

ABSTRAK

Glaukoma merupakan penyebab utama kebutaan permanen di dunia yang merusak saraf optik secara progresif. Deteksi dini sangat krusial, tetapi pemeriksaan manual terhadap citra fundus retina oleh dokter spesialis memakan waktu lama dan sering kali bersifat subjektif. Sistem Computer-Aided Diagnosis (CAD) berbasis *deep learning* menawarkan solusi penapisan massal yang cepat dan konsisten. Selama ini, arsitektur Convolutional Neural Network (CNN) seperti ResNet-50 mendominasi dengan bias induktif lokalnya, tetapi berpotensi melewatkan anomali tekstur global pada retina. Sebagai alternatif mutakhir, arsitektur Vision Transformer (ViT) hadir memecahkan batasan tersebut melalui mekanisme *self-attention* untuk menangkap hubungan spasial secara utuh, meskipun menuntut beban komputasi yang sangat tinggi. Perbedaan karakteristik mendasar ini memunculkan permasalahan mengenai arsitektur mana yang memiliki efektivitas terbaik antara akurasi prediktif klinis dan efisiensi waktu. Oleh karena itu, penelitian ini bertujuan mengevaluasi dan membandingkan secara komprehensif kinerja kedua arsitektur tersebut dalam mendeteksi glaukoma.

Metode penelitian diawali dengan akuisisi dataset publik citra fundus retina yang dibagi secara terstratifikasi dengan rasio 80:20 untuk pengalokasian data latih dan data uji. Teknik augmentasi data kombinasi diterapkan secara khusus pada data latih guna memperkaya variasi fitur. Tahap pra-pemrosesan citra mengintegrasikan metode kontras adaptif (CLAHE), penyesuaian dimensi menjadi 224×224 piksel, transformasi geometris, serta normalisasi standar dari ImageNet. Eksperimen diuji secara teliti melalui 16 skenario yang memvariasikan *batch size* (16 dan 32), fungsi optimasi (Adam dan AdamW), serta *learning rate* (1×10^{-5} dan 1×10^{-4}), dengan *early stopping*. Sistem evaluasi kemudian diimplementasikan ke dalam *dashboard* antarmuka interaktif berbasis Streamlit.

Hasil pengujian membuktikan bahwa arsitektur ViT-B/16 pada skenario 10 (*batch size* 16, fungsi optimasi Adam, dan *learning rate* 1×10^{-4}) berhasil menjadi model terbaik dengan capaian akurasi 95,55%, presisi 95,74%, sensitivitas 95,55%, F1-Score 95,56%, dan nilai AUC 0,9872. Sementara itu, kelompok ResNet-50 mencapai performa optimalnya pada skenario 6 dengan akurasi 94,74%. Terdapat kompromi waktu komputasi operasional yang sangat tajam di antara keduanya. ResNet-50 terbukti sangat efisien dengan waktu pelatihan di bawah satu jam dan inferensi kurang dari satu menit, sedangkan ViT-B/16 membutuhkan hingga 674 menit pelatihan dan inferensi 3-4 menit. Penelitian ini berkontribusi memberikan panduan komprehensif bagi implementasi penapisan glaukoma yang akurat, responsif, dan dapat diandalkan.

Kata Kunci: Glaukoma, ResNet-50, Vision Transformer, Citra Fundus Retina, *Deep Learning*

ABSTRACT

Glaucoma is a leading cause of permanent blindness worldwide, progressively damaging the optic nerve. Early detection is highly crucial, but manual examination of retinal fundus images by specialists is time-consuming and often subjective. Deep learning-based Computer-Aided Diagnosis (CAD) systems offer a fast and consistent mass screening solution. Thus far, Convolutional Neural Network (CNN) architectures like ResNet-50 have dominated with their local inductive bias, but they potentially miss global texture anomalies in the retina. As a cutting-edge alternative, the Vision Transformer (ViT) architecture emerges to break these boundaries through a self-attention mechanism to capture spatial relationships entirely, although demanding a very high computational load. These fundamental characteristic differences raise a problem regarding which architecture possesses the best effectiveness between clinical predictive accuracy and time efficiency. Therefore, this study aims to comprehensively evaluate and compare the performance of both architectures in detecting glaucoma.

The research method began with the acquisition of public retinal fundus image datasets, which were partitioned using a stratified split with an 80:20 ratio for allocating training and testing data. Combined data augmentation techniques were specifically applied to the training data to enrich feature variations. The image preprocessing stage integrated the adaptive contrast method (CLAHE), dimension resizing to 224×224 pixels, geometric transformations, and ImageNet standard normalization. The experiments were meticulously evaluated through 16 scenarios varying the batch size (16 and 32), optimization function (Adam and AdamW), and learning rate (1×10^{-5} dan 1×10^{-4}), managed by early stopping controls. The evaluation system was then implemented into an interactive Streamlit-based dashboard interface.

The test results proved that the ViT-B/16 architecture in scenario 10 (batch size 16, Adam optimization function, and *learning rate* 1×10^{-4}) successfully became the best model, achieving an accuracy of 95.55%, precision of 95.74%, sensitivity of 95.55%, F1-Score of 95.56%, and an AUC value of 0.9872. Meanwhile, the ResNet-50 group reached its optimal performance in scenario 6 with an accuracy of 94.74%. There is a very sharp operational computational time trade-off between the two. ResNet-50 proved to be highly efficient with a training time of under one hour and an inference time of less than one minute, whereas ViT-B/16 required up to 674 minutes of training and 3 to 4 minutes of inference. This research contributes to providing a comprehensive guide for the implementation of accurate, responsive, and reliable glaucoma screening.

Keywords: Glaucoma, ResNet-50, Vision Transformer, Retinal Fundus Image, Deep Learning