

Ekstraksi dan Klasifikasi Sinyal MIT-BIH Arrhythmia Database dengan Model Random Forest

M. RIZKY

*Advanced and Creative Networks
Research Center
Universitas Logistik dan Bisnis Internasional
Bandung, Indonesia
1194021@std.poltekpos.ac.id*

Roni Andarsyah

*Advanced and Creative Networks
Research Center
Universitas Logistik dan Bisnis Internasional
Bandung, Indonesia
roniandarsyah@ulbi.ac.id*

Abstract—Di era modern ini, banyak sekali kegiatan – kegiatan operasional maupun kegiatan lainnya yang melibatkan atau menggunakan Artificial Intelligence (AI). Tidak bisa dipungkiri lagi bahwa teknologi yang berkembang pesat seperti sekarang ini tentu dibuat untuk mempermudah pekerjaan manusia bahkan menggantikan peran manusia. Salah satu bagian dari Artificial Intelligence (AI) adalah Machine Learning. Machine Learning sebagai metode dalam sistem kecerdasan buatan yang mampu mengklasifikasikan data yang dimasukkan untuk keperluan dan kebutuhan masing - masing. Banyak aplikasi yang menerapkan Machine Learning untuk keperluan klasifikasi data, memprediksi hubungan antar data, membaca pola data, dan banyak implementasi lainnya. Dalam laporan ini akan dibahas penggunaan keluaran struktur dari Machine Learning digunakan untuk mendeteksi sinyal ECG apakah sudah sesuai dengan data yang sudah di tetapkan pada anotasi. Berdasarkan hasil penelitian menunjukkan pemodelan struktur dari Machine Learning dapat digunakan untuk mendeteksi kesesuaian dari sinyal baik pemodelan secara terpisah ataupun pemodelan secara gabungan.

Index Terms—Artificial Intelligence, Machine Learning, Model, Klasifikasi, ECG

I. LATAR BELAKANG

Tidak bisa dipungkiri lagi bahwa perkembangan Artificial Intelligence begitu sangat cepat. Seiring berkembangnya teknologi yang sangat cepat, banyak sekali jenis - jenis teknologi yang bermunculan untuk membantu bahkan menggantikan pekerjaan manusia salah satunya di bidang kesehatan. Salah satu cabang Artificial Intelligence yang sekarang banyak sekali diminati adalah Machine Learning dan sampai saat ini masih terus berkembang pesat di kalangan programmer atau khususnya di dunia IT. Machine Learning sendiri terdiri atas 2 bagian yaitu Machine Learning Supervised dan Machine Learning Unsupervised.

Machine Learning Supervised adalah struktur dari suatu data yang hendak dianalisis telah ditentukan dahulu dan Machine Learning mencari data di struktur tersebut, sedangkan Machine Learning Unsupervised struktur dari suatu data dicari oleh Machine Learning itu sendiri [1]. Salah satu algoritma Supervised Learning yang sering digunakan pada proses klasifikasi adalah algoritma Random Forest (RF). RF adalah teknik bagging yang memiliki karakteristik signifikan

yang berjalan efisien pada dataset besar. Random forest dapat menangani ribuan variabel masukan tanpa penghapusan variabel dan memperkirakan fitur penting untuk klasifikasi [2].

Di dalam dunia medis, teknologi - teknologi banyak sekali diterapkan untuk memenuhi kebutuhan medis itu sendiri seperti AI pendeteksi pasien positif Covid atau tidak dengan memanfaatkan hembusan nafas dari pasien tersebut. Dan masih banyak lagi hal - hal yang bisa kita manfaatkan untuk mengembangkan teknologi di bidang kesehatan salah satunya adalah dengan memanfaatkan sinyal EKG atau Electrocardiography untuk mengklasifikasi penyakit gagal jantung pada pasien. EKG merupakan sebuah informasi sinyal yang digambarkan dalam bentuk diagram yang menampilkan informasi penting mengenai keadaan jantung manusia. Electrocardiography atau EKG adalah rekaman aktivitas listrik yang dihasilkan melalui siklus detak jantung [1]. Pada kegiatan internship yang penulis lakukan akan berfokus pada Ekstraksi dan klasifikasi sinyal MIT-BIH Arrhythmia menggunakan model Random Forest.

II. LANDASAN TEORI

2.1. MIT-BIH Arrhythmia Database

MIT-BIH Arrhythmia database adalah rangkaian uji standar yang umumnya tersedia untuk mengevaluasi aritmia deteksi. Sejak 1980, database ini telah digunakan untuk dasar penelitian untuk dinamika jantung di sekitar 500 lokasi di seluruh dunia. Basis data ini sebagian besar digunakan untuk tujuan medis dan penelitian dari deteksi dan analisis aritmia jantung yang berbeda. Basis data ini mencoba menyediakan informasi yang tepat untuk mendeteksi aritmia ventrikel [3]. Aritmia adalah perubahan detak jantung yang tidak normal karena detak jantung yang tidak tepat yang menyebabkan kegagalan dalam pemompaan darah. Aritmia dapat menyebabkan kematian jantung mendadak. Gejala aritmia yang umum adalah denyut prematur, jantung berdebar, pusing, kelelahan, dan pingsan. Aritmia lebih sering terjadi pada orang yang menderita tekanan darah tinggi, diabetes dan arteri koroner penyakit [4]. Sinyal

Electrocardiogram yang akan digunakan pada kegiatan internship ini diambil dari MIT-BIH Arrhythmia database.

2.2. Electrocardiogram (ECG)

Electrocardiogram (ECG) adalah tes medis yang mengukur aktivitas listrik jantung. ECG digunakan untuk mendiagnosis dan memantau berbagai kondisi jantung, seperti serangan jantung, aritmia, dan gagal jantung. ECG akan merekam aktivitas listrik kecil yang dihasilkan oleh jantung selama periode waktu tertentu dengan menempatkan elektroda pada tubuh pasien [5]. Rekaman ECG berisi noise dan amplitudo yang bervariasi dari setiap orang sehingga sulit dalam proses mendiagnosis [6].

