Cirebon Annual Multidisciplinary International Conference (CAMIC 2024)

Ensemble Learning Model for Fruit and Vegetable Classification

1st Putri Rizqiyah
Informatics Engineering, Faculty of Information Technology
Universitas Catur Insan Cendekia
Cirebon, Indonesia
putri.rizqiyah@cic.ac.id

2nd Kusnadi
Informatics Engineering, Faculty of Information Technology
Universitas Catur Insan Cendekia
Cirebon, Indonesia
kusnadi@cic.ac.id

3rd Petrus Sokibi
Informatics Engineering, Faculty of Information Technology
Universitas Catur Insan Cendekia
Cirebon, Indonesia
petrus.sokibi@cic.ac.id

4th Ferri Krisdiantoro
Informatics Engineering, Faculty of Information Technology
Universitas Catur Insan Cendekia
Cirebon, Indonesia
ferri.krisdiantoro.ti.22 @cic.ac.id

5th Suwandi
Informatics Engineering, Faculty of Information Technology
Universitas Catur Insan Cendekia
Cirebon, Indonesia
suwandi@cic.ac.id

***Abstract*—Nutritional information about fruits and vegetables is vital for promoting healthy eating patterns and combating malnutrition. This research presents the development of a web application for fruit and vegetable image classification, utilizing ensemble learning with a stacking technique. The model combines Swin Transformer and ResNet as base learners, with Support Vector Machine (SVM) serving as the meta-learner. Trained on a dataset encompassing 32 fruit and vegetable classes, the model achieved an impressive 98% accuracy, along with consistently high precision, recall, and F1-score. The application was implemented with Flask for the backend and ReactJS for the frontend and is hosted on PythonAnywhere. Beyond image classification, the application provides users with detailed nutritional information, including energy content and vitamin composition, in a quick and user-friendly manner. This study highlights the effectiveness of ensemble learning in enhancing classification accuracy. Future work will focus on expanding the dataset and transitioning to more robust hosting platforms to improve performance and user experience.**

Keywords— ***Image Classification; Ensemble Learning; Swin Transformer; Resnet; Web Application***

I. INTRODUCTION

Nutritional information about fruits and vegetables is crucial to meeting daily dietary needs and reducing malnutrition. A 2021 report by UNICEF/WHO/World Bank highlighted the high prevalence of malnutrition in Southeast Asia, including Indonesia, where over 30% of children under five years old suffer from stunting [1]. The SEANUTS II (2021) study revealed that 70% of children in the region do not meet daily calcium requirements, and 84% lack sufficient vitamin D [2]. Furthermore, promoting the benefits of fruits and vegetables can increase their consumption by up to 30%, according to FAO (2021) [3]. However, significant challenges remain in identifying different types of fruits and vegetables and accessing their nutritional information quickly and accurately.

Machine learning has been widely applied to image classification, including a study by Steinbrener et al. (2019), which used Convolutional Neural Networks (CNN) for fruit and vegetable classification with an accuracy of 92.23% [4]. This research developed a stacking ensemble learning-based classification model combining Swin Transformer and ResNet as base models with SVM as a meta-learner. Swin Transformer excels in handling complex images, while ResNet effectively addresses vanishing gradient issues [5]. This combination aims to enhance classification accuracy and stability compared to single methods.

The model was trained on a dataset of 32 fruit and vegetable classes and evaluated using accuracy, precision, recall, and F1-score metrics. The research will observe the superiority of the stacking approach compared to other methods, such as CNN and transfer learning [6], [7]. This study also explored the effectiveness of stacking techniques in improving classification performance and their application in various image classification needs.

LITERATURE REVIEW

Ensemble Learning and Stacking Techniques

Ensemble learning is a machine learning method that combines multiple models to improve prediction accuracy and stability. This approach allows individual model errors to offset each other, yielding better predictions. A popular technique is stacking, where several initial models make predictions that are combined by a meta-learner to produce the final, more accurate result. Stacking enables the use of diverse models, leveraging their strengths to enhance overall performance [8].

Base Model and Meta-Learner

Swin Transformer and ResNet were selected as base models for their strengths in visual feature extraction and stability in training deeper networks [9], [10]. SVM serves as the meta-learner to combine predictions from these base models, ensuring improved classification accuracy and generalization [11].

