BAB 2
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian ini didasari dari beberapa penelitian yang sudah ada sebelumnya dapat dilihat melalui Tabel 2.1

Tabel 2.1. Penelitian Terdahulu

Penulis	**Judul**	**Topik**	**Hasil**
Rüstem Yilmaz, Fatma Hilal Yağın	*Early Detection of Coronary Heart Disease Based on Machine Learning Methods* [6]	Algoritma *Random Forest*, *Support Vector Machine*, *Logistic Regression* dan Rasio *Split* Data *Train:Test* yaitu 80:20	Penelitian ini membandingkan tiga algoritma *Random Forest* (RF), *Support Vector Machine* (SVM), dan *Logistic Regression* (LR) untuk prediksi penyakit jantung dengan dataset bersumber dari IEEE*DataPort database* berjumlah 11 fitur dan 1 target. Hasil menunjukkan bahwa RF memiliki performa terbaik dengan akurasi 92,9%, diikuti oleh LR dengan akurasi 86.1%, dan SVM dengan akurasi 89.7%.
Lanjut pada halaman berikutnya			

Tabel 2.1 Penelitian Pendahulu (Lanjutan)

Penulis	**Judul**	**Topik**	**Hasil**
Charles Bemando, Eka Miranda, Mediana Aryuni	*Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms* [8]	Algoritma *Naïve Bayes* dan *Random Forest*	Penelitian menggunakan dataset dari *Cleveland database UCI repository of coronary heart disease patients* dengan 13 fitur dan 1 target. Hasil penelitian menunjukkan bahwa *Gaussian Naïve Bayes* dan *Bernoulli Naïve Bayes* mencapai akurasi 85% dan *Random Forest* memiliki akurasi 75%.
Nur Halizah Alfajr, Sofi Defiyanti	Prediksi Penyakit Jantung Menggunakan Metode *Random Forest* dan Penerapan *Principal Component Analysis* (PCA) [9]	Algoritma *Random Forest*	Dataset penelitian yang digunakan berasal dari UCI *Machine Learning Repository* dengan 13 fitur dan 1 target serta sampel sebanyak 1.026 pasien. Hasil penelitian menunjukkan bahwa algoritma *Random Forest* dalam prediksi penyakit jantung memiliki akurasi 98% dengan fitur *thal* yang sangat mempengaruhi akurasi model.

Lanjut pada halaman berikutnya

Tabel 2.1 Penelitian Pendahulu (Lanjutan)

Penulis	**Judul**	**Topik**	**Hasil**
Agung Rachmat Raharja, Jayadi, Angga Pramudianto, Yoki Muchsam	Penerapan Algoritma Decision Tree dalam Klasifikasi Data "Framingham" Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung dalam 10 Tahun Mendatang [10]	Algoritma *Decision Tree* dan *Dataset Framingham*	Dataset penelitian menggunakan dataset *framingham* dengan 15 fitur dan 1 target. Penelitian dengan algoritma *Decision Tree* menunjukkan hasil akurasi sebesar 74% diikuti dengan *precision* sebesar 84% untuk kelas 0 dan 21% untuk kelas 1, *recall* sebesar 84% untuk kelas 0 dan 22% untuk kelas 1, serta f1-*score* sebesar 84% untuk kelas 0 dan 22% untuk kelas 1.
Agus Fajar Riany, Gusmelia Testiana	PENERAPAN DATA MINING UNTUK KLASIFIKASI PENYAKIT JANTUNG KORONER MENGGUNAKAN ALGORITMA NAÏVE BAYES [11]	Algoritma *NAÏVE BAYES* *Dataset Framingham*	Dataset penelitian menggunakan dataset *framingham* dengan 15 fitur dan 1 target serta berjumlah 4.115 data. Penelitian dengan algoritma *NAÏVE BAYES* menunjukkan hasil akurasi sebesar 79.1% diikuti dengan *precision* sebesar 87.63% untuk kelas 0 dan 32.81% untuk kelas 1, *recall* sebesar 87.63% untuk kelas 0 dan 32.81% untuk kelas 1.
Lanjut pada halaman berikutnya			

