Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1835~1845
e-ISSN: 2548-6861

Enhancing Eye Diseases Classification Using Imbalance Training & Machine Learning

Muhammad Azrul Ihwan [1]*, Ajie Kusuma Wardhana [2]*
* Informatika, Universitas Amikom Yogyakarta
muhammadazrulihwan@students.amikom.ac.id [1], ajiekusuma@amikom.ac.id [2],

Article Info

Article history:

Received 2025-07-13
Revised 2025-07-21
Accepted 2025-07-30

Keyword:

Eye Disease Classification,
XGBoost,
LightGBM,
SMOTE.

ABSTRACT

This research aims to evaluate the effectiveness of various machine learning algorithms in classifying eye diseases based on retinal images. The dataset comprises four categories of eye diseases: Cataract, Diabetic Retinopathy, Glaucoma, and Normal. The feature extraction method employed a transfer learning approach using ResNet50, followed by SMOTE for data balancing, PCA for dimensionality reduction, and normalization for scaling data consistently. Eleven machine learning models were evaluated, including basic algorithms, ensemble methods, and neural networks. The evaluation utilized metrics such as accuracy, precision, recall, and F1-score. K-Fold Cross Validation is also employed to observe all models' generalisation. The results revealed that the XGBoost algorithm achieved the highest performance with an accuracy of 92.03%, followed by LightGBM 91.88% and MLP 91.50%. K-Fold Validation also improved the MLP performance, which achieved an average accuracy of 91.94% with a standard deviation of 0.0178. This study successfully enhanced classification accuracy compared to previous studies and shows significant potential for clinical applications in resource-limited environments.

I. PENDAHULUAN

Penyakit mata seperti katarak, glaukoma, dan retinopati diabetik merupakan penyebab utama gangguan penglihatan dan kebutaan yang memengaruhi kualitas hidup manusia secara global diperkirakan sebanyak 596 juta orang di seluruh dunia mengalami gangguan penglihatan jarak jauh pada tahun 2020, dengan sekitar 43 juta kasus di antaranya merupakan kebutaan[1]. Diagnosis umumnya dilakukan oleh dokter mata melalui pemeriksaan langsung atau analisis citra retina. Namun, dalam beberapa tahun terakhir, teknologi machine learning mulai dimanfaatkan untuk mendukung proses diagnosis[2]. Sejumlah penelitian menunjukkan bahwa metode berbasis machine learning mampu mengidentifikasi pola dan karakteristik visual dalam citra medis yang sulit dideteksi oleh mata manusia, serta meningkatkan kecepatan dan akurasi untuk penyakit mata[3], [4].

Namun, Sebagian besar studi sebelumnya yang menggunakan teknik deep learning cenderung memerlukan dataset besar serta daya komputasi tinggi, dan sebagian besar hanya berfokus pada deteksi satu jenis penyakit mata seperti Diabetic Retinopaty, Cataract atau Glaucoma secara terpisah.[5]. Saat ini, belum banyak yang mengeksplorasi klasifikasi multi-kelas untuk membedakan berbagai bentuk gangguan mata dalam satu system klasifikasi[6]. Meskipun efektivitas deep learning telah menunjukkan performa tinggi dalam analisis citra medis, model deep learning masih dianggap sebagai “black box” karena kurangnya kejelasan dalam mekanisme pengambilan Keputusan dan keterbatasan pada aspek interpretabilitasnya[7], [8]. Oleh karena itu, diperlukan metode alternatif yang lebih ringan dan fleksibel yang dapat memanfaatkan dataset terbatas secara efektif sambil tetap menjaga tingkat akurasi yang tinggi[9].

Metode berbasis ekstraksi fitur dari citra retina menawarkan solusi alternatif yang lebih ringan dibanding pendekatan deep learning end-to-end. Dengan mengekstraksi elemen-elemen penting seperti tekstur, warna, dan bentuk, informasi visual dari retina dapat di representasikan dalam bentuk vector numerik. Hasil vector kemudian dapat dimanfaatkan oleh algoritma machine learning untuk melakukan klasifikasi penyakit mata[10]. Metode ini lebih efisien dalam penggunaan sumber daya komputasi, serta lebih

lebih mudah dipahami cara kerjanya dan masing-masing fitur terhadap hasil klasifikasi. [11].

Penelitian ini menggunakan dataset gambar retina yang berisi empat kelas kondisi mata yaitu Normal, Diabetic Retinopathy, Glaucoma, dan Cataract. Proses diawali dengan melakukan ekstraksi fitur menggunakan model pretrained ResNet50 untuk memperoleh hasil data numerik dari gambar retina. Untuk mengatasi ketidakseimbangan data antar kelas, diterapkan teknik Over-sampling Minoritas Sintetis (SMOTE), kemudian Principal Component Analysis (PCA) digunakan untuk mereduksi dimensi fitur guna meningkatkan efisiensi komputasi dan mencegah overvitting, Data hasil ekstraksi kemudian dinormalisasi sebelum digunakan dalam proses klasifikasi menggunakan algoritma machine learning.

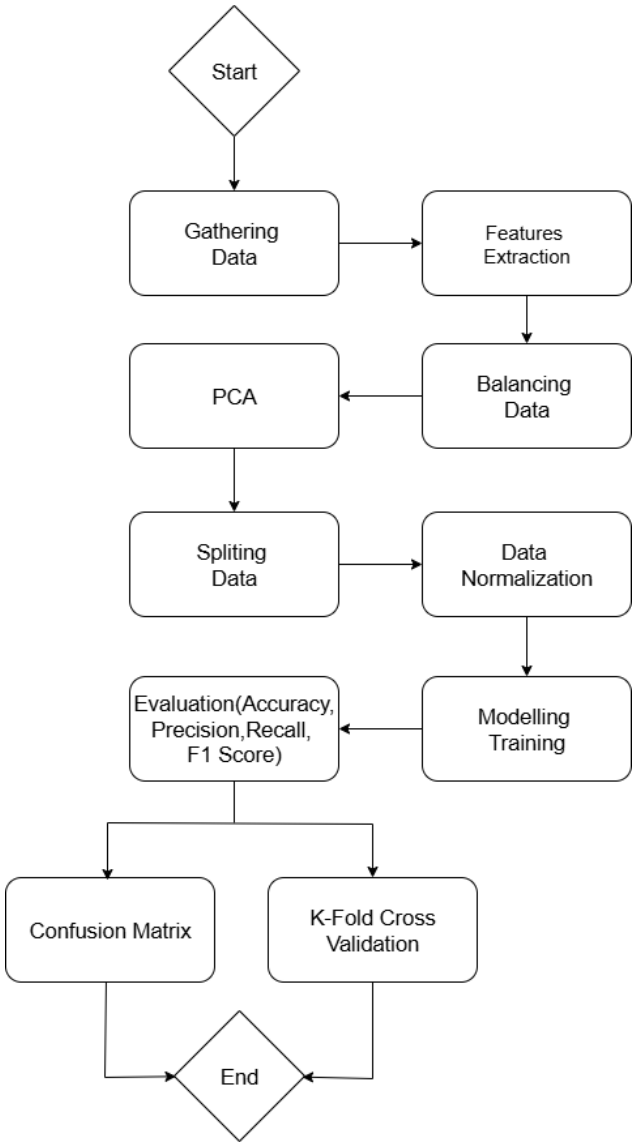
Tujuan penelitian adalah meningkatkan akurasi klasifikasi penyakit mata menggunakan metode machine learning dan mengidentifikasi sejauh mana metode ini mampu meningkatkan performa secara akurat. Dan membandingkan algoritma machine learning yang berbeda sehingga bisa melihat model mana yang terbaik untuk melakukan klasifikasi dengan dataset penyakit mata.