Data Augmentation

Data augmentation is a crucial technique in deep learning, as it is used to enhance the performance of a model. Augmentation involves introducing small changes or distortions to the original data, such as rotations, cropping, or color adjustments [12]. The purpose of this technique is to expand and modify values in a neural network, particularly smaller values in the Jacobian matrix, without altering its primary direction. Data augmentation helps models learn more complex patterns, prevents overfitting, and improves the model’s generalization ability. As a result, the model can better recognize patterns even with a limited variety of training data

Transfer Learning

Transfer learning is a technique where a model trained on one dataset, task, or domain, such as image recognition with text supervision, is used to initiate training on another dataset or task, which may be similar or different. In the context of CLIP, the model employs text supervision to generate visual representations that can be applied to various image recognition or classification tasks without requiring extensive retraining (fine-tuning) for each task. CLIP trains the model in a way that enables transfer learning from existing large labeled datasets, such as images and text, to new tasks. This reduces the need for task-specific annotation data and enhances its flexibility in handling a range of image recognition applications [13].

Framework Deployment

Flask is a Python backend framework used to implement applications as web services. ReactJS is a JavaScript library for building user interfaces. This combination enables the development of responsive and dynamic web applications.

Related Works

Research on image recognition using Convolutional Neural Networks (CNN) successfully classified 30 fruit classes with 94% accuracy on 971 images, demonstrating CNN's effectiveness in vision-based control systems due to its ability to extract hierarchical features directly from image data [14]. Another study employed Random Forest (RF) to classify apples, strawberries, and oranges using features like shape, color, and SIFT. Evaluated on 178 images, RF outperformed K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), showcasing its robustness and accuracy when implemented in MATLAB [15]. These studies highlight the strengths of CNN and RF in image classification, with CNN excelling in automated feature extraction and RF offering a simpler, feature-based alternative for smaller datasets. Stacking ensemble learning has been applied in several studies[16], [17], [18] [1]–[3]. Research conducted by Zhang et al.[16] utilized ensemble learning techniques with machine learning models such as Support Vector Machine (SVM), Random Forest, K-Nearest Neighbor, and others, demonstrating that the ensemble model performed well with an AUC of 0.928. Another study [19] combined CNN with ResNet50, EfficientNetB4, and Xception, achieving 88.12% accuracy, and Sarkar et al. [18] reported 93.24% accuracy. This research also applies the Swin Transformer as a base model. Swin Transformer, a recent method for image classification, has shown excellent performance, achieving 98% accuracy in classifying leaf diseases [20] and 97.85% accuracy in classifying plastic waste [19].

II. METHOD

Dataset Collection

The dataset used in this research was sourced from Kaggle under the title "Fruit and Vegetable Image Recognition" provided by Kritik Seth .

This dataset contains a total of 36 classes, each representing a type of fruit or vegetable, such as apples, grapes, carrots, tomatoes, etc. The images are in RGB and RGBA formats with varying resolutions. Details of the dataset are as follows:

1. Fruit and Vegetable Types in the Dataset:
   - Fruits: Banana, Apple, Pear, Grape, Orange, Kiwi, Watermelon, Pomegranate, Pineapple, Mango
   - Vegetables: Cucumber, Carrot, Capsicum, Onion, Potato, Lemon, Tomato, Radish, Beet, Cabbage, Lettuce, Spinach, Soybean, Cauliflower, Bell Pepper, Chili, Turnip, Corn, Sweet Corn, Sweet Potato, Paprika, Jalapeno, Ginger, Garlic, Peas, Eggplant
2. Dataset Folder Structure:
   - train: Contains up to 100 images per category
   - test: Contains 10 images per category
   - valid: Contains 10 images per category

Data Preprocessing

To simplify classification, similar categories were merged. For example, corn and sweet corn were combined into a single class, and chili, jalapeno, and paprika were merged into one category. This reduced the number of classes from 36 to 32.

The dataset was then preprocessed, including converting RGBA images (with transparency channels) to RGB format to align with the requirements of deep learning models like Swin Transformer and ResNet. This conversion was automated using Python to ensure all images had three primary color channels (red, green, blue).