Tabel 2.1 Penelitian Pendahulu (Lanjutan)

Penulis	**Judul**	**Topik**	**Hasil**
R. Valarmathi, T. Sheela	*Heart disease prediction using hyper parameter optimization (HPO) tuning* [12]	*Hyperparameter Tuning*	Dataset penelitian menggunakan *Cleveland Heart disease Dataset* dengan jumlah 13 fitur dan 1 target serta 303 data. Penelitian menggunakan algoritma *Random Forest* dan *hyperparameter tuning* menghasilkan akurasi 91.32% untuk metode *Grid Search*, akurasi 95.04% untuk metode *Randomized Search* dan 97.52% untuk metode *Genetic Programming*.
Muhammad Waqar, Hassan Dawood, Hussain Dawood, Nadeem Majeed, Ameen Banjar, Riad Alharbey	*An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction* [13]	SMOTE	Dataset penelitian menggunakan dataset UCI Dataset dengan jumlah 13 fitur dan 303 data. Penelitian menggunakan algoritma *Random Forest* dan SMOTE menghasilkan akurasi sebesar 85%.
Lanjut pada halaman berikutnya			

Tabel 2.1 Penelitian Pendahulu (Lanjutan)

Penulis	**Judul**	**Topik**	**Hasil**
Atta Ur Rahman, Yousef Alsenani, Adeel Zafar, Kalim Ullah, Khaled Rabie, Thokozani Shongwe	*Enhancing heart disease prediction using a self-attention-based transformer model* [14]	*Validation Set* sebesar 10%	Dataset penelitian menggunakan dataset UCI ML *repository Cleveland* dengan jumlah 13 fitur dan 303 data. Penelitian menggunakan *validation set* sejumlah 10% untuk validasi dan menghasilkan akurasi sebesar 96.51%.

Pada penelitian pertama dengan judul *Early Detection of Coronary Heart Disease Based on Machine Learning Methods* oleh Rüstem Yilmaz, Fatma Hilal Yağın dilakukan perbandingan antara tiga algoritma *machine learning* untuk klasifikasi penyakit jantung koroner yaitu *Random Forest*, *Support Vector Machine*, dan *Logistic Regression* dengan dataset berasal dari *IEEEDataPort database*. Model ini menunjukkan hasil bahwa model *Random Forest* memiliki performa terbaik dengan akurasi 92,9%, diikuti oleh *Logistic Regression* dengan akurasi 86.1%, dan SVM dengan akurasi 89.7% [6].

Pada penelitian kedua dengan judul *Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms* oleh Charles Bemando, Eka Miranda, Mediana Aryuni meneliti menggunakan *supervised machine learning* yaitu Gaussian *Naïve Bayes*, Bernoulli *Naïve Bayes*, dan *Random Forest*, dengan menggunakan data dari *Cleveland Database* UCI. Hasil penelitian menunjukkan bahwa *Gaussian Naïve Bayes* dan *Bernoulli Naïve Bayes* mencapai akurasi 85%, dan *Random Forest* memiliki akurasi 75% [8].

Pada penelitian ketiga dengan judul ”Prediksi Penyakit Jantung Menggunakan Metode *Random Forest* dan Penerapan *Principal Component Analysis* (PCA)” oleh Nur Halizah Alfajr, Sofi Defiyanti mengembangkan model

prediksi penyakit jantung dengan algoritma *Random Forest* menggunakan dataset dari UCI *Machine Learning Repository* yang terdiri dari 1.026 sampel pasien. Hasil penelitian menunjukkan bahwa model dengan 100 pohon keputusan mencapai akurasi 98.23%, presisi 100%, *recall* 96%, dan F1-*score* 98% [9].

