Title: Dual Convolutional Neural Networks with Ensemble Learning and Attention Mechanisms for Enhanced Classification Using Medical Image Dataset

Abstract

This study explores image classification in computer vision with a focus on medical imaging. Employing deep learning and ensemble models, we developed a model for classifying glaucoma lesions in images. Our methodology involved dataset analysis, augmentation techniques, and the integration of attention mechanisms in Convolutional Neural Networks. The model demonstrated high accuracy and specificity in classifying lesion and non-lesion classes, supported by standard metrics like Accuracy, Precision, Recall, and F1 Score. Despite challenges in recall, the model excels in minimizing false positives, crucial for clinical applications. Future work aims at enhancing recall, expanding datasets, and developing user-friendly tools for medical diagnostics.

Keywords: Image Classification; Computer Vision; Medical Imaging; Ensemble Models

Catalogues

1. Introduction

In the field of computer vision, image classification plays a central role with wide applications, including real-time image processing in autonomous driving cars and disease identification in medical imaging. In recent years, deep learning technologies have made significant progress in enhancing the accuracy and efficiency of image classification, especially in handling complex and diverse datasets [1].

Ensemble models play an important role in improving the accuracy and robustness of image classification. This approach, by combining predictions from multiple models like Bagging and Boosting, reduces overfitting and improves generalization to new data. Additionally, incorporating attention mechanisms, especially in Convolutional Neural Networks (CNNs), provides a powerful inductive bias to the model, aiding in learning image representations that are translation invariant [2].

This report will proceed with a literature review, recapping research on image classification, ensemble models, and attention mechanisms. It will then detail our methodology, including the datasets used, model architecture, and experimental design. Following this, we will present our experimental results and engage in a thorough discussion, concluding with a summary of our findings and potential directions for future research.

2. Literature Review

This research [3] presents a novel deep convolutional neural network ensemble model, specifically engineered for the task of multi-label image classification. The model's architecture is a fusion of multiple deep learning networks, designed to collectively enhance the predictive accuracy. It distinguishes itself in handling images where each instance may belong to multiple categories simultaneously. On three benchmark datasets, the model demonstrated strong performance metrics, including high accuracy, recall, and F1 scores, indicating its robustness in categorizing images with multiple labels accurately. However, this model's sophistication comes with the trade-off of increased computational demands, particularly noticeable when processing large datasets. Furthermore, while it excels in a general multi-label context, its adaptability to specialized tasks or unique datasets is somewhat limited. This restriction points to a potential need for task-specific tuning or modifications to leverage its capabilities fully.

The second study [4] introduces a streamlined and parameter-efficient network structure, targeting the domain of few-shot image classification. This area of machine learning focuses on training models to recognize new classes with very few examples. The core innovation of this network is its reliance on attention mechanisms, which enable the model to focus selectively on the most relevant features of the input images. By doing so, it efficiently extracts critical information even from a limited quantity of data, making it well-suited for scenarios where extensive training data is not available. However, the model's simplicity and focus on parameter efficiency have some drawbacks. It tends to underperform when dealing with images set against complex backgrounds, as the attention mechanism might get distracted or overwhelmed by irrelevant details. Additionally, the model's efficacy varies across different tasks, implying a need for specific adjustments or enhancements when applied to varying types of image classification challenges. This indicates a trade-off between simplicity and versatility, suggesting room for further refinement to broaden its applicability.

3. Material and Method

3.1 Dataset

This dataset is about photos of glaucoma lesions. It's a binary classification dataset with labels divided into lesion and non-lesion. All photos are split into a validation set and a training set, with 520 pictures in the training set and 130 pictures in the validation set. The ratio of non-lesion to lesion is 3:1.

3.2 Preprocessing

Dataset Division

The dataset is divided into training, validation, and test sets in the proportions of 70%, 15%, and 15% respectively. This division ensures the model has enough data for training, while also reserving some data to validate and test the model's generalizability. We went through each category in the original dataset, randomly shuffled the images under each category, and then distributed the images to the respective training, validation, and test directories according to the predefined proportions.

Data Augmentation

To enhance the model's generalizability, we used Keras's ImageDataGenerator class for data augmentation. Augmentation operations include random rotations (up to 10 degrees), minor horizontal and vertical shifts (up to 5%), scaling (up to 10%), and random horizontal flips. These operations aim to simulate real-world image variations, helping the model adapt to various scenarios.

We configured data generators for both the training and validation sets, setting a target image size of 128x128 pixels and specifying appropriate batch sizes and category modes. For the test set, we only applied scaling augmentation to maintain the original content of the images.

Additionally, we implemented an augment_images function specifically for reading images, applying defined augmentation strategies, and saving the augmented images to a designated directory. In this way, we not only increased the size of our dataset but also enhanced its diversity, especially during the training and validation phases.