Nutritional Data Collection

Nutritional data was collected from the website to label each fruit and vegetable with its respective nutritional values. The integration process involved the following steps:

1. Data Collection: Nutritional information (e.g., energy, protein, carbohydrates, vitamins) from nilaigizi.com was compiled and formatted appropriately.
2. Data Storage: The data was stored in a MySQL table with columns for fruit/vegetable names and their nutritional values.
3. Data Retrieval: After image classification, the application accessed MySQL to retrieve nutritional information using SQL queries based on the classification results.
4. Integration with the Application: Retrieved nutritional data was displayed on the user interface through an API connecting the Flask backend and React frontend.

Model Architecture

The research employed an ensemble model with a stacking technique, combining two base models, Swin Transformer and ResNet, and using Support Vector Machine (SVM) as a meta-learner. Swin Transformer efficiently extracts image features, while ResNet addresses vanishing gradient issues in deep networks. Predictions from these base models were combined by SVM to produce more accurate and stable final predictions.

Model Training

During model training, the two base models, Swin Transformer and ResNet, were fine-tuned using the preprocessed fruit and vegetable dataset. Both models utilized pre-trained weights and were optimized using AdamW and Cross-Entropy loss to enhance performance for the classification task.

A. Training Parameters for Base Models:

1. Swin Transformer:
   - Pre-trained weights: Swin-Base Patch4 Window7 224 from Hugging Face Transformers
   - Optimizer: AdamW with a learning rate of 1e-5
   - Scheduler: StepLR, with a learning rate reduction of 0.1 every 5 epochs
   - Loss function: Cross-Entropy Loss
   - Epochs: 10
2. ResNet:
   - Pre-trained weights: ResNet-50 from PyTorch
   - Optimizer: AdamW with a learning rate of 1e-5
   - Scheduler: StepLR, with a learning rate reduction of 0.1 every 5 epochs
   - Loss function: Cross-Entropy Loss
   - Epochs: 10

B. Meta-Learner (SVM):

After training the base models, features from each model were extracted and combined for training the meta-learner using the stacking approach. The process involved:

1. Feature extraction from Swin Transformer and ResNet for training and validation data.
2. Combining feature vectors from both models into a single feature matrix.
3. Training an SVM with a linear kernel using the combined feature matrix to predict labels.

C. Evaluation Metrics:

The model was trained using 80% of the data for training and 10% for validation. The evaluation metrics included:

- Precision: Proportion of correct positive predictions.
- Recall: Proportion of actual positives correctly identified.
- F1-Score: Harmonic mean of precision and recall, measuring balance.
- Confusion Matrix: For analyzing classification errors in each class.

The final evaluation on the test data (10%) demonstrated the effectiveness of the stacking model with SVM as the meta-learner, showing improved accuracy compared to individual base models.

Model Performance Evaluation

Evaluating the model's performance is critical to ensuring prediction accuracy and reliability. In this research, the following evaluation metrics were used:

1. Accuracy: Measures model performance during training, calculated by comparing the number of correct predictions with the total data for each epoch in both training and validation.
2. Evaluation Metrics: Post-training, the model was evaluated on the test data using precision, recall, and F1-score. Precision assessed the correctness of positive predictions, recall measured the model’s sensitivity, and F1-score evaluated the balance between precision and recall.
3. Confusion Matrix: Provided detailed insights into the model's performance for each class, highlighting patterns of errors, such as misclassifications between similar classes..

III. RESULTS AND DISCUSSION

Model Evaluation

In this experiment, three main models were tested: Swin Transformer, ResNet, and SVM (as a meta-learner in the ensemble stacking technique). The training and evaluation results of each model showed significant performance differences, which encouraged the use of the ensemble technique to improve classification accuracy.

A. Evaluation Metrics for Swin Transformer and ResNet

The evaluation results showed that the Swin Transformer model achieved 99% accuracy with precision, recall, and F1-score values nearing perfection across almost all classes, as shown in Table 4.1. In contrast, ResNet achieved 97% accuracy, with some classes, such as "Apple," "Chili," and "Potato," having metrics below 0.9, indicating certain classification errors (Table 2).