Electrocardiogram (ECG) memberikan informasi penting tentang berbagai kondisi manusia [2]. Untuk melakukan ECG, petugas kesehatan akan menempelkan tambalan kecil dan lengket yang disebut elektroda ke dada, lengan, dan kaki pasien. Elektroda terhubung ke mesin ECG, yang merekam sinyal listrik yang dihasilkan oleh jantung saat bergerak ke seluruh tubuh. Mesin tersebut menghasilkan jejak aktivitas listrik jantung, yang disebut strip ECG, yang kemudian diinterpretasikan oleh petugas kesehatan.

Machine Learning dapat diterapkan pada analisis data ECG untuk meningkatkan akurasi dan efisiensi diagnosis dan pengobatan. Misalnya, algoritma pembelajaran mesin dapat dilatih untuk mengenali pola dalam data ECG yang menunjukkan kondisi jantung tertentu. Algoritma ini kemudian dapat digunakan untuk menganalisis data ECG secara otomatis dan memberikan rekomendasi diagnosis atau pengobatan. Algoritma pembelajaran mesin dapat dilatih untuk mengklasifikasikan data ECG ke dalam kategori yang berbeda, seperti normal atau abnormal, atau untuk mengidentifikasi kondisi jantung tertentu.

Biasanya, klasifikasi sinyal ECG memiliki empat fase: preprocessing, segmentasi, ekstraksi fitur, dan klasifikasi. Fase preprocessing terutama ditujukan untuk mendeteksi dan melemahkan frekuensi sinyal ECG yang terkait dengan artefak, yang juga biasanya melakukan normalisasi dan peningkatan sinyal. Setelah preprocessing, segmentasi akan membagi sinyal menjadi segmen yang lebih kecil, yang dapat mengekspresikan aktivitas listrik jantung dengan lebih baik [6].

2.3. Machine Learning

Pembelajaran Mesin atau Machine Learning merupakan kemajuan teknologi yang penting karena dapat membantu dalam mengambil keputusan dengan mekanisme prediksi dan klasifikasi berdasarkan data yang ada [7]. Berfokus pada performance yang tinggi, teknik pembelajaran mesin atau machine learning diterapkan pada bisnis dengan data yang berkembang pesat. Karena pendekatan desain cocok

untuk komunikasi komputasi paralel dan terdistribusi yang berevolusi atau data bisnis yang dinamis dan berkembang kedalam model Machine Learning [8].

Teknologi berbasis komputer modern banyak yang telah menggunakan Pembelajaran Mesin atau Machine Learning. Machine Learning merupakan cabang dari Kecerdasan Buatan atau Artificial Intelligence yang luas dan sudah berkembang pesat saat ini yang memungkinkan komputer untuk belajar dan berkembang secara otomatis tanpa harus diprogram secara eksplisit. Teknologi ini berasal dari mempelajari pengenalan pola dan teori pembelajaran komputasi. Secara umum, metode pembelajaran yang umum digunakan oleh Machine Learning dapat diklasifikasikan menjadi Supervised, Unsupervised, dan Reinforcement Learning [9].

2.3.1. Supervised Learning

Supervised Learning merupakan metode Machine Learning untuk menyimpulkan fungsi dari data train ada. Algoritma Supervised Learning biasanya berisi kumpulan sampel input (feature) dan label yang berkaitan dengan kumpulan data tersebut. Tujuan dari pengklasifikasian adalah untuk menemukan batas yang sesuai yang dapat memprediksi label yang benar pada data test. Secara singkat, dalam Supervised Learning memiliki setiap contoh data yang berpasangan yang terdiri dari objek masukan (input) dan objek keluaran (output) yang diinginkan. Algoritma Supervised Learning menganalisis data train dan menghasilkan fungsi (model) [10]. Beberapa contoh metode algoritma pada Supervised Learning:

2.3.1.1. Random Forest

Random Forest adalah metode Machine Learning yang diperkenalkan pada tahun 2001 oleh Leo Breiman. Metode ini menggunakan serangkaian besar dari Decision Tree dengan korelasi timbal balik yang rendah dan fitur yang dipilih secara acak menggunakan metode bagging (Bootstrap AGGREGATING) [10].

Random Forest merupakan salah satu metode pengklasifikasian terbaik dan banyak digunakan untuk regresi dan juga klasifikasi. Random Forest memiliki algoritma yang sederhana sehingga menjadi salah satu pilihan yang menarik untuk mengklasifikasi teks. Selain itu, Random Forest juga memiliki kemampuan untuk mengolah data berdimensi tinggi dan memiliki performa yang tinggi walaupun menggunakan data yang banyak sehingga menjadi salah satu keuntungan menggunakan model ini dibandingkan dengan model Machine Learning lainnya [11].

Pemilihan model ini didasarkan karena pada faktanya bahwa Random Forest secara luas dianggap sebagai salah satu metode Machine Learning yang paling sukses dan banyak digunakan hingga saat ini [12].

2.3.1.2. Naive Bayes

Naive Bayes adalah salah satu pengklasifikasi terkemuka yang telah banyak dikutip oleh banyak peneliti dan digunakan di banyak aspek karena kesederhanaannya dan kinerja dari klasifikasi yang nyata [13]. Diantara bermacam-macam teknik atau metode klasifikasi saat ini, pengklasifikasi Naive Bayes (NB) berperan penting karena kesederhanaan, traktabilitas dan efisiensinya [14].

