State-of-the-Art Deep Learning Strategies for Multi-Class Classification of Retinal OCT Images

Sharon G[1]
Computer Science and Engineering
KLE Technological University
Hubli, India
sharonhere777@gmail.com

Nagratna Yaligar[2]
Computer Science and Engineering
KLE Technological University
Hubli, India
nagratna.yaligar@kletech.ac.in

Konkathi Rithin Kumar[3]
Computer Science and Engineering
KLE Technological University
Hubli, India
rithinkumar.k111@gmail.com

Achyut Padaki[4]
Computer Science and Engineering
KLE Technological University
Hubli, India
achyutpadaki@gmail.com

Reemak Dawe[5]
Computer Science and Engineering
KLE Technological University
Hubli, India
dawereemak@gmail.com

Savita Katagi[6]
Computer Science and Engineering
KLE Technological University
Hubli, India
savita.katagi@kletech.ac.in

***Abstract*—This research introduces Optical Coherence Tomography (OCT), a non-invasive imaging method utilizing light rays to capture detailed cross-sectional eye structure images. Beyond retinal pathology, it explores neuro-ophthalmology, glaucoma, and uveitis. Employing deep learning, our models categorize OCT images into Choroidal Neo-Vascularization (CNV), Diabetic Macular Edema (DME), Drusen, and Normal. To handle class imbalance, various undersampling techniques are applied to medical image data. State-of-the-art deep learning models, including CNN, VGG-19, and DenseNet-121, are employed for multi-class classification. Training on both imbalanced and balanced datasets enables a comprehensive comparison of results. Ensemble learning is implemented to enhance prediction accuracy by combining trained models for improved performance. Achieving balanced accuracy scores of 92.89%, 83.78%, and 82.64% for CNN, VGG-19, and DenseNet-121, respectively. Keywords: Deep learning, Ensemble learning,Multi-class classification,DenseNet-121,VGG-19.**

I. Introduction

Retinal Optical Coherence Tomography (OCT) stands out as a revolutionary, non-invasive imaging technique in ophthalmology. By utilizing near-infrared light, it captures high-resolution cross-sectional images crucial for diagnosing conditions like glaucoma, AMD, diabetic retinopathy, CNV, and DME [1]. OCT enables early detection, precise visualization, and continuous monitoring of retinal layers, offering valuable insights into thickness, morphology, and structural changes in the retina [2]. Its role as a cornerstone in modern ophthalmic practice is further emphasized by its ability to facilitate personalized treatment strategies and monitor disease progression over time, significantly improving patient care and visual outcomes. [3]The integration of deep learning algorithms with OCT further enhances its diagnostic capabilities, underscoring its transformative potential in advancing medical diagnostics [4].

Beyond excelling in multiclass image classification, VGG-19 and DenseNet-121 demonstrate flexibility in managing varied datasets, enhancing their applicability in real-world scenarios. Moreover, their hierarchical feature extraction not only guarantees precise categorization but also aids in discerning subtle details essential for early and accurate breast cancer detection. This underscores the transformative impact of deep learning in the field of medical diagnostics.

At the outset, imbalanced datasets resulted in diminished accuracies, with CNN achieving 80.28%, VGG-19 at 65.49%, and DenseNet-121 at 68.21%. However, training on a balanced dataset significantly improved accuracy, with CNN reaching 92.59%, VGG-19 at 86.78%, and DenseNet-121 at 84.38%.

In section 2, our examination delves into prior research focused on the application of machine learning in the classification of Retinal OCT images.

II. BACKGROUND STUDY

In this section, we discuss the related works about the Classification of OCT Images using Deep Learning Technologies. Medical imaging (nuclear medicine, ultrasound, MRI, X-ray) aids diverse diagnoses (brain bleeds, fractures, vascular diseases).According to this article [5] X-ray mammography supports OCT treatment; MRI, CT, and dynamic CT vary in tumor detection. Neural networks, transfer learning, and CNN tuning show promise in OCT image classification. Patient identification, emphasized in the Privacy Policy, safeguards privacy. Medical fusion imaging enhances diagnosis through data fusion. Advanced learning, crucial in cancer detection, holds potential for complex diagnostics and predictions. The