Pada penelitian keempat dengan judul ”Penerapan Algoritma Decision Tree dalam Klasifikasi Data “Framingham” Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung dalam 10 Tahun Mendatang” oleh Agung Rachmat Raharja, Jayadi, Angga Pramudianto, Yoki Muchsam melakukan penelitian dengan dataset penelitian yaitu dataset *framingham* menggunakan 15 fitur dan 1 target. Penelitian dengan algoritma *Decision Tree* menunjukkan hasil akurasi sebesar 74% diikuti dengan metrik evaluasi yang cukup rendah pada kelas terkena penyakit jantung seperti *precision* sebesar 84% untuk kelas 0 dan 21% untuk kelas 1, *recall* sebesar 84% untuk kelas 0 dan 22% untuk kelas 1, serta f1-*score* sebesar 84% untuk kelas 0 dan 22% untuk kelas 1 [10].

Pada penelitian kelima dengan judul ”PENERAPAN DATA MINING UNTUK KLASIFIKASI PENYAKIT JANTUNG KORONER MENGGUNAKAN ALGORITMA NAÏVE BAYES” oleh Agus Fajar Riany dan Gusmelia Testiana melakukan hasil penelitian menggunakan dataset *framingham* dengan 15 fitur dan 1 target serta berjumlah 4.115 data. Penelitian dengan algoritma *NAÏVE BAYES* ini menunjukkan hasil akurasi sebesar 79.1% diikuti dengan metrik evaluasi yang rendah pada kelas terkena penyakit jantung seperti *precision* sebesar 87.63% untuk kelas 0 dan 32.81% untuk kelas 1, *recall* sebesar 87.63% untuk kelas 0 dan 32.81% untuk kelas 1 [11].

Pada penelitian keenam dengan judul *Heart disease prediction using hyper parameter optimization (HPO) tuning* oleh R. Valarmathi dan T. Sheela melakukan penelitian menggunakan dataset *Cleveland Heart disease Dataset* berjumlah 13 fitur dan 1 target serta 303 data. Penelitian menggunakan algoritma *Random Forest* dan melakukan *hyperparameter tuning* menghasilkan akurasi 91.32% untuk metode *Grid Search*, akurasi 95.04% untuk metode *Randomized Search* dan 97.52% untuk metode *Genetic Programming* [12].

Pada penelitian ketujuh dengan judul *An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction* oleh Muhammad Waqar, Hassan Dawood, Hussain Dawood, Nadeem Majeed, Ameen Banjar, Riad Alharbey melakukan penelitian dengan *dataset* penelitian UCI berjumlah 13 fitur dan 1 target serta 303 data. Penelitian menggunakan algoritma *Random Forest* dan dengan penggunaan SMOTE menghasilkan akurasi sebesar 85% [13].

Pada penelitian kedelapan dengan judul *Enhancing heart disease prediction using a self-attention-based transformer model* oleh Atta Ur Rahman, Yousef Alsenani, Adeel Zafar, Kalim Ullah, Khaled Rabie, Thokozani Shongwe melakukan penelitian menggunakan dataset UCI ML *repository Cleveland* dengan jumlah 13 fitur dan 1 target serta 303 data. Penelitian menggunakan *validation set* sejumlah 10% ini menghasilkan akurasi sebesar 96.51% [14].

2.2 Penyakit Jantung

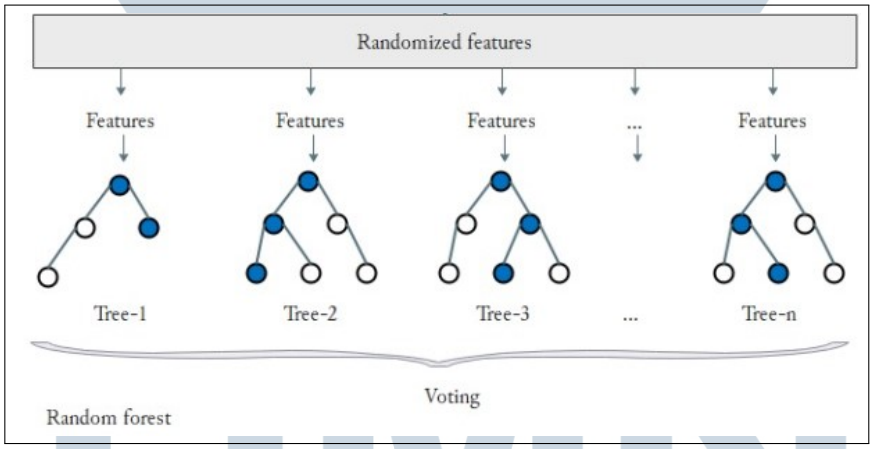
Penyakit kardiovaskular merupakan penyakit gangguan pada jantung dan pembuluh darah yang menjadi penyebab utama dalam kematian di dunia. Faktor resiko yang memicu terjangkitnya penyakit jantung seperti usia, jenis kelamin, genetik atau riwayat keluarga, pola hidup, dan faktor internal seperti tekanan darah dan kadar kolesterol [15]. Penyakit kardiovaskular menjadi masalah kesehatan global termasuk di Indonesia, hal ini dapat meningkatkan angka kesakitan, kecacatan, dan beban ekonomi.

