

Implementasi Tpot Automated Machine Learning dan Particle Swarm Optimization pada HistGradientBoostingClassifier untuk Klasifikasi Breast Cancer Wisconsin

Ika Wida Nuragustin^{*1}, Rosa Nur Aliana Sawafi²

^{1,2}Program Studi Sains Data, Universitas Telkom

Email: ¹ikawidanuragustin@student.telkomuniversity.ac.id, ²alinxyz@student.telkomuniversity.ac.id

Abstrak

Kanker payudara merupakan salah satu penyakit dengan tingkat kematian yang tinggi sehingga membutuhkan metode diagnosis yang cepat dan akurat. Namun, pemilihan model *machine learning* yang tepat serta penentuan *hyperparameter* yang optimal masih menjadi tantangan dalam proses klasifikasi data medis. Selain itu, proses pemilihan model secara manual cenderung memakan waktu dan tidak selalu menghasilkan performa terbaik. Penelitian ini bertujuan untuk menentukan model klasifikasi terbaik menggunakan TPOT (*Automated Machine Learning*), mengimplementasikan model *HistGradientBoostingClassifier* secara manual, mengoptimasi performa model menggunakan *Particle Swarm Optimization* (PSO), serta menganalisis perbandingan performa antara model *baseline*, model hasil TPOT, dan model hasil optimasi PSO pada dataset *Breast Cancer Wisconsin (Diagnostic)*. Metode yang digunakan meliputi tahapan *data understanding*, *exploratory data analysis (EDA)*, *preprocessing data*, pemodelan menggunakan TPOT, pembangunan model *baseline*, serta optimasi *hyperparameter* menggunakan PSO. Evaluasi model dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil penelitian menunjukkan bahwa TPOT berhasil menentukan *HistGradientBoostingClassifier* sebagai model terbaik. Model *baseline* menghasilkan akurasi sebesar 95,61%, sedangkan optimasi menggunakan PSO mampu meningkatkan akurasi menjadi 97,00% dan meningkatkan *recall* pada kelas Malignant dari 0,93 menjadi 0,99. Perbandingan ketiga model menunjukkan bahwa integrasi TPOT dan PSO memberikan hasil yang lebih efektif dan efisien dibandingkan pendekatan manual, terutama dalam meningkatkan sensitivitas deteksi kasus kanker ganas. Dengan demikian, penelitian ini membuktikan bahwa penggunaan *Automated Machine Learning* yang dikombinasikan dengan metode optimasi mampu meningkatkan performa klasifikasi pada data medis, khususnya dalam deteksi kanker payudara.

Kata kunci: *Breast Cancer, HistGradientBoostingClassifier, Machine Learning, Particle Swarm Optimization, TPOT*

Application of Tpot Automated Machine Learning and Particle Swarm Optimization to Histgradientboosting Classification for Wisconsin Breast Cancer Classification

Abstract

Breast cancer is a disease with a high mortality rate, necessitating rapid and accurate diagnostic methods. However, selecting the appropriate machine learning model and determining optimal hyperparameters remain challenges in the medical data classification process. Furthermore, the manual model selection process tends to be time-consuming and does not always yield the best performance. This study aims to determine the best classification model using TPOT (*Automated Machine Learning*), manually implement the *HistGradientBoostingClassifier* model, optimize model performance using *Particle Swarm Optimization* (PSO), and analyze the performance comparison between the baseline model, the TPOT-generated model, and the PSO-optimized model on the *Wisconsin Breast Cancer (Diagnostic)* dataset. The methods used include data understanding, exploratory data analysis (EDA), data preprocessing, modeling using TPOT, building a baseline model, and hyperparameter optimization using PSO. Model evaluation was performed using the accuracy, precision, recall, and F1-score metrics. The results of the study show that TPOT successfully identified the *HistGradientBoostingClassifier* as the best model. The baseline model achieved an accuracy of 95.61%, while optimization using PSO increased the accuracy to 97.00% and improved the recall of the Malignant class from 0.93 to 0.99. A comparison of the three models shows that the integration of TPOT and PSO yields more effective and efficient results compared to the manual approach, particularly in improving the sensitivity of malignant cancer detection. Thus, this study demonstrates that the use of *Automated Machine Learning* combined with optimization methods can improve classification performance on medical data, particularly in breast cancer detection.

Keywords: *Breast Cancer, HistGradientBoostingClassifier, Machine Learning, Particle Swarm Optimization, TPOT*

1. PENDAHULUAN

Kanker payudara merupakan salah satu penyakit dengan tingkat kejadian tertinggi di dunia dan menjadi penyebab utama kematian pada perempuan. Deteksi dini terhadap kanker payudara sangat penting untuk meningkatkan peluang kesembuhan serta menekan angka kematian [1]. Namun, proses diagnosis kanker payudara tidaklah sederhana karena melibatkan analisis data medis yang kompleks, seperti karakteristik sel tumor yang diperoleh dari hasil pemeriksaan laboratorium. Beberapa teknik AutoML telah dievaluasi secara komparatif untuk diagnosis kanker payudara dan terbukti mampu membedakan tumor jinak dan ganas secara akurat [2].

Seiring dengan perkembangan teknologi, *machine learning* telah banyak digunakan dalam bidang kesehatan untuk membantu proses pengambilan keputusan. Salah satu penerapannya adalah dalam klasifikasi data medis, yaitu untuk memprediksi apakah suatu tumor bersifat jinak (*benign*) atau ganas (*Malignant*) berdasarkan fitur-fitur tertentu [3]. Penggunaan model klasifikasi dalam *machine learning* mampu meningkatkan efisiensi dan akurasi dalam proses diagnosis dibandingkan dengan metode konvensional. Salah satu dataset yang digunakan dalam penelitian klasifikasi kanker payudara adalah *Breast Cancer Wisconsin (Diagnostic)*. Dataset ini memiliki karakteristik data yang terstruktur serta telah menjadi standar dalam pengujian berbagai algoritma klasifikasi [4]. Dengan adanya dataset ini, berbagai metode *machine learning* dapat diuji untuk memperoleh model yang mampu memberikan performa terbaik dalam proses klasifikasi.

