Stacked Decision-Level Fusion of Classical and Deep Learning Models for Image Classification

Arwa Basbrain
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
abasbrain@kau.edu.sa

Samar Alamri
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
Salamri0362@stu.kau.edu.sa

Talah Faloudah
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
tfaloudah@stu.kau.edu.sa

Jumanah Banabilah
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
jbanabilah@stu.kau.edu.sa

Rana AlZahrani
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
rabdullahalzahrani0003@stu.kau.edu.sa

***Abstract*—This paper evaluates the effectiveness of decision-level fusion for improving image classification performance by integrating heterogeneous learning models. Specifically, we employ three independent classifiers: a HOG-based ANN, a CNN trained from scratch, and a MobileNetV2 transfer learning model. A stacking strategy is used to fuse the probability outputs of these classifiers through a Logistic Regression meta-learner. Experimental results demonstrate that the stacked model consistently outperforms all individual base learners, achieving higher accuracy and improved robustness on a multi-class fish image dataset. These findings highlight the advantages of combining geometric, textural, and semantic representations through decision-level fusion for fine-grained image classification.**

I. Introduction

Image classification has been studied using both classical and deep learning approaches. Classical models, such as Histogram of Oriented Gradients (HOG) combined with shallow Artificial Neural Network (ANN), effectively capture local structural patterns. Deep learning models, including Convolutional Neural Network (CNN) and transfer learning architectures like MobileNet, learn high-level semantic representations directly from raw data. However, no single model consistently performs best across all image categories. To address this limitation, decision-level fusion is used to combine complementary model predictions. In particular, stacking enables a meta-classifier to learn optimal combinations of decisions from heterogeneous models, improving robustness and generalization.

In this work, we propose a stacking-based decision-level fusion framework that integrates three independent classifiers: HOG-based ANN, a CNN trained from scratch, and a MobileNet transfer learning model. The predictions of these models are fused at the decision level using stacking to produce the final classification output. Experimental results show that the proposed fusion approach outperforms individual models, demonstrating the effectiveness of decision-level fusion in hybrid classification systems.

II. Literature Review

A. Theoretical Foundations of Stacked Generalization (Wolpert & Breiman)

The tendency of high-capacity models to overfit on homogeneous or augmented datasets is a well-known phenomenon in learning theory. The conceptual framework for addressing this was established by Wolpert (1992) [1], who introduced "Stacked Generalization" (Stacking) not merely as a method for combining predictions, but as a strategy to deduce and correct the biases of generalizers. Wolpert demonstrated that a "Level-1" meta-learner could smooth out the Model-specific weaknesses of "Level-0" base learners, effectively acting as a regularizer against overfitting.

This framework was refined by Breiman (1996) [2] in *Stacked Regressions*. Breiman provided mathematical evidence that combining diverse predictors via non-negative constraints ensures stability, preventing the ensemble from degrading performance even when individual base learners are prone to high variance due to data augmentation. These foundational studies justify our use of stacking to ensure that our classification relies on genuine fish features rather than the synthetic variations introduced during augmentation.

B. Heterogeneous Ensembles in Bio-Image Analysis (Nanni et al.)

The risk of overfitting is A critical challenge in biological classification, where subjects share high morphological similarity. In Nanni et al. [3] research,They highlighted that while data augmentation increases dataset size, relying on a single architecture can still lead to bias toward specific texture patterns.

Crucially, Nanni et al. demonstrated that heterogeneous ensembles—combining diverse Architecture—are more robust than single models on constrained datasets. By fusing a classical HOG-ANN (which ignores background color) with

Fig. 1. Example images from the dataset

a deep MobileNetV2 (which learns high-level semantics), we decouple the fish classification from the static background, ensuring the model generalizes well despite the repetitive nature of the capture environment.

C. Theoretical Foundations of HOG (Dalal & Triggs)

This research [4] introduced HOG as a robust handcrafted feature descriptor for human detection, demonstrating that local object shape can be effectively represented using dense histograms of gradient orientations. Their study focuses on analyzing the impact of feature design choices, showing that fine-scale gradient computation, fine orientation binning, relatively coarse spatial cells, and strong local contrast normalization using overlapping blocks are critical for high discriminative power. A linear SVM is employed primarily as a baseline classifier to isolate and evaluate the effectiveness of the proposed features. The findings of this work establish HOG as a strong feature representation independent of the specific classifier used, making it suitable for integration with alternative learning models such as artificial neural networks.

