Restnet152	82.73	82.73	82.45	35.08	66.75
Learning Rate 0.001					
VGG19	86.23	86.23	86.73	41.78	71.58
MobileNetV2	88.05	88.05	87.77	44.37	73.40
Restnet152	85.00	85.00	83.38	37.86	68.75
Learning Rate 0.0001					
VGG19	85.05	85.05	85.39	40.13	70.19
MobileNetV2	86.91	86.91	86.23	44.18	72.44
Restnet152	85.31	85.31	81.19	35.65	67.38

From Table I, it is evident that MobileNetV2 achieved the highest overall performance with a Final Score of 73.40% at a learning rate of 0.001. In particular, MobileNetV2 achieved a Cohen's Kappa value of 44.37%, an accuracy and F1-score of 88.05%, and an AUC of 87.77%. MobileNetV2 outperformed VGG19 and ResNet152, demonstrating its efficacy for multi-label classification of eye disorders at this learning rate.

Although VGG19 demonstrated decent performance with a Final Score of 71.58% at a learning rate of 0.001, MobileNetV2's higher AUC and Kappa coefficient indicate better overall classification capability and agreement with the true labels. While the AUC indicates how well the model can differentiate between positive and negative classes, a greater Kappa indicates a higher level of agreement than chance. Thus, with a learning rate of 0.001, MobileNetV2 was found to be the optimal model for further fine-tuning.

B. *Fine-Tuning the Best Model*

Following the identification of MobileNetV2 as the top-performing model with a learning rate of 0.001, fine-tuning was carried out to further improve its performance. The MobileNetV2 model was fine-tuned to better fit the ODIR dataset, allowing for greater alignment with the unique features of the dataset.

Post-fine-tuning, MobileNetV2 achieved an accuracy and F1-score of 86.54%, an AUC of 87.88%, and a Cohen's Kappa coefficient of 44.62%, resulting in a Final Score of 73.01%, as shown in Table II.

TABLE II. FINE-TUNING RESULTS OF MOBILENETV2

Dataset	Accuracy (%)	F1 Score (%)	AUC (%)	Kappa (%)	Final (%)
Validation	87.60	87.60	88.65	44.87	73.71
Testing	86.54	86.54	87.88	44.62	73.01

Compared to the baseline performance, the accuracy and F1-score changed from 88.05% to 87.60%, and the AUC improved from 87.77% to 88.65%. The Kappa coefficient also increased from 44.37% to 44.87%, shifting the Final Score from 73.40% to 73.71%. This outcome suggests that while the model's discrimination capability improved, the overall agreement with the true labels remained nearly consistent.

While the AUC demonstrated an improvement after fine-tuning, the small increase in the Kappa coefficient from 44.37% to 44.87% indicates a marginally better agreement with the true labels. This subtle change suggests that the model maintains reliable performance.

V. CONCLUSION

Using fundus images from the ODIR dataset, this work successfully created a multi-label eye disease recognition system employing the deep learning architectures ResNet152, VGG19, and MobileNetV2 as feature extraction backbone. Prior to fine-tuning, the evaluation results showed that MobileNetV2, with a learning rate of 0.001, performed the best, with a Final Score of 73.40%, an accuracy and F1-score of 88.05%, an AUC of 87.77%, and a Cohen's Kappa factor of 44.37%. MobileNetV2's final score was 73.71% after the model was fine-tuned, with an accuracy and F1-score of 87.60%, an AUC of 88.65%, and a Cohen's Kappa value of 44.87%. The final configuration on the testing set yielded a Final Score of 73.01%, with an accuracy and F1-score of 86.54%, an AUC of 87.88%, and a Kappa coefficient of 44.62%.

Compared to VGG19 and ResNet152, MobileNetV2 delivered superior results, underscoring its capability for multi-label eye disease classification. Overall, the developed system demonstrates potential in streamlining eye disease detection, particularly in regions with limited ophthalmologist availability. However, the marginal performance gains after fine-tuning indicate that further optimization is necessary to enhance detection accuracy. Moreover, this study still faces challenges that require further research, including addressing data imbalance, enlarging the dataset for better generalization, and experimenting with diverse hyper-parameter configurations, such as exploring different optimizers not used in this study.

REFERENCES

[1] World Health Organization, "Blindness and vision impairment," WHO. Accessed: May 16, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[2] George Raden Mas Said, Vanny Leutualy, and Saleh Tualeka, "Karakteristik Pasien Penyakit Mata Di RSUD Masohi Maluku Tengah: Studi Deskriptif," vol. 3, pp. 1–6, 2021.

[3] S. M. Prawito, "Habis Gelap Terbitlah Terang," SehatNegeriKu. Accessed: May 16, 2024. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/blog/20170426/3120633/habis-gelap-terbitlah-terang/

[4] N. A. Wirawan, "Angka kebutaan dan gangguan penglihatan di Indonesia tembus 8 juta kasus," GoodStats Data. Accessed: Jan. 17, 2025. [Online]. Available: https://data.goodstats.id/statistic/angka-kebutaan-dan-gangguan-penglihatan-di-indonesia-tembus-8-juta-kasus-1buXL

[5] R. Bernardes, P. Serranho, and C. Lobo, "Digital Ocular Fundus Imaging: A Review," Ophthalmologica, vol. 226, no. 4, pp. 161–181, 2011, doi: 10.1159/000329597.

[6] J. Wang, L. Yang, Z. Huo, W. He, and J. Luo, "Multi-Label Classification of Fundus Images With EfficientNet," IEEE Access, vol. 8, pp. 212499–212508, 2020, doi: 10.1109/ACCESS.2020.3040275.

[7] M. Pektaş, "Performance Analysis of Efficient Deep Learning Models for Multi-Label Classification of Fundus Image," Artificial Intelligence Theory and Applications, vol. 3, no. 2, pp. 105–112, 2023.

[8] T. D. Nguyen, D.-T. Le, J. Bum, S. Kim, S. J. Song, and H. Choo, "Retinal Disease Diagnosis Using Deep Learning on Ultra-Wide-Field Fundus Images," Diagnostics, vol. 14, no. 1, p. 105, Jan. 2024, doi: 10.3390/diagnostics14010105.

[9] Keras Documentation, "Keras Documentation." Accessed: Nov. 29, 2024. [Online]. Available: https://keras.io/api/

[10] ODIR 2019 Challenge, “ODIR 2019 Challenge - Retinal Disease Classification.” Accessed: Nov. 29, 2024. [Online]. Available: https://odir2019.grand-challenge.org/

[11] S. M. Pizer et al., “Adaptive histogram equalization and its variations,” Comput Vis Graph Image Process, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[13] Melisa Bardhi, “Image Detection Using the VGG-19 Convolutional Neural Network,” Medium. Accessed: May 16, 2024. [Online]. Available: https://melisabardhi.medium.com/image-detection-using-convolutional-neural-networks-89c9e21fffa3

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[15] M.-L. Zhang and Z.-H. Zhou, “A Review on Multi-Label Learning Algorithms,” IEEE Trans Knowl Data Eng, vol. 26, no. 8, pp. 1819–1837, Aug. 2014, doi: 10.1109/TKDE.2013.39.

[16] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf Process Manag, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

[17] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” in 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), IEEE, Dec. 2014, pp. 844–848. doi: 10.1109/ICARCV.2014.7064414.

[18] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognit Lett, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.