

# Analisis Kinerja Model Transformer untuk Klasifikasi Atrial Fibrilasi Berbasis Representasi Statistik Sinyal Elektrokardiogram Time-Series

Berliana Sarlita Rahajeng<sup>1</sup>, Raditya Akmal<sup>2</sup>, Wijdan Razefi Al Hawwari<sup>3</sup>

Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo, Kec. Sukolilo, Kota Surabaya,  
Prov. Jawa Timur, 031-5994251-4 (1132)

E-mail : 5054241023@student.its.ac.id<sup>1</sup>, 5054241027@student.its.ac.id<sup>2</sup>,  
5054241044@student.its.ac.id<sup>3</sup>

**Abstract** - Cardiovascular diseases remain the leading cause of global mortality, with Atrial Fibrillation (AF) being one of the most prevalent cardiac arrhythmias and a major risk factor for stroke and cognitive decline. Early and reliable detection of AF from Electrocardiogram (ECG) signals is therefore critical. This study investigates and compares the performance of sequence-based deep learning models, namely Long Short-Term Memory (LSTM) and Transformer, for AF classification using statistically represented ECG time-series data. ECG signals from the MIMIC Perform AF dataset were segmented into one-second windows and transformed into sequential statistical features consisting of skewness, kurtosis, minimum, maximum, and standard deviation. Both end-to-end single models and hybrid models combining deep learning feature extractors with machine learning classifiers (SVM, XGBoost, and Random Forest) were evaluated. Furthermore, three hyperparameter tuning strategies—default configuration, Grid Search, and Optuna—were systematically compared, along with balanced and imbalanced data scenarios. Experimental results show that the single LSTM model achieved the highest recall, making it the most effective architecture for AF detection on this dataset. Conversely, Transformer demonstrated superior performance as a feature extractor in hybrid models, particularly when combined with Random Forest. These findings highlight the importance of model architecture selection, hyperparameter optimization, and data balancing in developing reliable ECG-based AF detection systems for clinical and wearable healthcare applications.

**Keywords** - Atrial Fibrillation, Electrocardiogram (ECG), Transformer, Long Short-Term Memory (LSTM), Deep Learning

## 1. INTRODUCTION

Penyakit kardiovaskular (*Cardiovascular diseases/CVD*) merupakan penyebab kematian tertinggi di dunia. Laporan World Health Organization (WHO) menunjukkan bahwa pada tahun 2022 sekitar 19,8 juta kematian disebabkan oleh CVD, atau setara dengan 32% dari total kematian global. Sebagian besar kematian tersebut disebabkan oleh komplikasi kardiovaskular seperti serangan jantung dan stroke. Kondisi ini semakin diperparah oleh fakta bahwa lebih dari tiga perempat kematian akibat CVD terjadi di negara berpenghasilan rendah dan menengah [1], sehingga deteksi dini dan pemantauan kondisi jantung menjadi sangat penting untuk menekan dampak klinis dan ekonomi jangka panjang.

Salah satu kondisi yang termasuk dalam spektrum penyakit kardiovaskular adalah Atrial Fibrilasi (AF), yaitu gangguan irama jantung (aritmia) yang ditandai oleh aktivitas listrik atrium yang abnormal. AF merupakan aritmia yang paling banyak diteliti selama satu abad terakhir dan menjadi jenis aritmia paling umum pada manusia, dengan prevalensi sekitar 2% pada populasi umum serta meningkat hingga 10–12% pada individu berusia 80 tahun ke atas. Secara global, AF diperkirakan memengaruhi hingga 33,5 juta orang, dengan prevalensi 2,5–3,5% di berbagai negara [2]. AF juga berperan sebagai faktor risiko terhadap penurunan fungsi kognitif dan

demensia, terutama pada pasien dengan riwayat stroke, sehingga menunjukkan besarnya dampak klinis kondisi ini terhadap kesehatan jangka panjang [3].

*Electrocardiogram* (ECG) merupakan modalitas utama dalam diagnosis AF karena mampu merekam aktivitas listrik jantung secara langsung. Sinyal ECG memiliki karakteristik time-series yang kompleks dengan pola temporal yang mencerminkan dinamika ritme jantung. Pada kondisi AF, EKG menunjukkan beberapa karakteristik khas, yaitu tidak tampaknya gelombang P yang normal, interval RR yang sangat ireguler, serta munculnya gelombang fibrilasi (F waves) yang acak. Laju ventrikel pada EKG umumnya berkisar 110–140 kali per menit, meskipun jarang melebihi 160–170 kali per menit. Selain membantu konfirmasi diagnosis, EKG juga diperlukan untuk memantau interval QT dan QRS pada pasien yang mendapat terapi antiaritmia. Pemeriksaan ini menjadi langkah dasar setelah anamnesis dan pemeriksaan fisik, terutama pada pasien dengan keluhan palpitasi, pusing, intoleransi aktivitas, atau gejala terkait gangguan irama jantung[4]. Tantangan utama dalam analisis sinyal ECG adalah bagaimana merepresentasikan informasi temporal tersebut secara efektif agar perbedaan antara kondisi AF dan non-AF dapat dikenali secara akurat oleh model komputasi.