III. DATASET

Experiments were conducted on the A Large-Scale Fish Dataset, which contains labeled images of nine fish species captured under realistic conditions with varying orientations and noisy backgrounds. The dataset includes Black Sea Sprat, Gilt-Head Bream, Horse Mackerel, Red Mullet, Red Sea Bream, Sea Bass, Shrimp, Striped Red Mullet, and Trout. Due to class imbalance in the original data, data augmentation (random rotations and reflections) was applied, resulting in 1,000 images per class. All experiments in this paper follow a consistent evaluation protocol based on 5-fold cross-validation. The reported results correspond to the best-performing fold.

IV. PROPOSED METHODOLOGY

This study employs a Heterogeneous Stacked Generalization framework that integrates geometric, textural, and semantic features. The system operates on a two-level hierarchy to Reduce the biases of individual classifiers.

A. System Architecture

The architecture is divided into two processing stages:

- **Level-0 (Base Learners):** Three diverse models (HOG-ANN, CNN, MobileNetV2) independently process the input image to generate class probability vectors.
- **Level-1 (Meta-Learner):** A Logistic Regression meta-classifier receives the concatenated probability vectors from Level-0 and learns an optimal weighting scheme to predict the final class.

B. Level-0 Base Classifiers

We utilize three distinct network Architecture to capture complementary visual cues.

1) Geometric Expert: HOG-ANN: To capture structural shape profiles, we utilize Histogram of Oriented Gradients (HOG) features fed into a shallow Artificial Neural Network (ANN).

- **Feature Extraction:**
  - Orientations: 9
  - Pixels per cell: 8×8
  - Cells per block: 2×2

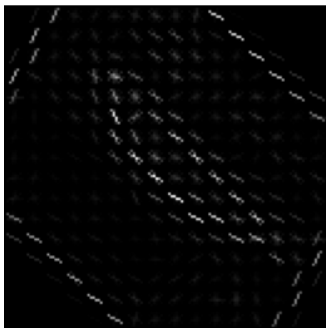

Fig. 2. Example of HOG feature Extraction

- **Classifier:** A feed-forward ANN with dense layers using ReLU activation to map gradient histograms to fish categories.

2) Texture Expert: CNN: A lightweight Convolutional Neural Network (CNN) was designed to capture low-level textural patterns (e.g., scales, skin reflections) specific to the marketplace environment.

- **Block 1:** Conv2D (32 filters, 3×3) $\rightarrow$ MaxPooling2D (2×2).
- **Block 2:** Conv2D (64 filters, 3×3) $\rightarrow$ MaxPooling2D (2×2).
- **Head:** Flatten $\rightarrow$ Dense (128 neurons, ReLU) $\rightarrow$ Softmax Output.

3) Semantic Expert: MobileNetV2: To leverage high-level semantic representations, we utilized the MobileNetV2 architecture pre-trained on ImageNet, adapted via Transfer Learning.

- **Backbone:** MobileNetV2 (ImageNet weights).
- **Fine-Tuning:** The last 20 layers were unfrozen to adapt to domain-specific features.
- **Custom Head:** GlobalAveragePooling2D $\rightarrow$ Dense (256 neurons) $\rightarrow$ BatchNormalization $\rightarrow$ Dropout (0.3) $\rightarrow$ Softmax Output.

C. Level-1 Stacked Generalization

To integrate the diverse predictions of the base learners, we employ a Stacked Generalization (Stacking) strategy. Unlike simple voting mechanisms (e.g., Majority Voting), Stacking trains a meta-learner to learn the optimal combination of base classifiers based on their confidence levels.

1) Mathematical Formulation: The stacking framework operates by treating the outputs of the three base models as inputs for a final meta-learner.

For any input image x, each base model (HOG, CNN, MobileNetV2) predicts a probability vector $P(x)$ for the 9 fish classes. We combine these predictions into a single meta-feature vector $\mathbf{Z}(x)$ using concatenation:

$$\mathbf{Z}(x) = [P_{HOG}(x), P_{CNN}(x), P_{Mob}(x)] \tag{1}$$

This combined vector $\mathbf{Z}(x)$ is then fed into the Level-1 Logistic Regression classifier. The final predicted class $\hat{y}$ is calculated by applying learned weights $\mathbf{W}$ to these probabilities:

$$\hat{y} = \text{argmax}\left(\text{softmax}(\mathbf{W}^T\mathbf{Z}(x) + \mathbf{b})\right) \tag{2}$$

Here, the weight matrix $\mathbf{W}$ allows the meta-learner to assign higher importance to the specific base model that is most reliable for a given fish species, effectively "correcting" the errors of weaker models.