Penyakit Jantung Koroner merupakan gangguan fungsionalitas jantung yang diakibatkan oleh otot jantung kekurangan aliran darah karena penyempitan pembuluh darah koroner. Penyakit ini ditandai dengan rasa nyeri pada dada atau rasa tidak nyaman di dada hingga terasa tertekan berat ketika sedang beraktivitas yang padat. Penyakit Jantung Koroner terdiri dari penyakit jantung koroner stabil tanpa gejala, angina pektoris stabil, dan sindrom koroner akut. Penyakit jantung koroner stabil tanpa gejala dapat diketahui melalui skrining kesehatan, sedangkan angina pektoris stabil dapat diketahui melalui gejala seperti nyeri pada dada akibat aktivitas sehari hari yang berat [16].

Analisis terhadap kualitas hidup pasien dengan Penyakit Jantung Koroner menjadi penting dalam memahami prediksi atau perkiraan dan pengaruhnya terhadap terapi serta rehabilitasi. Menurut penelitian *Assessment of the Quality of Life of Patients with Coronary Heart Disease* menunjukkan bahwa faktor seperti usia dan riwayat penyakit jantung koroner memiliki hubungan positif dengan kualitas hidup pada pasien Penyakit Jantung Koroner [17]. Kemudian, menurut penelitian dengan judul ”Gambaran faktor yang mempengaruhi kejadian penyakit jantung koroner (pjk) di pusat jantung terpadu (pjt)” faktor yang dapat mempengaruhi Penyakit Jantung Koroner antara lain, jenis kelamin, usia, riwayat merokok, obesitas, dan hipertensi [18].

2.3 Algoritma Random Forest

Algoritma *Random Forest* (RF) merupakan algoritma *machine learning* yang digunakan dalam klasifikasi dan pendeteksi. Algoritma ini dibangun dari beberapa *Decision Tree* dari subset data acak lalu digabungkan hasil dari setiap pohon untuk mengeluarkan prediksi akhir. Sehingga, RF dapat mengatasi masalah data yang tidak seimbang dan variabel dengan *missing value*. Algoritma RF memiliki kemampuan dalam menghindari masalah *overfitting* karena tiap pohon telah dilatih menggunakan subset data berbeda. Data pada algoritma RF memungkinkan pengenalan pola, pengambilan keputusan, pembuatan model prediktif, dan pemecahan masalah. Algoritma ini juga memungkinkan dalam menganalisis ribuan variabel input tanpa menghapus salah satu dari variabelnya [19]. Diagram alur dari proses kerja algoritma *Random Forest* dapat dilihat melalui Gambar 2.1.



Gambar 2.1. Diagram Alur Cara Kerja Algoritma Random Forest
Sumber: [20]

Pada Gambar 2.1 menunjukkan proses dari cara kerja algoritma *Random Forest* sebagai berikut [20].