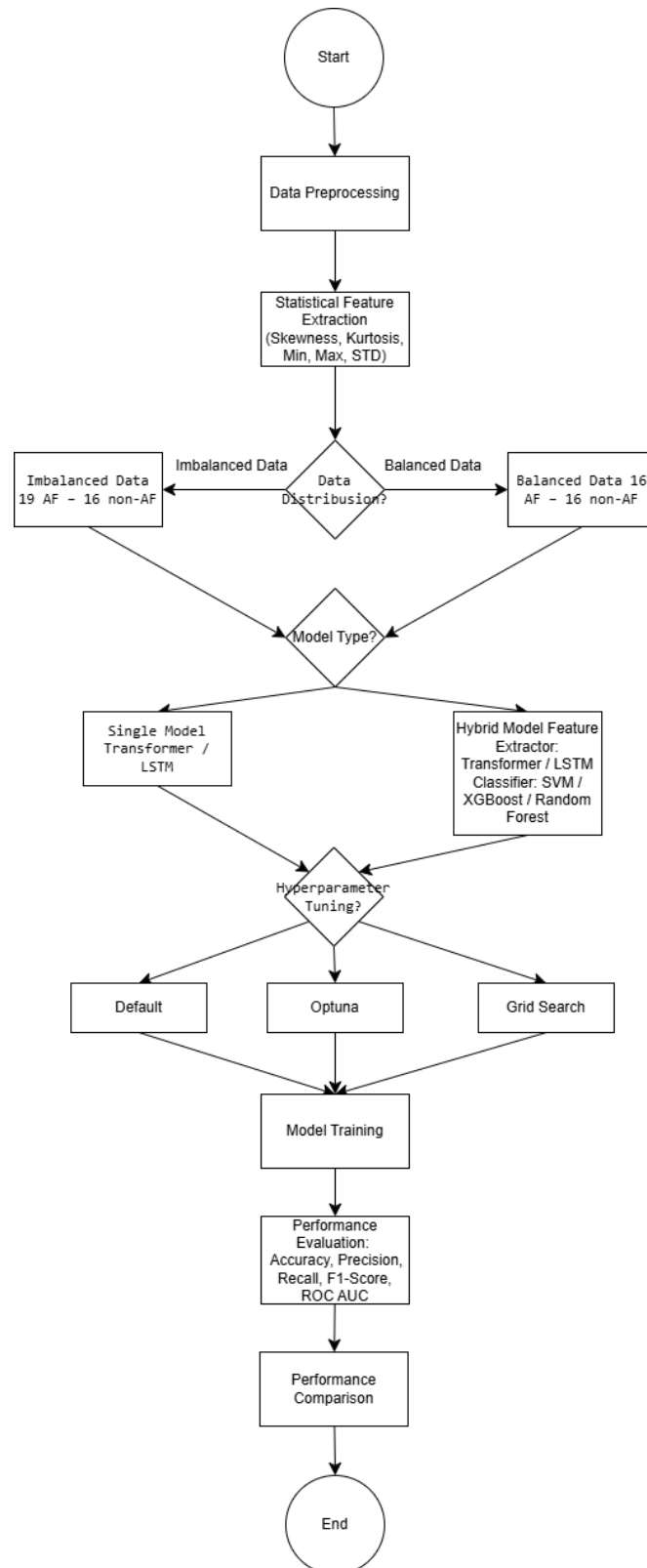
Dalam beberapa tahun terakhir, pendekatan *deep learning* banyak digunakan untuk analisis sinyal ECG karena kemampuannya dalam memodelkan dependensi temporal[5] [6]. Model berbasis sekuens seperti *Long Short-Term Memory* (LSTM) dan Transformer telah menunjukkan potensi yang signifikan[7][8], meskipun perbedaan arsitektur keduanya memungkinkan terciptanya performa yang berbeda ketika diterapkan pada representasi sinyal ECG.

Selain pendekatan *end-to-end*, representasi statistik sinyal ECG juga banyak digunakan untuk menyederhanakan sinyal berdimensi tinggi menjadi fitur yang lebih ringkas namun informatif[9]. Fitur-fitur statistik ini dapat disusun dalam bentuk sekuens sehingga tetap mempertahankan informasi temporal dan dapat digunakan baik oleh model *deep learning* maupun algoritma *machine learning* konvensional.

Pendekatan *hybrid* yang menggabungkan *deep learning* sebagai *feature extractor* dan algoritma *machine learning* sebagai *classifier* juga menunjukkan potensi dalam meningkatkan stabilitas dan performa klasifikasi[10][11]. Namun, studi yang secara sistematis membandingkan kinerja Transformer dan LSTM pada representasi statistik sinyal ECG, baik sebagai model tunggal maupun sebagai bagian dari pendekatan *hybrid*, masih relatif terbatas.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan performa model Transformer dan LSTM dalam klasifikasi Atrial Fibrilasi berbasis sinyal ECG time-series yang direpresentasikan secara statistik, dengan mengevaluasi model tunggal, model *hybrid*, strategi *hyperparameter tuning*, serta penanganan ketidakseimbangan data.

## 2. RESEARCH METHOD



Gambar 1. Alur Metodologi Penelitian

### 2.1. Research Design

Penelitian ini menggunakan pendekatan eksperimental untuk mengevaluasi kinerja berbagai model klasifikasi *Atrial Fibrillation* (AF) berbasis sinyal *Electrocardiogram* (ECG) time-series. Desain penelitian difokuskan pada analisis perbandingan antara model *deep learning* berbasis sekuens, yaitu Transformer dan *Long Short-Term Memory* (LSTM), baik sebagai model klasifikasi tunggal maupun sebagai *feature extractor* dalam pendekatan *hybrid deep learning-machine learning*. Eksperimen dilakukan dalam beberapa skenario pengujian, meliputi:

1. Perbandingan performa model Transformer dan LSTM sebagai model tunggal.
2. Evaluasi model *hybrid* dengan mengombinasikan Transformer dan LSTM sebagai *feature extractor* dengan algoritma *machine learning* (SVM/Support Vector Machine, XGBoost/Extreme Gradient Boosting, dan Random Forest) sebagai *classifier*.
3. Analisis pengaruh *hyperparameter tuning* Grid search dan Optuna dibandingkan dengan konfigurasi default.
4. Evaluasi dampak penanganan ketidakseimbangan data (*class imbalance*) terhadap performa model tunggal dan *hybrid*.