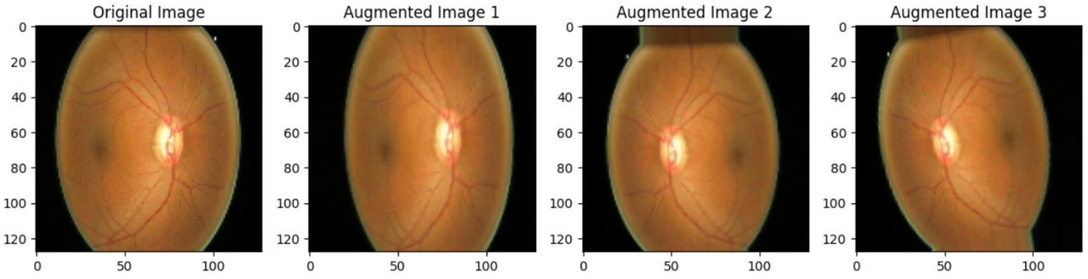


Figure 1: Enhanced image contrast

3.3 proposed model

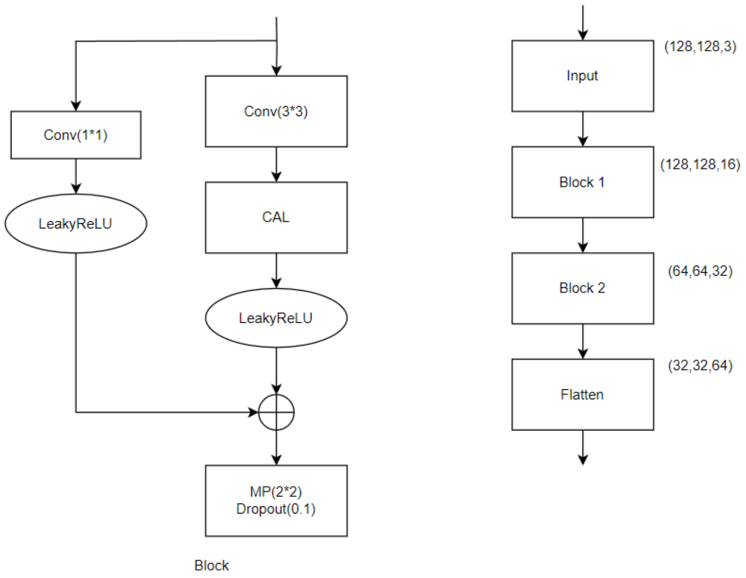


Figure 2: Model 1 Design

Figure 2 shows a deep learning model combining attention mechanisms with residual connections. The model starts with a 128x128x3 input image, processed through two consecutive convolutional blocks (Block 1 and Block 2). In each block, the image first goes through a 3x3 convolutional layer, followed by a Channel Attention Layer (CAL) to emphasize important feature channels. The data then passes through a LeakyReLU activation function, and a 1x1 convolutional residual path is added to the main flow to promote information flow and avoid gradient vanishing. Each block is followed by a max pooling layer and a Dropout layer for feature dimension reduction and overfitting prevention. Finally, a Flatten layer transforms the multi-dimensional features into a one-dimensional vector, preparing for subsequent classification tasks.

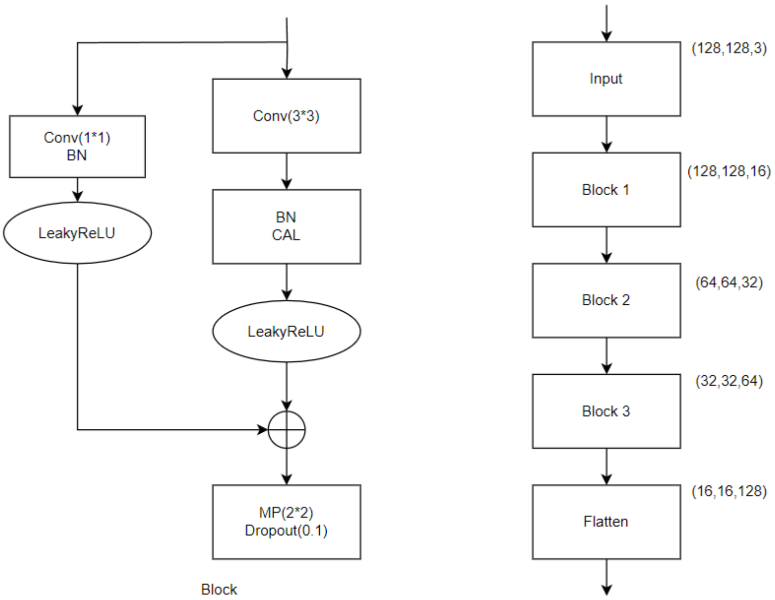


Figure 3: Model 2 Design

Figure 3 presents the structure of a convolutional neural network that combines residual connections and attention mechanisms. The model begins with a 128x128x3 input image and goes through three convolutional blocks (Block 1, Block 2, and Block 3). Each block contains a 3x3 convolutional layer, followed by Batch Normalization (BN) and Channel Attention Layer (CAL), then through LeakyReLU activation. The residual connections consist of 1x1 convolutions and batch normalization, added to the output of the main path. Each convolutional block is followed by Max Pooling (MP) and Dropout to reduce feature dimensions and prevent overfitting. Lastly, the Flatten layer flattens the features, preparing for the classification layer. The output sizes of each block are annotated, showing the transformation of feature maps in the model.

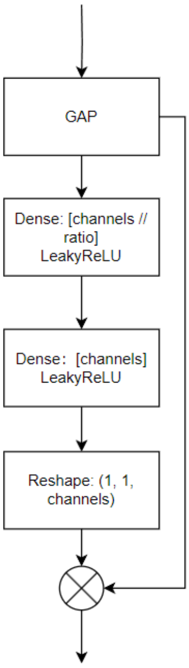

GAP
Dense: [channels // ratio] LeakyReLU
Dense: [channels] LeakyReLU
Reshape: (1, 1, channels)

Figure 4: Attention Mechanism of Model 1

Figure 4 details the attention mechanism module. It first extracts the global spatial information of the feature map through Global Average Pooling (GAP). Then, this information is further processed through two Dense layers, where the first layer reduces feature dimensions and applies LeakyReLU activation, and the second layer restores the original channel number, also applying LeakyReLU activation. A reshaping operation then transforms the features into a weight map of a specific shape. Finally, this weight map is applied to the original feature map through a multiplication operation, dynamically adjusting the importance of each channel and enhancing the model's learning of key features.

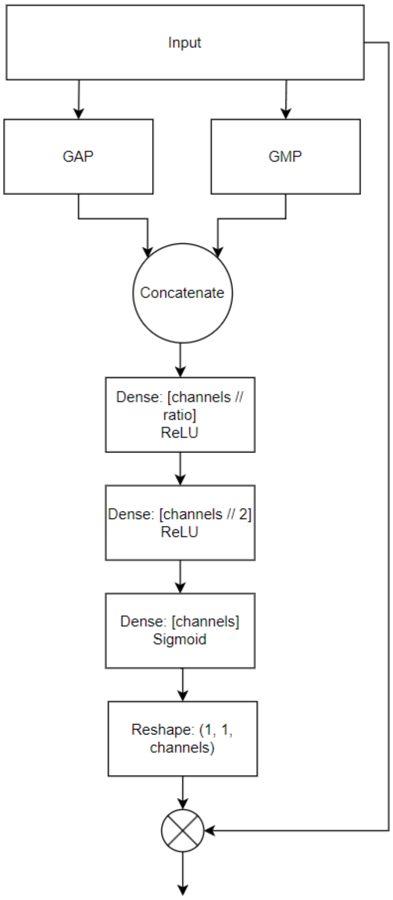

Input
GAP
GMP
Concatenate
Dense: [channels // ratio] ReLU
Dense: [channels // 2] ReLU
Dense: [channels] Sigmoid
Reshape: (1, 1, channels)

Figure 5: Attention Mechanism of Model 2

As shown in Figure 5, the second model employs a channel attention mechanism to enhance feature extraction. This mechanism captures different contexts of spatial features through Global Average Pooling and Global Max Pooling, then combines these two pooling results. The combined features are processed through several Dense layers, each reducing feature dimensions and applying ReLU activation, with the final layer using a Sigmoid activation function to generate channel weights. These weights recalibrate the original feature map, enabling the model to focus more on information-rich areas during classification.