study [6]trains a large convolutional neural network on the ImageNet LSVRC-2010 dataset, achieving top-1 and top-5 error rates of 37.5% and 17.0% . The network features five convolutional layers, three fully-connected layers, ReLUs, multiple GPU training, and local response normalization. Results showcase superior performance on various datasets, emphasizing the model's ability to recognize diverse objects. The study underscores model efficiency and suggests potential enhancements with faster GPUs for further improvements. The study [7] compares the deep-learning enhanced IDx-DR version X2.1 to the Iowa Detection Program (IDP) without deep learning for automated diabetic retinopathy (DR) detection. The deep-learning algorithm exhibits 96.8% sensitivity, 87.0% specificity, and a 99.0% negative predictive value using a consensus reference standard. Notably, its performance surpasses the non-deep learning version, emphasizing the potential of deep-learning systems to enhance efficiency in DR screening and prevent associated visual loss and blindness. In the study [9], logistic regression, support vector machines, K-nearest neighbor, random forest, and artificial neural networks (ANN) are examined for early diagnosis using the NEARMISS Dataset. Electronic equipment and DenseNet-121 both performed exceptionally well, achieving high F1 scores (CNV: 0.94, DME: 0.95, DRUSEN: 0.88, NORMAL: 0.92) and 92% accuracy. The findings underscore the importance of negative radiation in early macular edema detection, enhancing diagnosis and reducing mortality. The primary goal of this research, as stated in [10] is to compare several deep learning techniques (XGBoost, Random Forest, etc.) for OCT detection, aiming for early diagnosis. Using consistent datasets and feature analysis, they optimized model performance, with XGBoost excelling. Its F1 scores exceeding 0.95 demonstrated excellent recall, accuracy, and precision, suggesting its potential for effective OCT prediction in comparison to other models.

In section 3, we analyze the classifiers utilized in these investigations, the proposed system, characteristics of the dataset, methods for achieving balance, and ensemble techniques, specifically average ensemble and weighted average ensemble, are discussed.

III. Methodology

The approach consists of three primary components: addressing data imbalance through resampling techniques, training classification models (CNN, VGG-19, DenseNet-121) on both balanced and unbalanced data, and implementing ensemble techniques across all three models. Figure 1 shows an illustration of the proposed system.

A. Dataset Details

Retrospective OCT images (July 2013 - March 2017) from diverse institutions underwent a tiered grading system involving students, ophthalmologists, and senior specialists. Quality control at the initial tier excluded images

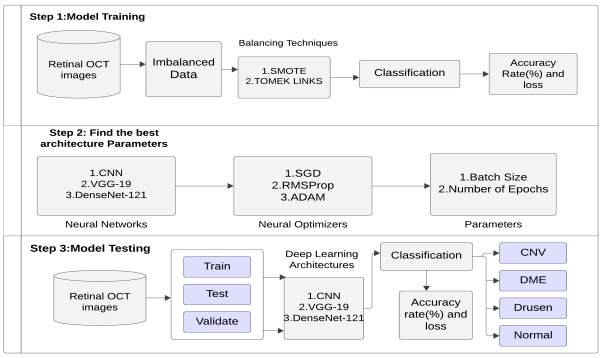

Figure 1. Proposed methodology of the system

with severe artifacts. Subsequent tiers independently assessed and confirmed labels for pathologies like choroidal neovascularization, macular edema, and drusen. A validation subset of 993 scans, graded separately, addressed potential human error. The dataset selection process is visually outlined in a CONSORT-style diagram, ensuring meticulous image annotation and minimizing grading discrepancies. The dataset is divided into five folders (train,

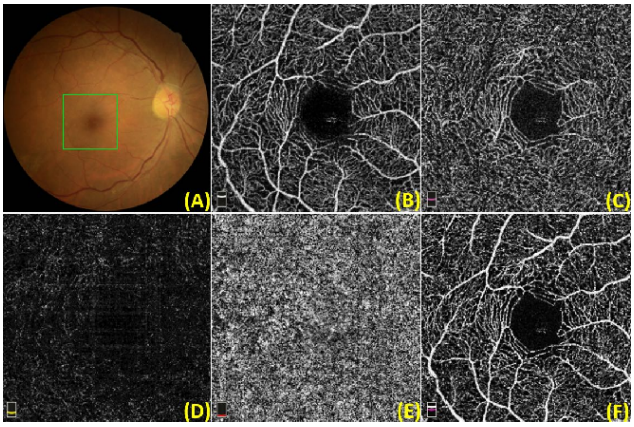

Figure 2. Retinal image with its classes

test, val) and categorized by subfolders for each image type (NORMAL, CNV, DME, DRUSEN). It consists of a total of 84,495 X-Ray images in JPEG format, spanning across four classifications (NORMAL, CNV, DME, DRUSEN).

