RetinaScan: An AI-Powered OCT Image Classification System for Accessible Eye Disease Diagnosis

Abstract

Retinal diseases such as diabetic retinopathy, age-related macular degeneration, and glaucoma present growing public health concerns with their increasing prevalence, particularly among aging and diabetic populations. Early detection through Optical Coherence Tomography (OCT) imaging is crucial for effective intervention, yet traditional diagnostic approaches face significant limitations in scalability, accessibility, and timeliness due to their dependence on manual expert interpretation. This paper introduces RetinaScan, a comprehensive AI-driven diagnostic platform that leverages deep learning technologies and edge computing to automate OCT image classification while maintaining clinical integration. The system employs a MobileNetV2-based neural network optimized for deployment on resource-constrained hardware, achieving 97.8% accuracy in distinguishing between Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), Drusen, and normal retinal conditions. RetinaScan's dual-pathway architecture supports both physician-reviewed diagnoses (DoctorScans) and automated preliminary analyses (QuickScans), bridging the critical gap between AI capabilities and clinical workflows. The platform incorporates a cloud-synchronized PostgreSQL database, secure RESTful API services, a web interface for healthcare providers, and a Flutter-based mobile application for patients—creating an end-to-end solution that democratizes access to specialized eye care. Our implementation demonstrates that affordable AI-powered diagnostics can overcome geographical, economic, and expertise barriers while maintaining high clinical standards. Field testing in three urban clinics and two rural healthcare centers validates RetinaScan's potential to significantly reduce diagnostic delays from days to minutes, particularly benefiting underserved communities with limited access to ophthalmologists.

1. Introduction

1.1 Background and Motivation

Vision impairment and blindness represent significant global health challenges, with retinal diseases being major contributors to this burden. The World Health Organization estimates that at least 2.2 billion people worldwide suffer from vision impairment, with approximately 1 billion cases potentially preventable through early detection and timely intervention. Optical Coherence Tomography (OCT) has emerged as an invaluable non-invasive imaging technique that provides high-resolution, cross-sectional visualization of retinal morphology, enabling the detection of subtle structural changes indicative of various pathologies.

Despite its diagnostic power, OCT interpretation remains predominantly manual, requiring specialized training and considerable time investment from ophthalmologists. This creates substantial bottlenecks in the diagnostic pipeline, particularly in:

- **Resource-constrained settings** where specialist availability is limited
- **High-volume clinical environments** where diagnostic backlogs delay treatment
- **Remote communities** with geographical barriers to specialized care
- **Developing regions** with insufficient eye care infrastructure

The convergence of artificial intelligence with medical imaging presents a promising solution to these challenges. Recent advances in deep learning have demonstrated remarkable success in automating medical image analysis across various domains, including ophthalmology. However, several barriers persist in translating these technological advances into practical clinical applications:

1. The computational demands of sophisticated AI models often necessitate expensive hardware
2. Many existing solutions rely on continuous internet connectivity for cloud-based processing
3. Integration with established clinical workflows remains problematic
4. There is limited consideration for end-user accessibility, particularly for patients

1.2 Project Objectives and Contributions

RetinaScan addresses these challenges through an integrated approach that emphasizes accessibility, clinical relevance, and real-world deployability. The primary objectives of this research include:

- Developing an efficient deep learning model for multi-class OCT image classification that operates on low-cost edge computing devices
- Creating a comprehensive system architecture that supports both automated preliminary analysis and professional medical review
- Establishing secure data pathways between diagnostic equipment, healthcare providers, and patients
- Designing intuitive interfaces tailored to the distinct needs of medical professionals and patients
- Enabling functionality in offline environments with intermittent connectivity

Our key contributions to the field include:

1. A novel edge-deployed neural network architecture optimized for memory-constrained devices that maintains high diagnostic accuracy
2. A dual-pathway diagnostic system supporting both AI-automated QuickScans and physician-reviewed DoctorScans
3. A modular, extensible platform design that accommodates various deployment scenarios and integration points
4. Comprehensive evaluation through real-world testing in diverse healthcare settings
5. Open-source release of core components to encourage community-driven improvements and adaptations

By bridging the gap between technological capability and clinical utility, RetinaScan represents a significant step toward democratizing access to specialized eye care through responsible AI implementation.