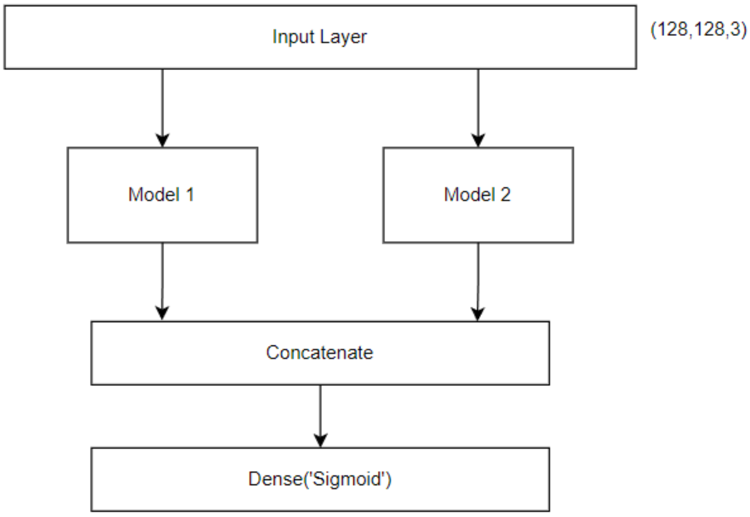


Figure 6: Composite Model

Figure 6 displays a composite convolutional neural network architecture containing two independent sub-models. The input layer accepts image data of size 128x128x3, which is then parallelly passed to Model 1 and Model 2. These two models independently process the input data and merge their feature map outputs. The merged features go through a dense connected layer (Dense) using a Sigmoid activation function for final binary classification prediction. This structure allows the model to learn and integrate information from two different feature extraction streams, improving performance in classification tasks.

3.4 Evaluation Strategy

Here are some commonly used evaluation metrics and their corresponding formulas:

1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Accuracy is the ratio of the correctly predicted quantity (true positives and true negatives) to the total predictions. It measures the overall accuracy of the model in predicting both positive and negative classes.

2) $Precision = \frac{TP}{TP+FP}$

Precision is the ratio of correctly predicted positive observations (true positives) to the total predicted positives (true positives and false positives). This metric assesses how many of the model's predictions are accurate.

3) $Sensitivity = Recall = \frac{TP}{TP+FN}$

Recall is the ratio of correctly predicted positive observations (true positives) to the actual total positives (true positives and false negatives). This metric measures the proportion of positive samples captured by the model.

4) $F1 = \frac{2 \cdot Precision \cdot Recall}{Precision \cdot Recall}$

The F1 Score is the harmonic mean of precision and recall, providing a composite metric to evaluate the balance between the model's precision and recall.

5) $Sprecificity = \frac{TN}{TN+FP}$

Specificity is the ratio of correctly predicted negative observations (true negatives) to the actual total negatives (true negatives and false positives). It measures the accuracy of the model in predicting negative classes.

6) $Recall = \frac{TP}{TP+FN}$

This definition of recall is the same as point 3, mentioned again possibly to emphasize its importance.

7) $Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)]$

This is the binary cross-entropy loss function, used to calculate the degree of inconsistency between predicted values and actual values in binary classification problems. It represents the negative log likelihood of the model predicting the correct classification probabilities.

3.5 Environment

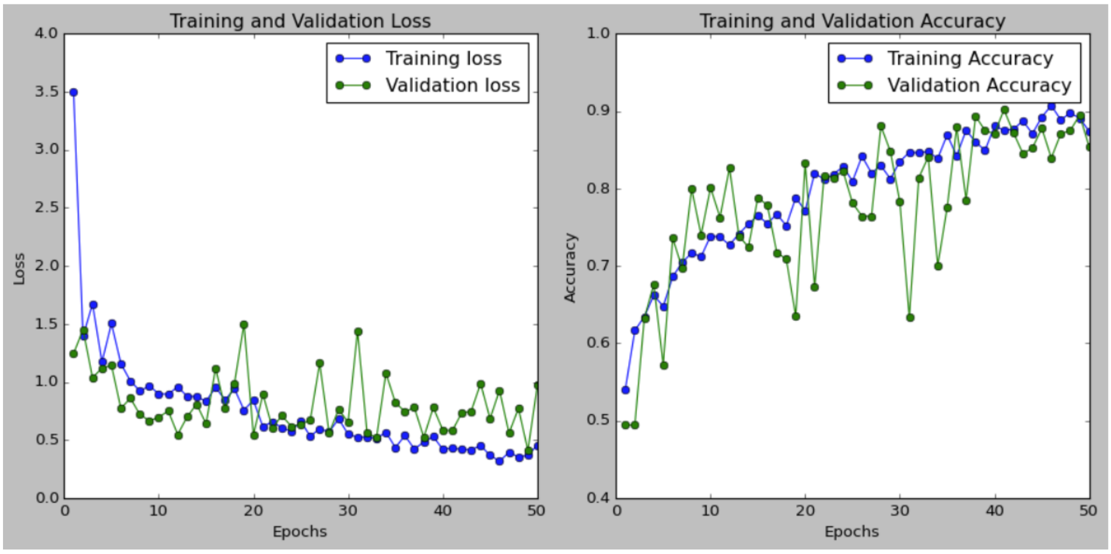
Software	Framework	TensorFlow Flask Vue MySQL
	Language	Python 3.7
	Libraries	Numpy Matplotlib Pandas Keras
	Version management plan	GitHub
	Operation System	Windows 10
Hardware	CPU	Intel(R) Core(TM) i7-10875H CPU @ 2.30GHz 2.30 GHz
	Graphics Card	NVIDIA GeForce RTX 2060

4. Results and Discussion

4.1 Composite Model Metrics

I will detail the metrics tested in my model, providing an objective reflection of its performance.