B. Balancing

In the realm of imbalanced multi-class data, SMOTE is employed to generate synthetic samples for the minority class, mitigating bias. Simultaneously, Tomek Links are utilized to eliminate noisy majority samples, refining decision boundaries. The amalgamation of these techniques serves to balance class sizes, optimize boundaries, and elevate classifier performance across all classes. This strategic approach effectively addresses the challenges associated with precise prediction for minority classes commonly encountered in imbalanced datasets.

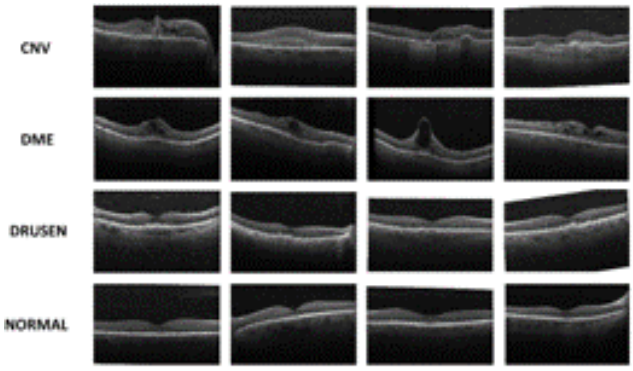

Figure 3. The four categories of OCT images for benign and malignant

C. Classifiers

In the context of multi-class classification, deep learning models like CNN, VGG-19, and DenseNet-121 typically utilize the softmax activation function in their output layers. Described by Formula 1, this function converts raw outputs into meaningful probabilities.In the equation, $P(class_i)$ denotes the probability for class i, e is Euler's number (exponential component), and z_i represents the raw score or logit for class i. The term N signifies the total number of classes, influencing the normalization process.The softmax function plays a crucial role in normalizing raw outputs, creating a probability distribution that enhances clarity in multi-class classification and guides the decision-making process of the model.

$$P(class_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{1}$$

1) CNN:Convolutional Neural Network: The CNN incorporates multiple convolutional layers (32 to 128 filters) with ReLU activation and batch normalization for feature extraction. To prevent overfitting, max-pooling (2x2 window) is applied, complemented by dropout layers for generalization. The architecture concludes with a dense layer of 128 neurons and softmax activation for eight-class multi-classification, as depicted in Figure 4. This specialized CNN model excels in image classification, extracting hierarchical representations for precise predictions across diverse classes.

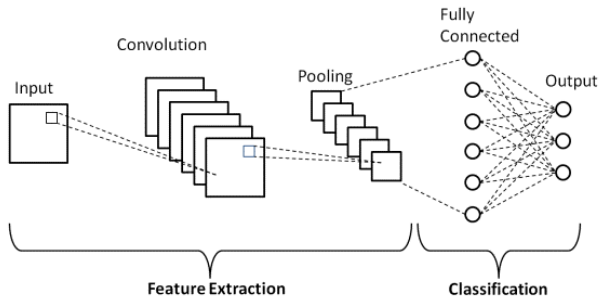

Figure 4. CNN Architecture

2) VGG-19: In VGG-19 transfer learning, pre-trained ImageNet weights are utilized, skipping initial layers for feature extraction. Custom classification layers, including Flatten, fully connected layers, ReLU activation, dropout, and batch normalization, are added. The final dense layer, employing softmax activation, produces class probabilities for eight-class classification. The model is compiled with RMSprop optimizer and categorical cross-entropy loss. See Figure 5 for an architectural overview.

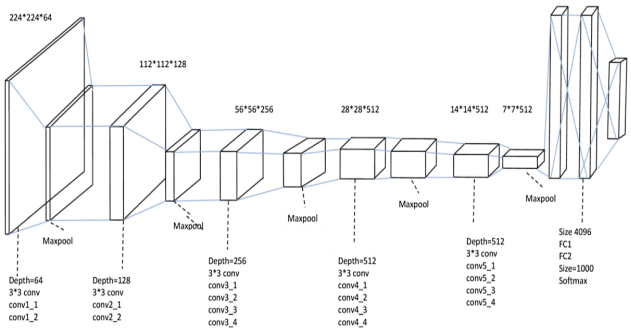

Figure 5. VGG-19 Architecture

3) DenseNet-121: DenseNet-121 employs ImageNet pre-trained transfer learning for an eight-class task, utilizing pre-trained weights, custom layers (Flatten, Dense with ReLU activation, Dropout), and batch normalization. The final Dense layer utilizes softmax activation. Compiled with Adam optimizer, a learning rate of 0.1, and categorical cross-entropy loss, this configuration maximizes accuracy. Figure 6 illustrates the architecture, combining transfer learning with customized top-layer adjustments for the classification task.