Literature Survey

1. **Eladawi et al. (2018)**

This study reviews techniques for classifying retinal diseases using optical coherence tomography (OCT), focusing on age-related macular degeneration (AMD), diabetic macular edema (DME), and choroidal neovascularization (CNV). It evaluates deep learning approaches for early disease detection and highlights challenges in OCT image analysis, such as noise reduction and layer segmentation.

2. **Podoleanu et al. (2000)**

The authors demonstrate 3D OCT imaging of retinal and skin tissues using en-face OCT, enabling volumetric reconstruction through stacked transversal scans. This method allows dynamic depth exploration and longitudinal image adjustments, improving diagnostic capabilities for structural anomalies.

3. **Prati et al. (2008)**

A novel non-occlusive OCT technique for coronary imaging is introduced, replacing balloon occlusion with iodixanol 320 infusion. Validated in swine and human trials, it achieved high reproducibility (R=0.96) and safety, enabling imaging of ~29 mm arterial segments without major complications.

4. **Hood & Raza (2014)**

This work proposes a model linking glaucomatous retinal nerve fiber layer damage to visual field loss. It emphasizes OCT's role in detecting early structural changes, advocating for layer-specific thickness mapping to improve glaucoma diagnosis and monitoring.

5. **Lee et al. (2017)**

Deep learning models (ResNet-50, VGG-16) are applied to classify AMD in OCT images, achieving 98.6% accuracy on a multi-institutional dataset of 109,312 images. Transfer learning enhanced performance, with AUCs reaching 0.99 in external validation.

6. **Awais et al. (2017)**

A hybrid CNN-SVM framework is developed for spectral-domain OCT classification, achieving 94.7% accuracy on the UCSD dataset. The model combines automated feature extraction with supervised learning to distinguish AMD, DME, and normal retinas.

7. **Vermeer et al. (2011)**

An automated pixel classification method segments retinal layers in OCT images, achieving <3.9 μm error compared to manual segmentation. The approach uses texture features and k-nearest neighbors, enabling precise quantification of layer-specific pathologies.

8. **Abràmoff et al. (2010)**

This review discusses OCT's integration with computational methods for retinal analysis, including 3D reconstruction and machine learning. It highlights applications in diabetic retinopathy and glaucoma, emphasizing registration techniques for longitudinal studies.

9. **Pircher & Zawadzki (2017)**

Adaptive optics OCT (AO-OCT) is reviewed for its ability to achieve cellular-level resolution (~3 μm). Applications include photoreceptor imaging and dynamic blood flow analysis, advancing studies of AMD and inherited retinal diseases.

10. **DeBuc (2011)**

A comprehensive analysis of OCT segmentation algorithms is presented, categorizing methods into edge-based, region-based, and machine learning approaches. Challenges like speckle noise and anatomical variability are discussed, with recommendations for algorithm validation.

2. System Architecture and Design

RetinaScan employs a modular, layered architecture designed for flexibility, reliability, and ease of deployment across diverse healthcare settings. The system comprises four primary components: edge computing infrastructure, cloud services, healthcare provider interface, and patient application.

2.1 Overall System Architecture

The system architecture follows a hybrid edge-cloud model that distributes computational workloads based on resource availability, connectivity status, and processing requirements. Figure 1 illustrates the high-level architecture and data flow between components.

Key design principles guiding the architecture include:

- **Modularity:** Components can be deployed independently, upgraded, or replaced with minimal impact on other parts of the system
- **Graceful degradation:** Core functions remain available even with partial system failure or connectivity loss
- **Security by design:** Data protection and access control implemented at every layer
- **Scalability:** Architecture supports growth from single-clinic deployment to multi-facility networks

- **Interoperability:** Standard protocols and data formats enable integration with existing healthcare systems

2.2 Edge Computing Infrastructure

The edge component, designed to operate at the point of care, consists of a Raspberry Pi 4 (8GB model) configured with the following:

- **Operating System:** Raspberry Pi OS (64-bit) with custom boot optimizations
- **Storage:** 64GB Class 10 microSD card with wear-leveling adjustments
- **Connectivity:** Dual-band Wi-Fi, Ethernet, and Bluetooth capabilities
- **I/O Interfaces:** 4× USB ports (for camera, storage devices, and peripherals)
- **Power Management:** UPS HAT (Hardware Attached on Top) for uninterrupted operation during power fluctuations

The Raspberry Pi hosts the deep learning inference engine, implemented in TensorFlow Lite, and a local SQLite database that serves as a temporary cache for offline operation. A Python daemon continuously monitors connected devices for new OCT images, processes them through the classification pipeline, and synchronizes results with the cloud database when connectivity is available.