Meskipun demikian, pemilihan algoritma *machine learning* yang tepat menjadi tantangan tersendiri. Terdapat berbagai algoritma klasifikasi yang dapat digunakan, seperti *Logistic Regression*, *Support Vector Machine*, *Random Forest*, dan metode berbasis *boosting*. Setiap algoritma memiliki karakteristik dan performa yang berbeda, sehingga pemilihan model terbaik secara manual memerlukan waktu yang cukup lama serta eksperimen yang kompleks [5]. Selain itu, tidak semua model memberikan hasil yang optimal pada setiap dataset, sehingga diperlukan pendekatan yang lebih efisien dalam proses pemilihan model.

Sebagai solusi, *Automated Machine Learning* (AutoML) hadir untuk mengotomatisasi proses pemilihan model dan *pipeline machine learning*. Salah satu framework AutoML yang banyak digunakan adalah TPOT (*Tree-based Pipeline Optimization Tool*), yang memanfaatkan pendekatan genetic programming untuk mencari kombinasi model dan parameter terbaik secara otomatis. Pendekatan ini telah diterapkan dalam klasifikasi kanker payudara dengan mengombinasikan TPOT dan PCA sehingga menghasilkan pipeline yang lebih efisien dan akurat [6]. Dengan menggunakan TPOT, proses eksplorasi model menjadi lebih sistematis dan efisien dibandingkan dengan pendekatan manual. Berdasarkan hasil eksplorasi menggunakan TPOT, diperoleh bahwa model yang memberikan performa terbaik pada dataset *Breast Cancer Wisconsin (Diagnostic)* adalah *HistGradientBoostingClassifier*. Model ini merupakan pengembangan dari metode *gradient boosting* yang menggunakan pendekatan histogram untuk meningkatkan efisiensi komputasi serta mempercepat proses pelatihan [7]. Oleh karena itu, model ini dipilih untuk diimplementasikan dalam penelitian ini.

Namun demikian, performa model *machine learning* sangat dipengaruhi oleh pemilihan *hyperparameter* yang tepat [8]. Tanpa proses optimasi yang baik, model yang digunakan belum tentu mencapai performa maksimal. Oleh karena itu, diperlukan metode optimasi yang mampu meningkatkan kinerja model secara optimal. Salah satu metode optimasi yang dapat digunakan adalah *Particle Swarm Optimization* (PSO), yaitu algoritma berbasis populasi yang terinspirasi dari perilaku kawanan dalam mencari solusi terbaik. PSO dikenal efektif dalam melakukan pencarian global pada ruang solusi yang kompleks serta mampu menemukan kombinasi *hyperparameter* yang optimal. Hal ini telah dibuktikan pada analisis data medis berskala besar, di mana PSO berhasil meningkatkan performa model *machine learning* secara signifikan dibandingkan dengan penggunaan parameter default [9].

Pada bidang klasifikasi kanker payudara, beberapa penelitian terdahulu telah mengeksplorasi penggunaan AutoML maupun optimasi hyperparameter secara terpisah. Penelitian [13] membandingkan tujuh teknik AutoML pada delapan dataset berbeda untuk diagnosis kanker payudara dan menemukan bahwa beberapa teknik AutoML seperti Lazy Predict dan MATLAB Classification Learner memberikan performa terbaik untuk tugas klasifikasi biner. [14] membandingkan tiga framework AutoML, yaitu TPOT, H2O AutoML, dan MLJAR, pada empat dataset varian genetik kanker payudara dan melaporkan bahwa H2O AutoML mencapai akurasi tertinggi hingga 99,99 persen pada dataset yang paling representatif, sementara TPOT juga menunjukkan generalisasi yang baik. [15] menerapkan TPOT yang dikombinasikan dengan PCA pada dataset Breast Cancer Wisconsin yang sama dengan penelitian ini dan memperoleh akurasi sebesar 98,6 persen. Selain itu, [10] menunjukkan bahwa penggunaan AutoML dalam mengoptimasi pemilihan dataset mampu meningkatkan akurasi prediksi

patogenesis varian genetik kanker payudara secara signifikan, yang mengindikasikan bahwa AutoML memiliki potensi yang luas dalam berbagai pendekatan analisis kanker payudara berbasis data. Pada kasus kanker jenis lain, [16] juga menunjukkan bahwa TPOT yang digabungkan dengan SVM mampu meningkatkan akurasi klasifikasi kanker paru dari citra CT. Hasil-hasil ini menunjukkan bahwa TPOT dan framework AutoML lain secara konsisten memberikan performa tinggi pada berbagai tugas klasifikasi kanker. Sejalan dengan itu, [11] mengembangkan model hybrid TPOT_SVM untuk klasifikasi kanker paru dari citra CT scan dengan tahapan preprocessing menggunakan Adaptive Median Filter, segmentasi menggunakan M-SegNet, dan ekstraksi fitur GLCM, sehingga membuktikan bahwa kombinasi TPOT dengan algoritma klasifikasi lain dapat memberikan hasil yang lebih baik dibandingkan penggunaan masing-masing metode secara terpisah.

Pada sisi optimasi hyperparameter, [17] menerapkan PSO untuk menentukan hyperparameter model machine learning pada data medis berskala besar terkait COVID-19 dan menunjukkan bahwa pendekatan ini efektif meningkatkan akurasi prediksi dibandingkan penggunaan parameter default. Namun demikian, penelitian-penelitian tersebut umumnya menerapkan AutoML dan PSO secara terpisah, dan belum banyak yang mengintegrasikan kedua pendekatan ini dalam satu pipeline yang sama. Penelitian ini berusaha mengisi kekosongan tersebut dengan mengintegrasikan TPOT sebagai metode pemilihan model terbaik dan PSO sebagai metode optimasi parameter lanjutan pada model yang sama, yaitu HistGradientBoostingClassifier, sehingga kebaruan penelitian ini terletak pada kombinasi dua pendekatan otomatis yang saling melengkapi pada satu pipeline klasifikasi kanker payudara.