1. *Randomized Features* dengan memilih fitur dari dataset secara acak untuk setiap pohon keputusan
2. Pembuatan beberapa pohon keputusan dengan menggunakan *subset* fitur yang telah dipilih secara acak
3. Proses *voting* atau *Ensemble Learning* ketika setelah semua pohon keputusan terbentuk, setiap pohon memberikan prediksi sendiri terhadap suatu data

input

4. Pengambilan keputusan akhir atau *Final Prediction Output* ditentukan berdasarkan mayoritas *voting* dari semua pohon keputusan yang ada

Dalam konteks *data mining*, RF dengan pendekatan deskriptif bertujuan dalam memberikan deskripsi dan ringkasan data, dan RF dengan pendekatan prediktif menggunakan data historis dalam menemukan tren atau pola yang dapat diprediksi di masa mendatang. Proses *data mining* prediktif pada umumnya menggunakan metode *Supervised Learning* atau pembelajaran terawasi, dan *data mining* deskriptif mayoritas menggunakan *Unsupervised Learning* atau pembelajaran tidak terawasi. RF dapat menangani data dalam skala besar dan mengakomodir banyak variabel tanpa harus dihapus. Maka, algoritma ini sering digunakan dalam pengkategorian otomatis, analisis data prediktif, dan pembelajaran yang diawasi [19].

Dalam meningkatkan akurasi prediksi pada algoritma *Random Forest* dapat melakukan proses *parameter tuning*. Algoritma *Random Forest* menggunakan konsep *Wisdom of Crowds* yaitu memiliki banyak model yang saling tidak bergantung bertindak sebagai satu kelompok. Cara ini dapat meminimalisir kesalahan sehingga dapat menghasilkan prediksi lebih akurat. Kelebihan algoritma RF dapat dibagi menjadi 2 aspek, antara lain [21].

1. Keunggulan Komputasional

   RF secara umum dikenal dengan algoritma yang fleksibel dan memiliki akurasi tinggi. Algoritma RF dapat meng*handle* tugas regresi dan klasifikasi bahkan *multiclass*. Pada proses pelatihan dan prediksi RF relatif cepat walau diterapkan pada dataset besar dengan noise atau memiliki dimensi tinggi.

2. Keunggulan Statistik

   RF tidak bergantung pada asumsi statistik tertentu dan pra pemrosesan yang mudah. RF memiliki varians lebih rendah dan *resist* terhadap *overfitting* dibandingkan dengan satu *decision tree*. Selain itu, RF lebih kuat daripada metode *boosting* karena dapat mengatasi masalah dengan bobot kelas berbeda, deteksi outlier, dan mengisi *missing value*.

Sehingga, *Random Forest* merupakan metode *ensemble learning* yang membangun banyak pohon keputusan (*n_estimators*) secara independen untuk klasifikasi. Proses ini dimulai dengan pembentukan setiap pohon menggunakan

subset data pelatihan yang diambil secara acak dengan pengembalian (teknik *bagging*), menciptakan keragaman pada data yang digunakan. Untuk menambah keragaman dan mengurangi *overfitting*, pada setiap *node* di setiap pohon, algoritma tidak mempertimbangkan semua fitur yang tersedia, melainkan hanya memilih subset acak dari fitur (*max_features*) sebagai kandidat untuk mencari titik pemisahan (*split*) terbaik. Pemisahan ini dilakukan berdasarkan kriteria yang dipilih (*criterion*), dengan tujuan memaksimalkan kemurnian setiap subset data yang dihasilkan. Dalam menentukan titik ambang batas (*threshold*) pada fitur numerik, algoritma akan mengurutkan nilai-nilai fitur yang unik dan mengevaluasi setiap titik di antara nilai-nilai tersebut. Titik yang menghasilkan pemisahan paling murni akan dipilih. Proses pemisahan ini berulang secara rekursif dari *node* akar hingga mencapai *node* daun, setiap *node* daun merepresentasikan prediksi kelas. Setelah semua pohon selesai dibangun, *Random Forest* mengumpulkan prediksi dari setiap pohon dan mengambil keputusan akhir berdasarkan suara terbanyak (*majority vote*).