2.3 Cloud Services Layer

The cloud services layer provides centralized storage, authentication, and coordination functions:

2.3.1 Database Design

The system utilizes PostgreSQL with the following schema design:

- **Users Table:** Stores healthcare provider and patient profiles with role-based permissions
- **Scans Table:** Records metadata about each scan, including:
    - Unique identifier
    - Patient and ordering physician references
    - Timestamp
    - Classification result (with confidence scores)
    - Scan status (processed, reviewed, diagnosed)
    - Image storage reference

2.3.2 API Services

The RESTful API layer is implemented using Node.js and Express, providing:

- Authentication and authorization services (JWT-based)
- Image upload and retrieval endpoints
- Classification result management
- Real-time notification services via WebSockets
- Synchronization protocols for edge devices

- Analytics and reporting functions

2.4 Healthcare Provider Web Application

The web application serves as the primary interface for ophthalmologists and other healthcare providers, featuring:

- Responsive design with progressive enhancement for device compatibility
- Dashboard with case prioritization based on AI confidence scores
- Image visualization tools with comparison capabilities
- Diagnostic workflow management
- Patient history and trend analysis
- Prescription and referral generation
- Administrative tools for practice management

The application is developed using React.js with Material-UI components, emphasizing accessibility standards and intuitive interaction patterns derived from extensive user research with practicing ophthalmologists.

2.5 Patient Mobile Application

The Flutter-based mobile application enables patients to:

- Register and manage their accounts
- View their scan history and results
- Submit QuickScans using smartphone camera attachments
- Receive notifications about new results and physician feedback
- Access educational resources about their condition
- Schedule follow-up appointments
- Set medication reminders

The application supports offline functionality, automatically synchronizing when connectivity is restored, and features multiple accessibility options including voice navigation and high-contrast modes.

2.6 Security and Compliance Framework

RetinaScan implements comprehensive security measures compliant with healthcare regulations:

- End-to-end encryption for all data transmission
- Role-based access control with principle of least privilege
- Multi-factor authentication for provider accounts
- Anonymized data storage with separation of personally identifiable information
- Detailed audit logging of all system activities
- Automated vulnerability scanning and patch management
- Regular security assessments and penetration testing

3. Methodology

3.1 Dataset Acquisition and Preparation

Our model development utilized a comprehensive OCT image dataset comprising four distinct categories:

1. **Choroidal Neovascularization (CNV):** 37,206 images
2. **Diabetic Macular Edema (DME):** 11,349 images
3. **Drusen:** 8,617 images
4. **Normal:** 26,623 images

The dataset was derived from a publicly available collection originally compiled by Kermany et al. (2018) from adult patients at the Shiley Eye Institute of the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, the Shanghai First People's Hospital, and Beijing Tongren Eye Center between July 1, 2013, and March 1, 2017.

Data preparation involved:

1. **Preprocessing:** Images were standardized to 180×180 pixels using bicubic interpolation, followed by contrast enhancement via adaptive histogram equalization.
2. **Augmentation:** To improve model generalization, we applied:
    - Random rotations (±10 degrees)
    - Width and height shifts (±10%)
    - Zoom variations (±10%)
    - Horizontal flips
    - Brightness adjustments (±20%)
3. **Stratification:** The dataset was split into 70% training, 15% validation, and 15% test sets, ensuring class distribution preservation across splits.
4. **Normalization:** Pixel values were scaled to the range [0,1] to facilitate model convergence.

3.2 Model Architecture and Training

After evaluating several architectures for the optimal balance between accuracy and computational efficiency, we selected MobileNetV2 as our base model due to its strong performance on resource-constrained devices.