2. METODE PENELITIAN

2.1. Dataset

Penelitian ini menggunakan dataset *Breast Cancer Wisconsin (Diagnostic)* yang diperoleh dari *UCI Machine Learning Repository*. Dataset ini digunakan untuk permasalahan klasifikasi biner dalam menentukan apakah suatu tumor bersifat jinak (*benign*) atau ganas (*Malignant*). Dataset terdiri dari 569 data observasi dengan 30 variabel fitur numerik dan 1 variabel target. Variabel target memiliki dua kelas, yaitu nilai 0 untuk *Malignant* (ganas) dan nilai 1 untuk *benign* (jinak), sesuai dengan label asli pada dataset *Breast Cancer Wisconsin (Diagnostic)* dari *UCI Machine Learning Repository*. Seluruh fitur pada dataset berupa data numerik kontinu yang merepresentasikan karakteristik inti sel kanker payudara, seperti *radius*, *texture*, *perimeter*, *area*, *smoothness*, *compactness*, *concavity*, *symmetry*, dan *fractal dimension*. Dataset ini dipilih karena memiliki struktur data yang baik dan banyak digunakan sebagai standar pengujian algoritma klasifikasi pada penelitian *machine learning*.

2.2. Eksplorasi dan Persiapan Data

Tahap awal penelitian dilakukan melalui *Exploratory Data Analysis* (EDA) untuk memahami karakteristik dataset. Proses EDA meliputi pemeriksaan struktur data, analisis statistik deskriptif, distribusi kelas, distribusi fitur, dan korelasi antar fitur. Setelah proses eksplorasi data, dilakukan tahap *preprocessing* untuk mempersiapkan data sebelum digunakan pada proses pemodelan. Tahapan *preprocessing* meliputi pemisahan variabel fitur (X) dan variabel target (y), pemeriksaan *missing value*, pembagian data menjadi data latih (*training data*) sebesar 80% dan data uji (*testing data*) sebesar 20%, serta standarisasi data menggunakan metode *StandardScaler*. Proses standarisasi dilakukan untuk menyamakan rentang nilai antar fitur sehingga dapat meningkatkan performa model klasifikasi. Standarisasi data dilakukan menggunakan Persamaan (1).

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Keterangan:

x = nilai data,

μ = rata-rata data,

σ = standar deviasi.

2.3. Pemodelan Menggunakan TPOT

Pada penelitian ini digunakan pendekatan *Automated Machine Learning* (AutoML) menggunakan TPOT (*Tree-based Pipeline Optimization Tool*) untuk menentukan model klasifikasi terbaik secara otomatis [17]. AutoML merupakan pendekatan dalam *machine learning* yang bertujuan untuk mengotomatisasi proses perancangan dan optimasi model, termasuk pemilihan algoritma, *preprocessing* data, serta penentuan *hyperparameter* [13]. TPOT merupakan salah satu *framework AutoML* yang memanfaatkan pendekatan *genetic*

programming untuk mencari kombinasi algoritma dan parameter terbaik berdasarkan performa model yang dihasilkan [14]. Dalam prosesnya, TPOT merepresentasikan *pipeline machine learning* dalam bentuk struktur pohon (*tree-based pipeline*), di mana setiap node merepresentasikan operator tertentu seperti *preprocessing*, seleksi fitur, maupun algoritma klasifikasi [14].

Pada penelitian ini, proses TPOT dijalankan dengan konfigurasi jumlah generasi (*generations*) sebanyak 15, ukuran populasi (*population_size*) sebanyak 15 pipeline per generasi, validasi silang (cv) sebanyak 5 fold, metrik evaluasi (*scoring*) berupa *accuracy*, dan batas waktu maksimum (*max_time_mins*) selama 15 menit, dengan *random_state* sebesar 42 untuk menjaga reproduktibilitas hasil. Proses TPOT dilakukan dengan membangkitkan sejumlah *pipeline machine learning* secara otomatis, kemudian mengevaluasi setiap *pipeline* menggunakan metrik akurasi. *Pipeline* dengan performa terbaik akan dipilih melalui proses seleksi, *crossover*, dan *mutation* hingga diperoleh model optimal [14]. Pada penelitian ini, TPOT digunakan untuk mengeksplorasi berbagai kombinasi algoritma klasifikasi dan parameter secara otomatis pada dataset *Breast Cancer Wisconsin (Diagnostic)*. Model terbaik yang diperoleh dari proses TPOT selanjutnya digunakan pada tahap implementasi dan optimasi lebih lanjut.

Penggunaan *StandardScaler* pada tahap *preprocessing manual* dan *MinMaxScaler* pada pipeline hasil TPOT bukan merupakan kontradiksi. *StandardScaler* diterapkan pada tahap awal sebagai persiapan data sebelum proses pencarian pipeline oleh TPOT dijalankan. TPOT kemudian secara independen mengeksplorasi langkah *preprocessing* tambahan di dalam *pipeline* yang dibangunnya sendiri, dan ditemukan bahwa penambahan *MinMaxScaler* pada data yang telah distandarisasi memberikan kompatibilitas yang lebih baik dengan langkah selanjutnya, yaitu Binarizer yang membutuhkan rentang nilai mutlak.

2.4. Implementasi HistGradientBoostingClassifier

HistGradientBoostingClassifier merupakan algoritma berbasis *ensemble learning* yang merupakan pengembangan dari metode *Gradient Boosting* [15]. Algoritma ini menggunakan pendekatan histogram untuk meningkatkan efisiensi komputasi dan mempercepat proses pelatihan model [15]. Pada penelitian ini, *HistGradientBoostingClassifier* diimplementasikan sebagai model *baseline* menggunakan parameter bawaan (*default parameter*). Model *baseline* digunakan sebagai acuan awal untuk membandingkan performa model sebelum dan sesudah dilakukan optimasi *hyperparameter*.