II. METODE

Penelitian ini menggunakan gambar retina sebagai dasar untuk mengklasifikasikan berbagai jenis penyakit mata melalui beberapa tahap yang melibatkan ekstraksi fitur dan pembelajaran mesin. Proses ini dimulai dengan pengumpulan dataset dari kaggle yang terdiri dari gambar retina yang telah dikelompokkan ke dalam empat kategori, yaitu Normal, Diabetic Retinopathy, Glaucoma, dan Cataract. Selanjutnya, gambar-gambar tersebut diproses menggunakan metode ekstraksi fitur berdasarkan arsitektur ResNet50. Hasil dari proses ini adalah vektor numerik yang mewakili karakteristik penting dari setiap gambar retina.

Langkah berikutnya dalam proses ini adalah menyeimbangkan jumlah data antara kelas-kelas dengan menggunakan Teknik Over-sampling Minoritas Sintetis (SMOTE), yang bertujuan untuk mengatasi ketidakseimbangan distribusi kelas dalam dataset. Setelah data diseimbangkan, pengurangan dimensi dilakukan menggunakan metode Analisis Komponen Utama (PCA) untuk memilih fitur-fitur yang paling berpengaruh pada proses klasifikasi. Kumpulan data yang telah direduksi kemudian dibagi menjadi dua bagian, 70% untuk pelatihan dan 30% untuk pengujian. Sebelum proses pelatihan dimulai, semua fitur dinormalisasi terlebih dahulu agar semua nilai berada pada skala yang konsisten.

Beberapa algoritma machine learning digunakan dalam proses klasifikasi untuk memperoleh hasil akurasi optimal dengan menggunakan Scikit-learn. Library ini digunakan dalam tahapan preprocessing, pembuatan berbagai model machine learning, serta evaluasi kinerja model. Setelah model mencapai tingkat akurasi yang melebihi hasil studi sebelumnya, penelitian dilanjutkan ke tahap akhir penarikan kesimpulan, di mana hasil dan temuan penelitian dirangkum berdasarkan kinerja model dan uji coba yang telah dilakukan. Seperti yang ditunjukkan pada Gambar 1, alur penelitian mencakup seluruh proses, mulai dari pengumpulan dataset hingga penarikan kesimpulan.



Gambar 1 Research Flow

A. Pengumpulan Dataset

Dataset yang digunakan dalam studi ini diperoleh dari platform Kaggle, yang dikenal sebagai Eye Diseases Dataset.

TABEL I
CLASS DATASET

No	Class	Size
1	Cataract	1038
2	Diabetic Retinopathy	1098
3	Glaucoma	1007
4	Normal	1074

Setiap gambar dalam data set ini mengandung informasi yang relevan tentang retina mata, yang digunakan untuk analisis dan prediksi terkait penyakit mata dan kondisi normal. Tabel 1 menjelaskan informasi gambar dalam data set ini.

B. Ekstraksi Fitur

Ekstraksi fitur dilakukan untuk menyederhanakan data dengan mengurangi kompleksitas gambar asli menjadi representasi numerik yang lebih informatif, efisien, dan relevan untuk klasifikasi[10]. Dalam penelitian ini, setiap gambar retina diubah ukurannya menjadi 224×224 piksel agar sesuai dengan dimensi input standar dari model ResNet50. Ukuran tersebut dipilih karena merupakan dimensi default yang digunakan ResNet50 yang telah dilatih sebelumnya (pretrained) pada dataset ImageNet. Proses ekstraksi fitur ini dilakukan menggunakan layer konvolusi terakhir dari ResNet50 yang disertai Global Average Pooling, sehingga menghasilkan vektor fitur numerik berdimensi 2048 yang mencerminkan karakteristik visual penting dari gambar retina[12].

C. Balancing Data SMOTE

SMOTE secara luas diakui sebagai teknik yang efektif untuk menangani ketidakseimbangan data. Dengan secara sintetis menciptakan sampel baru dalam kelas minoritas, metode ini membantu meningkatkan kinerja model dalam skenario data yang tidak seimbang. Fleksibilitasnya menjadikan SMOTE pilihan utama di berbagai bidang aplikasi, karena memungkinkan algoritma pembelajaran mesin bekerja lebih stabil dan adil tanpa harus terlalu bergantung pada proses augmentasi data yang berlebihan[13]. Berikut ini rumus dari SMOTE :

$$x_{\text{new}} = x_i + \delta \cdot (x_{zi} - x_i)$$

Keterangan :
x_i = Fitur vector dari sampel kelas minoritas yang dipilih.
x_{zi} = Salah satu tetangga terdekat dari x_i dalam kelas minoritas.
$\delta \cdot$ = Bilangan acak antara 0 dan 1, digunakan untuk interpolasi linier.
x_{new} = Data sintesis baru dari interpolasi.

D. PCA

Principal Component Analysis (PCA) yaitu sebuah metode untuk mengurangi dimensi data dengan menyoroti informasi paling penting dari data berdimensi tinggi. Teknik ini bekerja dengan mentransformasi data ke dalam koordinat baru yang disebut komponen utama. Komponen-komponen ini merupakan kombinasi linier dari fitur asli, saling orthogonal satu sama lain, dan diurutkan berdasarkan seberapa besar mereka menjelaskan variasi dalam data, mulai dari yang terbesar hingga yang terkecil[14]. Penelitian ini menggunakan PCA untuk mengurangi dimensi fitur yang dihasilkan oleh model ResNet50, yang awalnya memiliki dimensi 2048. Proses pengurangan ini dilakukan untuk menyederhanakan kompleksitas model, mencegah overfitting, dan mempercepat proses pelatihan. Dengan PCA, proses klasifikasi dapat lebih fokus pada fitur-fitur yang memiliki pengaruh signifikan terhadap variansi data tanpa mengorbankan informasi penting. Dengan memilih sejumlah komponen utama yang dapat mempertahankan sebagian besar variansi, kinerja klasifikasi tetap terjaga meskipun jumlah fitur yang digunakan berkurang secara signifikan.

E. Normalisasi Data

Normalisasi data yaitu sebuah proses transformasi fitur numerik agar berada dalam skala yang seragam, biasanya dalam rentang tertentu seperti 0 dan 1[15]. Teknik ini dipilih agar semua fitur memiliki rata-rata nol dan satu, sehingga model machine learning dapat bekerja secara optimal tanpa terpengaruh oleh perbedaan skala nilai. Proses penskalaan dilakukan berdasarkan data pelatihan, kemudian diterapkan pada data uji untuk menjaga konsistensi. Semua model menjalani tahap preprocessing yang sama untuk memastikan evaluasi kinerja yang baik.

F. Modelling

Setelah melalui proses feature extraction, balancing data, PCA, dan normalisasi, data tersebut kemudian digunakan untuk melatih berbagai algoritma machine learning dalam upaya mengklasifikasikan jenis penyakit mata berdasarkan gambar retina. Pemilihan algoritma yang beragam bertujuan untuk membandingkan kinerja masing-masing model dalam mengenali gambar retina yang diproses secara konsisten, dengan tujuan menemukan metode klasifikasi paling efektif untuk membedakan antara empat kategori: Normal, Retinopati Diabetes, Glaukoma, dan Katarak. Proses pelatihan dilakukan pada 70% data, sementara 30% sisanya digunakan untuk pengujian. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, F1-score, dan matriks kebingungan untuk menilai seberapa baik model dapat mengklasifikasikan gambar retina secara keseluruhan dan per kelas.