2.4 Confusion Matrix

Confusion Matrix merupakan alat untuk melakukan evaluasi kinerja model. Alat ukur ini memperlihatkan perbandingan antara hasil prediksi model dengan data aktual dari kelas yang sama. *Confusion Matrix* ini bertujuan untuk mengidentifikasi kesalahan dalam klasifikasi yang telah dibuat oleh model dan memberikan pengetahuan lebih mengenai model yang perlu diperbaiki. *Confusion Matrix* terdiri dari dua sumbu yaitu sumbu vertikal atau baris sebagai perwakilan untuk *actual class* dari data, sedangkan sumbu horizontal atau kolom sebagai perwakilan untuk *predicted class* dari model. Komponen *Confusion Matrix* terdiri dari: [22].

1. *True Positive*

   True Positive yaitu data positif yang diprediksi dengan benar sebagai positif.

2. *False Positive*

   False Positive yaitu data negatif yang diprediksi secara salah sebagai positif atau *Type I Error*.

3. *True Negative*

   True Negative yaitu data negatif yang diprediksi dengan benar sebagai negatif.

4. *False Negative*

False Negative yaitu data positif yang diprediksi secara salah sebagai negatif atau *Type II Error*.

Berikut merupakan metode evaluasi menggunakan metrik evaluasi, antara lain.

1. Akurasi (*Accuracy)*

   Akurasi yaitu akurasi mengukur proporsi prediksi yang benar terhadap total prediksi.

   Rumus 2.1 menunjukkan cara perhitungan *Accuracy*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

2. Presisi (*Precision*)

   Presisi yaitu presisi mengukur seberapa akurat model dalam memprediksi kelas positif. Rumus 2.2 menunjukkan cara perhitungan *Precision*.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

3. Recall

   Recall yaitu untuk mengukur seberapa baik model dalam mendeteksi semua kasus positif.

   Rumus 2.3 menunjukkan cara perhitungan *Recall*.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

4. F1-*Score*

   F1-*Score* yaitu rata rata harmonis dari presisi dan *recall* ketika distribusi kelas tidak seimbang.

   Rumus 2.4 menunjukkan cara perhitungan *F1 Score*.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

2.5 SMOTE

Distribusi data pada kelas yang tidak seimbang terjadi ketika jumlah dari satu kelas jauh lebih sedikit dibandingkan kelas yang lain. Kondisi ini memungkinkan untuk memberikan dampak penurunan pada akurasi prediksi terhadap suatu data yang memiliki kelas minoritas. Hal ini dikarenakan model lebih terfokus untuk mempelajari pada pola kelas mayoritas yang mendominasi. Dalam mengatasi hal ini, SMOTE atau *Synthetic Minority Oversampling Technique* digunakan sebagai salah satu metode untuk *oversampling* dalam penelitian dengan kasus ketidakseimbangan data pada suatu kelas. Cara kerja SMOTE yaitu dengan cara membuat data sintetis atau data baru berdasarkan sampel dari kelas minoritas untuk menghindari *overfitting*. Proses SMOTE ini dilakukan untuk menyeimbangkan jumlah data antar kelas tanpa perlu menggandakan data. SMOTE ini dapat memperjelas batas keputusan antara kelas mayoritas dan kelas minoritas [23].

2.6 Random Search

Random Search merupakan metode dalam mencari kombinasi hyperparameter terbaik secara acak dari distribusi tertentu. Metode ini dapat digunakan dengan banyak parameter dan lebih efisien. Metode ini mempercepat proses pencarian dengan memfokuskan pada parameter-parameter yang paling berpengaruh terhadap kinerja model [24]. Random Search menjadi salah satu metode optimasi hyperparameter dengan mengeksplorasi ruang parameter secara acak dari sejumlah kombinasi parameter untuk diuji. Pendekatan ini memungkinkan eksplorasi yang lebih luas dan efisien terhadap ruang pencarian. Nilai parameter yang dicoba pada Random Search dapat disesuaikan untuk menemukan kombinasi yang memberikan kinerja terbaik. Sehingga, Random Search dapat menjadi solusi efisien dalam optimasi model terutama ketika jumlah hyperparameter yang cukup banyak [25].