TABLE I. EVALUATION METRICS SWIN MODEL

Class	Precision	Recall	F1-Score	Support
Grapes	1	1	1	9
Apple	1	0.8	0.89	10
Onion	1	1	1	10
Garlic	1	1	1	10
Spinach	1	1	1	10
Beetroot	1	1	1	10
Chili	0.95	1	0.97	18
Pomegranate	1	1	1	10
Corn	1	1	1	20
Ginger	1	1	1	10
Orange	0.9	1	0.95	9
Peas	1	1	1	10
Soybean	1	1	1	10
Cauliflower	1	1	1	10
Potato	1	0.9	0.95	10
Kiwi	1	1	1	10
Cabbage	1	1	1	10
Lemon	1	1	1	10
Radish	1	1	1	9
Mango	0.91	1	0.95	10
Pineapple	1	1	1	10
Paprika	1	0.97	0.98	29
Pear	0.83	1	0.91	10
Banana	1	0.89	0.94	9
Lettuce	1	1	1	9
Watermelon	1	1	1	10
Eggplant	1	1	1	10
Cucumber	1	1	1	10
Tomato	1	1	1	10
Turnip	1	1	1	10
Sweet Potato	1	1	1	10
Carrot	1	1	1	9
Accuracy			**0.99**	**351**
Macro avg	**0.99**	**0.99**	**0.99**	**351**
Weighted avg	**0.99**	**0.99**	**0.99**	**351**

TABLE II. EVALUATION METRICS RESNET MODEL

Class	Precision	Recall	F1-Score	Support
Grapes	1	1	1	9
Apple	0.89	0.8	0.84	10
Onion	1	1	1	10
Garlic	1	1	1	10
Spinach	1	1	1	10
Beetroot	1	1	1	10
Chili	0.86	1	0.92	18
Pomegranate	1	1	1	10
Corn	1	1	1	20
Ginger	1	1	1	10
Orange	0.9	1	0.95	9
Peas	1	1	1	10
Soybean	1	1	1	10
Cauliflower	1	1	1	10
Potato	0.89	0.8	0.84	10
Kiwi	1	1	1	10
Cabbage	1	1	1	10
Lemon	0.91	1	0.95	10
Radish	1	1	1	9
Mango	1	1	1	10
Pineapple	1	1	1	10
Paprika	0.93	0.93	0.93	29
Pear	1	1	1	10
Banana	1	0.78	0.88	9
Lettuce	1	1	1	9
Watermelon	1	1	1	10
Eggplant	1	1	1	10
Cucumber	1	1	1	10
Tomato	1	1	1	10
Turnip	1	1	1	10
Sweet Potato	1	0.9	0.95	10
Carrot	0.89	0.89	0.89	9
Accuracy			**0.97**	**351**
Macro avg	**0.98**	**0.97**	**0.97**	**351**
Weighted avg	**0.97**	**0.97**	**0.97**	**351**

B. Confusion Matrix for Swin Transformer and ResNet

Figure 4.1 presents the confusion matrix for the Swin Transformer model, which successfully classified nearly all classes with high accuracy and minimal errors. Conversely, Figure 4.2 depicts the confusion matrix for ResNet, which showed several classification errors, particularly in classes like "Apple," "Chili," and "Potato," though its overall performance remained reasonably good.