G. Evaluasi

Evaluasi model dilakukan untuk mengukur seberapa baik setiap algoritma mengklasifikasikan gambar retina ke dalam empat kategori penyakit mata, yaitu Normal, Retinopati Diabetes, Glaukoma, dan Katarak. Penilaian ini menggunakan 30% dari seluruh data yang sudah uji, yang sebelumnya telah melalui tahap-tahap seperti feature extraction, balancing data, PCA, dan normalisasi. Beberapa metrik digunakan dalam evaluasi, termasuk akurasi, presisi, recall, F1-score, dan confusion matrix. Akurasi digunakan untuk melihat persentase keseluruhan prediksi yang benar, sementara presisi dan recall digunakan untuk menilai akurasi dan cakupan prediksi untuk setiap kelas. F1-score, yang merupakan rata-rata harmonik dari presisi dan recall, memberikan gambaran keseluruhan kinerja model, terutama saat menangani data yang tidak seimbang. Confusion matrix digunakan untuk mengamati pola kesalahan prediksi antar

kelas dan menilai seberapa baik model dapat membedakan antara setiap jenis gangguan mata.

H. Cross Validation

Cross Validation dilakukan pada metode ini untuk mengukur kinerja model secara objektif dengan memanfaatkan data yang tersedia secara maksimal. Salah satu metode cross validation yang paling umum digunakan yaitu k-fold cross validation. Dalam penelitian ini, metode cross validation yang diterapkan adalah 10-Fold Cross Validation. Metode ini dilakukan dengan membagi dataset menjadi 10 bagian yang sama besar secara acak. Hasil akhir dari corss validation ini dihitung sebagai rata-rata dari performa model pada setiap iterasi, dan memberikan Gambaran yang lebih reliabel mengenai kinerja model secara keseluruhan. Berikut adalah rumus umum untuk menghitung performa metode K-Fold Cross Validation:

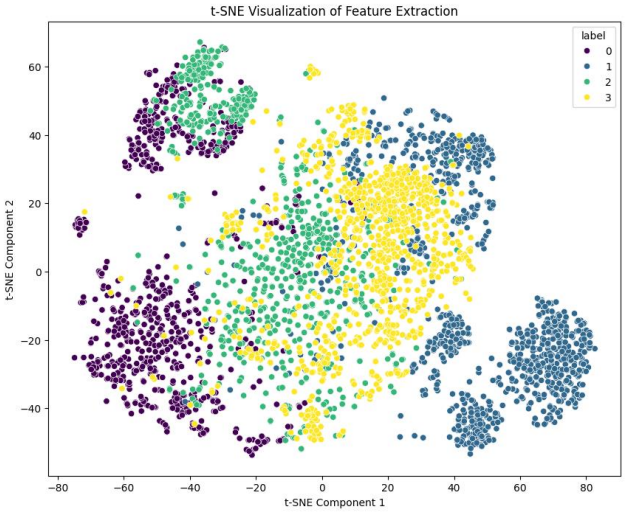
1. Dataset dibagi secara acak menjadi K bagian (fold) dengan ukuran yang sama:
$$D = D_1 \cup D_2 \cup \ldots \cup D_K$$
2. Pada setiap iterasi i, dengan i = 1, 2, …, K, digunakan bagian dataset Di sebagai data pengujian dan bagian dataset sisanya, yaitu $D - Di$ sebagai data pelatihan untuk mengevaluasi performa model, sehingga diperoleh nilai evaluasi Ei.
3. Hitung rata-rata nilai evaluasi untuk mendapatkan estimasi performa secara keseluruhan:
$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$
4. Hitung standar deviasi untuk melihat variabilitas dari performa model:
$$SD = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(E_i - E)^2}$$

Keterangan :
SD = Nilai standar deviasi performa model antar fold.
K = Jumlah fold.
E_i = Hasil evaluasi pada iterasi ke -i.
E = Rata-rata hasil evaluasi dari semua fold.

III. HASIL DAN PEMBAHASAN

A. Ekstraksi Fitur

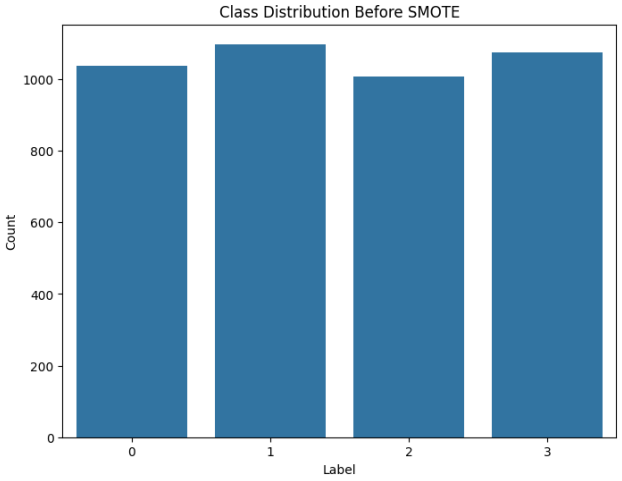
Sebagai bagian dari tahap awal klasifikasi, seluruh gambar retina pada dataset telah diproses melalui model ResNet50 yang telah dilatih sebelumnya untuk menghasilkan representasi numerik berdimensi tetap. Proses ekstraksi ini mengubah citra dua dimensi menjadi vektor fitur berdimensi 2048 yang mengandung informasi penting terkait tekstur, pola, dan karakteristik visual retina. Hasil ekstraksi dataset disimpan dalam format file csv. Hasil ini disimpan untuk analisis dan klasifikasi lebih lanjut. Hasil ekstraksi fitur secara keseluruhan divisualisasikan pada Gambar 2 untuk memudahkan pemahaman dan referensi.



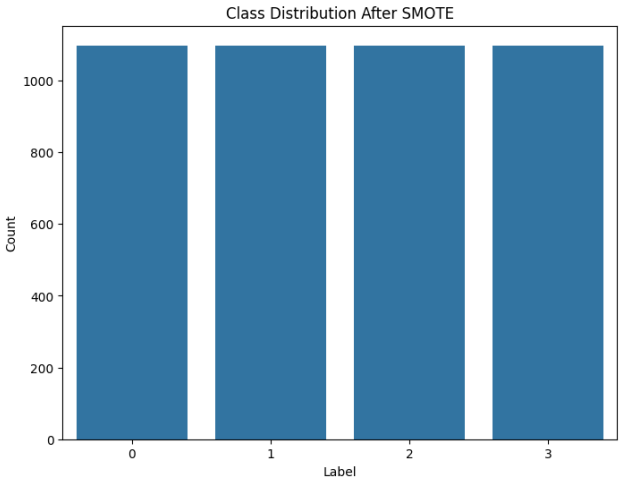
Gambar 2 Visualisasi Feature Extraction

B. Balancing Data SMOTE

Sebelum proses pelatihan model dimulai, dilakukan analisis awal terhadap distribusi label pada dataset yang menunjukkan adanya ketimpangan jumlah sampel antar kelas. Jumlah data keseluruhan pada dataset sebelum balancing dapat dilihat pada tabel 1. Ketidakseimbangan ini ditandai oleh selisih 91 gambar antara kelas terbanyak Diabetic Retinopathy dan kelas terkecil Glaukoma kondisi ini berisiko membuat model cenderung bias, yaitu lebih akurat dalam memprediksi kelas yang dominan dan kurang efektif dalam mengenali kelas dengan jumlah sampel terbatas. Hal ini dapat berdampak pada menurunnya kemampuan generalisasi model terhadap data baru. Untuk mengatasi permasalahan tersebut, digunakan metode Synthetic Minority Over-sampling Technique (SMOTE). Teknik ini menghasilkan sampel sintetis baru dengan melakukan interpolasi antar data minoritas yang ada, sehingga distribusi dataset menjadi seimbang tanpa perlu menggandakan data asli secara langsung. Gambar 3 dan Gambar 4 berikut menyajikan visualisasi distribusi label sebelum dan sesudah penerapan SMOTE. Setelah balancing, keempat kelas memiliki jumlah data yang sama yaitu 1.098 gambar per kelas, memungkinkan algoritma machine learning untuk belajar secara adil tanpa dominasi kelas tertentu.