Naive Bayes juga merupakan pengklasifikasi yang sangat kompeten dalam banyak aplikasi di dunia nyata. Meskipun Naive Bayes telah menunjukkan akurasi klasifikasi yang luar biasa, namun output yang dihasilkan jarang benar dalam kenyataan [15]. Terlepas dari hal itu, pada kenyataannya Naive Bayes bekerja dengan baik diimplementasikan di dunia ini seperti memprediksi waktu, pemfilteran spam, prakiraan cuaca, dan diagnosis medis [16].

Pengklasifikasian Naive Bayes didasarkan pada kombinasi Teorema Bayes dan asumsi independensi atribut. Pengklasifikasi Naive Bayes didasarkan pada asumsi yang disederhanakan bahwa nilai atribut bersifat independen secara kondisional, berdasarkan asumsi nilai target yang diberikan. Pendekatan Bayes untuk klasifikasi kasus baru terdiri dari penetapan nilai target yang paling mungkin, dengan asumsi bahwa ada [17].

2.3.1.3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) merupakan salah satu algoritma klasifikasi yang paling stabil dalam kelompok algoritma klasifikasi supervised. Dikarenakan kesederhanaan dan implementasi algoritma yang mudah. K-Nearest Neighbor (KNN) merupakan salah satu metode klasifikasi nonparametric. KNN menjadi terkenal karena algoritmanya yang luas dan yang paling mudah. KNN dapat menyimpan semua masalah atau studi kasus yang ada dan mengklasifikasikan berdasarkan kesamaannya. Secara umum, KNN menggunakan jarak Euclidean untuk menemukan data yang paling mirip dengan kelompoknya [18].

Pada metode ini, nilai yang hilang dari variabel tertentu diganti dengan nilai rata-rata atau nilai mean dari KNN terdekat dari pengamatan variabel yang sama. Fungsi jarak yang berbeda dapat digunakan untuk memilih tetangga yang memungkinkan metode untuk menyertakan variabel numerik dan kategori. Keuntungan utama KNN adalah tidak memerlukan spesifikasi model prediktif apa pun [19].

2.3.1.4. Support Vector Machine (SVM)

Support Vector Machine membuktikan bahwa salah satu algoritma yang memiliki performance yang powerful selama beberapa dekade terakhir dan mengandalkan prinsip SRM. Support Vector Learning (SVM) umumnya digunakan untuk

masalah klasifikasi dan regresi.

Support Vector Machine (SVM) bekerja dengan membangun hyperplane yang memisahkan sampel berdasarkan pendekatan margin yang maksimum. Berbeda dengan Artificial Neural Network (ANN) yang memiliki kelemahan local minimal. Support Vector Machine memberikan solusi dengan menyelesaikan masalah optimasi dengan konveks [6].

Metode Support Vector Machine juga disebut model klasifikasi biner. Dalam ruang dua dimensi, garis lurus menjadikan garis segmentasi yang paling cocok di tengah 2 kelas data, dan untuk kumpulan data berdimensi tinggi, ini untuk menetapkan bidang keputusan yang optimal sebagai tolak ukur klasifikasi. Prinsip dasar Support Vector Machine (SVM) mensyaratkan bahwa ketika masalah klasifikasi diselesaikan, jarak dari titik sampel terdekat ke permukaan keputusan adalah yang terbesar, yaitu jarak minimum memaksimalkan dua kelas titik sampel untuk memisahkan tepi [20].

2.3.2. Unsupervised Learning

Unsupervised Learning hanya dapat digunakan untuk tugas pengelompokan (clustering). Banyak pendekatan menggunakan Unsupervised Learning untuk mendukung tugas klasifikasi. Misalnya, algoritma pengelompokan (clustering) dapat meningkatkan kinerja tugas klasifikasi dengan mengelompokkan objek data ke dalam kelompok yang lebih homogen.

Unsupervised Learning banyak digunakan untuk preprocessing data seperti ekstraksi fitur, pemilihan fitur, dan resampling. Namun, ada banyak juga kasus penggunaan pembelajaran tanpa pengawasan sebagai algoritma pilihan untuk klasifikasi dengan kinerja yang sebanding dan mungkin lebih baik dibandingkan Supervised Learning [21].

Pada algoritma Unsupervised Learning yang mampu memisahkan data tanpa sebuah pengetahuan yang dalam tentang berbagai jenis peristiwa meningkatkan efisiensi analisis secara luar biasa, dan memungkinkan analisis hilir untuk berkonsentrasi pada upaya penyesuaian hanya pada peristiwa yang menarik. Selain itu, algoritma pengelompokan memungkinkan lebih banyak eksplorasi data, berpotensi memungkinkan jenis reaksi baru dan tak terduga [22].

2.3.2.1. K-Means

Di era big data, sejumlah besar sumber daya data dikumpulkan dari kehidupan orang sehari-hari, ditransfer ke dalam internet, dan disimpan pada pusat data [23]. Pengelompokan data (Clustering), sebagai bagian penting dari data mining, dan sudah dianggap sebagai tugas penting dalam Unsupervised Learning.

Untuk kumpulan data tertentu, clustering akan membaginya menjadi beberapa kelompok atau cluster yang sedemikian rupa sehingga objek data dalam kelompok atau cluster yang sama berupa satu sama lain [24]. K-Means adalah pengelompokan masalah yang dipelajari dengan baik yang menghasilkan aplikasi di banyak bidang dan merupakan bagian dari Unsupervised Learning. K-Means merupakan salah satu masalah paling mendasar dalam ilmu komputer [25].

III. METODOLOGI PENELITIAN

3.1. Ruang Lingkup

Penelitian ini akan membahas tentang ekstraksi dan klasifikasi MIT-BIH Arrhythmia Database dengan menggunakan salah satu model Machine Learning yaitu Random Forest. Sumber data yang akan digunakan berasal dari situs PhysioNet yang merupakan Database Complex Physiologic Signals. Data yang akan diambil pada penelitian ini adalah MIT-BIH Arrhythmia Database dengan 48 record dan masing-masing durasi yang tersedia. Dari data-data tersebut akan digabungkan lalu dilakukan ekstraksi dan klasifikasi sinyal dengan Random Forest.