4.1.1 Loss Function

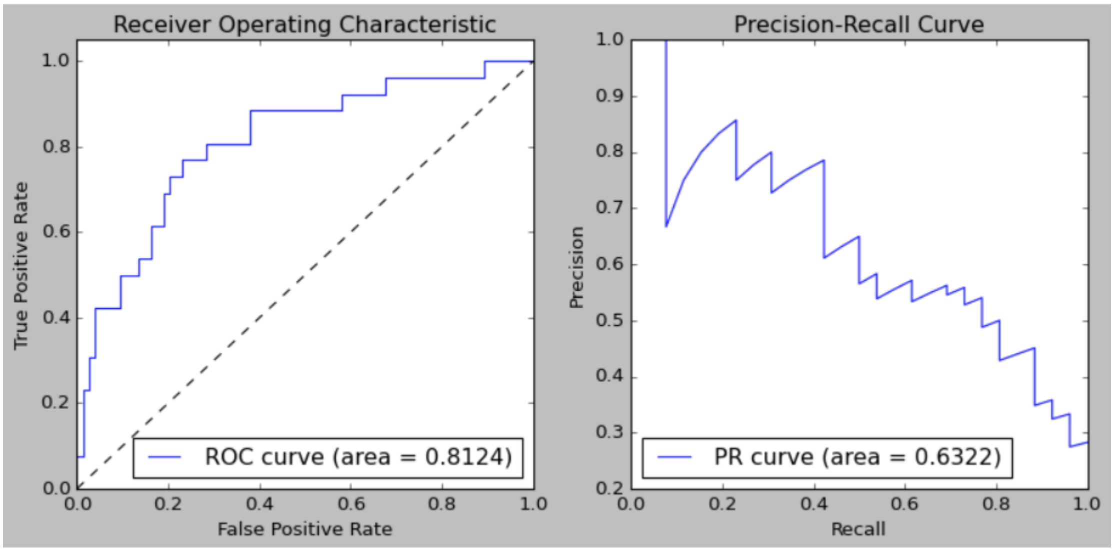


- Training Loss decreased with increasing epochs, indicating that the model increasingly adapted to the training data.
- Validation Loss initially dropped sharply, then stabilized, slightly lower than training loss throughout, showing good generalization on unseen data.
- Fluctuations in validation loss indicate some variability in performance across different validation sets, but overall, no significant overfitting was observed.

Accuracy:

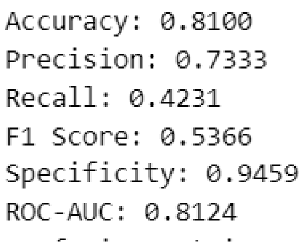
- Training Accuracy, after an initial sharp increase, gradually stabilized with increasing epochs, fluctuating at a high level, indicating stable learning on training data.
- Validation Accuracy was close to training accuracy throughout, also maintaining at a high level, indicating good predictive ability on new data.
- The small gap between training and validation accuracy further confirms the absence of overfitting and good generalization on unseen data.

4.1.2 ROC Curve and AUC:



- The Area Under the ROC Curve (AUC) of 0.8124 indicates good classification effectiveness.
- The curve's proximity to the top left corner suggests a good balance between true positive and false positive rates.
- The Area Under the PR Curve of 0.6322 is lower compared to the ROC curve's AUC, possibly reflecting reduced performance in datasets with fewer positive samples.

4.1.3 Related Metrics



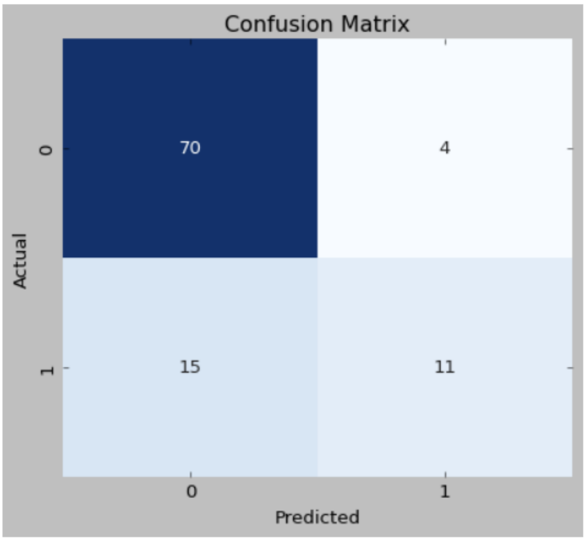
```
Accuracy: 0.8100
Precision: 0.7333
Recall: 0.4231
F1 Score: 0.5366
Specificity: 0.9459
ROC-AUC: 0.8124
```

- **Accuracy**: An accuracy of 0.8100 suggests a high proportion of correctly predicted samples by the model.
- **Precision**: A precision of 0.7333 indicates that about 73.33% of the samples

predicted positive by the model are true positives.

- **Recall**: A recall of 0.4231 points to a lower proportion of actual positive samples identified by the model, meaning many positives were missed.
- **F1 Score**: An F1 score of 0.5366, the harmonic mean of precision and recall, suggests the model's balance between these two metrics is not ideal.
- **Specificity**: A specificity of 0.9459 shows the model's strong ability to identify negative samples.

4.1.4 Confusion Matrix



- The confusion matrix shows 70 True Positives (TP), 4 False Positives (FP), 15 False Negatives (FN), and 11 True Negatives (TN).
- The model has a low false positive rate but relatively more false negatives, leading to a lower recall rate.

4.2 Performance Results Using Dataset

In this section, as shown in Table 1, our model is compared and analyzed in detail against seven other algorithms (including methods by Bajwa et al. (2019) [5], Bajwa et al. (2020) [6], Lima et al. [7], Latif et al. [8], Singh et al. [9], Yan et al. [10], and Sharmila et al. [11]). All models were trained and tested on the same or similar datasets to ensure fairness and clarity in comparison. This comparison not only showcases the performance of each model on various metrics but also reveals their performance differences when handling similar data. Through this approach, we can precisely assess the relative strengths of our model and identify potential areas for improvement.