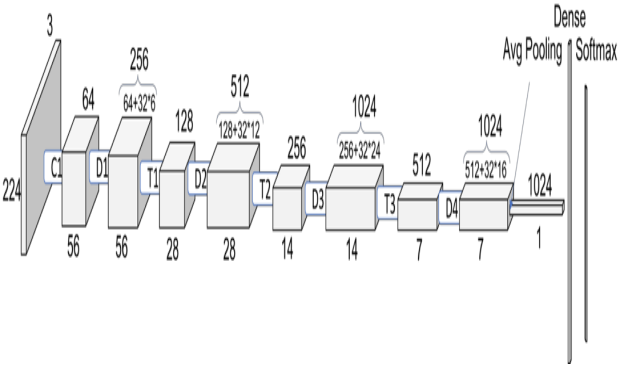

Figure 6. DenseNet-121 Architecture

D. Ensembling

Ensembling with CNN, VGG-19, and DenseNet-121 amalgamates diverse predictions utilizing average and weighted average methods, with the latter assigning weights based on individual model strengths. This nuanced approach aims to improve overall predictive performance.

1) Average Ensemble: The average ensemble merges results from several models to create a joint prediction determined by the highest sum.The ensemble prediction, denoted as $\hat{y}_{\text{avg\_ensemble}}$, is derived by finding the argument that maximizes the sum of individual predictions $\hat{y}_i$, where i ranges from 1 to N, representing the total number of individual predictions. This approach, encapsulated by Formula 2, improves accuracy by combining diverse model outputs, consolidating their contributions into a unified decision.

$$\hat{y}_{\text{avg\_ensemble}} = \arg\max\left(\sum_{i=1}^{N} \hat{y}_i\right) \tag{2}$$

2) Weighted Average Ensemble: The ensemble with weighted averages integrates predictions with specified weights: 0.6 for CNN and 0.2 each for VGG-19 and DenseNet-121.The computation, delineated by Formula 3, derives the prediction $\hat{y}_{\text{weighted\_avg\_ensemble}}$ by summing the products of individual model predictions $\hat{y}_i$ and their corresponding weights w_i, where i ranges from 1 to N, The ensemble using weighted averages, assigning a greater weight to CNN, integrates perspectives from various models, representing the overall predictions.

$$\hat{y}_{\text{weighted\_avg\_ensemble}} = \arg\max\left(\sum_{i=1}^{N} w_i \cdot \hat{y}_i\right) \tag{3}$$

In section 4, we examine the distribution of classes before and after image balancing, evaluate the effectiveness of deep learning models on both balanced and imbalanced data and assess the application of ensemble techniques to enhance prediction outcomes.

IV. Results and Discussions

A. Balancing

The dataset analysis showed a notable imbalance, especially with CNV class having a larger image volume than others (Figure 7). This may impact model effectiveness.

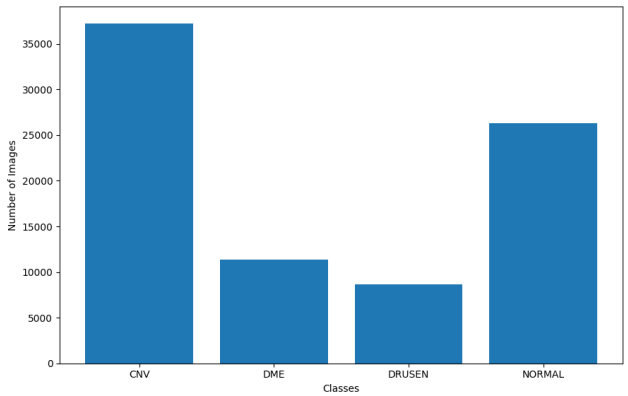

Figure 7. Distribution of different classes before balancing

SMOTE and Tomek links balanced the dataset (Figure 8), ensuring equitable class distribution for robust, unbiased learning in subsequent model training and evaluation.