3.2. Alur Metodologi Penelitian

Pada penelitian ini, metode akan menjadi hal penting dalam melakukan ekstraksi hingga klasifikasi sinyal karena akan mempengaruhi hasil klasifikasi atau output dari model tersebut. Oleh karena itu, alur metodologi yang tepat untuk mendapatkan hasil yang terbaik.

3.3. Indikator Capaian Penelitian

Berdasarkan gambar diatas, terdapat beberapa indikator capaian yaitu sebagai berikut:

No	Tahapan	Indikator Pencapaian
1	Studi Literatur	Uraian Teori - teori
2	Pengumpulan Data	Dataset sinyal MIT-BIH Arrhythmia
3	Pra-Pemrosesan Data	Ekstraksi Fitur, Denoising, Normalisasi
4	Pemodelan	Model Random Forest untuk Klasifikasi Sinyal
5	Evaluasi Model	Melakukan perbandingan dengan beberapa model

Tabel 3.1 Indikator Capaian Penelitian

IV. HASIL PENELITIAN

4.1. Data

Data yang akan digunakan diambil langsung dari website PhysioNet yang merupakan Research Resource for Complex Physiological Signals untuk melakukan penelitian dan

pendidikan biomedis. Data yang dikumpulkan merupakan data hasil rekaman detak jantung dari beberapa orang. Bentuk data yang digunakan merupakan data CSV dengan nilai yang membentuk grafik dari sebuah detak jantung atau beat. Data tersebut memiliki annotation nya masing-masing disetiap beat yang bisa dijadikan label untuk proses klasifikasi.

4.2. Pra-Pemrosesan Data

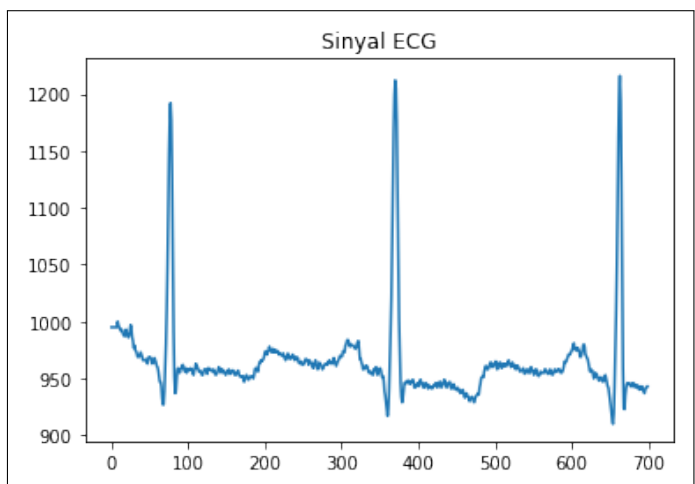
Pada tahap ini, kita akan mempersiapkan data dengan mengolahnya agar data tersebut dapat digunakan untuk proses akhir atau pemodelan. Beberapa langkah Pra-Pemrosesan data yang akan dilakukan seperti Split Extension File, Record Signal Dataset, Denoising, Normalisasi, Merge Annotation, Rebalancing Dataset, Split Dataset.

4.2.1. Split Extension File

Dataset yang digunakan merupakan kumpulan beberapa record dan annotation, untuk itu perlu dilakukan split untuk untuk memisahkan file record dan annotation masing-masing record. Terdapat kondisional yang memisahkan jenis extension file “.csv” dan memasukkan ke dalam sebuah array. Terdapat 2 extension yang di pisah yaitu “.csv” untuk dataset attribute dan “.txt” untuk label.

4.2.2. Record Signal Dataset

Tahap ini dilakukan untuk proses menyiapkan bentuk nilai dari record dataset untuk dilakukan plotting apakah hasil plotting sudah sesuai dengan bentuk sinyal EKG atau tidak. Dengan membaca file CSV pada parameter dengan mengambil nilai MLII pada CSV untuk dijadikan attribute pada proses plot sinyal maupun proses klasifikasi. Nilai tersebut akan disimpan pada sebuah array yang nantinya akan diplot menggunakan array tersebut. Untuk melakukan plot sinyal menggunakan library matplotlib dari python.



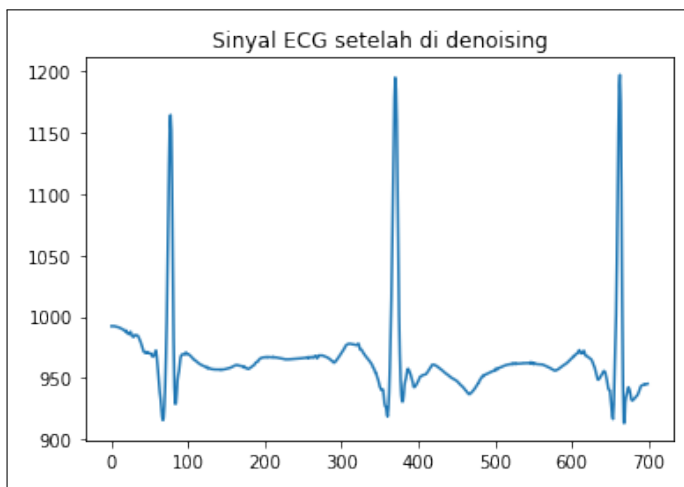
Gambar 4.1 Contoh Sinyal EKG

Hasil plotting membentuk sinyal EKG atau Detak Jantung

terdapat 3 beat yang dihasilkan dari hasil plotting dengan mengambil nilai mulai dari baris ke 1 sampai baris ke 700. Tetapi pada sinyal EKG tersebut masih memiliki noise yang merupakan sinyal gangguan. Untuk itu, perlu dilakukan Denoising pada dataset untuk menghilangkan noise tersebut.