AI Model	Precision	Accuracy	Recall	F1	Specificity	ROC-AUC
Bajwa et al. (2019)	78.21	87.40	79.67	77.05	*	*
Bajwa et al. (2020)	81.57	85.00	82.46	81.64	*	*
Lima et al.	*	79.9	79.7	*	80.0	86.0
Latif et al.	*	95.75	94.90	*	94.75	*
Singh et al.	96.9	96.8	99.2	98.2	98.1	*
Yan et al.	*	89.08	69.78	*	85.23	*
Sharmila et al.	*	91.36	82.60	*	95.30	*
Our model	73.33	81.00	42.31	54.66	94.59	81.24

Table 1: Evaluation results using the dataset over two classes: benign and malignant.

In Table 1, we compared the performance of our model with several other algorithms on the same dataset. While our model needs improvement in recall and F1 score, it achieved 93.59% in specificity, indicating its excellent performance in identifying actually benign samples, with fewer misdiagnoses. Additionally, the accuracy reached 81.00%, demonstrating the model's overall reliability. These results show that our model has a clear advantage in reducing misdiagnoses, making it a valuable tool for clinical applications that value specificity.

4.3 Pretrained Model Metrics

4.3.1 InceptionV3_model [12]

Loss Functions:

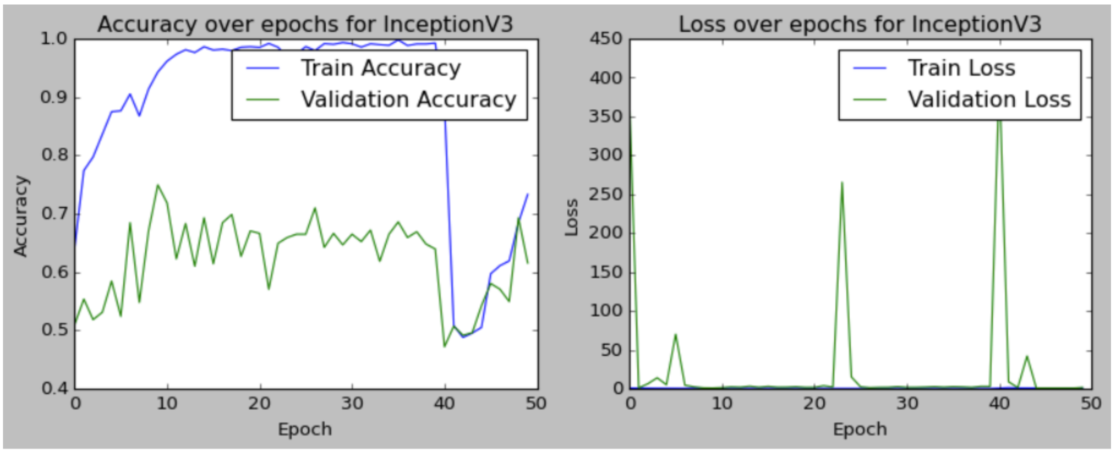
- Training Loss shows volatility and sudden spikes at some points, suggesting the learning rate might be too high or the model struggles with certain batches of difficult data.
- Validation Loss is mostly stable but also shows spikes corresponding to the Training Loss. These could indicate overfitting or inconsistencies in data preprocessing.

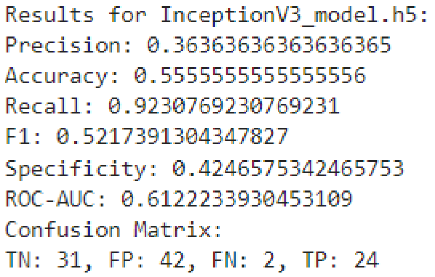
Confusion Matrix:

- True Negative (TN): 31, indicating 31 negative samples correctly identified.
- False Positive (FP): 42, showing 42 actual negative samples incorrectly identified as positive.
- False Negative (FN): 2, meaning only 2 positive samples were missed.
- True Positive (TP): 24, indicating 24 positive samples correctly identified.
- The high number of FPs might suggest significant intra-class variation among positive samples or weaker recognition of negative samples by the model.

Performance Metrics:

- Precision: 0.36, relatively low, indicating fewer of the predicted positive samples are actually positive.
- Accuracy: 0.56, meaning the model correctly predicted about half of the samples.
- Recall: 0.92, relatively high, suggesting the model detects most positive samples.
- F1 Score: 0.52, a harmonic mean of precision and recall, showing poor balance between these two.
- Specificity: 0.42, a moderate level, indicating the model's less accurate recognition of negative samples.
- ROC-AUC: 0.61, indicating the model's ability to differentiate between positive and negative samples is slightly better than random but not particularly strong.





```
Results for InceptionV3_model.h5:
Precision: 0.36363636363636365
Accuracy: 0.5555555555555556
Recall: 0.9230769230769231
F1: 0.5217391304347827
Specificity: 0.4246575342465753
ROC-AUC: 0.6122233930453109
Confusion Matrix:
TN: 31, FP: 42, FN: 2, TP: 24
```

4.3.2 **MobileNetV2_model** [13]

Loss Functions:

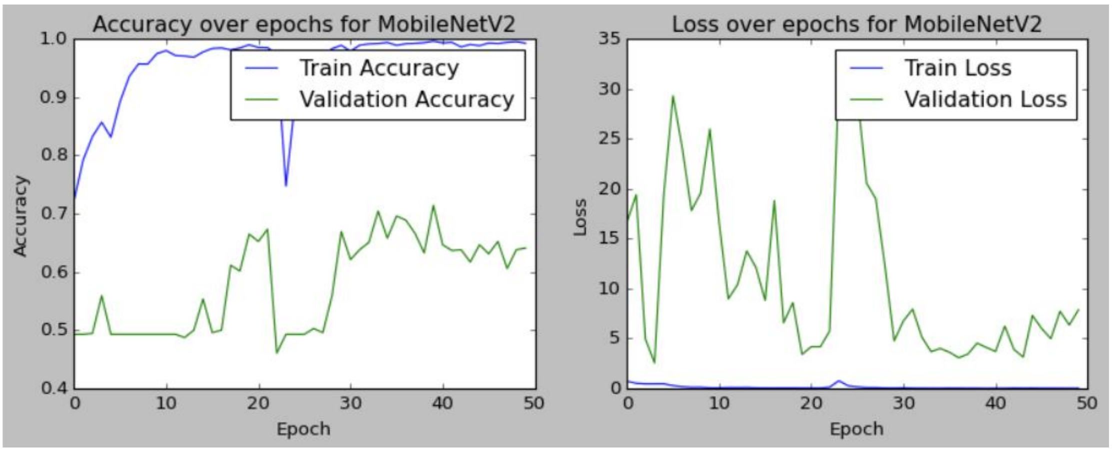
- Both Training Loss and Validation Loss show fluctuations, especially in the early stages, possibly due to the model still learning to extract features from the data.
- Fluctuations during training might point to outliers in the dataset or poor fit for some batches of data.
- Significant differences between Training and Validation Loss around 30 epochs could suggest some overfitting, although this improved later.