### 2.2. Dataset

Dataset yang digunakan dalam penelitian ini adalah MIMIC Perform AF Dataset[12] yang diekstraksi dari MIMIC-III Waveform Database[13], yaitu dataset sinyal fisiologis yang dirancang untuk penelitian deteksi *Atrial Fibrillation* (AF). Dataset ini berisi rekaman sinyal *Electrocardiogram* (ECG), *Photoplethysmography* (PPG), dan respirasi dengan durasi 20 menit per subjek.

Dataset mencakup 35 subjek dewasa, yang terdiri dari 19 subjek dengan kondisi *Atrial Fibrillation* (AF) dan 16 subjek dengan irama sinus normal (non-AF). Seluruh data direkam dari pasien dewasa kritis selama perawatan klinis rutin menggunakan *bedside patient monitor* dengan frekuensi sampling 125 Hz. Label kondisi AF dan non-AF diperoleh melalui proses anotasi manual berdasarkan referensi klinis yang tersedia pada publikasi terkait dataset.

Meskipun dataset menyediakan sinyal ECG, PPG, dan respirasi, penelitian ini hanya menggunakan sinyal ECG sebagai sumber data utama. Sinyal PPG dan respirasi dihapus pada tahap awal preprocessing untuk menjaga fokus penelitian pada analisis sinyal ECG serta menghindari pengaruh informasi multimodal di luar ruang lingkup eksperimen yang dilakukan.

### 2.3. Data Preprocessing

Tahap *preprocessing* dilakukan untuk menyiapkan sinyal ECG agar dapat digunakan secara efektif oleh model klasifikasi. Langkah pertama adalah pembagian dataset menjadi data latih dan data uji dengan rasio 65% untuk data latih dan 35% untuk data uji. Pembagian ini dilakukan untuk memastikan evaluasi performa model dilakukan pada data yang tidak digunakan selama proses pelatihan.

Selanjutnya, dilakukan normalisasi fitur menggunakan metode *Standard Scaler*, yang mengubah distribusi data sehingga memiliki nilai rata-rata nol dan standar deviasi satu. Normalisasi ini bertujuan untuk mengurangi pengaruh perbedaan skala antar fitur serta meningkatkan stabilitas dan konvergensi proses pelatihan model, baik pada model *deep learning* maupun *machine learning*[14].

Sinyal ECG kemudian disegmentasi menggunakan teknik *windowing* dengan frekuensi sampling 125 Hz dan ukuran window 125 sampel, yang merepresentasikan durasi 1 detik sinyal. Untuk menangkap dinamika temporal yang lebih panjang, *window* tersebut disusun menjadi sekuens dengan *sequence length* sebanyak 10 window, sehingga setiap sampel input merepresentasikan informasi sinyal ECG selama 10 detik. Pada tahap ini, kolom sinyal PPG dan respirasi dihapus, sehingga seluruh proses analisis dan pemodelan difokuskan secara eksklusif pada sinyal ECG.

#### 2.4. Feature Extraction

Ekstraksi fitur dilakukan secara statistik pada setiap *window* sinyal ECG untuk mereduksi dimensi data sekaligus mempertahankan karakteristik penting sinyal[9]. Lima fitur statistik yang diekstraksi pada setiap *window* adalah *Skewness*(1), *kurtosis*(2), Nilai minimum(3), Nilai maksimum(4) dan Standar deviasi/STD(5).

$$Skewness = \frac{1}{N} \frac{\sum_{i=1}^N (x_i - mean)^3}{STD^3} \quad (1)$$

$$Kurtosis = \frac{1}{N} \frac{\sum_{i=1}^N (x_i - mean)^4}{STD^4} \quad (2)$$

$$Min = \min (x_i) \quad (3)$$

$$Max = \max (x_i) \quad (4)$$

$$STD = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - mean)^2} \quad (5)$$

Parameter statistik ini digunakan untuk merepresentasikan karakteristik distribusi amplitudo sinyal ECG dalam setiap *window*, termasuk bentuk distribusi, tingkat variasi, dan rentang nilai sinyal. Pendekatan ini bertujuan untuk menyederhanakan representasi sinyal ECG berdimensi tinggi menjadi fitur yang lebih ringkas dan terstruktur.

#### 2.5. Model Architecture

##### 2.5.1. Single Model

Model Transformer dan LSTM digunakan sebagai model klasifikasi tunggal untuk mempelajari hubungan temporal antar *window* dalam satu sekuens. Kedua model menerima input berupa sekuens fitur statistik hasil ekstraksi dan menghasilkan prediksi kelas AF atau non-AF secara langsung.

LSTM memodelkan dependensi temporal melalui mekanisme *gating*, termasuk *forget*, input, dan output *gate*, yang dirancang untuk mempertahankan informasi jangka panjang dan mengatasi masalah *vanishing gradient* dalam RNNs. Sebaliknya, Transformer mengandalkan mekanisme *self-attention* untuk mempelajari hubungan antar semua elemen dalam sekuens secara paralel, memungkinkan model menangkap dependensi temporal global tanpa ketergantungan urutan komputasi rekuren[15]. Pendekatan ini telah diadaptasi pada sinyal klinis seperti ECG dalam kajian terkini.