V. PREPROCESSING

Preprocessing was designed to ensure stable learning and fair comparison across all models. Images were resized to a fixed resolution, and for the HOG-based and CNN models, pixel intensities were normalized to the range [0,1] to stabilize feature extraction and training. For the handcrafted HOG pipeline, features were extracted using a fixed configuration to ensure consistent representation and prevent scale-related bias during learning.

The same preprocessing strategy was applied consistently across experiments, while normalization was intentionally omitted for the MobileNetV2 model to maintain compatibility with it pretrained weights. This ensured that performance differences are attributable to the model architecture rather than preprocessing variations.

VI. RESULTS AND DISCUSSION

This section analyzes the performance of the individual base classifiers (HOG-ANN, CNN, MobileNetV2) and evaluates the effectiveness of the proposed Stacked Generalization framework.

A. Performance Analysis of HOG-ANN Base Learner

The HOG-ANN model, which relies on handcrafted geometric features, demonstrated strong overall performance but exhibited distinct behavior in distinguishing species with similar body contours.

1) Numerical Metrics: The final model metrics achieved a test accuracy of **93.33%**. The weighted averages for the performance metrics are:

- **Precision:** 0.93
- **Recall:** 0.93
- **F1-score:** 0.93

While the overall performance is high, the class-wise breakdown reveals notable variance. The *Red Sea Bream* achieved the highest recall (98.50%), indicating distinct geometric features. In contrast, the *Horse Mackerel* recorded the lowest precision (86.88%), suggesting that its shape descriptors frequently overlap with those of other classes.

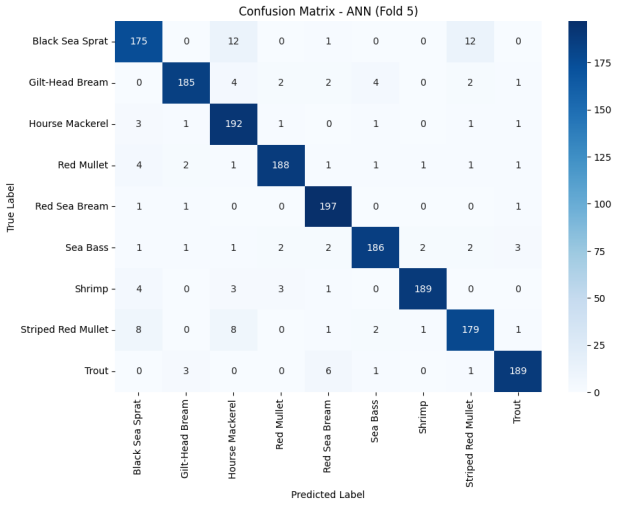

Fig. 3. Confusion Matrix for HOG-ANN Base Learner.

2) Confusion Matrix and Training Curves: As shown in Fig. 3, the matrix confirms the specific confusion cluster involving the **Black Sea Sprat**. Specifically, 12 samples of Black Sea Sprat were misclassified as **Striped Red Mullet**, and 12 were misclassified as **Horse Mackerel**. This error pattern aligns with the lower precision scores for these classes, indicating that HOG descriptors, which encode edge gradients—struggled to differentiate these species due to their similar elongated profiles.

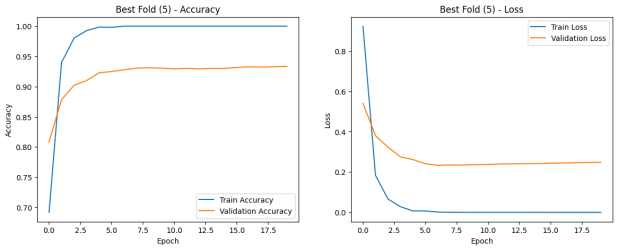

Fig. 4. Training and Validation Curves for HOG-ANN.