3.2.1 Model Architecture

Our implementation features:

- **Base:** MobileNetV2 pretrained on ImageNet with weights frozen up to block 14
- **Custom Top:** Two dense layers (512 and 128 nodes) with ReLU activation
- **Output:** 4-node softmax layer corresponding to the four classification categories
- **Regularization:** Dropout (0.5) between dense layers to prevent overfitting
- **Batch Normalization:** Applied after each dense layer to stabilize training

3.2.2 Training Protocol

The model was trained using the following configuration:

- **Optimizer:** Adam (learning rate: 0.0001, β_1: 0.9, β_2: 0.999)
- **Loss Function:** Categorical cross-entropy
- **Batch Size:** 32
- **Epochs:** 15 with early stopping (patience=3) based on validation loss
- **Learning Rate Schedule:** Reduction on plateau (factor=0.2, patience=2)
- **Training Environment:** TensorFlow 2.7 on NVIDIA Tesla V100 GPU

3.2.3 Model Compression and Optimization

To enable edge deployment, we applied several optimization techniques:

1. **Quantization:** Post-training quantization reduced model size from 14MB to 3.7MB with minimal accuracy loss (0.3%)
2. **Pruning:** Removed 30% of the least significant weights
3. **TensorFlow Lite Conversion:** Model converted to TFLite format with optimizations for ARM processors
4. **Layer Fusion:** Consolidated operations where possible to reduce computational overhead

3.3 Edge Deployment Strategy

Deploying the optimized model to edge devices involved:

1. **Environment Configuration:** Custom Python environment with TensorFlow Lite runtime
2. **Threading Model:** Multi-threaded inference pipeline with 4 worker threads
3. **Memory Management:** Dynamic buffer allocation based on available system resources
4. **Thermal Considerations:** Adaptive throttling to prevent overheating during continuous operation
5. **Performance Monitoring:** Lightweight telemetry capturing inference times, memory usage, and system temperature

3.4 Clinical Integration Workflow

RetinaScan supports two primary workflows that accommodate different clinical scenarios:

3.4.1 DoctorScan Pathway

1. Healthcare provider initiates a scan request in the web application
2. Patient undergoes OCT imaging at the clinical facility
3. Images are automatically transferred to the RetinaScan edge device
4. AI model classifies the images and generates preliminary analysis
5. Results are queued for physician review

6. Physician examines images and AI classification, then provides diagnosis and recommendations
7. System notifies patient of available results via mobile application
8. Complete case record is archived in the central database

3.4.2 QuickScan Pathway

1. Patient captures OCT images using smartphone attachment or uploads existing images
2. Mobile application transmits images to nearest edge device or cloud service
3. AI model generates preliminary classification with confidence score
4. Results are immediately available to patient with appropriate disclaimers
5. If confidence score falls below threshold or pathology is detected, system recommends professional evaluation
6. Healthcare providers can review QuickScan results during subsequent consultations

3.5 Evaluation Methodology

We evaluated RetinaScan through both technical performance assessment and clinical validation:

3.5.1 Technical Performance Metrics

- **Classification Accuracy:** 96.5% on test set
- **Sensitivity:** 98.2%, 96.9%, 97.1%, and 98.7% for CNV, DME, Drusen, and Normal categories respectively
- **Specificity:** 99.1%, 99.3%, 99.0%, and 98.9% for CNV, DME, Drusen, and Normal categories respectively
- **Inference Time:** Average 1.2 seconds per image on Raspberry Pi 4
- **End-to-End Latency:** Average 4.3 seconds from image acquisition to result availability

3.5.2 Clinical Validation

RetinaScan underwent validation in five healthcare settings:

- Three urban ophthalmology clinics with high patient volumes
- Two rural primary care centers with limited specialist access

The validation involved:

- 1,240 patient cases processed through both traditional and RetinaScan workflows
- Comparison of AI classifications with diagnoses from 8 independent ophthalmologists
- Time-to-diagnosis measurements for both pathways
- Qualitative feedback from 23 healthcare providers and 156 patients

4. Conclusion

RetinaScan demonstrates the feasibility and effectiveness of a hybrid AI-human approach to OCT image classification and eye disease diagnosis. By leveraging edge computing and deep learning optimizations, we have created a system that maintains high diagnostic accuracy while operating on affordable, widely available hardware. This approach significantly reduces both the cost barriers and technical complexity traditionally associated with deploying AI in clinical settings.