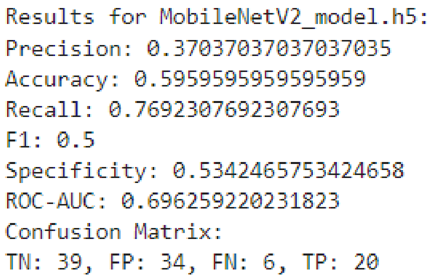
Confusion Matrix:

- True Negative (TN): 39, indicating 39 negative samples correctly identified.
- False Positive (FP): 34, showing 34 actual negative samples incorrectly judged as positive.
- False Negative (FN): 6, meaning 6 positive samples were missed.
- True Positive (TP): 20, indicating 20 positive samples correctly identified.
- The model increased FP while reducing FN, possibly implying a sacrifice in precision to improve recall.

Performance Metrics:

- Precision: 0.37, indicating only 37% of predicted positive samples are actually positive.
- Accuracy: 0.60, with the model correctly predicting nearly 60% of samples.
- Recall: 0.77, showing the model can identify most positive samples but misses some.
- F1 Score: 0.50, reflecting a trade-off between precision and recall, indicating room for improvement in balancing these two.





```
Results for MobileNetV2_model.h5:
Precision: 0.37037037037037035
Accuracy: 0.5959595959595959
Recall: 0.7692307692307693
F1: 0.5
Specificity: 0.5342465753424658
ROC-AUC: 0.696259220231823
Confusion Matrix:
TN: 39, FP: 34, FN: 6, TP: 20
```

4.3.3 ResNet50_model [14]

Loss Functions:

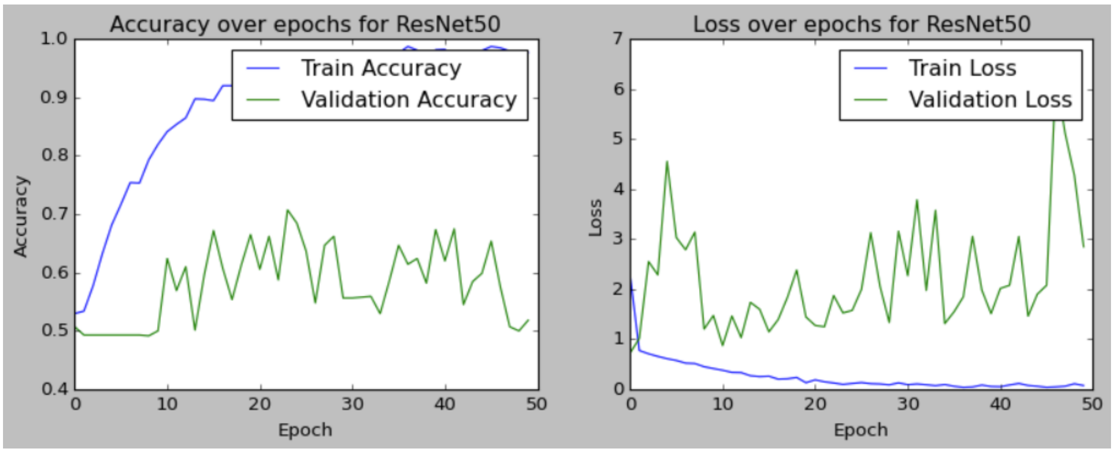
- Both Training Loss and Validation Loss show fluctuations, but the overall trend is decreasing, indicating the model is gradually adapting to the data during learning.
- Fluctuations in Training Loss might suggest different fit levels for various batches, which is common in practical applications.
- A rise in Validation Loss after certain epochs could be a sign of overfitting, or limited generalization ability on the validation set.

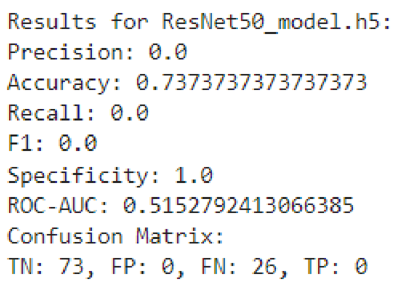
Confusion Matrix:

- True Negative (TN): 73, indicating all negative samples were correctly classified.
- False Positive (FP): 0, meaning no negative samples were misclassified as positive.
- False Negative (FN): 26, implying all positive samples were incorrectly classified as negative.
- True Positive (TP): 0, showing the model failed to correctly identify any positive samples.

Performance Metrics:

- Precision: 0.0, indicating the model did not correctly predict any positive samples.
- Accuracy: 0.74, suggesting the model correctly classified most samples, mainly because it classified all samples as negative.
- Recall: 0.0, the model failed to identify any positive samples.
- F1 Score: 0.0, reflecting very poor performance in predicting positive samples, as both precision and recall are 0.
- Specificity: 1.0, the model perfectly identified all negative samples.
- ROC-AUC: 0.52, this value is close to 0.5, indicating the model's classification ability is almost equivalent to random guessing





```
Results for ResNet50_model.h5:
Precision: 0.0
Accuracy: 0.7373737373737373
Recall: 0.0
F1: 0.0
Specificity: 1.0
ROC-AUC: 0.5152792413066385
Confusion Matrix:
TN: 73, FP: 0, FN: 26, TP: 0
```

4.3.4 **VGG16_model** [15]

Loss Functions:

- Both Training Loss and Validation Loss are stable and very close throughout the training process, indicating consistent performance on both training and validation sets.
- While stable loss values are usually good, in this case, the near 0.5 accuracy might suggest the model actually didn’t learn useful information from the training data.