4.2.3. Denoising

Denoising merupakan tahapan yang sangat penting dilakukan untuk menghilangkan noise pada sinyal. Jika noise tersebut tidak dihilangkan akan mengganggu dan menurunkan performa dari hasil modeling nanti karena banyak bentuk sinyal yang hanya mengganggu dari hasil sinyal yang sebenarnya. Untuk melakukan Denoising menggunakan library “pywt” dengan mengkalkulasikan Wavelet untuk mengelompokkan nilai menjadi sebuah Wavelet tunggal. Untuk nilai Threshold yang digunakan adalah “0.04”. Setelah sinyal dilakukan Denoising, maka akan membentuk sinyal yang sempurna dan tidak terdapat banyak noise.

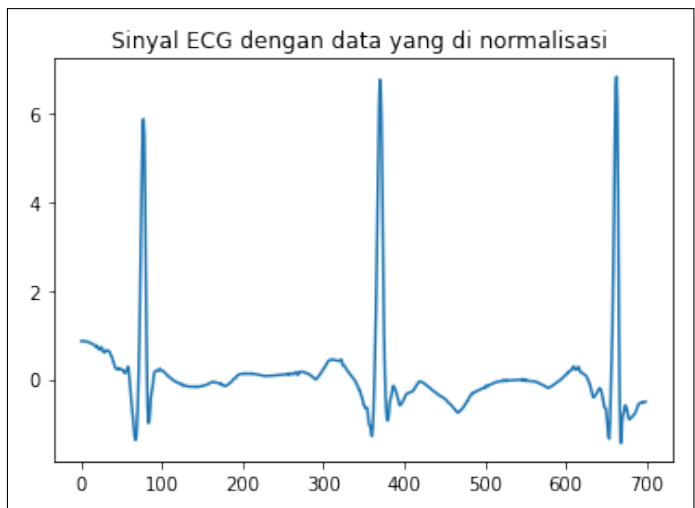


Gambar 4.2 Contoh Sinyal EKG setelah Denoising

Sinyal yang di plotting sudah bersih dari noise dan memiliki bentuk sinyal yang lebih sempurna daripada sinyal yang sebelumnya belum dilakukan Denoising. Karena sinyal EKG memiliki nilai yang berbeda-beda tiap recordnya, maka perlu dilakukan Normalisasi pada dataset tersebut untuk menyamakan skalanya.

4.2.4. Normalisasi

Normalisasi merupakan tahapan yang perlu dilakukan untuk mengubah nilai dengan tipe data Numerik pada himpunan data agar skala dari himpunan data tersebut sama. Akan Tetapi, tidak semua dataset bisa dilakukan normalisasi tergantung dari algoritma atau model yang akan dipakai untuk proses klasifikasi. Setelah dilakukan normalisasi maka data sinyal tersebut bisa langsung digunakan untuk klasifikasi. Untuk proses normalisasi ini akan mengubah nilai dari MLII pada data sebelumnya menjadi nilai yang sudah dinormalisasikan tentunya.



Gambar 4.3 Contoh Sinyal EKG setelah Normalisasi

Nilai Y dari hasil plotingan diatas berubah yang awalnya bernilai 900an menjadi nilai dengan range -1 sampai 7 tetapi tetap membentuk sinyal atau detak jantung yang sama seperti sebelumnya. Sebelum ke proses selanjutnya harus dilakukan penggabungan annotation pada setiap detak jantung atau beat.

4.2.5. Merge Annotation

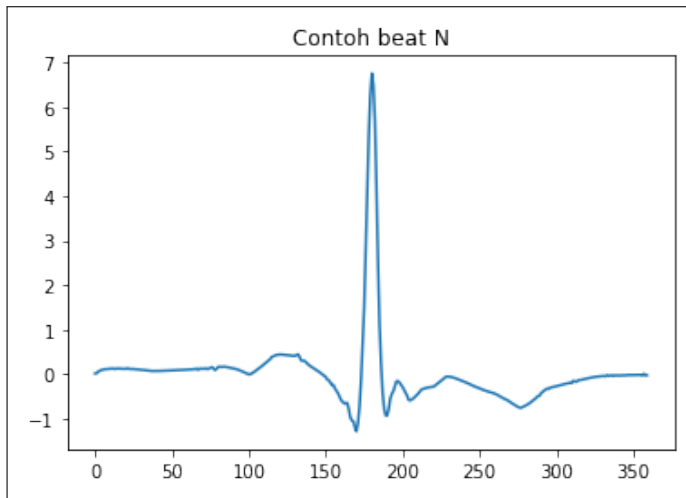
Pada tahapan ini, menggabungkan annotation atau label pada data yang akan kita lakukan klasifikasi sangat penting, karena annotation inilah yang akan menentukan bahwa model yang dibuat berjalan dengan baik atau tidak nantinya. Pada proses ini juga akan dilakukan plotting dari masing-masing class atau label pada dataset tersebut. Pada masing-masing class berdasarkan nilai pada variabel classes yang terdiri dari 5 class seperti berikut:

classes = ['N', 'L', 'R', 'A', 'V']

Keterangan:

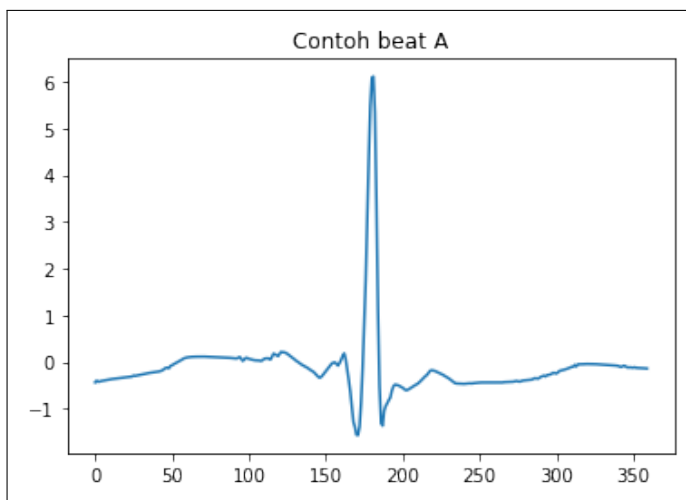
- N: Normal beat (displayed as “.” by the PhysioBank ATM, LightWAVE, pschart, and psfd)
- L: Left bundle branch block beat
- R: Right bundle branch block beat
- A: Atrial premature beat
- V: Premature ventricular contraction

Dari keterangan diatas merupakan arti dari masing-masing label yang akan dipakai untuk proses klasifikasi. N merupakan beat normal manusia dan diikuti dengan A dan V untuk beat yang prematur.



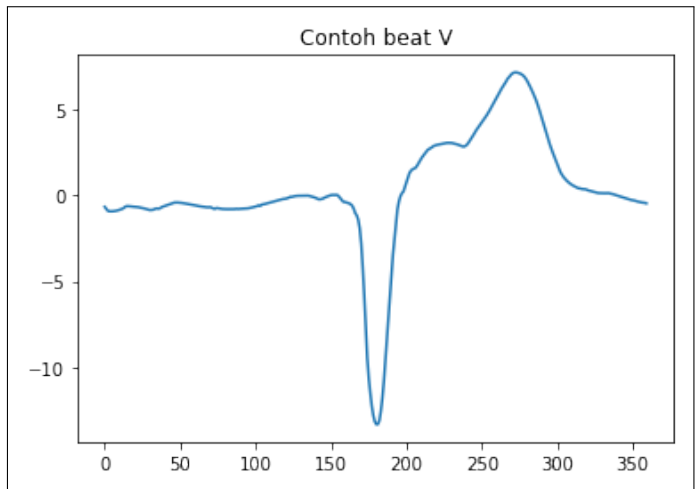
Gambar 4.4 Contoh Sinyal Anotasi N

Hasil plot diatas merupakan merupakan detak jantung Normalnya manusia yang digambarkan dengan satu gelombang tengah.



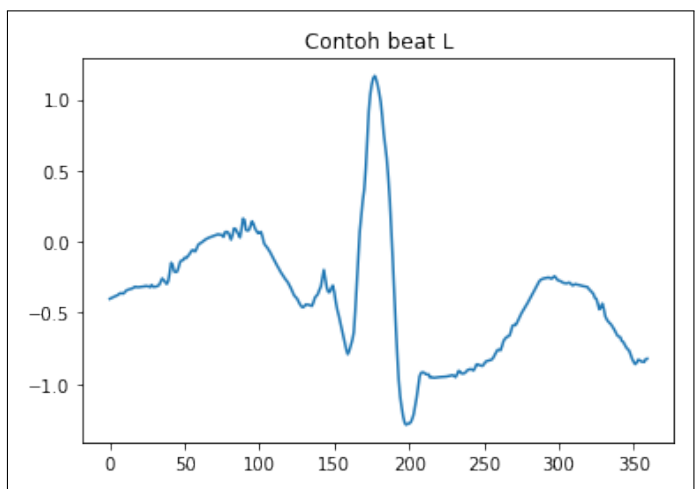
Gambar 4.5 Contoh Sinyal Anotasi A

Terdapat sedikit perbedaan bentuk plot dari detak jantung dari class N dan class A. Perbedaan terletak pada sisi kiri gelombang dengan adanya lekukan kebawah.



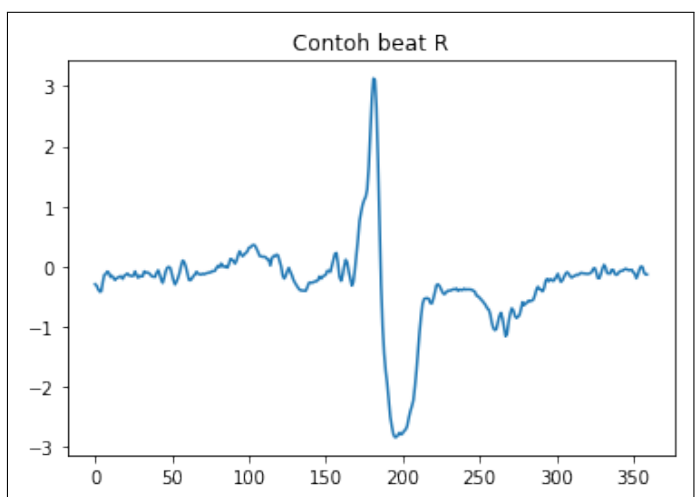
Gambar 4.6 Contoh Sinyal Anotasi V

Perbedaan yang sangat menonjol adalah berbentuk gelombang terbalik dari bentuk beat dengan class N.



Gambar 4.7 Contoh Sinyal Anotasi L

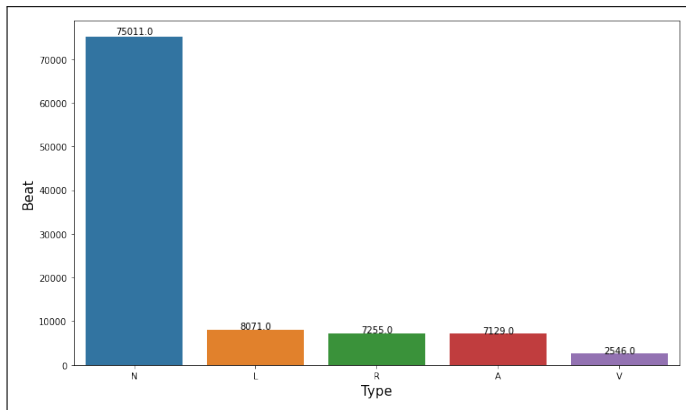
Terlihat bahwa bentuk plot dari detak jantung dengan class L sangat berbeda dengan bentuk beat dengan class N.