2.5. Optimasi Hyperparameter Menggunakan Particle Swarm Optimization

Optimasi *hyperparameter* dilakukan menggunakan metode *Particle Swarm Optimization* (PSO). PSO merupakan algoritma optimasi berbasis populasi yang terinspirasi dari perilaku kawanan burung dan ikan dalam mencari solusi terbaik [16]. Dalam PSO, setiap individu disebut sebagai partikel yang bergerak dalam ruang solusi untuk menemukan posisi optimal [17]. Parameter yang dioptimasi pada model *HistGradientBoostingClassifier* meliputi *learning_rate*, *max_iter*, *max_depth*, dan *max_bins*. Proses optimasi dilakukan melalui beberapa tahapan, yaitu inisialisasi partikel, perhitungan nilai *fitness*, penentuan *personal best* (*pbest*) dan *global best* (*gbest*), serta pembaruan posisi dan kecepatan partikel secara iteratif [18]. Pembaruan kecepatan partikel dilakukan menggunakan persamaan (2) berikut [19]:

$$v_k^{(i)} = w \cdot v_{k-1}^{(i)} + c_1 r_1 (Pbest^{(i)} - x_{k-1}^{(i)}) + c_2 r_2 (Gbest - x_{k-1}^{(i)}) \quad (2)$$

Sedangkan pembaruan posisi partikel dilakukan menggunakan persamaan berikut [19]:

$$x_k^{(i)} = x_{k-1}^{(i)} + v_k^{(i)} \quad (3)$$

Keterangan

- $v_k^{(i)}$ = kecepatan partikel ke-i pada iterasi ke-k
- $x_{k-1}^{(i)}$ = posisi partikel ke-i pada iterasi sebelumnya
- $v_{k-1}^{(i)}$ = kecepatan partikel ke-i pada iterasi sebelumnya
- $x_k^{(i)}$ = posisi partikel ke-i pada iterasi ke-k
- w = bobot inersia
- c_1, c_2 = koefisien akselerasi
- r_1, r_2 = bilangan random (0,1)
- $Pbest$ = personal best
- $Gbest$ = global best

Tujuan optimasi menggunakan PSO adalah untuk memperoleh kombinasi *hyperparameter* terbaik sehingga performa model klasifikasi dapat meningkat secara optimal. Pada penelitian ini, optimasi PSO dijalankan dengan jumlah partikel (*swarmsize*) sebanyak 10 dan jumlah iterasi maksimum (*maxiter*) sebanyak 20. Ruang pencarian *hyperparameter* ditentukan dengan batas bawah [0,01, 50, 3, 20, 0,0] dan batas atas [0,3, 500, 15, 100, 1,0] secara berurutan untuk parameter *learning_rate*, *max_iter*, *max_depth*, *min_samples_leaf*, dan *l2_regularization*.

2.6. Evaluasi Model

Evaluasi model dilakukan menggunakan *confusion matrix* dan beberapa metrik evaluasi, yaitu *accuracy*, *precision*, *recall*, dan *F1-score* [20]. *Confusion matrix* merupakan metode evaluasi yang digunakan untuk mengukur performa model klasifikasi dengan membandingkan hasil prediksi model terhadap nilai aktual [20]. Matriks ini terdiri dari empat komponen utama, yaitu *True Positive* (TP), *True Negative* (TN), *False positive* (FP), dan *False negative* (FN) [20]. *True Positive* (TP) menunjukkan jumlah data positif yang berhasil diprediksi dengan benar sebagai positif oleh model, sedangkan *True Negative* (TN) menunjukkan jumlah data negatif yang berhasil diprediksi dengan benar sebagai negatif [20]. *False positive* (FP) merupakan data negatif yang salah diprediksi sebagai positif, sementara *False negative* (FN) merupakan data positif yang salah diprediksi sebagai negatif [20]. Dalam konteks klasifikasi data medis, kesalahan *false negative* perlu diperhatikan karena dapat menyebabkan kasus penyakit tidak terdeteksi [21]. Bentuk *confusion matrix* yang digunakan dalam penelitian ini ditunjukkan pada Tabel 1.

Tabel 1. *Confusion Matrix*

	<i>Predicted Positive</i>	<i>Predicted Negative</i>
<i>Actual Positive</i>	TP	FN
<i>Actual Negative</i>	FP	TN

Perhitungan nilai *accuracy* dilakukan menggunakan persamaan (4) berikut [20]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Metrik *accuracy* digunakan untuk mengukur tingkat ketepatan model dalam melakukan prediksi terhadap seluruh data [20].

Sedangkan *precision*, *recall*, dan *F1-score* dihitung menggunakan persamaan berikut [20]:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Precision digunakan untuk mengukur tingkat ketepatan model dalam memprediksi kelas positif, sedangkan *recall* digunakan untuk mengukur kemampuan model dalam mendeteksi seluruh data positif [20]. Sementara itu, *F1-score* merupakan rata-rata harmonik antara *precision* dan *recall* yang digunakan untuk memberikan evaluasi performa model secara lebih seimbang [20]. Hasil evaluasi dibandingkan pada tiga skenario, yaitu model *baseline*, model hasil TPOT, dan model hasil optimasi menggunakan PSO.

3. HASIL DAN PEMBAHASAN

3.1. Hasil *Exploratory Data Analysis* (EDA)

Tahap *Exploratory Data Analysis* (EDA) dilakukan untuk memahami karakteristik data sebelum proses pemodelan dilakukan. Analisis meliputi pengecekan *missing value*, distribusi kelas target, distribusi fitur, serta korelasi antar fitur terhadap variabel target. Pengecekan *missing value* dilakukan menggunakan fungsi

isnull().sum(). Hasil pemeriksaan menunjukkan bahwa seluruh fitur pada dataset tidak memiliki nilai hilang, sehingga data dapat langsung digunakan pada tahap pemodelan tanpa proses imputasi.