##### 2.5.2. Hybrid Deep Learning–Machine Learning Models

Pada pendekatan *hybrid*, model Transformer dan LSTM digunakan sebagai *feature extractor*. Representasi laten yang dihasilkan dari lapisan terakhir sebelum lapisan klasifikasi pada model Transformer dan LSTM digunakan sebagai fitur masukan bagi algoritma *machine learning*, yaitu SVM, XGBoost, dan Random Forest. Pendekatan ini bertujuan untuk menggabungkan kemampuan *deep learning* dalam mempelajari dependensi temporal dengan kestabilan dan kemampuan generalisasi algoritma *machine learning* dalam proses klasifikasi.

#### 2.6. Hyperparameter Tuning Strategy

Dalam penelitian ini, dilakukan perbandingan tiga strategi *hyperparameter tuning*, yaitu penggunaan *hyperparameter default* (standar), Grid Search, dan Optuna. Pendekatan *default hyperparameters* digunakan sebagai *baseline* dengan memanfaatkan nilai bawaan dari masing-masing model tanpa proses optimasi tambahan. Grid Search diterapkan untuk mengeksplorasi kombinasi hyperparameter secara sistematis berdasarkan ruang pencarian yang telah ditentukan

sebelumnya, sehingga memungkinkan evaluasi menyeluruh terhadap berbagai konfigurasi model meskipun dengan biaya komputasi yang relatif tinggi. Sementara itu, Optuna digunakan sebagai metode optimasi berbasis *Bayesian optimization* yang melakukan pencarian hyperparameter secara adaptif dan efisien berdasarkan hasil iterasi sebelumnya[16]. Perbandingan ketiga strategi ini bertujuan untuk mengevaluasi efektivitas dan efisiensi masing-masing pendekatan dalam meningkatkan performa klasifikasi *Atrial Fibrillation*, baik pada model tunggal maupun model *hybrid*.

## 2.7. Handling Data Imbalance

Dalam penelitian ini, dilakukan evaluasi terhadap pengaruh ketidakseimbangan kelas (*class imbalance*) terhadap performa klasifikasi *Atrial Fibrillation* (AF). Ketidakseimbangan kelas sering dijumpai pada dataset medis karena distribusi beberapa kondisi penyakit yang lebih jarang dibandingkan kondisi normal. Dalam kondisi data yang tidak seimbang, algoritma *machine learning* cenderung bias terhadap kelas mayoritas sehingga kurang mampu mengenali kelas minoritas, yang dalam konteks diagnosis medis dapat berdampak buruk karena kasus positif mungkin tidak terdeteksi[17].

Dataset MIMIC Perform AF yang digunakan dalam penelitian ini terdiri dari 35 subjek, dengan 19 subjek AF dan 16 subjek non-AF. Komposisi ini digunakan sebagai skenario imbalanced data, yang merepresentasikan distribusi kelas asli dataset. Pada skenario ini, seluruh data digunakan tanpa melakukan penyesuaian jumlah subjek antar kelas, sehingga model dievaluasi pada kondisi distribusi data yang mendekati situasi nyata.

Sebagai pembandingan, dilakukan skenario *balanced data* dengan menyeimbangkan jumlah subjek pada masing-masing kelas. Pada skenario ini, digunakan 16 subjek AF dan 16 subjek non-AF, sehingga kedua kelas memiliki proporsi yang sama. Penyeimbangan dilakukan pada tingkat subjek untuk menghindari potensi data leakage akibat pemilihan *window* secara acak dari satu subjek yang sama. Pendekatan ini memastikan bahwa distribusi kelas seimbang tanpa mengubah karakteristik temporal sinyal ECG secara signifikan.

## 2.8. Model Evaluation

Performa model dievaluasi menggunakan data uji dengan metrik evaluasi yang umum digunakan pada masalah klasifikasi biner, yaitu *accuracy*(6), *precision*(7), *recall*(8), *F1-score*(9), dan *Area Under the Curve* (AUC). Penggunaan beberapa metrik evaluasi bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai kemampuan model dalam mendeteksi *Atrial Fibrillation*, khususnya pada kondisi data yang tidak seimbang.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

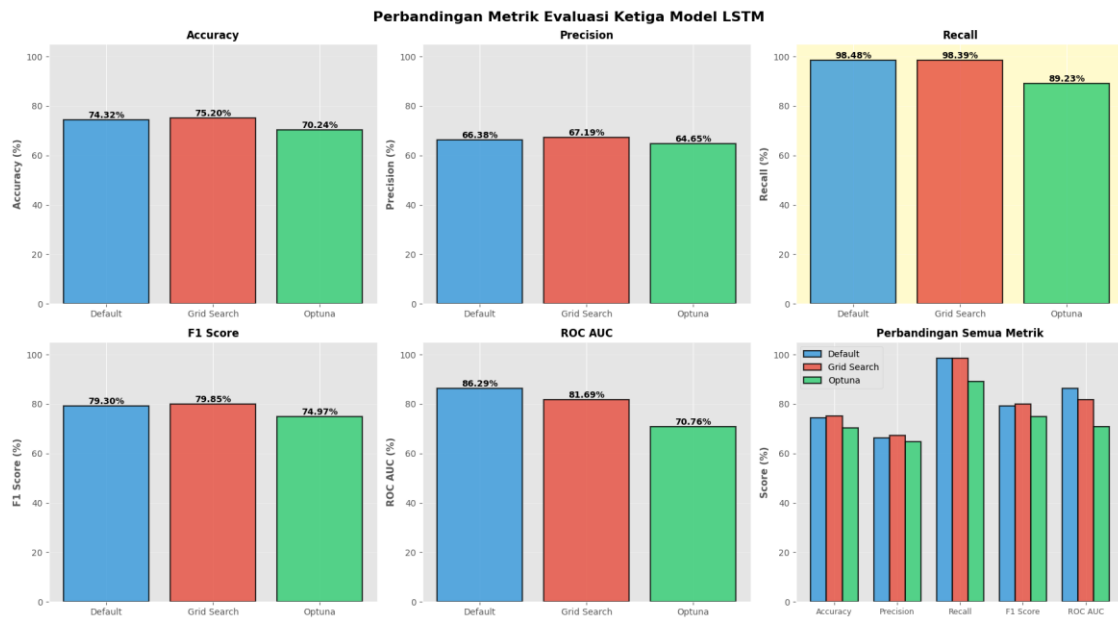
$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (9)$$