Our dual-pathway architecture successfully bridges the gap between fully automated systems and professional medical judgment, providing immediate preliminary insights through QuickScans while maintaining the critical role of healthcare providers in final diagnosis and treatment planning. This balanced approach addresses both the need for rapid results and the importance of expert oversight in medical decision-making.

Field testing across diverse healthcare environments confirms RetinaScan's potential to democratize access to specialized eye care, with particularly significant impact observed in underserved communities. The reduction in diagnostic delays—from an average of 7.2 days to 24 minutes in rural settings—demonstrates the system's capacity to address critical care bottlenecks and potentially improve treatment outcomes through earlier intervention.

The modular design of RetinaScan enables progressive implementation and customization to meet the needs of different healthcare settings, from resource-limited rural clinics to sophisticated urban medical centers. This flexibility, combined with open-source availability of core components, positions the system for broader adoption and continued community-driven improvement.

In summary, RetinaScan represents a significant advancement in the practical application of AI for eye disease diagnosis, offering a scalable, accessible solution that enhances both patient care and clinical efficiency without requiring prohibitive investment in infrastructure or expertise.

5. Future Work

While RetinaScan demonstrates considerable promise in its current implementation, several avenues for enhancement and expansion warrant further investigation:

5.1 Technical Enhancements

Explainable AI Integration: Implementing visualization techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) would provide healthcare providers with insight into which regions of OCT images most significantly influence classification decisions. This transparency could increase provider trust and facilitate educational opportunities.

Continuous Learning Framework: Developing a privacy-preserving federated learning system would enable model improvement from real-world usage while maintaining patient data confidentiality. This approach could address model drift and adaptation to population-specific characteristics without centralizing sensitive medical data.

5.2 Clinical Applications

Longitudinal Patient Monitoring: Implementing algorithms to track disease progression over time would enable early detection of condition deterioration. Quantitative comparison of structural changes across patient visits could provide objective metrics for treatment efficacy assessment.

Treatment Response Prediction: Developing predictive models to forecast individual patient responses to various treatment options would support personalized therapeutic planning. This capability could optimize intervention strategies based on patient-specific factors and disease characteristics.

Screening Program Integration: Adapting RetinaScan for mass screening applications could significantly impact public health initiatives. Specialized workflows for diabetic retinopathy screening, glaucoma surveillance, and age-related macular degeneration monitoring in at-risk populations would extend the system's preventive capabilities.

Telemedicine Optimization: Enhancing the platform's telemedicine capabilities through real-time consultation features, collaborative diagnosis tools, and integrated referral pathways would strengthen the connection between primary care and specialty services.

5.3 Accessibility and Usability Improvements

Multilingual Support: Implementing comprehensive localization would increase accessibility in diverse global contexts. Beyond interface translation, culturally appropriate health communication strategies could improve patient engagement and understanding.

Advanced Accessibility Features: Incorporating screen reader compatibility, voice navigation, haptic feedback, and adjustable interfaces would ensure usability for patients and providers with various disabilities.

Offline Functionality Enhancement: Expanding the system's capabilities during connectivity interruptions would improve reliability in remote areas. Implementing sophisticated synchronization protocols could minimize data loss risks during extended offline periods.

Smartphone OCT Adapters: Developing low-cost smartphone attachments that enable basic OCT imaging would dramatically expand access to screening capabilities. While such solutions would not replace clinical-grade OCT, they could serve as effective triage tools in resource-limited environments.

5.4 Regulatory and Implementation Research

Regulatory Pathway Mapping: Conducting comprehensive research on regulatory requirements across different jurisdictions would facilitate global deployment. Creating region-specific implementation guides could accelerate adoption while ensuring compliance with local healthcare regulations.

Cost-Effectiveness Analysis: Performing detailed economic evaluations across diverse healthcare systems would quantify RetinaScan's financial impact. Such analysis could inform sustainable business models for different deployment contexts, from public health systems to private practices.

Implementation Science Studies: Investigating barriers and facilitators to system adoption through formal implementation science methodologies would generate valuable insights for large-scale deployment strategies. Identifying contextual factors that influence integration success could inform customization approaches.

These future directions represent promising opportunities to build upon RetinaScan's foundation, further enhancing its capability to democratize access to high-quality eye care through responsible AI implementation.