Gambar 4.8 Contoh Sinyal Anotasi R

Terlihat bahwa bentuk plot dari detak jantung dengan class L sangat tidak beraturan dan masih memiliki banyak noise.

Kemudian, langkah selanjutnya adalah melakukan plotting diagram semua beat pada dataset dengan mengelompokkan berdasarkan jumlah dari class yang akan digunakan.



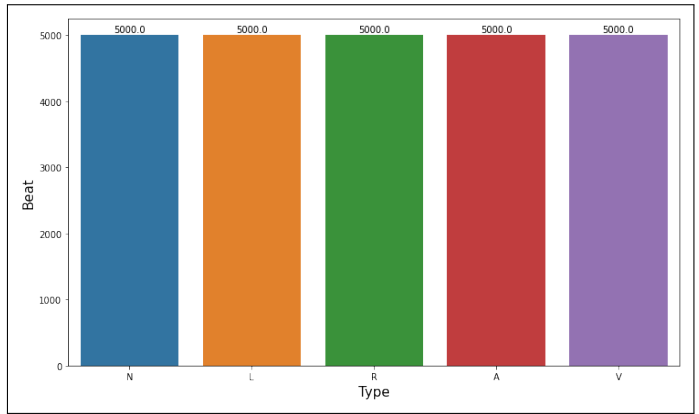
Gambar 4.9 Diagram jumlah beat masing-masing class

Pada diagram diatas terlihat bahwa beat dengan class N mencapai 75.011 beat sedangkan class lainnya memiliki jumlah beat kurang dari 10.000 beat. Oleh karena itu, data tersebut perlu dilakukan penyetaraan data sesuai dengan class nya atau Rebalancing Dataset.

4.2.6. Rebalancing Dataset

Tahap Rebalancing Dataset merupakan proses penyeimbangan data sesuai dengan class masing-masing. Rasio data yang tidak seimbang sesuai dengan class perlu dilakukan rebalancing data agar data yang dilakukan train nanti menjadi proporsional.

Tahapan rebalancing ini mengambil 5.000 sampel atau beat masing-masing class nya. Sehingga model mempelajari data yang seimbang dengan masing-masing class tersebut. Setelah dilakukan rebalancing, maka bentuk diagram dari dataset tersebut akan seimbang. Untuk mengetahui apakah data sudah seimbang atau tidak, perlu dilakukan plotting untuk mengetahui bentuk diagramnya.



Gambar 4.10 Diagram dataset setelah di Rebalancing

Terlihat dari diagram tersebut sudah memiliki masing-masing beat sebanyak 5.000 beat dengan jumlah class yang sama. Sehingga dataset tersebut sudah bisa dilakukan split untuk dilakukan train dan test pada tahap pemodelan.

4.2.7. Split Dataset

Pada tahap ini akan membagi 2 dataset untuk dilakukan train dan test untuk dilakukan pemodelan. Setelah dilakukan split dataset, maka akan ada data train dan data test. Data train berfungsi untuk model melakukan pembelajaran terhadap data yang telah disiapkan. Setelah model melakukan pembelajaran terhadap data tersebut, maka model tersebut akan melakukan proses pengecekan terhadap data test yang sudah mempunyai hasil atau outputnya untuk divalidasi apakah hasil klasifikasi atau output dari model tersebut tepat atau tidak.

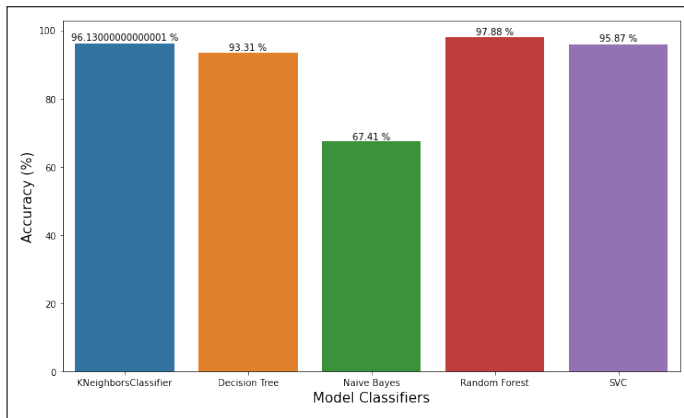
4.3. Pemodelan

Tahap modeling merupakan tahap yang paling penting dari sebuah Machine Learning dikarenakan tahap modeling akan menghasilkan sebuah model yang memiliki output yang dapat mengklasifikasi sebuah masalah atau objek serta memiliki akurasi dalam pembuatannya.

Dari hasil modeling tersebut mendapatkan akurasi yang tinggi yaitu sebesar 97.88%. Akurasi tersebut sudah termasuk akurasi yang tinggi dikarenakan tidak termasuk dalam model yang overfitting maupun underfitting.

4.4. Evaluasi Model

Untuk memastikan model yang dibuat adalah model yang memiliki akurasi yang paling tinggi, maka perlu dilakukan evaluasi model. Pada evaluasi ini akan membandingkan hasil klasifikasi dari beberapa model seperti K-Neighbors Classifier, Decision Tree, Naive Bayes, dan Support Vector Machine.



Gambar 4.11 Perbandingan Akurasi Model

Setelah dilakukan plot terlihat pada gambar diatas bahwa Random Forest memiliki akurasi yang paling tinggi dengan akurasi 97.88% diantara model yang ada dan diikuti dengan model Naive Bayes yang memiliki akurasi yang paling rendah dengan akurasi 67.41%.