Gambar 3 Sebelum Dilakukan SMOTE



Gambar 4 Setelah Dilakukan SMOTE

C. PCA

Setelah tahap penyeimbangan data selesai dilakukan, langkah berikutnya adalah melakukan reduksi dimensi terhadap vektor fitur hasil ekstraksi menggunakan metode Principal Component Analysis (PCA). Hasil ekstraksi fitur dari ResNet50 menghasilkan vektor berdimensi 2048 untuk setiap citra retina, yang berpotensi menimbulkan beban komputasi yang tinggi serta meningkatkan risiko overfitting jika digunakan secara langsung dalam proses pelatihan model. Untuk mengatasi permasalahan ini, PCA diterapkan dengan menetapkan jumlah komponen utama sebanyak 100, sehingga hanya fitur-fitur paling informatif yang dipertahankan berdasarkan kontribusinya terhadap variasi data secara keseluruhan, yaitu sebanyak 100 fitur utama yang mampu mempertahankan sekitar 90,17% dari total variasi data asli. Proses reduksi ini bertujuan untuk menyederhanakan kompleksitas data, mempercepat waktu pelatihan, dan tetap menjaga kinerja model dalam melakukan klasifikasi. Hasil dari transformasi PCA menunjukkan bahwa reduksi dimensi fitur yang dilakukan tidak menyebabkan kehilangan informasi yang berarti, yang tercermin dari tetap stabilnya akurasi model pada tahap evaluasi.

D. Splitting Data

Setelah melalui proses balancing dengan SMOTE serta reduksi dimensi menggunakan PCA, data kemudian dibagi ke dalam dua kelompok utama, yaitu data untuk pelatihan dan data untuk pengujian. Komposisi pembagian ditetapkan sebesar 70% untuk melatih model dan 30% sisanya untuk menguji performa model. Agar distribusi label pada masing-masing kelas tetap seimbang, proses pemisahan data dilakukan dengan tetap memperhatikan proporsi kelas secara keseluruhan. Hasil pembagian data train dan test dapat dilihat pada Tabel 2.

TABEL 2
TRAIN/TEST SPLIT

Information	**Training Data**	**Test Data**
Proportion	**70%**	**30%**
Cataract	768	330
Diabetic Retinopathy	769	329
Glaucoma	768	330
Normal	769	329

E. Normalisasi Data

Setelah data dibagi ke dalam set pelatihan dan pengujian, tahap selanjutnya adalah melakukan normalisasi pada fitur numerik. Langkah ini dilakukan untuk memastikan bahwa seluruh fitur berada pada skala yang sebanding, sehingga tidak ada fitur tertentu yang mendominasi proses pembelajaran model. Nilai-nilai fitur disesuaikan agar memiliki rata-rata nol dan standar deviasi satu, dengan tujuan menciptakan distribusi yang konsisten dan stabil, baik saat pelatihan maupun pengujian. Normalisasi dilakukan berdasarkan parameter statistik yang diperoleh dari data pelatihan, dan selanjutnya diterapkan pada data pengujian. Strategi ini penting untuk menjaga objektivitas evaluasi model serta mencegah terjadinya data leakage, yaitu penggunaan informasi dari data uji dalam proses pelatihan.

F. Evaluasi Model

Evaluasi terhadap berbagai algoritma machine learning sangat penting untuk menentukan solusi optimal dalam klasifikasi penyakit mata berdasarkan gambar retina. Untuk menilai efektivitas algoritma, dilakukan evaluasi secara sistematis terhadap dua belas model machine learning mulai dari algoritma dasar hingga metode *ensemble* tingkat lanjut. Dievaluasi secara sistematis untuk menentukan keefektifan pada dataset. Untuk memberikan performa yang menyeluruh Evaluasi dilakukan menggunakan beberapa metrik utama seperti akurasi (accuracy), presisi (precision), recall, dan F1-score untuk menunjukkan kelebihan dan kekurangan masing-masing model dalam menangani kompleksitas dataset penyakit mata yang digunakan. Hasil evaluasi model ditunjukkan dalam Tabel 3.

TABEL 3
HASIL PERBANDINGAN MODEL

Model	Result			
	Accuracy	**Precission**	**Recall**	**F1 Score**
Logistic Regression	90.74	90.77	90.75	90.75
Decission Tree	78.76	78.77	78.76	78.76
Random Forest	89.30	89.52	89.31	89.32
Gradient Boosting	91.27	91.51	91.28	91.29
Support Vector Machine	90.74	91.09	90.75	90.73
AdaBoost	77.47	79.37	77.47	77.91
Naive Bayes	75.42	78.60	75.42	75.74
K-Nearest Neighbors	87.71	88.20	87.71	87.77
Multi Layer Perceptron	**91.50**	**91.61**	**91.51**	**91.52**
XGBoost	**92.03**	**92.23**	**92.04**	**92.03**
LightGBM	**91.88**	**92.08**	**91.89**	**91.87**

Secara keseluruhan, hasil menunjukkan bahwa algoritma seperti XGBoost, LightGBM, dan MLP memiliki generalisasi dan akurasi yang lebih baik dibandingkan model lainnya. XGBoost dan LightGBM, sebagai metode ensemble, sangat efektif dalam mengelola data berdimensi tinggi dan menangkap fitur kompleks. Di sisi lain, MLP sebagai metode neural network mampu menangkap pola-pola nonlinear dalam data yang tidak dapat dilakukan oleh algoritma linear sederhana.

Walaupun Gradient Boosting, Support Vector Machine, Logistic Regression, Random Forest, dan K-Nearest Neighbors menunjukkan performa yang cukup baik, meskipun belum sebaik tiga model teratas. Gradient Boosting memperlihatkan kapabilitas tinggi dalam mengenali pola kompleks, meskipun secara komputasi masih sedikit tertinggal dibandingkan dengan XGBoost dan LightGBM. Support Vector Machine dan Logistic Regression keduanya memperlihatkan kemampuan dalam mengklasifikasikan data dengan batas Keputusan yang jelas dan optimal. Sementara itu, Random Forest juga menunjukkan performa yang stabil, diikuti oleh K-Nearest Neighbors, ini menegaskan bahwa algoritma-algoritma tersebut meskipun memiliki performa baik, masih memiki keterbatasan dalam mengatasi kompleksitas fitur pada dataset yang diuji.

Hasil evaluasi juga memperlihatkan bahwa algoritma-algoritma seperti Decision Tree, AdaBoost, dan Naïve Bayes memiliki tingkat akurasi paling rendah dibandingkan model-model lainnya. Decission Tree meskipun mudah dipahami secara interpretasi, model ini mengalami kesulitan dalam mengelola kompleksitas data yang menyebabkan performa klasifikasi menjadi kurang optimal. AdaBoost mengindikasikan bahwa penggunakan weak learner berbasis pohon Keputusan sederhana masih kurang efektif dalam menangkap pola-pola kompleks dalam dataset ini. Begitu juga Naïve Bayes yang memperlihatkan bahwa asumsi dasar algoritma ini tentang independensi antar fitur tidak sesuai dengan karakteristik data yang dianalisis, sehingga performanya menjadi paling rendah dibandingkan algoritma lainnya dalam penelitian ini.

Analisis mendalam menggunakan confusion matrix untuk model terbaik, XGBoost, menunjukkan kinerja yang baik dalam mengklasifikasikan semua kategori penyakit mata, khususnya Cataract dan Diabetic Retinopathy. Namun, masih ditemukan beberapa misprediksi antara kategori Glaucoma dan Normal. Hal ini menunjukkan perlunya studi lebih lanjut untuk meningkatkan akurasi dalam membedakan kedua kelas ini.