The training curves (Fig. 4) indicate rapid convergence, with training accuracy reaching near 100% by epoch 5. The validation accuracy stabilizes around 93%, and the validation

loss flattens quickly. The gap between training and validation performance suggests a degree of overfitting, which is expected given that HOG features are rigid and may not generalize to pose variations as effectively as learned features.

B. Performance Analysis of CNN Base Learner

The CNN, trained from scratch on fish data, demonstrated a significant performance improvement over the handcrafted HOG features, particularly in distinguishing species with complex textural patterns.

1) Numerical Metrics: The final model metrics achieved a test accuracy of **97.18%**, surpassing the HOG-ANN model. The weighted averages for the performance metrics are:

- **Precision:** 0.97
- **Recall:** 0.97
- **F1-score:** 0.97

This high performance across all metrics indicates that the deep convolutional layers successfully learned robust feature representations that capture both the shape and the fine-grained surface details of the fish.

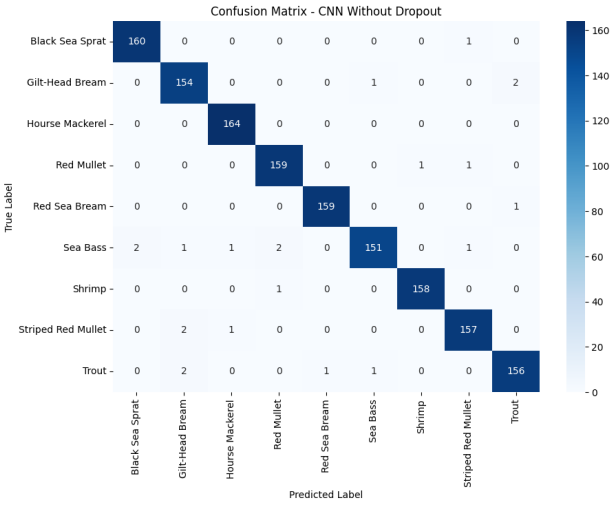

Fig. 5. Confusion Matrix for CNN Base Learner.

2) Confusion Matrix and Training Curves: The confusion matrix (Fig. 5) reveals a critical strength of this model: it completely resolved the confusion between **Black Sea Sprat** and **Horse Mackerel** observed in the HOG model. The CNN achieved 100% classification accuracy for the Black Sea Sprat (160/160 correct) and Horse Mackerel (164/164 correct). This confirms our hypothesis that learned textural features are superior to geometric gradients for distinguishing these similarly shaped species.

However, a new, minor confusion pattern emerged regarding the **Sea Bass**. While still high performing, the Sea Bass had the highest variance in misclassifications, being occasionally confused with the *Red Mullet* and *Black Sea Sprat*. This suggests that while the CNN excels at texture, it may occasionally struggle with species that share very similar scale patterns under specific lighting conditions.

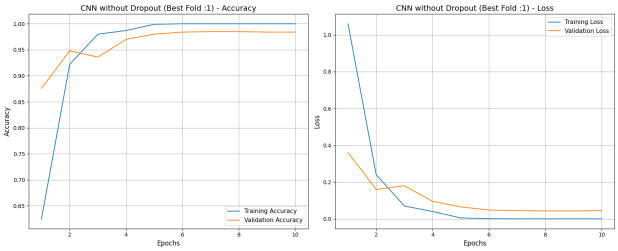

Fig. 6. Training and Validation Curves for CNN.

The training dynamics (Fig. 6) demonstrate exceptional stability. The model converged rapidly, reaching near-optimal performance within just 6 epochs. The validation loss stabilized around 0.05, and the gap between training and validation accuracy is minimal.The CNN architecture was well-regularized and avoided the severe overfitting often seen in deep models trained on homogeneous backgrounds.

C. Performance Analysis of MobileNetV2 Base Learner

The MobileNetV2 model utilized transfer learning to leverage high-level semantic features pre-learned from ImageNet.

1) Numerical Metrics: As detailed in the classification report, the model achieved a test accuracy of **96.72%**. The weighted averages for the performance metrics are:

- **Precision:** 0.97
- **Recall:** 0.97
- **F1-score:** 0.97

Notably, the model achieved perfect precision (1.00) for the *Red Mullet* (Class 3), *Shrimp* (Class 6), and *Trout* (Class 8). This indicates that for species with distinct semantic features (like the unique body shape of the Shrimp), the transfer learning approach provides exceptional discriminative power.