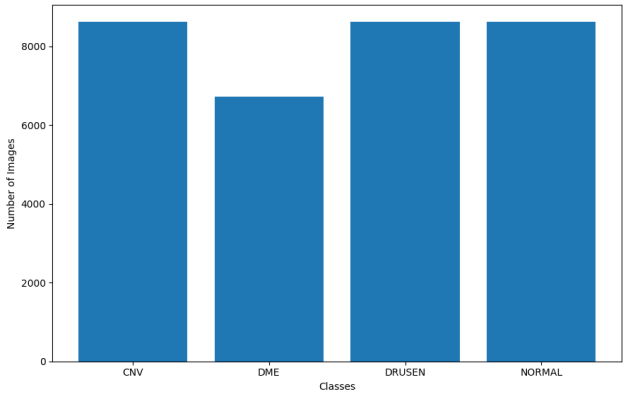

Figure 8. Distribution of different classes after balancing

B. Individual Model Performance Analysis

1) Convolutional Neural Network: :Following 50 epochs, the CNN achieved an impressive accuracy of 91.29% in both training and validation, as evidenced by the confusion matrix. This underscores its proficiency in discerning intricate patterns for reliable classifications. The well-trained CNN guarantees dependable classifications and predictions, instilling increased confidence through exceptional pattern comprehension.

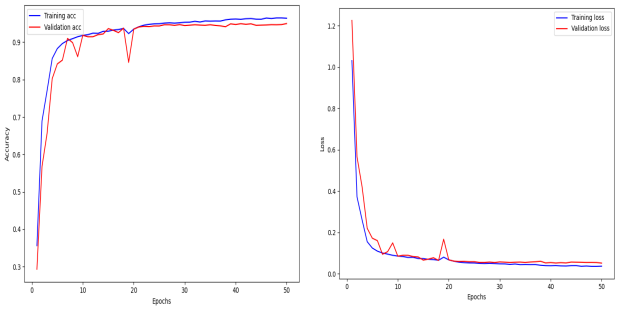

Figure 9. CNN training and validation accuracy and loss graph

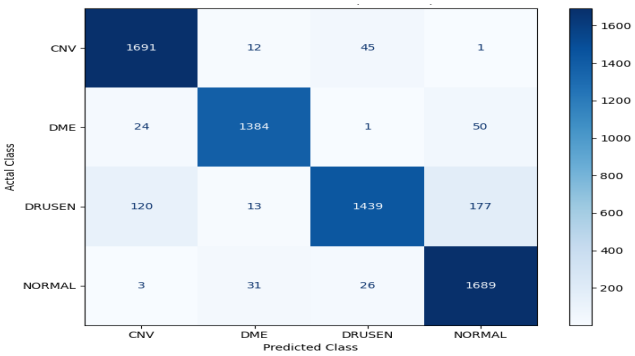

Figure 10. CNN Confusion Matrix

2) VGG-19: Following 50 training cycles, VGG-19 exhibited an impressive accuracy of 81.78%, showcasing its versatility in addressing both validation and unseen data scenarios, as indicated by the confusion matrix. This underscores its strong generalization and precision, emphasizing its dependability and resilience in effectively managing a variety of data challenges.