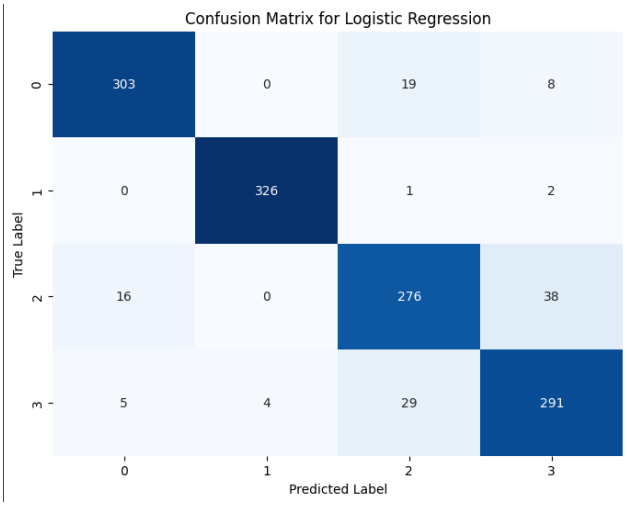
Jika dibandingkan dengan penelitian sebelumnya [16] yang mencapai akurasi sekitar 90% dengan metode ekstraksi fitur menggunakan histogram, penelitian ini menunjukkan peningkatan signifikan sebesar 2,03% dengan ekstraksi pre-trained transfer learning ResNet50. Hasil perbandingan penelitian ditunjukkan dalam Tabel 4

TABEL 4
HASIL PERBANDINGAN PENELITIAN

Peneliti	Preprocessing	Ekstraksi Fitur	Akurasi
Ramanathan et al [16]	PCA	Histogram	90%
Penlitian ini	PCA, SMOTE, Normalisasi	Pre-trained	92,03%

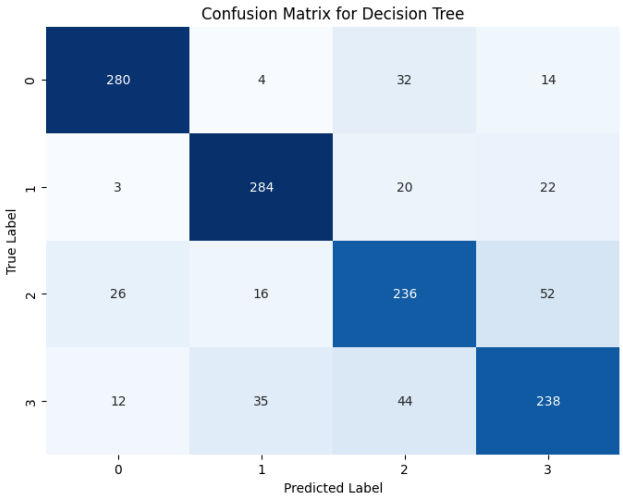
Namun demikian, terdapat beberapa keterbatasan dalam penelitian ini, seperti jumlah dataset yang terbatas dan optimasi hyperparameter yang belum maksimal. Oleh karena itu, penelitian lanjutan direkomendasikan untuk memperbesar dataset, melakukan optimasi hyperparameter lebih mendalam, serta menguji model pada dataset yang lebih bervariasi untuk mengukur generalisasi secara optimal.

Untuk memperoleh gambaran yang lebih menyeluruh mengenai kinerja klasifikasi tiap algoritma terhadap masing-masing jenis penyakit mata, digunakan confusion matrix sebagai alat bantu visual. Confusion matrix menyajikan distribusi prediksi benar dan salah untuk setiap kelas, sehingga mempermudah identifikasi pola kesalahan dan kekuatan model dalam membedakan kategori yang mirip. Berdasarkan visualisasi confusion matrix yang dibuat, Dimana model XGBoost sebagai model dengan performa terbaik, guna memperlihatkan sejauh mana model mampu mengklasifikasikan empat jenis penyakit mata secara akurat dan direpresentasikan dengan: 0 untuk label Cataract, 1 untuk Diabetic Retinopathy, 2 untuk Glaucoma, dan 3 untuk Normal.



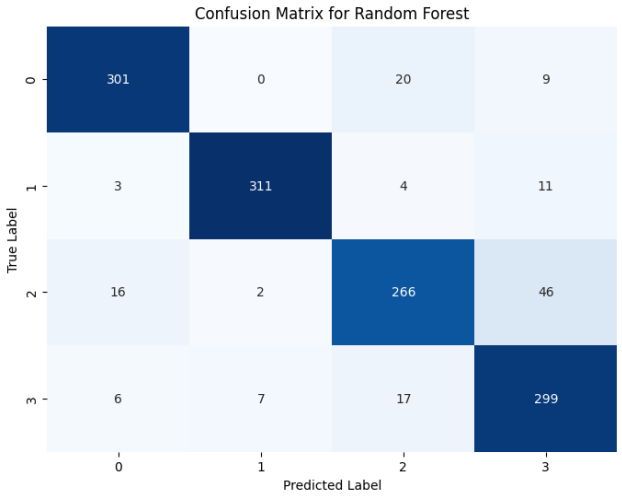
Gambar 5 Model Logistic Regression

Confusion matrix pada gambar 5 menunjukkan kemampuan logistic regression dapat memprediksi Diabetic Retinopathy dengan jumlah prediksi 326. Namun, model masih kesulitan dalam membedakan antara Glaucoma dan Normal, serta sebagian kasus Cataract diartikan sebagai Glaucoma. Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Normal.



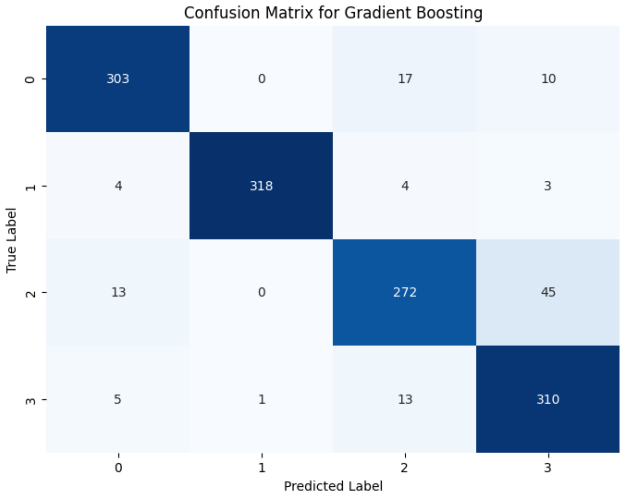
Gambar 6 Model Decision Tree

Confusion matrix pada gambar 6 menunjukkan kemampuan decision tree dapat memprediksi Diabetic Retinopathy dengan jumlah prediksi 284. Namun, model masih kesulitan dalam membedakan antara Glaucoma dan Normal, Hal ini dapat dilihat dari tingginya jumlah kesalahan klasifikasi pada kedua kelas, Dimana sebanyak 52 sampel Cataract diprediksi sebagai Glaucoma, sementara 44 sampel Glaucoma diprediksi sebagai Cataract. Selain itu, sejumlah kasus Normal juga salah diprediksi, yaitu sebanyak 32 sampel diprediksi sebagai Cataract dan 14 sebagai Glaucoma. Kesalahan prediksi lainnya terjadi pada Normal yang Sebagian kecil 35 sampel diprediksi sebagai Diabetic Retinopathy. Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki kesalahan prediksi pada kasus Glaucoma dan Normal.