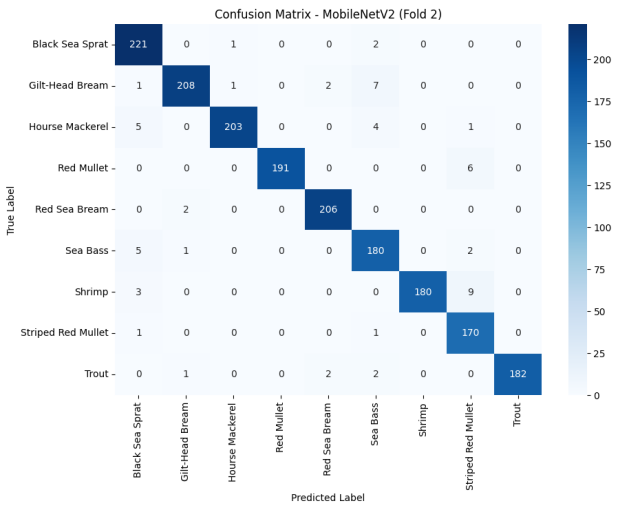

Fig. 7. Confusion Matrix for MobileNetV2 Base Learner.

2) Confusion Matrix and Training Curves: The confusion matrix (Fig. 7) shows a clean diagonal but highlights a specific

confusion cluster different from the previous models. The *Gilt-Head Bream* had 7 samples misclassified as *Sea Bass*. This suggests that while MobileNetV2 understands general object features, it may occasionally miss the subtle, domain-specific texture differences between these two silvery fish that the CNN captured more effectively.

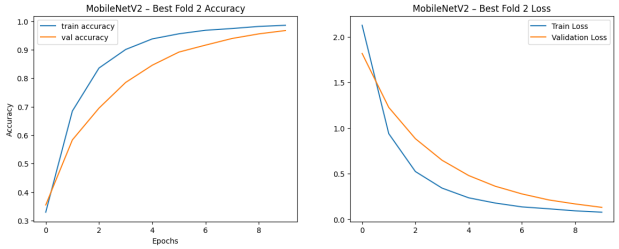

Fig. 8. Training and Validation Curves for MobileNetV2.

The training curves (Fig. 8) exhibit a classic transfer learning profile. The accuracy starts lower than the CNN but climbs steadily as the pre-trained weights adapt to the new domain. Because the validation and training loss curves stayed close together, we know the model did not overfit. This is likely because it started with high-quality features learned from the large ImageNet dataset.

D. Performance Analysis of Stacked Generalization Framework

The proposed Heterogeneous Stacked Ensemble demonstrated superior performance, achieving a near-perfect test accuracy of **99.50%**. This result validates the core hypothesis of this research: that fusing diverse feature representations (geometric, learned texture, and semantic) effectively cancels out the specific weaknesses of each individual model.

1) Numerical Metrics: As shown in the classification report, the ensemble achieved weighted averages of **1.00** for Precision, Recall, and F1-score. Notably, the model achieved perfect precision (1.00) and recall (1.00) for Classes 2 (Horse Mackerel), 3 (Red Mullet), 4 (Red Sea Bream), and 6 (Shrimp). This represents a substantial improvement over the HOG-ANN model, which had previously struggled with the Horse Mackerel (86% precision).

2) Confusion Matrix Analysis: The confusion matrix (Fig. 9) exhibits an exceptionally clean diagonal. Critically, the ensemble successfully resolved the "confusing pairs" identified in the base learners:

- **Black Sea Sprat:** The Stack correctly identified 197/200 samples, effectively retaining the texture-based improvements of the CNN while discarding the shape-based errors of the HOG model.
- **Red Sea Bream:** The model achieved 200/200 correct classifications, correcting the minor misclassifications observed in the MobileNetV2 model.

The few remaining errors (e.g., 3 samples of Black Sea Sprat misclassified as Sea Bass) are negligible and likely attributable to extreme occlusion or labeling noise in the dataset.