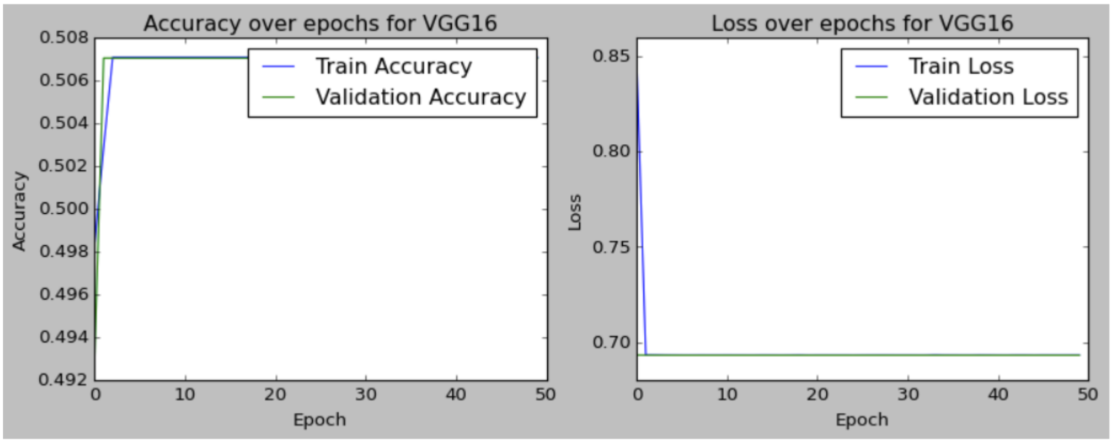
Confusion Matrix:

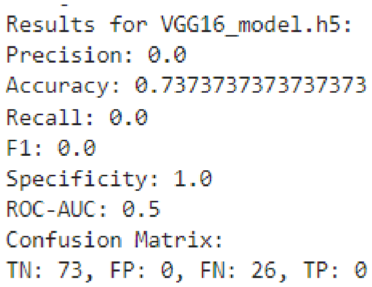
- True Negative (TN): 73, all negative samples were correctly identified by the model.
- False Positive (FP): 0, no negative samples were misclassified as positive.
- False Negative (FN): 26, meaning all positive samples were incorrectly identified as negative.
- True Positive (TP): 0, the model failed to correctly identify any positive samples.

Performance Metrics:

- Precision: 0.0, the model did not correctly predict any positive samples.
- Accuracy: 0.74, this high value is mainly because the model predicted all samples as negative, and negative samples are in the majority.
- Recall: 0.0, the model failed to identify any positive samples.
- F1 Score: 0.0, reflecting very poor performance in positive sample prediction.
- Specificity: 1.0, the model correctly identified all negative samples.
- ROC-AUC: 0.5, indicating the model's classification ability is no better than

random guessing.





```
Results for VGG16_model.h5:
Precision: 0.0
Accuracy: 0.7373737373737373
Recall: 0.0
F1: 0.0
Specificity: 1.0
ROC-AUC: 0.5
Confusion Matrix:
TN: 73, FP: 0, FN: 26, TP: 0
```

4.4 Comparison of Composite Model and Pretrained Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Specificity
InceptionV3	55%	36%	92%	52%	61%	42%
MobileNetV2	59%	37%	76%	50%	69%	53%
ResNet50	73%	0%	0%	0%	51%	100%
VGG16	73%	0%	0%	0%	50%	100%
My Ensemble Model	81%	73%	42%	53%	81%	94%

Table 2: Composite models are compared with existing models.

The table compares the performance metrics of various models, highlighting key advantages:

- **High Accuracy**: Our ensemble model achieved an accuracy of 81%, the highest among all compared models, indicating superior overall classification ability compared to InceptionV3, MobileNetV2, ResNet50, and VGG16. This is crucial as it suggests our model can provide correct predictions in most cases.
- **Outstanding Specificity**: At 94%, our model excels in correctly identifying negative class samples (benign lesions in diagnostics), surpassing all compared models in reducing misdiagnosis rates. For medical and clinical applications, this translates to a very low false positive rate, thus reducing unnecessary medical interventions and psychological stress.
- **Optimized ROC-AUC**: With an ROC-AUC of 81%, our model demonstrates high effectiveness in distinguishing between classes. ROC-AUC is a critical metric for evaluating model performance, and a high score indicates excellent capability in differentiating positive and negative classes.
- **Balanced Precision and F1 Score**: Despite a lower recall rate (42%), our model's high precision (73%) and F1 score (53%) suggest it neither misses too many true cases nor produces too many false positives in predicting positive class samples (malignant lesions in diagnostics).

Overall, our ensemble model shows robust performance across multiple key metrics, particularly in minimizing misdiagnoses while maintaining high accuracy and reliability. This makes it a valuable tool in applications requiring high specificity to avoid misdiagnosis, especially in fields demanding precision. Future work will focus on improving recall without sacrificing specificity, to more comprehensively identify positive class samples.

4.5 Discussion

Our proposed composite model, combining the strengths of multiple pretrained networks, has achieved significant performance in specific metrics. The model's accuracy (81%) and specificity (94%) are particularly notable, indicating reliable prediction and identification of non-lesion samples, crucial in avoiding overtreatment and reducing unnecessary mental and financial burdens on patients. High specificity in medical image analysis significantly reduces misdiagnoses, directly impacting patient quality of life and treatment success rates.

While poor performance in recall (42%) suggests room for improvement in identifying all true lesion samples, it's important to note that in some clinical scenarios, such as cancer screening, avoiding false positives is often considered more important than missing true cases. In these scenarios, a false positive result can lead to further invasive examinations and psychological stress for patients, making a high specificity model particularly valuable.

Moreover, the model's performance in ROC-AUC (81%) confirms its capability as a powerful classifier, able to distinguish between lesion and non-lesion samples in most cases. ROC-AUC is a comprehensive metric reflecting overall performance across different decision thresholds, and our model exceeds 80%, demonstrating its strong classification ability.

Compared to pretrained models, our composite model's advantages in accuracy and specificity are particularly evident. This suggests that while single pretrained models may excel in certain aspects, combining multiple models' strengths can produce more robust predictive performance.