2.7 Log Loss Validation

Log loss adalah metrik evaluasi yang digunakan untuk mengukur performa model klasifikasi. *Log loss* mengevaluasi seberapa dekat probabilitas prediksi dengan label yang sebenarnya, dan semakin rendah nilai *log loss*, maka semakin baik performa model tersebut. *Log loss* dapat digunakan tidak hanya untuk

mengevaluasi akurasi prediksi, tetapi juga sebagai indikator untuk mendeteksi *overfitting*, melalui analisis kurva pelatihan yang mencakup nilai *training loss* dan *validation loss*. Model dikatakan *overfitting* jika hanya menyesuaikan diri dengan data pelatihan dan gagal mempelajari pola umum dari data. *Overfitting* dapat dikenali dari perbedaan (*gap*) yang signifikan antara nilai *training loss* dan *validation loss* [26].

2.8 Splitting Rasio Data

Pada penelitian ini, salah satu *splitting* rasio data *training* dan *testing* yang digunakan yaitu 80:20 seperti pada penelitian dengan judul *Early Detection of Coronary Heart Disease Based on Machine Learning Methods* oleh Rüstem Yilmaz, Fatma Hilal Yağın dilakukan klasifikasi penyakit jantung koroner dengan *Random Forest* memiliki performa terbaik dengan akurasi 92,9% [6]. Selain itu, penelitian juga melakukan uji coba dengan skenario *splitting* rasio data lainnya yang didasari dari beberapa penelitian yang telah ada seperti 90:10, 85:15, 75:25, 70:30, dan 65:35. Menurut penelitian dengan judul *Max Depth Impact on Heart Disease Classification: Decision Tree and Random Forest* oleh Rian Oktafiani, Arief Hermawan, Donny Avianto menunjukkan hasil penelitian terbaik dari algoritma *Random Forest* ketika *splitting* rasio data 90:10 memiliki hasil akurasi sebesar 99.29%. Kemudian, pada penelitian berjudul *A Comparative Study for Time-to-Event Analysis and Survival Prediction for Heart Failure Condition using Machine Learning Techniques* oleh Saurav Mishra menunjukkan hasil penelitian menggunakan algoritma *Random Forest* dengan perbandingan rasio data 85:15 menghasilkan akurasi sebesar 80%. Selanjutnya, penelitian berjudul *Enhanced cardiovascular disease prediction model using random forest algorithm* oleh Kellen Sumwiza, Celestin Twizere, Gerard Rushingabigwi, Pierre Bakunzibake, dan Peace Bamurigire menunjukkan hasil penelitiannya menggunakan algoritma *Random Forest* dengan rasio data 75:25 memiliki hasil akurasi sebesar 96% sebelum melakukan *feature selection* dan akurasi sebesar 99% setelah melakukan *feature selection*. Berikutnya, penelitian dengan judul IMPACT OF DATA SPLITTING ON PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR PREDICTING GESTATIONAL DIABETES oleh Ikechukwu Okechi KAMALU menunjukkan hasil penelitiannya menggunakan algoritma *Random Forest* dengan rasio data 70:30 menghasilkan akurasi sebesar 72.1%. Selain itu, penelitian berjudul LABORATORY TESTS TO DETECT COVID-19 IN

SOUTH AFRICA USING STATISTICAL ANALYSIS AND RANDOM FOREST MODELLING oleh Mark Strickett dan Dr. Farai Mlambo menunjukkan hasil bahwa penelitian pada penyakit COVID-19 menggunakan algoritma *Random Forest* dengan rasio data 65:35 menghasilkan akurasi sebesar 81.26%. Pada skripsi dengan judul ”Rancang Bangun Sistem Klasifikasi Tingkat Obesitas dengan Algoritma Random Forest Classifier” oleh Tesalonika Abigail melakukan penelitian dengan algoritma *Random Forest* dan rasio data 80:20 menggunakan *Hyperparameter Tuning* yaitu metode *RandomizedSearchCV* menghasilkan akurasi model sebesar 96.8% [6, 27, 28, 29, 30, 31, 32].