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (10)$$

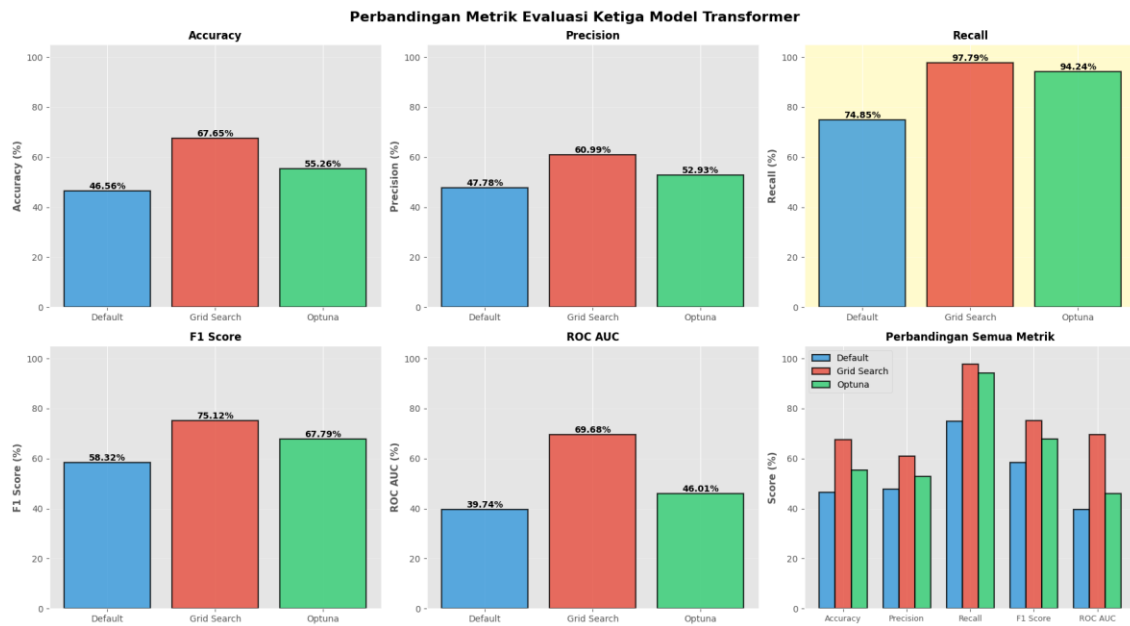
### 3. RESULTS AND DISCUSSION

Hasil performa setiap model diukur menggunakan lima metrik evaluasi, yaitu *recall/sensitivity*, *accuracy*, *precision*, *F1-Score*, dan *ROC (Receiver Operator Characteristic) AUC (Area Under the Curve)*. Dengan *recall* sebagai metrik utama.

#### 3.1 Hasil Perbandingan Single Model



Gambar 2. Grafik Perbandingan Metrik Evaluasi Model LSTM *end-to-end* dengan Beberapa Metode *Hyperparameter Tuning*



Gambar 2. Grafik Perbandingan Metrik Evaluasi Model Transformer *end-to-end* dengan Beberapa Metode *Hyperparameter Tuning*

Tabel 1. Metrik Evaluasi Model Tunggal (*End-to-End*).

Metrik	Transformer			LSTM		
	Standar	Grid Search	Optuna	Standar	Grid Search	Optuna
<i>Recall</i>	74,85%	<b>97,79%</b>	94,24%	<b>98,48%</b>	98,39%	89,23%
<i>Accuracy</i>	46,56%	<b>67,65%</b>	55,26%	74,32%	<b>75,20%</b>	70,24
<i>Precision</i>	47,78%	<b>60,99%</b>	52,93%	66,38%	<b>67,19%</b>	64,65%
<i>F1-Score</i>	58,32%	<b>75,12%</b>	67,79%	79,30%	<b>79,85%</b>	74,97%
ROC AUC	39,74%	<b>69,68%</b>	46,01%	<b>86,29%</b>	81,69%	70,76%

Setiap model dilatih dengan parameter standar dan parameter yang disesuaikan menggunakan Optuna dan Grid Search. Kelima metrik evaluasi model Transformer meningkat setelah disesuaikan menggunakan Optuna, dan lebih tinggi lagi ketika disesuaikan dengan Grid Search. Berbeda dengan Transformer, tidak semua metrik meningkat pada model LSTM setelah dilakukan *hyperparameter tuning*. Hanya *accuracy*, *precision*, dan *F1-Score* yang menunjukkan sedikit peningkatan setelah disesuaikan dengan Grid Search. Optuna memiliki angka yang terendah di semua metrik. Secara keseluruhan, Grid Search lebih optimal dalam mengatur dan memilih parameter yang paling sesuai pada skenario model tunggal.