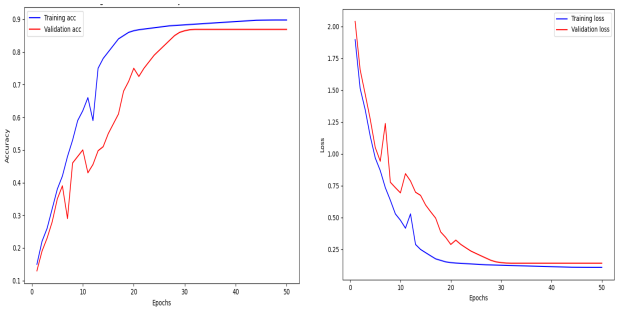

Figure 11. VGG-19 Model Training and Validation accuracy and loss graph

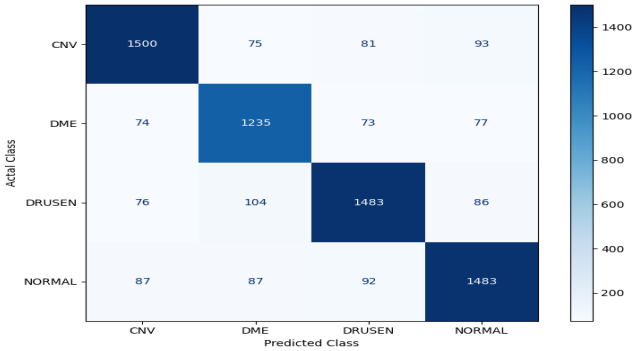

Figure 12. VGG-19 Confusion Matrix

3) DenseNet-121: After 50 training cycles, DenseNet-121 demonstrated notable accuracy at 84.38%, highlighting its adaptability in handling validation and unseen data scenarios, as evidenced by the confusion matrix. This underscores its robust generalization and precision, accentuating its reliability and resilience in effectively navigating diverse data challenges.

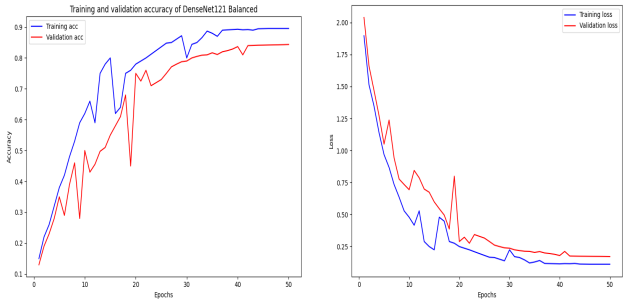

Figure 13. DenseNet-121 Model Training and Validation accuracy and loss graph

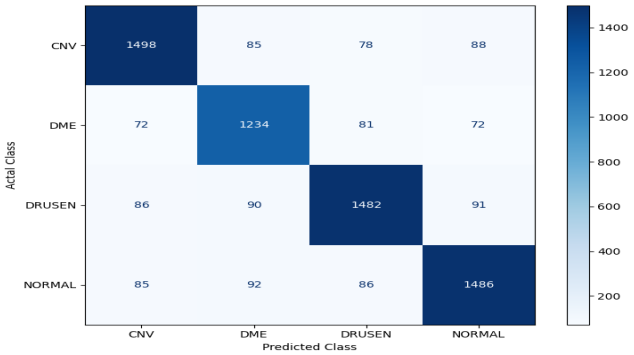

Figure 14. DenseNet-121 Confusion Matrix

C. Comparitive Analysis of all Three models:

Initially, models trained on imbalanced datasets exhibited lower accuracy (CNN 80.28%, VGG-19 65.49%, DenseNet-121 68.21%). However, after training on a balanced dataset, there was a substantial improvement in accuracy (CNN 92.59%, VGG-19 86.78%, DenseNet-121 84.38%). This shift emphasizes the pivotal role of balanced datasets and associated techniques in augmenting model performance for reliable classification. Achieving higher accuracy on balanced data highlights the significance of addressing class imbalances, ensuring models are well-equipped to handle diverse scenarios and produce more accurate and robust predictions. Figure 15 visually reinforces the impact of balanced training on the models' classification capabilities.

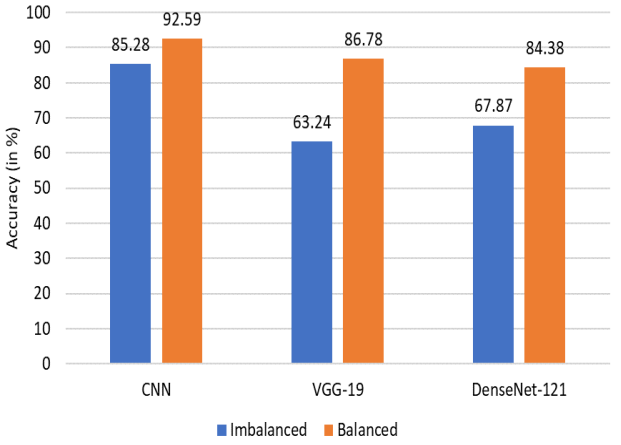

Figure 15. Accuracy of different models on imbalanced and balanced data

1) Ensembling: Ensemble techniques, applied to balanced data, resulted in an accuracy of 92.91% for the average ensemble and a slightly higher accuracy of 93.20% for the weighted average. This demonstrates the efficacy of ensemble methods in improving predictive performance. Further insights into the performance of individual models and ensembles are provided in Table I, offering a comprehensive overview of the achieved accuracies.