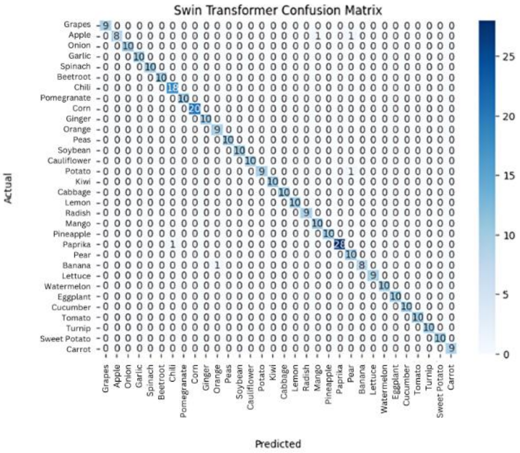


Figure I. Confusion Matrix Swin Model

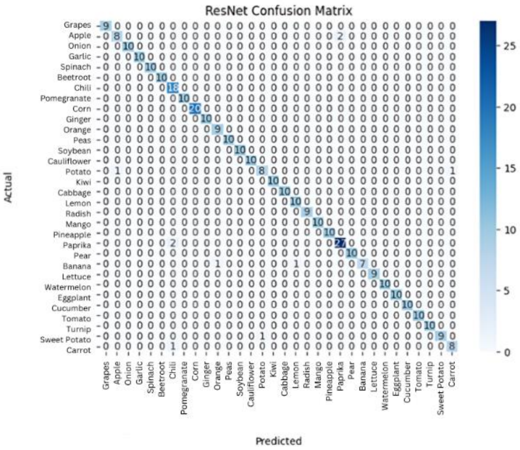


Figure II. Confusion Matrix ResNet Model

C. Evaluation Metrics and Confusion Matrix for SVM

The evaluation results of the SVM meta-learner, which combined Swin Transformer and ResNet, are presented in Figure 4.3 The confusion matrix indicates excellent classification performance across most classes, with minor errors in "Apple" and "Potato" classes. The metrics evaluation in Table 4.3 showed an overall accuracy of 98%, with average precision, recall, and F1-score values reaching 0.98 across all classes. The meta-learner approach consistently improved classification performance compared to individual models.

TABLE III. EVALUATION METRICS SVM MODEL

Class	Precision	Recall	F1-Score	Support
Grapes	1	1	1	9
Apple	0.89	0.8	0.84	10
Onion	1	1	1	10
Garlic	0.91	1	0.95	10
Spinach	1	1	1	10
Beetroot	1	1	1	10
Chili	0.95	1	0.97	18
Pomegranate	1	1	1	10
Corn	1	1	1	20
Ginger	1	1	1	10
Orange	1	1	1	9
Peas	1	1	1	10
Soybean	1	1	1	10
Cauliflower	1	1	1	10
Potato	1	0.9	0.95	10
Kiwi	0.91	1	0.95	10
Cabbage	1	1	1	10
Lemon	0.91	1	0.95	10
Radish	1	1	1	9
Mango	0.91	1	0.95	10
Pineapple	1	1	1	10
Paprika	1	0.97	0.98	29
Pear	1	1	1	10
Banana	1	0.78	0.88	9
Lettuce	1	1	1	9
Watermelon	1	1	1	10
Eggplant	1	1	1	10
Cucumber	1	1	1	10
Tomato	1	1	1	10
Turnip	1	1	1	10
Sweet Potato	1	1	1	10
Carrot	1	1	1	9
Accuracy			**0.98**	**351**
Macro avg	**0.98**	**0.98**	**0.98**	**351**
Weighted avg	**0.98**	**0.98**	**0.98**	**351**

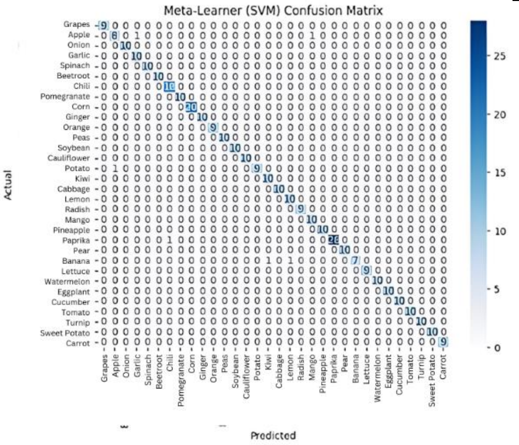
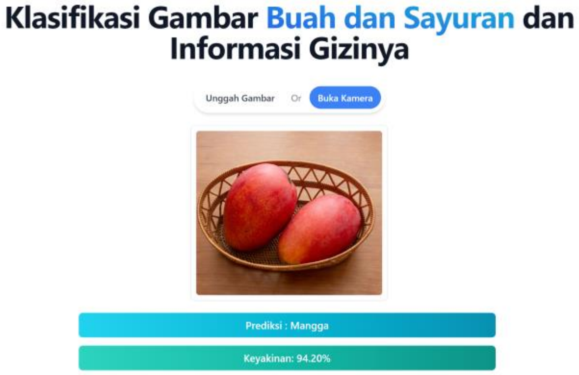


Figure III. Evaluation Metrics SVM Model

Web Application Evaluation

The ensemble learning model in this project combined Swin Transformer and ResNet as base models, with SVM as the meta-learner using the stacking technique. Swin Transformer excelled in recognizing complex visual patterns, while ResNet leveraged residual architecture for accurate predictions. Their combination through SVM resulted in more precise predictions for the dataset, comprising 32 fruit and vegetable classes, such as mango, grapes, and carrots, trained using fine-tuning for optimal performance.