Model LSTM menghasilkan *recall* sebesar 98,48%, sedikit lebih tinggi dari Transformer di angka 97,79%. LSTM juga terlihat lebih unggul di semua metrik lain. Ini mungkin menunjukkan kelebihan *inductive bias* yang dimiliki LSTM sebagai model *Recurrent Neural Network* (RNN). Sinyal fisiologis seperti ECG memiliki karakteristik sekuensial yang kuat, di mana status saat ini sangat bergantung pada status waktu sebelumnya (*temporal dependency*). Ideal bagi LSTM.

### 3.2 Hasil Perbandingan Hybrid Model.

Tabel 2. Metrik Evaluasi Model Hibrida.

Feature Extractor	Classifier	Metode Tuning	Recall	Accuracy	Precision	F1-Score	ROC AUC
Transformer	SVM	Standar	<b>56.01%</b>	<b>59.74%</b>	<b>60.49%</b>	<b>58.17%</b>	<b>58.42%</b>
		Grid Search	52,52%	57,47%	58,26%	55,24%	51,60%
		Optuna	53,37%	59,17%	60,34%	56,64%	57,02%
	XGBoost	Standar	61,22%	<b>60.32%</b>	<b>63.66%</b>	62,41%	<b>63.72%</b>
		Grid Search	68,81%	52,41%	51,79%	59,10%	48,14%
		Optuna	<b>69.02%</b>	59,84%	58,29%	<b>63.20%</b>	62,63%
	Random Forest	Standar	57,41%	51,55%	51,36%	54,22%	50,89%
		Grid Search	74,26%	54,86%	53,48%	62,18%	<b>60.69%</b>
		Optuna	<b>74.79%</b>	<b>59.33%</b>	<b>57.10%</b>	<b>64.76%</b>	55,51%
LSTM	SVM	Standar	66,17%	<b>51.69%</b>	<b>51.29%</b>	57,78%	<b>51.33%</b>
		Grid Search	66,83%	51,44%	51,08%	57,90%	51,22%
		Optuna	<b>70.20%</b>	51,46%	51,04%	<b>59.11%</b>	49,99%
	XGBoost	Standar	52,77%	49,54%	49,54%	51,10%	43,22%
		Grid Search	69,59%	53,70%	52,78%	60,03%	<b>57.33%</b>
		Optuna	<b>70.66%</b>	<b>55.72%</b>	<b>54.38%</b>	<b>61.46%</b>	52,52%
	Random Forest	Standar	<b>67.47%</b>	<b>50.71%</b>	<b>50.51%</b>	<b>57.77%</b>	<b>51.31%</b>
		Grid Search	55,33%	44,53%	45,48%	49,92%	43,12%
		Optuna	59,44%	45,57%	46,51%	52,18%	48,39%

*Hyperparameter tuning* pada model *hybrid* dengan Transformer sebagai *feature extractor* hanya memperlihatkan peningkatan yang cukup signifikan ketika dipasangkan bersama Random Forest dengan metode *tuning* Optuna, disusul oleh Grid Search. Ketika dipasangkan dengan XGBoost, hanya *recall* dan *F1-Score* yang meningkat. Dan ketika dipasang dengan SVM, tidak ada metrik yang menunjukkan peningkatan.

Adapun LSTM, kombinasi dengan Random Forest lebih baik ketika dibiarkan tanpa *tuning*. Semua metrik meningkat pada XGBoost, terutama dengan Optuna, kecuali ROC AUC

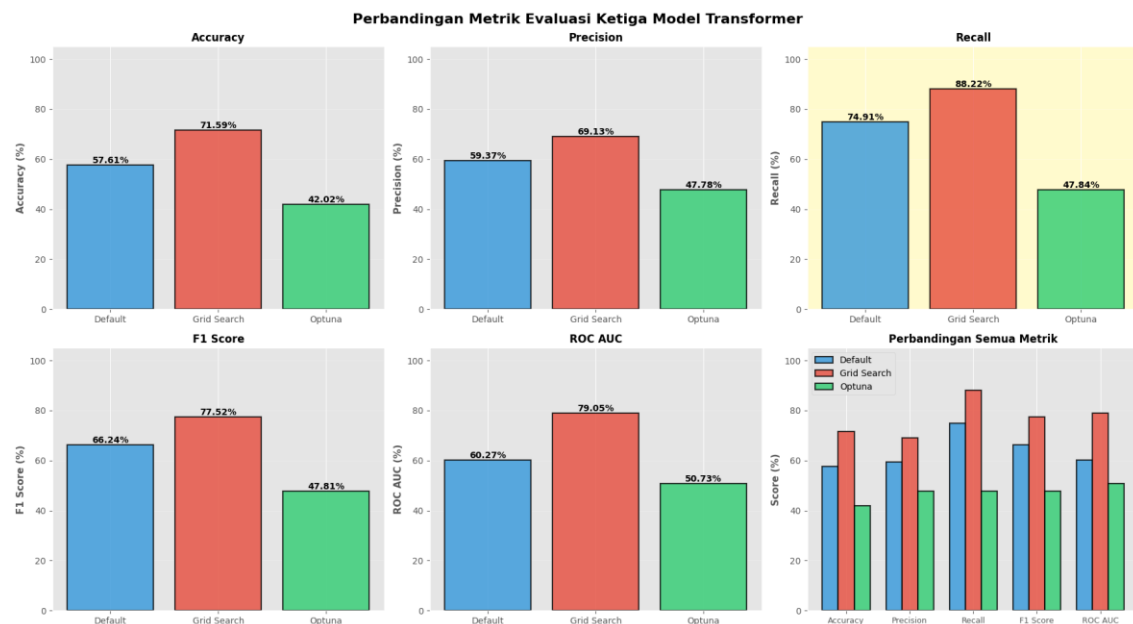


nya. Untuk SVM, hanya *recall* dan *F1-Score* yang meningkat. Berbeda dengan model tunggal, Optuna lebih optimal dalam mengatur dan memilih parameter yang paling sesuai pada skenario model *hybrid*.

