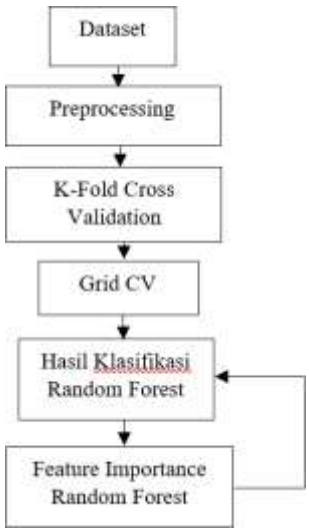
Tabel menunjukkan proses skalar tidak mempengaruhi algoritma Random Forest dalam proses klasifikasinya. Proses *balancing* menaikkan nilai akurasi latih model menjadi 99.231%. Proses sistemasi klasifikasi dibangun menggunakan bantuan framework PyCharm dan library pada python yang telah tersedia.

4. Pengujian

Pada penelitian ini Random Forest akan dibangun dengan bantuan K-Fold Cross Validation dan GridCV. Parameter yang digunakan dalam GridCV adalah:

a. *'n_estimators'*: [50, 100, 300],
b. *'max_features'*: [2, 3, 4],
c. *'max_depth'*: [3, 5, 7, 9],
d. *'min_samples_split'*: [2, 5, 8]

Diagram alur penelitian dapat dilihat pada Gambar 2.2 dibawah. Diagram alur menggambarkan langkah proses penelitian yang dilakukan pada penelitian ini.



Gambar 2.2 Diagram Alur Penelitian

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Dataset yang dibagi dengan persentase 0.15 terbagi menjadi 780 entri data untuk data latih dan 138 entri data uji. Normalisasi data diputuskan menggunakan metode *scaling* karena terbukti dapat meningkatkan sedikit akurasi data latih dari 0.98974 menjadi 0.99231. Metode *scaling* yang digunakan memanfaatkan teknik minmax. Hasil minmax diperoleh dengan mensubtrasi nilai data dengan nilai terkecil pada fitur lalu membaginya dengan jarak (selisih antara nilai terbesar dengan terkecil pada fitur).

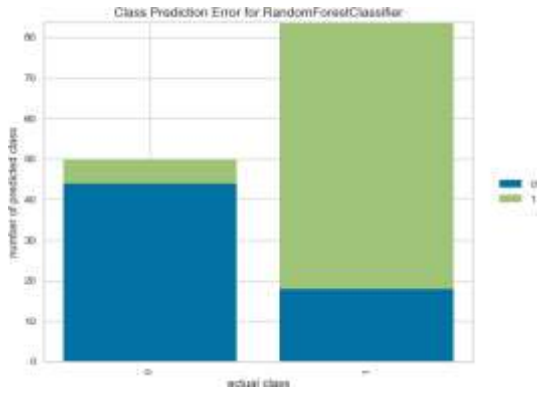
Tabel 3.1 menunjukkan hasil klasifikasi terhadap dataset EKG dimana pasien terdeteksi mempunyai penyakit jantung (1) atau tidak (0) dan Tabel 3.2 menjabarkan secara detail nilai peforma Random Forest.

Tabel 3.1. Confusion Matrix Hasil Klasifikasi Random Forest.

True/Predicted	0	1
0	44	18
1	6	70

Tabel 3.2. Peforma Hasil Klasifikasi Random Forest.

	Precision	Recall	F1-Score	Support
0	0.88	0.71	0.79	62
1	0.88	0.92	0.85	76
accuracy			0.83	138
macro avg	0.84	0.82	0.82	138
weighted avg	0.83	0.83	0.82	138



Gambar 3.1. Class Prediction Error Hasil Klasifikasi Random Forest

Gambar 3.1 lebih jauh melakukan visualisasi terhadap *confusion matrix* yang didapat. Diagrambatang pertama menunjukan data 0 sebenernya ada sebagian kecil terprediksi sebagai 1 (area yang diberiwarna hijau) dan data 1 sebenernya ada sebagian juga terprediksi sebagai 0 pada diagram batang kedua (area yang diberi warna biru).

Hasil klasifikasi kemudian dioptimalkan lebih lanjut dengan metode K-Fold dan GridSearchCV untuk mencari parameter terbaik bagi model Random Forest yang digunakan terhadap dataset. Tabel 3.3 menunjukan daftar kemungkinan parameter yang diuji menggunakan GridSearchCV [9,10].