Gambar 7 Model Random Forest

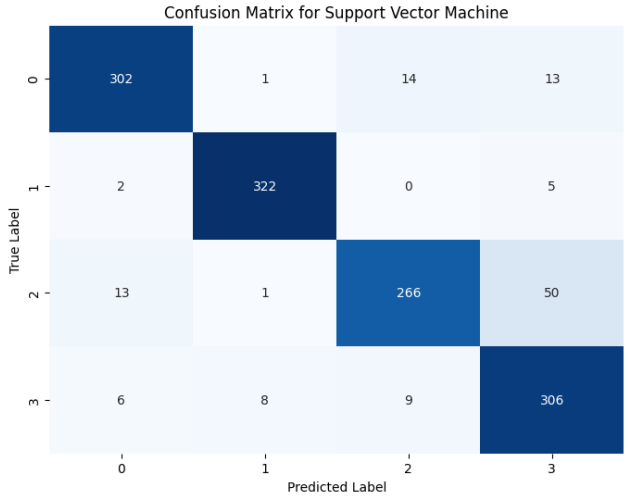
Confusion matrix pada gambar 7 menunjukkan kemampuan random forest dapat memprediksi Diabetic Retinopathy dengan jumlah prediksi 311. Namun, model masih kesulitan dalam membedakan antara Glaucoma Dimana 46 sampel Glaucoma diprediksi sebagai Normal dan 16 lainnya sebagai Cataract. Kesalahan serupa juga terjadi pada kelas Cataract dengan 17 sampel yang diprediksi sebagai Glaucoma, Sebagian kecil kasus Normal juga mengalami salah prediksi, terutama Ctaract dan Glaucoma. Secara keseluruhan model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Normal.



Gambar 8 Model Gradient Boosting

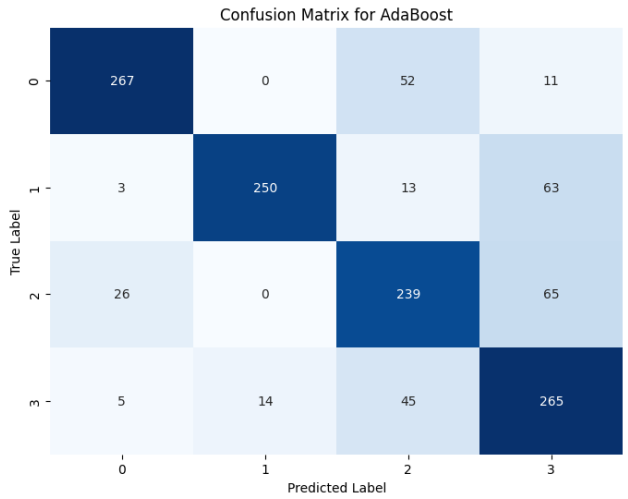
Confusion matrix pada gambar 8 menunjukkan kemampuan gradient boosting dapat memprediksi Diabetic Retinopathy dan Normal dengan jumlah prediksi 318 dan 310.

Namun, model masih kesulitan dalam membedakan Glaucoma Dimana 45 sampel salah diprediksi sebagai Normal dan 13 sebagian Cataract. Sebagian kecil kasus Cataract juga salah prediksi sebagai Glaucoma dan Normal. Secara keseluruhan, model menunjukkan kinerja yang sangat baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma.



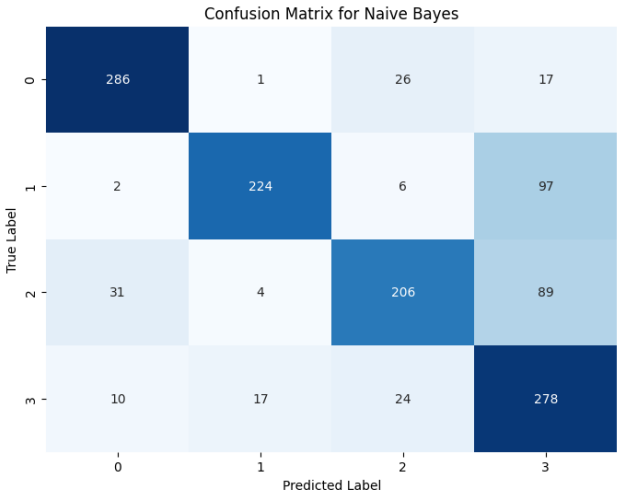
Gambar 9 Model Support Vector Machine

Confusion matrix pada gambar 9 menunjukkan kemampuan Support Vector Machine dapat memprediksi Diabetic Retinopathy dan Normal dengan jumlah prediksi 322 dan 306. Model juga cukup akurat dalam memrediksi Cataract dengan jumlah prediksi 302, namun model masih kesulitan dalam membedakan Glaucoma dimana 50 sampel salah diprediksi sebagai Normal dan 13 lainnya sebagai Cataract. Secara keseluruhan, model menunjukkan kinerja yang sangat baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma.



Gambar 10 Model AdaBoost

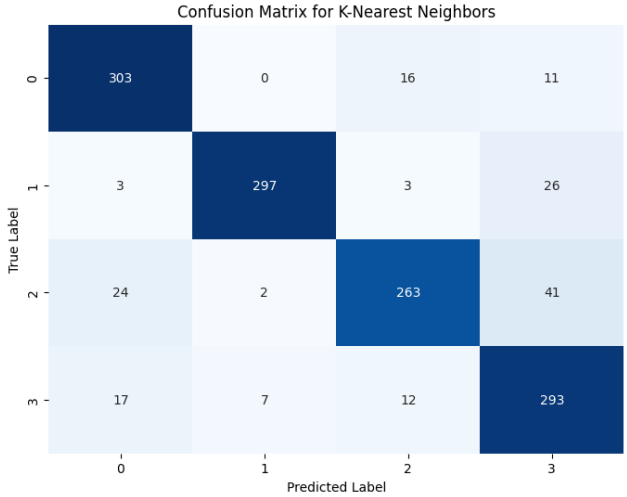
Confusion matrix pada gambar 10 menunjukkan kemampuan AdaBoost dapat memprediksi Cataract dengan jumlah 267. Namun model masih kesulitan dalam membedakan Diabetic Retinopathy dan Glaucoma, dimana Diabetic Retinopathy salah prediksi sebanyak 63 sampel sebagai Normal. Hal serupa terjadi pada kelas Glaucoma dimana sebanyak 65 sampel juda diprediksi sebagai normal. Model juga menunjukkan kesalahan pada kelas Normal dengan 45 kasus salah prediksi sebagai Glaucoma dan 14 sebagai Diabetic Retinopathy. Selain itu pada Cataract sebanyak 52 sampel diprediksi sebagai Glaucoma. Secara keseluruhan, model menunjukkan kinerja yang sangat baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Diabetic Retinopathy.



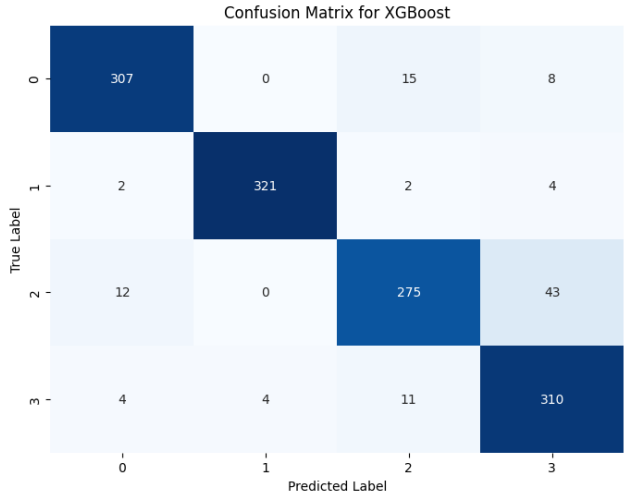
Gambar 11 Model Naïve Bayes

Confusion matrix pada gambar 11 menunjukkan kemampuan Naïve Bayes dapat memprediksi Cataract dengan jumlah prediksi 286. Namun, model masih kesulitan dalam membedakan antara Diabetic Retinopathy dan Glaucoma, Pada kelas Diabetic Retinopathy sebanyak 97 sampel salah diprediksi sebagai Normal, sedangkan pada kelas Glaucoma sebanyak 89 sampel juga diprediksi sebagai Normal dan 31 sebagai Ctaract. Kesalahan prediksi lainnya juga terjadi pada kelas Normal dimana 24 sampel diprediksi sebagai Glaucoma dan 17 sebagai Diabetic Retinopathy. Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Diabetic Retinopathy dan Glaucoma.