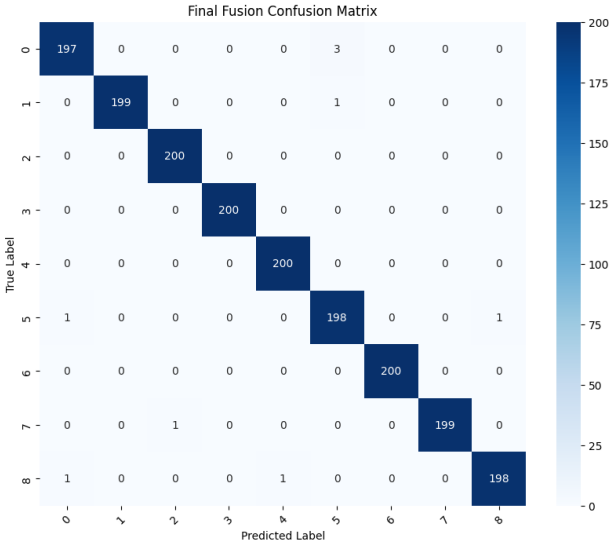

Fig. 9. Confusion Matrix for Stacked Ensemble.

E. Comparative Discussion

Table I summarizes the quantitative comparison of all developed models.

TABLE I
COMPARISON OF MODEL PERFORMANCE

Model	Accuracy	Precision	Recall	F1-Score
HOG-ANN	93.33%	0.93	0.93	0.93
MobileNetV2	96.72%	0.97	0.97	0.97
CNN	97.18%	0.97	0.97	0.97
Stacked Ensemble	**99.50%**	**1.00**	**1.00**	**1.00**

The comparative analysis reveals a clear hierarchy of performance. The **HOG-ANN** model, while robust to lighting changes, was limited by its inability to capture fine-grained texture, resulting in the lowest accuracy (93.33%). The **MobileNetV2** and **CNN** models significantly improved upon this by leveraging learned features, achieving 96.72% and 97.18% respectively.

However, the **Stacked Ensemble** outperformed all single models by a margin of roughly 2.3%. This improvement is attributed to the meta-learner's ability to dynamically weight the contributions of each expert. For instance, in cases where the HOG model was confused by shape similarity (e.g., Horse Mackerel), the meta-learner leveraged the high confidence of the CNN's texture analysis to correct the decision. Conversely, where the CNN might have overfitted to background artifacts, the generalization power of MobileNetV2 provided a safeguard. This confirms that retail fish classification benefits significantly from using multiple feature viewpoints that jointly encode geometry, texture, and high-level semantics.

VII. SUMMARY OF APPROACH

Our experimental pipeline consists of three stages: data preprocessing, heterogeneous feature extraction, and stacked

generalization. We utilized a dataset of 9,000 augmented fish images to train three distinct base learners: a HOG-ANN model for geometric analysis, a CNN for fine-grained texture recognition, and a pre-trained MobileNetV2 for high-level semantic classification. The probability outputs of these diverse experts were concatenated and fed into a Logistic Regression meta-learner. Across all experiments, this stacked ensemble consistently outperformed individual models, achieving a peak accuracy of 99.50%, confirming that integrating geometric, textural, and semantic cues provides superior discriminative power in constrained environments.

VIII. CONCLUSION

This work introduced a heterogeneous stacked generalization framework for fish species classification that combines geometric (HOG-ANN), textural (CNN), and semantic (MobileNetV2) representations. By fusing the predictions of these diverse base learners using a Logistic Regression meta-classifier, the proposed approach effectively mitigates the individual weaknesses of each model. Experimental results demonstrate that the stacked ensemble consistently outperforms all standalone classifiers, achieving a peak accuracy of **99.50%**. These findings confirm that decision-level fusion is an effective strategy for improving robustness and classification performance in visually similar and constrained image domains.

IX. FUTURE WORK

- Evaluate the proposed framework on larger and more diverse fish datasets, including underwater imagery and real-time video streams, to assess robustness under varying illumination and occlusion conditions.
- Explore a Mixture of Experts (MoE) architecture and perform a comparative analysis against the stacking-based fusion strategy.
- Expanding experiments to include additional classification models in order to increase ensemble diversity and further improve generalization performance.

REFERENCES

[1] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[2] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[3] L. Nanni, S. Ghidoni, and S. Brahnam, "Ensemble of convolutional neural networks for bioimage classification," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 19–35, Jan. 2021.

[4] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 2005, pp. 886–893.

[5] O. Ulucan, D. Karakaya, and M. Turkan, "A large-scale dataset for fish segmentation and classification," in *Proc. Conf. Innovations in Intelligent Systems and Applications (ASYU)*, 2020.