Tabel 3.3. List Parameter Model

Parameter	Nilai
n_estimators	50, 100, 300
max_features	2, 3, 4
max_depth	3, 5, 7, 9
min_samples_split	2, 5, 8

Dengan *running time* 22.5 detik hasil parameter terbaik didapatkan pada: {'max_depth': 5, 'max_features': 4, 'min_samples_split': 2, 'n_estimators': 300} dengan model algoritma akhir terbaik:

RandomForestClassifier(bootstrap=True,class_weight=None,criterion='gini',max_depth=5,max_features=4,max_leaf_nodes=None,min_impurity_decrease=0.0,min_impurity_split=None,min_samples_leaf=1,min_samples_split=2,min_weight_fraction_leaf=0.0,n_estimators=300,n_jobs=None,oob_score=False,random_state=101, verbose=0, warm_start=False)

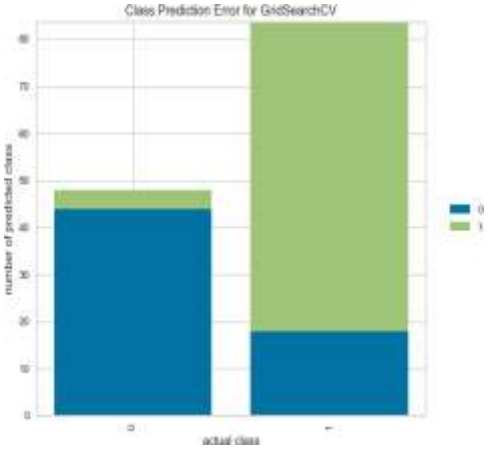
Parameter terbaik dan hasil model *estimator* terbaik kemudian dijalankan ulang untuk melakukan proses klasifikasi. Hasil klasifikasi tersebut dijabarkan dengan Tabel 3.3 dan Tabel 3.4. Gambar 3.2 menunjukan visualisasi hasil RF dengan K-Fold dan GridSearchCV

Tabel 3.3. Confusion Matrix RF dengan K-Fold dan GridSearchCV.

True/Predicted	0	1
0	44	18
1	4	72

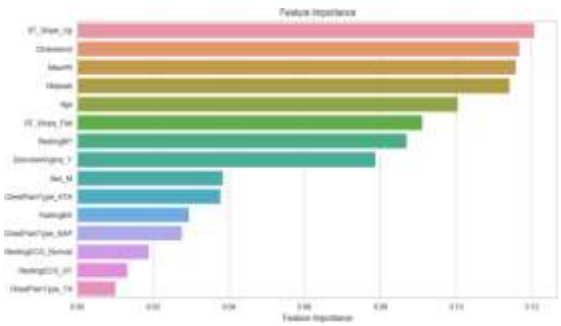
Tabel 3.4. Peforma Hasil Klasifikasi RF dengan K-Fold dan GridSearchCV.

	Precision	Recall	F1-Score	Support
0	0.92	0.71	0.80	62
1	0.80	0.95	0.87	76
accuracy			0.84	138
macro avg	0.86	0.83	0.83	138
weighted avg	0.85	0.84	0.84	138



Gambar 3.2. Class Prediction Error dengan K-Fold dan GridSearchCV

Gambar 3.2. diatas menunjukan optimasi parameter telah sukses meningkatkan peforma klasifikasi dan berhasil menekan angka prediksi yang salah. Selanjutnya, untuk dapat mengoptimasi pengumpulan data kedepannya penting juga untuk dicari *feature importance* tiap fitur untuk mengukur kontribusi tiap fitur. Random forest merupakan bagian algoritma turunan dari *tree* sehingga pengukuran fitur mengunakan *permutation tree*. Hasil permutasi dapat dilihat pada Gambar 3.3.