Kombinasi Transformer sebagai model *feature extractor* dan Random Forest sebagai model *classifier* memperoleh *recall* dan *F1-Score* tertinggi (74,79% dan 64,76%). Sedangkan *accuracy*, *precision*, dan ROC AUC tertinggi diperoleh melalui kombinasi Transformer dan XGBoost (60,32%, 63,66%, 63,72% berturut-turut). Dari ketiga model *classifier*, Transformer hanya unggul di Random Forest. Sementara itu, LSTM unggul di SVM dan XGBoost. Meskipun LSTM unggul pada dua model sebagai *feature extractor*, Transformer tetap memiliki performa tertinggi ketika dipasangkan dengan Random Forest.

Alasan Transformer dengan Random Forest lebih unggul di model *hybrid* mungkin disebabkan oleh mekanisme *self-attention* pada Transformer dan kompatibilitas Random Forest dengan hasil fitur yang diekstrak. *Self-attention* pada Transformer mampu menangkap dependensi global dan pola morfologi kompleks dari keseluruhan sinyal secara paralel, menghasilkan representasi fitur yang jauh lebih kaya dan informatif dibandingkan ekstraksi sekuensial konvensional. Kompatibilitas ini semakin diperkuat oleh karakteristik Random Forest sebagai algoritma *ensemble learning* non-linear, yang efektif dalam memetakan hubungan rumit pada fitur berdimensi tinggi tersebut. Selain itu juga lebih resisten terhadap *noise* dibandingkan dengan *dense layer* sederhana, sehingga mampu membentuk batas keputusan (*decision boundary*) yang lebih presisi untuk klasifikasi.

### 3.3 Hasil Perbandingan Jumlah Data Seimbang dan Tidak Seimbang



Gambar 3. Grafik Perbandingan Metrik Evaluasi Model Transformer *end-to-end* dengan Beberapa Metode *Hyperparameter Tuning* pada data yang tidak seimbang.

Tabel 3. Perbandingan Metrik Evaluasi *Balanced* dengan *Imbalanced Data*.

Jumlah Data	Metode Tuning	Recall	Accuracy	Precision	F1-Score	ROC AUC
Seimbang	Standar	74,85%	46,56%	47,78%	58,32%	39,74%
	Grid Search	97,79%	67,65%	60,99%	75,12%	69,68%
	Optuna	94,24%	55,26%	52,93%	67,79%	46,01%
	Standar	74,91%	57,61%	59,37%	66,24%	60,27%

Tidak Seimbang	Grid Search	88.22%	71.59%	69.13%	77.52%	79.05%
	Optuna	47.84%	42.02%	47.78%	47.81%	50.73%

Skenario pengujian dampak distribusi data ini diterapkan secara khusus pada model Transformer *end-to-end*, yang berfungsi sebagai model *baseline* dalam penelitian ini. Berbeda dengan eksperimen utama pada model *single* dan *hybrid* yang secara konsisten menggunakan data seimbang, pengujian ini membandingkan performa model pada kondisi data asli dan data yang telah disesuaikan. Untuk mencapai keseimbangan kelas (*balanced data*), teknik *undersampling* diterapkan dengan mengurangi jumlah sampel pada kelas pasien AF yang semula berjumlah 19 orang menjadi 16 orang, menyetarakan jumlahnya dengan kelas pasien non-AF yang juga berjumlah 16 orang.

Hasil evaluasi menunjukkan adanya *trade-off* performa yang kontras antara kedua skenario tersebut. Penggunaan proporsi data yang seimbang terbukti lebih efektif dalam mendongkrak nilai *recall*, yang mengindikasikan bahwa model menjadi lebih sensitif dalam mendeteksi seluruh kasus positif tanpa banyak melewatkan pasien sakit. Di sisi lain, ketika dijalankan menggunakan dataset asli yang tidak seimbang (*imbalanced*), model mencatatkan peningkatan pada metrik *accuracy*, *precision*, *F1-Score*, dan ROC AUC. Hal ini menyiratkan bahwa meskipun data *imbalanced* memberikan akurasi keseluruhan yang lebih tinggi, strategi penyeimbangan data tetap memegang peranan vital untuk memprioritaskan sensitivitas deteksi pada kasus medis.

#### 4. CONCLUSION

Berdasarkan hasil eksperimen, disimpulkan bahwa *single model* LSTM merupakan arsitektur terbaik untuk dataset ini, mencatatkan performa tertinggi dengan nilai *recall* sebesar 98.39%. Keunggulan LSTM ini mengkonfirmasi kompatibilitasnya yang kuat terhadap data sekuensial jangka pendek dan keterbatasan jumlah data dibandingkan arsitektur Transformer yang membutuhkan volume data masif.

Analisis mendalam terhadap arsitektur *hybrid* menunjukkan bahwa meskipun kalah dalam model tunggal, Transformer terbukti lebih unggul dibandingkan LSTM sebagai *feature extractor*. Representasi fitur global yang dihasilkan mekanisme *Self-Attention* pada Transformer terbukti sangat efektif ketika dipadukan dengan *classifier* non-linear, di mana Random Forest muncul sebagai *classifier* terbaik. Kombinasi Transformer dan Random Forest menjadi model *hybrid* dengan performa tertinggi. Namun, secara keseluruhan, pendekatan *end-to-end single model* tetap mengungguli seluruh variasi model *hybrid* dalam hal sensitivitas deteksi penyakit. Selain itu, penerapan strategi penyeimbangan data (*balanced data*) terbukti krusial. Skenario ini menghasilkan peningkatan Recall yang signifikan dibandingkan skenario *imbalanced*, menjadikan model jauh lebih reliabel untuk kepentingan medis yang memprioritaskan minimnya *False Negative*.