Tabel 2. Hasil Pengecekan *Missing Value*

Keterangan	Hasil
Jumlah fitur	30
<i>Missing value</i>	0
Data duplikat	0
Kondisi data	Bersih

Berdasarkan Tabel 2, seluruh fitur pada dataset berada dalam kondisi lengkap tanpa adanya *missing value* maupun data duplikat. Kondisi ini menunjukkan bahwa kualitas dataset sudah baik dan siap digunakan pada tahap *preprocessing* serta pemodelan *machine learning*. Analisis distribusi kelas dilakukan untuk mengetahui proporsi jumlah data pada masing-masing kelas target.

Tabel 3 Distribusi Kelas Target

Kelas	Label	Jumlah Data	Persentase
0	<i>Malignant</i>	212	37,3%
1	<i>Benign</i>	357	62,7%
Total	-	569	100%

Berdasarkan hasil pada Tabel 3 distribusi kelas, dataset memiliki jumlah data *Benign* lebih banyak dibandingkan *Malignant*. Meskipun distribusi kelas tidak sepenuhnya seimbang, proporsi data masih cukup representatif untuk dilakukan proses klasifikasi. Analisis korelasi dilakukan menggunakan *Pearson Correlation* untuk mengetahui fitur yang memiliki hubungan paling kuat terhadap variabel target. Pendekatan seleksi fitur berbasis korelasi ini terbukti efektif dalam meningkatkan akurasi diagnosis kanker payudara ketika dikombinasikan dengan metode lain seperti mutual information dan model ensemble [12]. Berdasarkan Tabel 3.2, jumlah data kelas *Benign* lebih banyak dibandingkan kelas *Malignant*. Dataset memiliki distribusi yang cukup tidak seimbang, namun proporsi kelas masih tergolong representatif untuk dilakukan proses klasifikasi. Dominasi kelas *Benign* menunjukkan bahwa model perlu memiliki kemampuan yang baik dalam mendeteksi kelas minoritas, yaitu *Malignant*. Dalam konteks medis, kesalahan klasifikasi pada kelas *Malignant* sangat penting untuk diperhatikan karena dapat menyebabkan keterlambatan diagnosis kanker payudara. Selain distribusi kelas, dilakukan juga analisis korelasi fitur menggunakan *Pearson Correlation* untuk mengetahui fitur yang memiliki hubungan paling kuat terhadap variabel target.

Tabel 4. Fitur dengan Korelasi Tertinggi terhadap Target

No	Nama Fitur	Nilai Korelasi
1	<i>worst concave points</i>	0,793
2	<i>worst perimeter</i>	0,783
3	<i>mean concave points</i>	0,776
4	<i>worst radius</i>	0,776
5	<i>mean perimeter</i>	0,743
6	<i>worst area</i>	0,734
7	<i>mean radius</i>	0,730
8	<i>mean area</i>	0,709
9	<i>mean concavity</i>	0,696
10	<i>worst concavity</i>	0,660

Berdasarkan Tabel 3.3, fitur *worst concave points* memiliki nilai korelasi tertinggi terhadap variabel target sebesar 0,793. Hal ini menunjukkan bahwa fitur tersebut memiliki hubungan yang sangat kuat terhadap diagnosis kanker payudara. Selain itu, fitur-fitur lain seperti *worst perimeter*, *mean concave points*, *worst radius*, dan *mean perimeter* juga menunjukkan nilai korelasi yang tinggi. Secara umum, fitur yang berkaitan dengan ukuran, bentuk, dan tingkat cekungan inti sel memiliki pengaruh yang signifikan terhadap klasifikasi tumor. Hasil analisis korelasi menunjukkan bahwa karakteristik morfologi inti sel memiliki peranan penting dalam membedakan tumor *benign* dan *Malignant*. Oleh karena itu, fitur-fitur dengan korelasi tertinggi digunakan sebagai fitur utama dalam proses pemodelan.

4. Hasil Pemodelan TPOT AutoML

Pada tahap ini dilakukan pemodelan menggunakan TPOT (*Tree-based Pipeline Optimization Tool*). TPOT merupakan *framework Automated Machine Learning* (AutoML) yang menggunakan *genetic programming* untuk mencari kombinasi *pipeline machine learning* terbaik secara otomatis. Proses optimasi dilakukan menggunakan beberapa konfigurasi parameter seperti jumlah generasi, ukuran populasi, serta *cross-validation*.

Tabel 5. Konfigurasi TPOT

Parameter	Nilai
<i>generations</i>	15
<i>population_size</i>	15
<i>cv</i>	5
<i>scoring</i>	<i>accuracy</i>
<i>max_time_mins</i>	15

Berdasarkan Tabel 5, TPOT melakukan proses evolusi *pipeline* selama 15 generasi dengan jumlah populasi sebanyak 15 *pipeline* pada setiap generasi. Evaluasi *pipeline* dilakukan menggunakan *5-fold cross-validation* dengan metrik *accuracy*. Setelah proses optimasi selesai, TPOT menghasilkan *pipeline* terbaik yang terdiri dari beberapa tahapan *preprocessing* dan model klasifikasi.