Gambar 3.3. Feature Importance RF

3.2 Pembahasan

Berdasarkan hasil uji coba dan optimasi terhadap peforma model RF, dapat disimpulkan bahwa model berhasil dioptimasi dengan K-Fold dan GridSearchCV dari akurasi awal 82,6087% menjadi 84,058%. *Feature importance* juga menunjukkan selain faktor generic seperti usia dan tingkat kolesterol informasi yang diambil dari hasil analisa EKG berperan penting seperti tingkat detak jantung per menit tertinggi dan detak jantung per menit dalam kondisi diam atau istirahat. Gelombang ST menjadi fitur terpenting dalam penyaluran kontribusi terbanyak terhadap dataset dalam dapat melakukan proses klasifikasi dengan baik. Informasi diatas yang diperoleh dapat kemudian digunakan untuk penelitian-penelitian tahap berikutnya hingga membangun dataset yang efisien tersendiri dengan afiliasi atas nama UNPRI baik untuk keperluan akademik maupun akreditasi.

4. PENUTUP

4.1 Kesimpulan

Berdasarkan analisa yang dilakukan akan hasil penelitian ini, berikut adalah beberapa hal yang dapat disimpulkan:

1. Proses implementasi algoritma Random Forest yang digunakan sukses dengan meraih tingkat akurasi sebesar 82,6087% yang kemudian dioptimasi dengan teknik K-Fold dan GridSearchCV menjadi 85,058%.
2. Sistem klasifikasi dengan Random Forest yang dibuat dapat dimplementasikan kedalam rangkaian alat untuk menciptakan alat deteksi aritmia EKG otomatis portable. Hubungan antara informasi yang disajikan oleh alat juga dengan sukses dibuktikan kontribusi pentingnya terhadap diagnosa positif negatifnya seseorang mempunyai penyakit jantung terutama melalui Gelombang ST.
3. Penelitian ini dengan sukses menjembatani penelitian sebelumnya (deteksi fitur) dengan penelitian lanjutannya (alat jadi siap dipasarkan hingga pembuatan dataset medis tersendiri).

4.2 Saran

Pada penelitian ini masih terdapat beberapa kelemahan, sehingga dapat dilakukan perbaikan pada penelitian selanjutnya. Adapun saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Atas dasar informasi yang didapatkan dari penelitian ini membangun dataset tersendiri bagi UNPRI untuk dapat dijadikan bahan penelitian lanjutan bagi mahasiswa/i kedepannya.
2. Mewujudkan alat portable deteksi EKG dengan klasifikasi otomatis yang siap untuk dipasarkan.

DAFTAR PUSTAKA

[1] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques," IEEE Access, vol. 9, pp. 19304–19326, 2021.

[2] A. M. Alaa, T. Bolton, E. Di Angelantonio, J. H. Rudd, and M. van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants," PLOS ONE, vol. 14, no. 5, 2019.

[3] Fedesoriano. (Juni 2021). Body Fat Prediction Dataset. Retrieved [29 December] from,https://www.kaggle.com/fedesoriano/body-fat-prediction-dataset).

[4] C. M. Yeşilkanat, "Spatio-temporal estimation of the daily cases of covid-19 in worldwide using Random Forest Machine Learning algorithm," Chaos, Solitons & Fractals, vol. 140, p. 110210, 2020.

[5] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," PeerJ, vol. 6, 2018.

[6] C. Wijaya, A. Andrian, M. Harahap, C. Christnatalis, M. Turnip, and A. Turnip, "Abnormalities state detection from P-wave, QRS complex, and T-wave in noisy ECG," Journal of Physics: Conference Series, vol. 1230, no. 1, p. 012015, 2019.

[7] Martiningsih and Abdul haris, "Cardiovascular Disease Risk Program for Participants Management of Chronic Disease (Prolanis) in Bima City Puskesmas: Correlation with Ankle Brachial Index and Obesity" vol. 22, no. 3, 2019.

[8] Z. Jin, H. Chan, J. Ning K. Lu, and D. Ma, "The Role of Hydrogen Sulfide in Pathologies of The Vital Organs and Its Clinical Application" Chandra Wijaya, Andrian, Mawaddah Harahap, Christnatalis, Madi Turnip, and Arjon Turnip, "Abnormalities State Detection from P-Wave, QRS Complex, and T-Wave in Noisy ECG" vol 66, no.2, pp 169-179, 2015.

[9] P. Pirjatullah, D. Kartini, D. T. Nugrahadi, M. Muliadi, and A. Farmadi, "Hyperparameter tuning using GRIDSEARCHCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers," 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), 2021.

[10] G. S. Ranjan, A. Kumar Verma, and S. Radhika, "K-nearest neighbors and grid search CV based Real Time Fault Monitoring System for Industries," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019.