Untuk pengembangan selanjutnya, penelitian ini memiliki prospek yang menjanjikan dalam implementasi pada perangkat kesehatan portabel (*wearable devices*). Mengingat efisiensi dan tingginya *recall* pada model tunggal LSTM, arsitektur ini potensial untuk ditanamkan pada sistem pemantauan *real-time* berdaya rendah. Penelitian masa depan disarankan untuk mengeksplorasi penggunaan dataset berskala lebih besar untuk memaksimalkan potensi arsitektur Transformer, serta menguji ketahanan model terhadap *noise* sinyal yang lebih bervariasi di lingkungan klinis nyata.

## REFERENCES

- [1] World Health Organization, “Cardiovascular diseases (CVDs),” [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
- [2] M. Sagris, E. P. Vardas, P. Theofilis, A. S. Antonopoulos, E. Oikonomou, and D. Tousoulis, “Atrial Fibrillation: Pathogenesis, Predisposing Factors, and Genetics,” *Int J Mol Sci*, vol. 23, no. 1, p. 6, Dec. 2021, doi: 10.3390/ijms23010006.
- [3] C. A. Papanastasiou *et al.*, “Atrial Fibrillation Is Associated with Cognitive Impairment, All-Cause Dementia, Vascular Dementia, and Alzheimer’s Disease: a Systematic Review and Meta-Analysis,” *J Gen Intern Med*, vol. 36, no. 10, pp. 3122–3135, Oct. 2021, doi: 10.1007/s11606-021-06954-8.
- [4] G. A. Andika, Asep Sukohar, and Ade Yonata, “Management of Arrhythmia: Atrial Fibrillation,” *Medical Profession Journal of Lampung*, vol. 11, no. 3, pp. 247–252, Oct. 2021.
- [5] B. Hou, J. Yang, P. Wang, and R. Yan, “LSTM-Based Auto-Encoder Model for ECG Arrhythmias Classification,” *IEEE Trans Instrum Meas*, vol. 69, no. 4, pp. 1232–1240, Apr. 2020, doi: 10.1109/TIM.2019.2910342.
- [6] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, “A Deep Learning Approach for Atrial Fibrillation Classification Using Multi-Feature Time Series Data from ECG and PPG,” *Diagnostics*, vol. 13, no. 14, p. 2442, Jul. 2023, doi: 10.3390/diagnostics13142442.
- [7] C. Qiu, H. Li, C. Qi, and B. Li, “Enhancing ECG classification with continuous wavelet transform and multi-branch transformer,” *Heliyon*, vol. 10, no. 5, p. e26147, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26147.
- [8] M. Alghieth, “DeepECG-Net: a hybrid transformer-based deep learning model for real-time ECG anomaly detection,” *Sci Rep*, vol. 15, no. 1, p. 20714, Jul. 2025, doi: 10.1038/s41598-025-07781-1.
- [9] F. BALCI, “A Hybrid Attention-based LSTM-XGBoost Model for Detection of ECG-based Atrial Fibrillation,” *Gazi University Journal of Science Part A: Engineering and Innovation*, vol. 9, no. 3, pp. 199–210, Sep. 2022, doi: 10.54287/gujsa.1128006.
- [10] A. D. Waberi, R. W. Mwangi, and R. M. Rimiru, “Advancing Type II Diabetes Predictions with a Hybrid LSTM-XGBoost Approach,” *Journal of Data Analysis and Information Processing*, vol. 12, no. 02, pp. 163–188, 2024, doi: 10.4236/jdaip.2024.122010.
- [11] A. A. Nafea, M. AL-Mahdawi, K. M. A. Alheeti, M. S. I. Alsumaidaie, and M. M. AL-Ani, “A Hybrid Method of 1D-CNN and Machine Learning Algorithms for Breast Cancer Detection,” *Baghdad Science Journal*, p. 3333, Oct. 2024, doi: 10.21123/bsj.2024.9443.
- [12] PhysioNet, “MIMIC PERform AF Dataset,” [https://ppg-beats.readthedocs.io/en/latest/datasets/mimic\\_perform\\_af/](https://ppg-beats.readthedocs.io/en/latest/datasets/mimic_perform_af/).
- [13] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci Data*, vol. 3, no. 1, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.
- [14] scikit-learn developers, “StandardScaler — scikit-learn documentation,” [https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?utm\\_source=](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?utm_source=).
- [15] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, “A Transformer-based Framework for Multivariate Time Series Representation Learning,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Aug. 2021, pp. 2114–2124. doi: 10.1145/3447548.3467401.
- [16] T. Kee and W. K. O. Ho, “Optimizing Machine Learning Models for Urban Sciences: A Comparative Analysis of Hyperparameter Tuning Methods,” *Urban Science*, vol. 9, no. 9, p. 348, Aug. 2025, doi: 10.3390/urbansci9090348.

- [17] I. Araf, A. Idri, and I. Chairi, “Cost-sensitive learning for imbalanced medical data: a review,” *ArtifIntell Rev*, vol. 57, no. 4, p. 80, Mar. 2024, doi: 10.1007/s10462-023-10652-8.