Tabel 6. Pipeline Terbaik Hasil TPOT

Tahapan Pipeline	Fungsi
<i>MinMaxScaler</i>	Normalisasi fitur
<i>SelectFwe</i>	Seleksi fitur statistik
<i>FeatureUnion</i>	Kombinasi transformasi fitur
<i>PCA</i>	Reduksi dimensi
<i>HistGradientBoostingClassifier</i>	Model klasifikasi

Berdasarkan Tabel 6, TPOT menghasilkan *pipeline* yang cukup kompleks dengan kombinasi *preprocessing* dan transformasi fitur secara otomatis. Penggunaan *MinMaxScaler* menunjukkan bahwa normalisasi data ke rentang [0,1] memberikan performa yang lebih baik dibandingkan standarisasi biasa. Selain itu, *SelectFwe* digunakan untuk melakukan seleksi fitur berdasarkan uji statistik sehingga hanya fitur-fitur signifikan yang dipertahankan. TPOT juga menggunakan *PCA* dan *FeatureUnion* untuk menghasilkan representasi fitur yang lebih optimal. Kombinasi beberapa teknik *preprocessing* ini menunjukkan bahwa TPOT mampu melakukan eksplorasi *pipeline* secara menyeluruh tanpa intervensi manual. Hasil eksplorasi TPOT menunjukkan bahwa *HistGradientBoostingClassifier* menjadi algoritma terbaik untuk dataset *Breast Cancer Wisconsin (Diagnostic)*. Selain menentukan *pipeline* terbaik, TPOT juga menghasilkan kombinasi *hyperparameter* optimal secara otomatis.

Tabel 7. Hyperparameter Optimal TPOT

Hyperparameter	Nilai
<i>learning_rate</i>	0,2422
<i>max_iter</i>	100
<i>max_leaf_nodes</i>	1147
<i>min_samples_leaf</i>	39
<i>l2_regularization</i>	0,2033
<i>max_features</i>	0,1161

Berdasarkan Tabel 7, TPOT menghasilkan nilai *max_leaf_nodes* yang sangat tinggi dibandingkan parameter *default*. Hal ini menunjukkan bahwa model menggunakan struktur pohon yang lebih kompleks untuk meningkatkan kemampuan klasifikasi. Selain itu, TPOT juga menerapkan regularisasi melalui parameter *l2_regularization* untuk mengurangi kemungkinan *overfitting*. Penggunaan *max_features* yang relatif kecil menunjukkan bahwa model hanya menggunakan sebagian fitur pada setiap proses *split* sehingga generalisasi model dapat ditingkatkan. Evaluasi model TPOT dilakukan menggunakan *classification report*.

Tabel 8. Hasil Evaluasi Model TPOT

Metrik	Nilai
<i>Accuracy</i>	95,61%
<i>Precision</i>	98,55%
<i>Recall</i>	94,44%
<i>F1-Score</i>	96,45%

Berdasarkan Tabel 8, model TPOT menghasilkan *accuracy* sebesar 95,61% dengan *precision* tertinggi sebesar 98,55%. Nilai *precision* yang sangat tinggi menunjukkan bahwa model sangat baik dalam mengurangi *false positive*. Hal ini berarti sebagian besar data yang diprediksi sebagai *Benign* benar-benar merupakan kelas *Benign*. Namun demikian, nilai *recall* model TPOT lebih rendah dibandingkan *precision*. Kondisi ini menunjukkan bahwa masih terdapat beberapa kasus kanker ganas yang belum berhasil dideteksi oleh model.

5. Hasil Pemodelan *Baseline*

Model *baseline* dibangun menggunakan *HistGradientBoostingClassifier* dengan parameter *default* tanpa optimasi *hyperparameter*. Evaluasi model dilakukan menggunakan data uji dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

Tabel 9. Hasil Evaluasi Model *Baseline*

Metrik	Nilai
<i>Accuracy</i>	95,61%
<i>Precision</i>	96%
<i>Recall</i>	96%
<i>F1-Score</i>	96%

Berdasarkan Tabel 9, model *baseline* menunjukkan performa yang cukup baik meskipun belum dilakukan optimasi parameter. Nilai *accuracy* sebesar 95,61% menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar. Selain itu, nilai *precision*, *recall*, dan *F1-score* yang relatif seimbang menunjukkan bahwa model *baseline* memiliki performa klasifikasi yang stabil pada kedua kelas. Analisis *confusion matrix* dilakukan untuk mengetahui distribusi kesalahan prediksi model.

Tabel 10. *Confusion Matrix Baseline*

Aktual / Prediksi	<i>Malignant</i>	<i>Benign</i>
<i>Malignant</i>	39	3
<i>Benign</i>	2	70

Berdasarkan Tabel 10, model *baseline* berhasil memprediksi 39 data *Malignant* dan 70 data *Benign* dengan benar. Namun demikian, masih terdapat 3 data *Malignant* yang diprediksi sebagai *Benign* (*false negative*). Dalam konteks diagnosis medis, *false negative* menjadi kesalahan yang cukup serius karena dapat menyebabkan kasus kanker tidak terdeteksi. Selain itu, terdapat 2 data *Benign* yang diprediksi sebagai *Malignant* (*false positive*). Meskipun demikian, *false positive* masih lebih dapat ditoleransi dibandingkan *false negative* karena pasien masih dapat menjalani pemeriksaan lanjutan.

6. Hasil Optimasi PSO

Tahap berikutnya adalah optimasi *hyperparameter* menggunakan *Particle Swarm Optimization* (PSO). Optimasi dilakukan untuk mencari kombinasi parameter terbaik pada model *HistGradientBoostingClassifier*.

Tabel 11. *Hyperparameter* Terbaik Hasil PSO

<i>Hyperparameter</i>	Nilai
<i>learning_rate</i>	0,2194
<i>max_iter</i>	477
<i>max_depth</i>	6

<i>min_samples_leaf</i>	45
<i>l2_regularization</i>	0,5433

Berdasarkan Tabel 11, PSO menghasilkan kombinasi parameter yang berbeda dibandingkan TPOT maupun *baseline*. Nilai *max_iter* yang tinggi menunjukkan bahwa model melakukan proses *boosting* dalam jumlah iterasi yang lebih banyak sehingga kemampuan pembelajaran model meningkat. Selain itu, nilai *l2_regularization* yang cukup besar menunjukkan bahwa model menerapkan regularisasi untuk mengontrol kompleksitas model dan mengurangi *overfitting*. Evaluasi model hasil optimasi PSO dilakukan menggunakan *classification report* dan *confusion matrix*.