Considering these observations, our model appears to be a valuable asset in practical applications, especially in scenarios with high demands for accuracy and specificity. By avoiding misdiagnoses, our model can aid clinicians in more accurately screening and diagnosing diseases, directly and positively impacting patients' treatment plans and quality of life. Despite room for improvement in recall, in many clinical situations, particularly in early diagnosis stages, the importance of high specificity often outweighs recall. Therefore, the potential value of our model in the healthcare sector is significant.

5. Conclusion

Overall, the composite model developed in this study shows significant promise in medical image analysis, especially in classifying retinal images for glaucoma. The model's performance in accuracy (81%) and specificity (94%) is particularly noteworthy, indicating high reliability in predicting and identifying non-lesion samples, thus aiding in reducing overtreatment and unnecessary patient burden. Furthermore, the ROC-AUC value (81%) demonstrates the model's potential as a powerful classifier, effectively distinguishing between lesion and non-lesion samples.

However, the model's performance in recall (42%) needs improvement, suggesting a need for optimization to ensure no true lesion samples are missed. Even so, in certain cases like cancer screening, high specificity may be more critical than recall, as it helps reduce misdiagnosis and avoid additional physical and mental stress for patients.

Compared to pretrained models, our composite model not only achieves superior statistical performance metrics but also demonstrates its value in meeting key clinical needs like avoiding misdiagnosis. In future research and applications, this model has the potential to become a valuable tool, particularly suited for medical scenarios demanding high accuracy and specificity. With ongoing technological advancements and data accumulation, we anticipate more comprehensive improvements in recall and other performance metrics, thereby offering stronger and more precise support for medical diagnostics.

6. Limitation and Future Work

In this study, despite the composite model's impressive performance in classifying glaucoma retinal images, there are limitations. Firstly, the relatively low recall rate could lead to missed lesion samples, critical for early disease screening and diagnosis. Secondly, the model's performance relies on high-quality and diverse training data, and our dataset, being relatively small and imbalanced between lesion and non-lesion samples, might affect the model's generalizability.

Future research will focus on:

- Data Expansion: Efforts will be made to collect more data and create a more balanced dataset to improve the model's recognition ability on different types and stages of lesion samples.

- Model Optimization: Improvements will be pursued through further parameter tuning, introducing advanced regularization techniques, and exploring more complex model architectures to enhance recall and other performance metrics.
- Algorithmic Enhancements: Plans include researching and implementing advanced ensemble learning strategies and attention mechanisms to boost the model's capability to recognize and learn key features in images.
- Clinical Validation: Testing the model in a broader clinical setting to assess its applicability and feasibility in real medical scenarios.
- Multi-task Learning: Exploring the model's performance in multi-task learning, such as simultaneous lesion classification and severity assessment.
- Interpretability Studies: To increase model transparency and interpretability, research will focus on understanding the significance of features in the model's decision-making process, aiding doctors in better comprehending the basis of model predictions.
- Software Tool Development: Plans to develop user-friendly software tools, enabling non-technical medical professionals to easily use our model for diagnostics.

Through these efforts, we aim to overcome the current model's limitations and ultimately develop a more accurate, robust, and user-friendly automatic image classification system, providing strong support for medical diagnostics.

Reference List

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012. Accessed: Dec. 12, 2023. [Online]. Available: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[3] Q. Wang, N. Jia, and T. P. Breckon, “A Baseline for Multi-Label Image Classification Using An Ensemble of Deep Convolutional Neural Networks.” arXiv, May 09, 2019. doi: 10.48550/arXiv.1811.08412.

[4] X. Meng, X. Wang, S. Yin, and H. Li, “Few-shot image classification algorithm based on attention mechanism and weight fusion,” *Journal of Engineering and Applied Science*, vol. 70, no. 1, p. 14, Mar. 2023, doi: 10.1186/s44147-023-00186-9.

[5] M. N. Bajwa *et al.*, “Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning,” *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 136, Jul. 2019, doi: 10.1186/s12911-019-0842-8.

[6] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed, “G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–7. doi: 10.1109/IJCNN48605.2020.9207664.

[7] A. C. de Moura Lima, L. Bezerra Maia, R. M. Pinheiro Pereira, G. B. Junior, J. D. Sousa de Almeida, and A. Cardoso de Paiva, “Glaucoma Diagnosis over Eye Fundus Image through Deep Features,” in *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Jun. 2018, pp. 1–4. doi: 10.1109/IWSSIP.2018.8439477.

[8] J. Latif, S. Tu, C. Xiao, S. Ur Rehman, A. Imran, and Y. Latif, “ODGNet: a deep learning model for automated optic disc localization and glaucoma classification using fundus images,” *SN Appl. Sci.*, vol. 4, no. 4, p. 98, Mar. 2022, doi: 10.1007/s42452-022-04984-3.

[9] L. K. Singh, M. Khanna, S. Thawkar, and R. Singh, “A novel hybridized feature selection strategy for the effective prediction of glaucoma in retinal fundus images,” *Multimed Tools Appl*, Oct. 2023, doi: 10.1007/s11042-023-17081-3.

[10] M. Yan, Y. Lin, X. Peng, and Z. Zeng, “mixDA: mixup domain adaptation for glaucoma detection on fundus images,” *Neural Comput & Applic*, Jul. 2023, doi: 10.1007/s00521-023-08572-3.

[11] C. Sharmila and N. Shanthi, “Retinal Image Analysis for Glaucoma Detection Using Transfer Learning,” in *Advances in Electrical and Computer Technologies*, T. Sengodan, M. Murugappan, and S. Misra, Eds., in Lecture Notes in Electrical Engineering. Singapore: Springer Nature, 2021, pp. 235–244. doi: 10.1007/978-

981-15-9019-1_21.

[12] X. Xia, C. Xu, and B. Nan, “Inception-v3 for flower classification,” in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Jun. 2017, pp. 783–787. doi: 10.1109/ICIVC.2017.7984661.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520. Accessed: Dec. 14, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html

[14] L. Wen, X. Li, and L. Gao, “A transfer convolutional neural network for fault diagnosis based on ResNet-50,” *Neural Comput & Applic*, vol. 32, no. 10, pp. 6111–6124, May 2020, doi: 10.1007/s00521-019-04097-w.

[15] Z. Omiotek and A. Kotyra, “Flame Image Processing and Classification Using a Pre-Trained VGG16 Model in Combustion Diagnosis,” *Sensors*, vol. 21, no. 2, Art. no. 2, Jan. 2021, doi: 10.3390/s21020500.