TABLE I
ACCURACY SCORES OF DIFFERENT MODELS

Sl. No.	Model	Accuracy Score
1	CNN	92.59 %
2	VGG-19	86.78 %
3	DenseNet-121	84.38 %
4	Average Ensemble	92.91 %
5	Weighted Average Ensemble	93.20 %

V. CONCLUSION AND FUTURE SCOPE

This study demonstrates that ensembling techniques, such as average ensemble and weighted average ensemble, outperform individual deep learning models like CNN, VGG-19, and DenseNet-121 in accuracy. Moreover, training on balanced data significantly improves accuracy compared to imbalanced data. In summary, resampling techniques are crucial for addressing class imbalance, and ensembling techniques contribute significantly to overall accuracy enhancement. Subsequent research endeavors could concentrate on enhancing ensemble methods through adaptive approaches tailored to different dataset complexities. The goal is to enhance model performance in various real-world scenarios.

REFERENCES

[1] M. D. Abramoff, M. K. Garvin and M. Sonka, "Retinal imaging and image analysis", IEEE Rev. Biomed. Eng., vol. 3, no. 1, pp. 169-208, Dec. 2010.

[2] Z. Zhang et al., "A survey on computer aided diagnosis for ocular diseases", BMC Med. Informat. Decision Making, vol. 14, no. 1, pp. 169-176, 2014.

[3] G. Quellec, K. Lee, M. Dolejsi, M. K. Garvin, M. D. Abramoff and M. Sonka, "Three-dimensional analysis of retinal layer texture: identification of fluid-filled regions in SD-OCT of the macula", IEEE Trans. Med. Imag., vol. 29, no. 6, pp. 1321-1330, Jun. 2010.

[4] X. Xu, K. Lee, L. Zhang, M. Sonka and M. D. Abràmoff, "Stratified sampling voxel classification for segmentation of intraretinal and subretinal fluid in longitudinal clinical OCT data", IEEE Trans. Med. Imag., vol. 34, no. 7, pp. 1616-1623, Jul. 2015.

[5] G. R. Wilkins, O. M. Houghton and A. L. Oldenburg, "Automated segmentation of intraretinal cystoid fluid in optical coherence tomography", IEEE Trans. Biomed. Eng., vol. 59, no. 4, pp. 1109-1114, Apr. 2012.

[6] K. Alsaih et al., "Classification of SD-OCT volumes with multi pyramids LBP and HOG descriptors: Application to DME detections", Proc. IEEE 38th Annu. Int. Conf. Eng. Med. Biol. Soc., pp. 1344-1347, 2016.

[7] S. Farsiu et al., "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography", Ophthalmology, vol. 121, no. 1, pp. 162-172, 2014.

[8] S. Farsiu et al., "Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography", Ophthalmology, vol. 121, no. 1, pp. 162-172, 2014.

[9] Y. Y. Liu, M. Chen, H. Ishikawa, G. Wollstein, J. S. Schuman and J. M. Rehg, "Automated macular pathology diagnosis in retinal OCT images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding", Med. Image Anal., vol. 15, no. 5, pp. 748-759, 2011.

[10] Z. H. Zhou, Ensemble Methods: Foundations and Algorithms, New York, NY, USA:Taylor and Francis, pp. 1-20, 2012.

[11] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs", JAMA, vol. 316, no. 22, pp. 2402-2410, 2016.

[12] C. S. Lee, D. M. Baughman and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration optical coherence tomography images", Ophthalmol. Retina, vol. 124, no. 8, pp. 1090-1095, 2017.

[13] P. Burlina, D. E. Freund, N. Joshi, Y. Wolfson and N. M. Bressler, "Detection of age-related macular degeneration via deep learning", Proc. IEEE Int. Symp. Biomed. Imag., pp. 184-188, 2016.

[14] S. P. K. Karri, D. Chakraborty and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration", Biomed. Opt. Express, vol. 8, no. 2, pp. 579-592, 2017.

[15] . M. D. Abramoff et al., "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning", Investigative Ophthalmol. Visual Sci., vol. 57, no. 13, pp. 5200-5206, 2016.

[16] N. Patton et al., "Retinal image analysis: Concepts applications and potential", Progress Retinal Eye Res., vol. 25, no. 1, pp. 99-127, 2006.

[17] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[18] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Proc. Adv. Neural Inf. Process. Syst., pp. 1097-1105, 2012.

[19] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 1, pp. 142-158, Jan. 2016.

[20] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3431-3440, 2015.