Tabel 12. Hasil Evaluasi Model PSO

Metrik	Nilai
<i>Accuracy</i>	97,00%
<i>Precision</i>	96%
<i>Recall</i>	97%
<i>F1-Score</i>	96%

Berdasarkan Tabel 12, model hasil optimasi PSO menghasilkan *accuracy* tertinggi sebesar 97%. Selain itu, model PSO juga menghasilkan nilai *recall* tertinggi dibandingkan model lainnya. Hal ini menunjukkan bahwa optimasi *hyperparameter* menggunakan PSO berhasil meningkatkan kemampuan model dalam mendeteksi kasus kanker payudara ganas. Nilai *recall* yang tinggi sangat penting dalam bidang medis karena menunjukkan bahwa sebagian besar kasus kanker berhasil terdeteksi dengan benar oleh model. Analisis *confusion matrix* model PSO ditunjukkan pada Tabel 13 berikut.

Tabel 13. *Confusion Matrix* Model PSO

Aktual / Prediksi	<i>Malignant</i>	<i>Benign</i>
<i>Malignant</i>	71	1
<i>Benign</i>	4	38

Berdasarkan Tabel 13, model PSO hanya menghasilkan 1 *false negative*. Hasil ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mendeteksi kasus kanker ganas. Meskipun jumlah *false positive* meningkat menjadi 4 data, kondisi ini masih lebih baik dibandingkan meningkatnya *false negative* karena pasien tetap dapat menjalani pemeriksaan medis lanjutan.

7. Analisis dan Pembahasan

Perbandingan performa dilakukan untuk mengetahui pengaruh penggunaan TPOT dan PSO terhadap peningkatan performa model klasifikasi kanker payudara.

Tabel 14. Perbandingan Performa Seluruh Model

Metrik	<i>Baseline</i>	TPOT	PSO
<i>Accuracy</i>	95,61%	95,61%	97,00%
<i>Precision</i>	96%	98,55%	96%
<i>Recall</i>	96%	94,44%	97%
<i>F1-Score</i>	96%	96,45%	96%

Berdasarkan Tabel 14, model *baseline* dan TPOT menghasilkan nilai *accuracy* yang sama, yaitu 95,61%. Namun demikian, TPOT menghasilkan *precision* tertinggi sebesar 98,55%. Nilai *precision* yang tinggi menunjukkan bahwa model TPOT sangat baik dalam meminimalkan *false positive*. Akan tetapi, nilai *recall* TPOT mengalami penurunan sehingga masih terdapat beberapa kasus kanker yang tidak berhasil dideteksi. Sementara itu, model PSO menghasilkan *accuracy* dan *recall* tertinggi dibandingkan model lainnya. Hal ini menunjukkan bahwa optimasi *hyperparameter* menggunakan PSO mampu meningkatkan sensitivitas model terhadap kasus kanker ganas. Perbandingan performa per kelas ditunjukkan pada Tabel 15 berikut.

Tabel 15. Perbandingan Performa per Kelas

Model	Kelas	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
-------	-------	------------------	---------------	-----------------

Baseline	<i>Malignant</i>	0.95	0.93	0.94
Baseline	<i>Benign</i>	0.96	0.97	0.97
TPOT	<i>Malignant</i>	0.98	0.94	0.96
TPOT	<i>Benign</i>	0.98	0.94	0.96
PSO	<i>Malignant</i>	0.95	0.99	0.97
PSO	<i>Benign</i>	0.97	0.90	0.94

Berdasarkan Tabel 15, model PSO menghasilkan *recall* tertinggi pada kelas *Malignant* sebesar 0,99. Hal ini menunjukkan bahwa hampir seluruh kasus kanker ganas berhasil dideteksi oleh model. Peningkatan *recall* pada model PSO menunjukkan bahwa optimasi *hyperparameter* berhasil meningkatkan sensitivitas model terhadap kelas positif. Pada klasifikasi kanker payudara, kemampuan mendeteksi kasus kanker ganas menjadi aspek yang sangat penting karena berhubungan langsung dengan proses diagnosis pasien. Selain itu, model TPOT menunjukkan *precision* tertinggi yang menandakan bahwa model lebih selektif dalam melakukan prediksi positif.

Peningkatan performa yang diperoleh PSO pada penelitian ini sejalan dengan temuan [17] yang menunjukkan bahwa optimasi *hyperparameter* berbasis PSO secara konsisten mampu meningkatkan akurasi model machine learning pada data medis dibandingkan penggunaan parameter default. Mekanisme di balik peningkatan ini dapat dijelaskan melalui kemampuan PSO dalam melakukan pencarian global pada ruang parameter yang kontinu, sehingga PSO tidak hanya bergerak menuju kombinasi parameter yang baik secara lokal, tetapi juga mampu menghindari jebakan local optimum yang sering terjadi pada pendekatan grid search atau penyetelan manual. Pada penelitian ini, peningkatan *max_iter* menjadi 477 dan penyesuaian nilai *l2_regularization* menjadi 0,5433 menunjukkan bahwa PSO menemukan keseimbangan baru antara kapasitas model dan kontrol overfitting yang tidak terjangkau melalui parameter default maupun hasil TPOT.