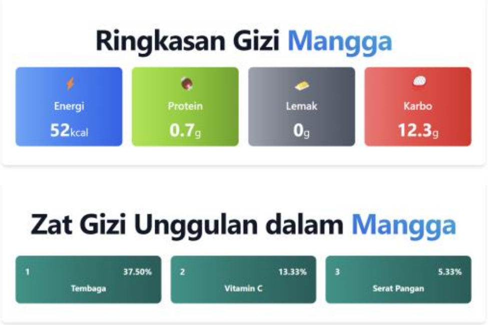


Klasifikasi Gambar Buah dan Sayuran dan Informasi Gizinya

Prediksi : Mangga

Keyakinan: 94.20%

Figure IV. Prediction Result



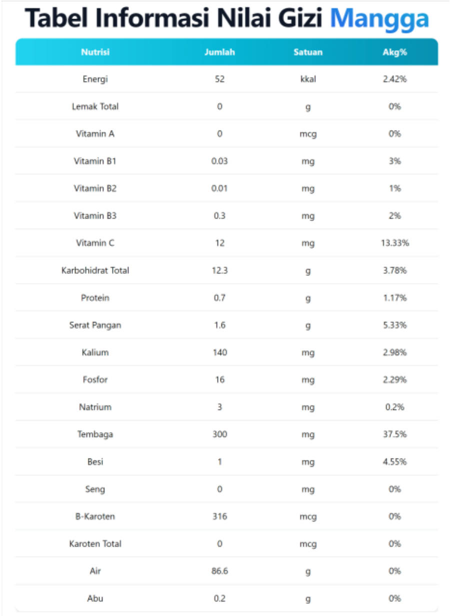
Ringkasan Gizi Mangga

Energi	Protein	Lemak	Karbo
52kcal	0.7g	0g	12.3g

Zat Gizi Unggulan dalam Mangga

1	2	3
37.50%	13.33%	5.33%
Tembaga	Vitamin C	Serat Pangan

Figure V. Nutrition Summary for Mango



Tabel Informasi Nilai Gizi Mangga

Nutrisi	Jumlah	Satuan	Akg%
Energi	52	kkal	2.42%
Lemak Total	0	g	0%
Vitamin A	0	mcg	0%
Vitamin B1	0.03	mg	3%
Vitamin B2	0.01	mg	1%
Vitamin B3	0.3	mg	2%
Vitamin C	12	mg	13.33%
Karbohidrat Total	12.3	g	3.78%
Protein	0.7	g	1.17%
Serat Pangan	1.6	g	5.33%
Kalium	140	mg	2.98%
Fosfor	16	mg	2.29%
Natrium	3	mg	0.2%
Tembaga	300	mg	37.5%
Besi	1	mg	4.55%
Seng	0	mg	0%
B-Karoten	316	mcg	0%
Karoten Total	0	mcg	0%
Air	86.6	g	0%
Abu	0.2	g	0%

Figure VI. Nutrition Table for Mango

This model was integrated into a web application using Flask (backend) and ReactJS (frontend). Users can upload fruit or vegetable images through features like "Upload Image" or "Open Camera." The application processes the images using the ensemble model and displays prediction results with confidence levels, such as "Mango (94.20%)." Additionally, the application provides comprehensive nutritional information, including energy, protein, vitamins, and minerals, to enhance users' understanding of the food's nutritional value.

The developed ensemble learning model demonstrated strong performance across multiple evaluation metrics, with the SVM meta-learner achieving an overall accuracy of 98% and high precision, recall, and F1-scores (average 0.98). The Swin Transformer model alone achieved 99% accuracy, excelling in most classes, while ResNet achieved 97%, with some errors observed in classes like "Apple," "Chili," and "Potato." These results highlight the capability of the Swin Transformer to handle complex visual patterns effectively, and the residual architecture of ResNet provided stability in feature extraction. However, the improvement in accuracy from the SVM meta-learner demonstrates the effectiveness of the stacking ensemble approach, which leverages the strengths of both base models to compensate for their individual weaknesses.