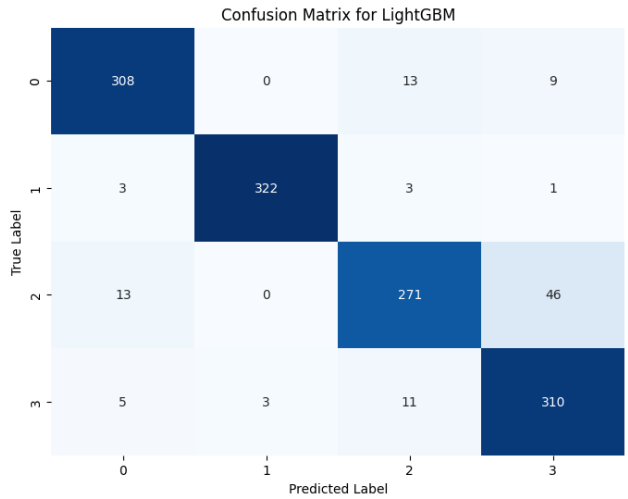
Confusion matrix pada gambar 12 menunjukkan kemampuan K-Nearest Neighbors dapat memprediksi Cataract dengan jumlah prediksi 303. Namun, model masih kesulitan dalam membedakan antara Glaucoma dan Normal, terlihat dari adanya 41 sampel Glaucoma yang salah diprediksi sebagai normal dan 24 sebagai Cataract. Kesalahan klasifikasi juga terjadi pada kelas Normal meskipun dalam jumlah yang relative kecil dimana beberapa sampel diprediksi sebagai Cataract atau Glaucoma. Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Normal.



Gambar 12 K-Nearest Neighbors



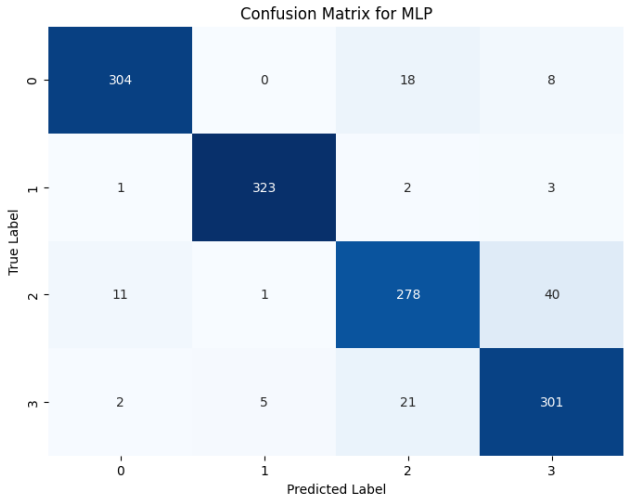
Gambar 13 Model XGBoost



Gambar 14 Model LightGBM

Confusion matrix pada gambar 13 menunjukkan kemampuan xgboost dapat memprediksi Diabetic Retinopathy dengan jumlah prediksi 321. Namun, model masih kesulitan dalam membedakan antara Glaucoma dan Normal, serta sebagian kasus Cataract diartikan sebagai Glaucoma. Dimana 43 kasus yang salah diprediksi sebagai normal dan 12 sebagai Cataract. Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Normal.

Confusion matrix pada gambar 14 menunjukkan kemampuan lightgbm dapat memprediksi Diabetic Retinopathy dengan jumlah prediksi 322. Namun, model masih kesulitan dalam membedakan antara Glaucoma dan Normal, serta sebagian kasus Cataract diartikan sebagai Glaucoma. Dimana 46 kasus yang salah diprediksi sebagai normal dan 13 sebagai Cataract. Terdapat 13 kasus yang diprediksi sebagai Glaucoma dan 9 sebagai Normal Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Normal.



Gambar 15 Model MLP

Confusion matrix pada gambar 15 menunjukkan kemampuan mlp dapat memprediksi Diabetic Retinopathy dengan jumlah prediksi 323. Namun, model masih kesulitan dalam membedakan antara Glaucoma dan Normal, serta sebagian kasus Cataract diartikan sebagai Glaucoma. Dimana 40 kasus yang salah diprediksi sebagai normal dan 11 sebagai Cataract. Terdapat 18 kasus yang diprediksi sebagai Glaucoma dan 8 sebagai Normal. Secara keseluruhan, model menunjukkan kinerja yang cukup baik meskipun masih memiliki sedikit kesalahan prediksi pada kasus Glaucoma dan Normal.

Berdasarkan hasil dari confusion matrix, sebagian besar model memiliki performa yang baik meskipun masih sedikit kesulitan dalam memprediksi Glaucoma dan Normal diakibatkan dari kemiripan data dibuktikan dengan fitur datanya saling berdekatan seperti pada Gambar 2.

Karena sebagian besar dari model sudah baik, uji untuk tingkat generalisasi model perlu untuk dilakukan dengan K-Fold Cross Validation. Metode ini tidak hanya memberikan

estimasi performa model yang lebih stabil, Hasil dari K-Fold Cross Validation juga dapat menjadi indicator bahwa model memiliki kinerja yang baik terhadatp variasi data, dan juga dapat menghindari model untuk mengalami overfitting atau underfitting.

G. Cross Validation

Hasil dari cross validation dari beberapa algoritma machine learning yang diuji menggunakan 10-Fold Cross validation. Metrik yang digunakan mencakup akurasi, precision, recall, dan F1-Score, masing-masing disertai dengan standar deviasi untuk menunjukkan stabilitas performa antar fold. Pada Tabel 5 menjadi dasar dalam mengevaluasi dan membandingkan efektivitas tiap metode dalam memodelkan data yang digunakan.

TABEL 5
PERBANDINGAN 10-FOLD CROSS VALIDATION

Model	***10-Fold Cross Validation***							
	Accuracy	***Standar Deviasi***	***Precission***	***Standar Deviasi***	***Recall***	***Standar Deviasi***	***F1 Score***	***Standar Deviasi***
Logistic Regression	91.42	0.0192	91.50	0.0194	91.42	0.0193	91.42	0.0192
Decission Tree	77.87	0.0256	78.06	0.0280	77.87	0.0257	77.88	0.0265
Random Forest	88.73	0.0135	88.94	0.0134	88.73	0.0135	88.76	0.0136
Gradient Boosting	89.89	0.0092	90.06	0.0098	89.89	0.0092	89.90	0.0095
Support Vector Machine	89.44	0.0124	89.69	0.0119	89.44	0.0124	89.42	0.0124
AdaBoost	80.53	0.0167	81.61	0.0139	80.53	0.0167	80.81	0.0159
Naive Bayes	75.64	0.0158	78.33	0.0156	75.64	0.0159	75.91	0.0161
K-Nearest Neighbors	88.87	0.0174	88.97	0.0173	88.86	0.0174	88.86	0.0174
Multi Layer Perceptron	**91.94**	**0.0178**	**92.00**	**0.0176**	**91.94**	**0.0179**	**91.91**	**0.0178**
XGBoost	**91.48**	**0.0177**	**91.64**	**0.0174**	**91.48**	**0.0177**	**91.46**	**0.0179**
LightGBM	**91.62**	**0.0145**	**91.76**	**0.0145**	**91.62**	**0.0145**	**91.61**	**0.0146**

Jika dibandingkan dengan hasil awal sebelum dilakukan Cross Validation terdapat beberapa model yang mengalami peningkatan performa hal ini biasanya disebabkan oleh kemampuan model dalam melakukan generalisasi terhadap data yang bervariasi. Model-model seperti Logistic Regression, AdaBoost, Naïve Bayes, K-Nearest Neighbors, dan Multi Layer Perceptron cenderung memperoleh manfaat dari evaluasi silang karena pendekatannya yang tidak terlalu kompleks, sehingga lebih tahan terhadap overfitting. Selain itu, pada model seperti KNN dan MLP, pelatihan pada data yang lebih bervariasi membantu model mengenali pola yang lebih umum, yang sebelumnya belum tergambarkan secara optimal dalam evaluasi awal. Dengan melibatkan seluruh bagian dataset secara bergantian, model memperoleh pengalaman yang lebih menyeluruh dalam mengenali struktur data.