V. KESIMPULAN

5.1. Kesimpulan

Berdasarkan hasil pemodelan yang dilakukan, didapatkan model dengan akurasi yang paling tinggi yaitu model Random Forest dengan nilai akurasi 97.88% dan di posisi paling rendah yaitu model Naive Bayes dengan nilai akurasi 67.41%.

5.2. Saran

Saran dari penelitian ini adalah dengan melakukan pengembangan terhadap model dan juga metode penelitian yang akan digunakan dimasa yang akan datang menambahkan proses ekstraksi nilai RR dan jarak antar sinyal atau beat. Adapun beberapa tujuannya adalah sebagai berikut:

- 1) Mendapatkan model dan metode penelitian yang lebih baik dan cocok.
- 2) Mendapatkan nilai akurasi yang lebih tinggi dari sebelumnya.

REFERENCES

- [1] M. Pojon, "Using machine learning to predict student performance," 2017.
- [2] M. Wasimuddin, K. Elleithy, A. S. Abuzneid, M. Faezipour, and O. Abuzaghlh, "Stages-based eeg signal analysis from traditional signal processing to machine learning approaches: A survey," *IEEE Access*, vol. 8, pp. 177782–177803, 2020.
- [3] J. Kuila, N. Dhandu, and S. Joardar, "Feature extraction and classification of mit-bih arrhythmia database," in *Proceedings of the 2nd International Conference on Communication, Devices and Computing*, pp. 417–427, Springer, 2020.
- [4] Z. F. M. Apandi, R. Ikeura, and S. Hayakawa, "Arrhythmia detection using mit-bih dataset: A review," in *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, pp. 1–5, IEEE, 2018.
- [5] T. Li and M. Zhou, "Ecg classification using wavelet packet entropy and random forests," *Entropy*, vol. 18, no. 8, p. 285, 2016.

- [6] B. Kumar and D. Gupta, "Universum based lagrangian twin bounded support vector machine to classify eeg signals," *Computer Methods and Programs in Biomedicine*, vol. 208, 9 2021.
- [7] N. K. Kamila, J. Frnda, S. K. Pani, R. Das, S. M. Islam, P. K. Bharti, and K. Muduli, "Machine learning model design for high performance cloud computing load balancing resiliency: An innovative approach," *Journal of King Saud University - Computer and Information Sciences*, 11 2022.
- [8] C. Renggli, S. Ashkboos, M. Aghagholzadeh, D. Alistarh, and T. Hoefler, "Sparcml: High-performance sparse communication for machine learning," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–15, 2019.
- [9] H. Bhatt, V. Shah, K. Shah, R. Shah, and M. Shah, "State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review," *Intelligent Medicine*, 10 2022.
- [10] M. Khanzadeh, S. Chowdhury, M. Marufuzzaman, M. A. Tschopp, and L. Bian, "Porosity prediction: Supervised-learning of thermal history for direct laser deposition," *Journal of manufacturing systems*, vol. 47, pp. 69–82, 2018.
- [11] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, pp. 2733–2742, 6 2022.
- [12] D. Makariou, P. Barrieu, and Y. Chen, "A random forest based approach for predicting spreads in the primary catastrophe bond market," *Insurance: Mathematics and Economics*, vol. 101, pp. 140–162, 11 2021.
- [13] S. H. Alizadeh, A. Hediehloo, and N. S. Harzevili, "Multi independent latent component extension of naive bayes classifier," *Knowledge-Based Systems*, vol. 213, 2 2021.
- [14] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for naïve bayes classification," *Computers and Operations Research*, vol. 135, 11 2021.
- [15] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve bayes algorithm," *Knowledge-Based Systems*, vol. 192, 3 2020.
- [16] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "Accurate detection of covid-19 patients based on distance biased naïve bayes (dbnb) classification strategy," *Pattern Recognition*, vol. 119, 11 2021.
- [17] M. Andrejiova and A. Grincova, "Classification of impact damage on a rubber-textile conveyor belt using naïve-bayes methodology," *Wear*, vol. 414–415, pp. 59–67, 11 2018.
- [18] N. Zamri, M. A. Pairan, W. N. A. W. Azman, S. S. Abas, L. Abdullah, S. Naim, Z. Tarmudi, and M. Gao, "River quality classification using different distances in k-nearest neighbors algorithm," *Procedia Computer Science*, vol. 204, pp. 180–186, 2022.
- [19] M. Cubillos, S. Wöhlk, and J. N. Wulff, "A bi-objective k-nearest-neighbors-based imputation method for multilevel data," *Expert Systems with Applications*, vol. 204, 10 2022.
- [20] H. Zhang, Y. Shi, X. Yang, and R. Zhou, "A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance," *Research in International Business and Finance*, vol. 58, p. 101482, 2021.
- [21] R. Jafari-Marandi, "Supervised or unsupervised learning? investigating the role of pattern recognition assumptions in the success of binary predictive prescriptions," *Neurocomputing*, vol. 434, pp. 165–193, 2021.
- [22] R. Solli, D. Bazin, M. Hjorth-Jensen, M. P. Kuchera, and R. R. Strauss, "Unsupervised learning for identifying events in active target experiments," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1010, 9 2021.
- [23] D. Zhao, X. Hu, S. Xiong, J. Tian, J. Xiang, J. Zhou, and H. Li, "k-means clustering and knn classification based on negative databases," *Applied Soft Computing*, vol. 110, 10 2021.
- [24] X. Wang, Z. Wang, M. Sheng, Q. Li, and W. Sheng, "An adaptive and opposite k-means operation based memetic algorithm for data clustering," *Neurocomputing*, vol. 437, pp. 131–142, 5 2021.
- [25] Z. Zhang, Q. Feng, J. Huang, Y. Guo, J. Xu, and J. Wang, "A local search algorithm for k-means with outliers," *Neurocomputing*, vol. 450, pp. 230–241, 8 2021.