Jika dibandingkan dengan penelitian terdahulu, akurasi 97,00 persen yang diperoleh pada penelitian ini sedikit lebih rendah dibandingkan hasil [14] yang mencapai akurasi hingga 99,99 persen menggunakan H2O AutoML pada dataset varian genetik kanker payudara, maupun hasil [15] yang mencapai 98,6 persen menggunakan TPOT dengan PCA pada dataset Breast Cancer Wisconsin yang sama dengan penelitian ini. Perbedaan ini kemungkinan disebabkan oleh perbedaan jumlah fitur yang digunakan, di mana penelitian ini hanya menggunakan 15 fitur hasil seleksi korelasi, sementara penelitian Jain dkk. menggunakan seluruh 30 fitur asli yang dikombinasikan dengan PCA. Meskipun demikian, peningkatan *recall* pada kelas *Malignant* dari 0,93 pada model baseline menjadi 0,99 pada model PSO menunjukkan bahwa pendekatan yang diusulkan pada penelitian ini tetap relevan secara klinis, karena kemampuan mendeteksi kasus kanker ganas menjadi prioritas yang lebih penting dibandingkan sekadar mengejar akurasi tertinggi semata.

DAFTAR PUSTAKA

- [1] E. Marfianti, "Peningkatan pengetahuan kanker payudara dan ketrampilan periksa payudara sendiri (SADARI) untuk deteksi dini kanker payudara di Semutan Jatimulyo Dlingo," *Jurnal Abdimas Madani dan Lestari (JAMALI)*, pp. 25–31, 2021.
- [2] A. E. E. Rashed, A. M. Elmorsy, and A. E. M. Atwa, "Comparative evaluation of automated machine learning techniques for breast cancer diagnosis," *Biomed. Signal Process. Control*, vol. 86, p. 105016, 2023, doi: 10.1016/j.bspc.2023.105016.
- [3] R. D. Destianti, M. D. Gultom, and G. Noviamanti, "Model klasifikasi tingkat keparahan tumor tulang menggunakan algoritma decision tree berbasis fitur klinis," *Journal of Computer Science and Artificial Intelligence (JCSAI)*, vol. 2, no. 3, 2025.
- [4] UCI Machine Learning Repository, "Breast Cancer Wisconsin (Diagnostic) Dataset," University of California, Irvine, 1995. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>
- [5] U. K. J. Dn and M. Rahardi, "Analysis of SMOTE and random search on machine learning algorithms for stroke disease diagnosis," *Journal of Applied Informatics and Computing*, vol. 10, no. 1, pp. 847–855, 2026.
- [6] C. Jain, S. S. Srinet, T. K. Goyal, and M. Karpagam, "AI4BCancer: Breast cancer classification using AutoML - TPOT with PCA," in *Proc. 2024 2nd DMIHER Int. Conf. Artif. Intell.*, 2024, pp. 136–141.
- [7] D. A. Murtiningsih, B. W. Sari, and I. N. Fajri, "Comparison of Light Gradient Boosting Machine, eXtreme Gradient Boosting, and CatBoost with balancing and hyperparameter tuning for hypertension

- risk prediction on clinical dataset,” *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2753–2763, 2025.
- [8] W. Nugraha and A. Sasongko, “Hyperparameter tuning on classification algorithm with grid search,” *Sistemasi: Jurnal Sistem Informasi*, vol. 11, no. 2, pp. 391–401, 2022.
- [9] H. S. Salem, M. A. Mead, and G. S. El-Taweel, “Particle swarm optimization-based hyperparameters tuning of machine learning models for big COVID-19 data analysis,” *J. Comput. Commun.*, vol. 12, no. 3, pp. 160–183, 2024.
- [10] R. M. Ahmad, N. AlDhaheri, M. S. Mohamad, and B. R. Ali, “Leveraging AutoML to optimize dataset selection for improved breast cancer variants pathogenicity prediction,” *Comput. Struct. Biotechnol. J.*, 2025, doi: 10.1016/j.csbj.2025.10.052.
- [11] N. N. Murthy and K. Thippeswamy, “TPOT with SVM hybrid machine learning model for lung cancer classification using CT image,” *Biomed. Signal Process. Control*, vol. 104, p. 107465, 2025, doi: 10.1016/j.bspc.2024.107465.
- [12] M. S. Hashim and A. A. Yassin, “Using Pearson correlation and mutual information (PC-MI) to select features for accurate breast cancer diagnosis based on a soft voting classifier,” *Iraqi J. Electr. Electron. Eng.*, vol. 19, no. 2, pp. 43–53, 2023.
- [13] M. Alfidyah, “Optimasi algoritma machine learning untuk prediksi kinerja sistem komputer,” *Jurnal Komputer dan Teknik Informatika*, vol. 1, no. 1, pp. 1–7, 2025.
- [14] R. S. Olson and J. H. Moore, “TPOT: A tree-based pipeline optimization tool for automating machine learning,” in *Proc. Workshop on Automatic Machine Learning*, 2016, pp. 66–74.
- [15] scikit-learn developers, “HistGradientBoostingClassifier — scikit-learn documentation,” 2026. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html>
- [16] N. A. Febriyanti, R. Goejantoro, and S. Prangga, “Optimasi fuzzy C-means menggunakan particle swarm optimization untuk pengelompokan kabupaten/kota di Pulau Kalimantan,” *Eksponensial*, vol. 14, no. 1, pp. 31–40, 2023.
- [17] I. Cholissodin and E. Riyandani, *Swarm intelligence*. Malang: Fakultas Ilmu Komputer Universitas Brawijaya, 2016.
- [18] T. S. Feri, “Analisis akurasi algoritma K-Nearest Neighbor berbasis Particle Swarm Optimization pada studi kasus diagnosa penyakit jantung,” Doctoral dissertation, Universitas Lampung, 2025.
- [19] N. Aini, “Penyelesaian sistem persamaan fully fuzzy nonlinear dengan menggunakan particle swarm optimization dan metode steepest descent,” Universitas Lampung, 2025.
- [20] S. Sathyanarayanan and B. R. Tantri, “Confusion matrix-based performance evaluation metrics,” *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, 2024.
- [21] N. P. Sari, “Analisis performa algoritma CNN dalam klasifikasi citra medis berbasis deep learning,” *Jurnal Komputer*, vol. 2, no. 2, pp. 87–92, 2024.