Sebaliknya, terdapat pula beberapa model yang mengalami penurunan performa setelah cross-validation. Penurunan ini umumnya terjadi pada model yang terlalu kompleks atau terlalu menyesuaikan diri dengan data pelatihan awal, seperti Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, XGBoost, dan LightGBM. Dalam evaluasi awal, model-model ini mungkin sangat baik dalam mengenali pola pada subset tertentu, namun saat diuji pada data yang belum pernah dilihat sebelumnya, performanya menurun karena ketidaksesuaian antara pola yang dipelajari dengan distribusi data yang baru. Hal ini mengindikasikan bahwa model-model tersebut lebih rentan terhadap overfitting, dan hasil evaluasi awal yang tinggi mungkin terlalu optimistik.

Dengan demikian, cross-validation berfungsi tidak hanya sebagai alat ukur akurasi model, tetapi juga sebagai indikator stabilitas dan kemampuan generalisasi model terhadap data yang bervariasi. Model dengan performa stabil dan tidak terlalu fluktuatif antar fold cenderung lebih dapat diandalkan untuk diterapkan pada data dunia nyata.

IV. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa pendekatan ekstraksi fitur menggunakan arsitektur ResNet50 yang dikombinasikan dengan berbagai algoritma machine learning mampu mengklasifikasikan penyakit mata dari gambar retina

secara efektif. Dari sebelas model yang diuji, XGBoost memberikan hasil terbaik dengan akurasi 92,03%, diikuti oleh LightGBM dengan akurasi 91,88% dan MLP 91,50%.

Ketika dilakukan K-Fold Cross Validation untuk menilai generalisasi data, XGBoost mendapatkan akurasi rata-rata dengan nilai 91,48% dengan standard deviasi rata-rata akurasi 0,0017 yang mengimplikasikan model dapat melakukan generalisasi dengan baik. Selain XGBoost, LightGBM juga memiliki generalisasi baik dengan akurasi rata-rata 91,62% dan standard deviasi rata-rata akurasi 0,0145. Pada model MLP, akurasi mengalami peningkatan dengan nilai rata-rata akurasi 91,94%, dan standard deviasi rata-rata akurasi 0,0178. Temuan ini memperkuat potensi metode machine learning sebagai solusi diagnosis berbasis citra retina, terutama dalam situasi dengan keterbatasan sumber daya komputasi.

Meskipun beberapa algoritma lain seperti AdaBoost dan Naive Bayes menunjukkan performa yang lebih rendah, keseluruhan performa menunjukkan kinerja yang menjanjikan dan dapat dikembangkan lebih lanjut. Keterbatasan seperti jumlah data yang terbatas dan belum dioptimalkannya hyperparameter dapat menjadi arah perbaikan pada penelitian selanjutnya. Dengan pengembangan lebih lanjut, performa dari setiap model berpotensi besar menjadi alat bantu diagnosis penyakit mata yang cepat, akurat, dan efisien.

DAFTAR PUSTAKA

[1] M. J. Burton *et al.*, "The Lancet Global Health Commission on Global Eye Health: vision beyond 2020," *Lancet Glob. Heal.*, vol. 9, no. 4, pp. e489–e551, 2021, doi: 10.1016/S2214-109X(20)30488-5.

[2] Y. Jeong, Y. J. Hong, and J. H. Han, "Review of Machine Learning Applications Using Retinal Fundus Images," *Diagnostics*, vol. 12, no. 1, pp. 1–27, 2022, doi: 10.3390/diagnostics12010134.

[3] M. Moannaei *et al.*, "Performance and limitation of machine learning algorithms for diabetic retinopathy screening and its application in health management: a meta-analysis," *Biomed. Eng. Online*, vol. 24, no. 1, pp. 1–15, 2025, doi: 10.1186/s12938-025-01336-1.

[4] J. H. Wu, T. Y. A. Liu, W. T. Hsu, J. H. C. Ho, and C. C. Lee, "Performance and limitation of machine learning algorithms for diabetic retinopathy screening: Meta-analysis," *J. Med. Internet Res.*, vol. 23, no. 7, pp. 1–15, 2021, doi: 10.2196/23863.

[5] S. Muchuchuti and S. Viriri, "Retinal Disease Detection Using Deep Learning Techniques: A Comprehensive Review," *J. Imaging*, vol. 9, no. 4, 2023, doi: 10.3390/jimaging9040084.

[6] Z. Li, M. Xu, X. Yang, and Y. Han, "Multi-Label Fundus Image Classification Using Attention Mechanisms and Feature Fusion," *Micromachines*, vol. 13, no. 6, 2022, doi: 10.3390/mi13060947.

[7] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable Deep Learning Methods in Medical Image Classification: A Survey," *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1–39, 2023, doi: 10.1145/3625287.

[8] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, pp. 1–24, 2022, doi: 10.1016/j.compbiomed.2021.105111.

[9] M. Avanzo, J. Stancanello, G. Pirrone, A. Drigo, and A. Retico, "The Evolution of Artificial Intelligence in Medical Imaging : From Computer Science to Machine and Deep Learning," *Cancers (Basel).*, vol. 16, no. 21, pp. 1–23, 2024.

[10] T. Sundeep, U. Divyasree, K. Tejaswi, U. R. Vinithanjali, and A. K. Kumar, "Feature Extraction of Ophthalmic Images Using Deep Learning and Machine Learning Algorithms †," *Eng. Proc.*, vol. 56, no. 1, 2023, doi: 10.3390/ASEC2023-15231.

[11] F. T. A. Sayyidul Laily, "Feature Extraction and Classification of Retinal Images Using Sobel Segmentation and Linear SVC," *Int. J. Artif. Intell. Med. Issues*, vol. 2, no. 2, pp. 136–149, 2024, doi: 10.56705/ijaimi.v2i2.153.

[12] K. He, "Deep Residual Learning for Image Recognition," *IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 770–778, 2016.

[13] M. A. Fikri, "Improving Osteosarcoma Detection through SMOTE-Driven Machine Learning Approaches," *IJID (International J. Informatics Dev.*, vol. 13, no. 2, pp. 517–529, 2025, doi: 10.14421/ijid.2024.4890.

[14] I. T. Jolliffe, J. Cadima, and J. Cadima, "Principal component analysis : a review and recent developments Subject Areas : Author for correspondence :," 2016.

[15] S. G. K. Patro and K. K. sahu, "Normalization: A Preprocessing Stage," *Iarjset*, pp. 20–22, 2015, doi: 10.17148/iarjset.2015.2305.

[16] G. Ramanathan, D. Chakrabarti, A. Patil, S. Rishipathak, and S. Kharche, "Eye Disease Detection Using Machine Learning," *2021 2nd Glob. Conf. Adv. Technol. GCAT 2021*, pp. 1–5, 2021, doi: 10.1109/GCAT52182.2021.9587740.