Despite its high accuracy, the model exhibited some limitations. The confusion matrix analysis revealed that certain classes with similar visual characteristics, such as "Apple" and "Potato," were occasionally misclassified. This suggests that while the ensemble model enhances overall performance, the feature differentiation for visually similar classes remains a challenge. Increasing dataset size and incorporating additional features, such as texture or contextual information, could improve classification for such challenging classes.

Another observed limitation is the dataset's size and scope, comprising only 32 fruit and vegetable classes with limited images per class (100 for training and 10 each for validation and testing). This restricted dataset limits the model's ability to generalize to unseen classes or suboptimal conditions, such as variations in lighting or image quality. Expanding the dataset and employing advanced augmentation techniques, such as synthetic image generation or domain-specific augmentation, would enhance the model's robustness and generalization capabilities.

Furthermore, while the model is relatively accurate in predicting cropped images, the accuracy significantly decreases for very small fragments. This indicates a need for refining the model to handle incomplete or fragmented inputs better. Techniques such as multi-scale feature extraction or attention mechanisms could improve performance in such cases.

The results of this research align with previous studies [16], [17], [18], where stacking ensemble learning demonstrated strong performance in image classification. Similarly, the Swin Transformer, used as a base model, also performed well in image classification, consistent with the findings in [19], [20].

From an application perspective, the deployment using the free PythonAnywhere hosting service posed operational challenges. The hosting environment requires manual reloads after inactivity and has limited uptime, which affects accessibility for end users. Migrating to a more stable hosting solution with features like automated scaling and higher uptime would ensure a smoother user experience and facilitate broader adoption of the application.

In conclusion, the ensemble learning model effectively leverages the strengths of Swin Transformer and ResNet through the SVM meta-learner, achieving high accuracy and demonstrating the potential for real-world applications. However, addressing dataset limitations, refining model performance for challenging conditions, and improving hosting infrastructure are essential steps for future development to maximize the model’s utility and scalability

IV. CONCLUSIONS

This research successfully demonstrates the effectiveness of a stacking ensemble learning approach for fruit and vegetable image classification. By combining Swin Transformer and ResNet as base models with SVM as the meta-learner, the model achieved 98% accuracy, surpassing the performance of individual models. The Swin Transformer excelled in extracting complex visual patterns, while ResNet provided stable feature extraction. The model was integrated into a web application that not only offers accurate classification but also provides detailed nutritional information, making it a practical tool for promoting healthy eating habits.

Despite its success, the study highlights areas for improvement. The limited dataset size and variety restrict the model’s generalization capabilities, and accuracy diminishes with fragmented images. Additionally, the use of a free hosting platform poses operational challenges for accessibility. Future research should focus on expanding the dataset, optimizing model robustness for challenging conditions, and adopting a more stable hosting infrastructure to enhance the application’s scalability and real-world impact.

REFERENCES

[1] UNICEF, WHO, and The World Bank Group, “Levels and trends in child malnutrition: UNICEF / WHO / The World Bank Group joint child malnutrition estimates: key findings of the 2021 edition,” Geneva, May 2021. Accessed: Jan. 02, 2025. [Online]. Available: https://iris.who.int/handle/10665/341135

[2] Najua Ismail, “Study: Children In Southeast Asia Suffer From Malnutrition,” https://ova.galencentre.org/study-children-in-southeast-asia-suffer-from-malnutrition/. Accessed: Jan. 02, 2025. [Online]. Available: https://ova.galencentre.org/study-children-in-southeast-asia-suffer-from-malnutrition/

[3] FAO and Ministry of Social Development and Family of Chile, Promoting safe and adequate fruit and vegetable consumption to improve health. FAO; Ministerio de Desarrollo Social y Familia de Chile MDSF ;, 2021. doi: 10.4060/cb7946en.

[4] J. Steinbrener, K. Posch, and R. Leitner, “Hyperspectral fruit and vegetable classification using convolutional neural networks,” Comput. Electron. Agric., vol. 162, pp. 364–372, Jul. 2019, doi: 10.1016/j.compag.2019.04.019.

[5] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” Eng. Appl. Artif. Intell., vol. 115, pp. 105–151, Oct. 2022, doi: 10.1016/j.engappai.2022.105151.

[6] I. D. Mienye and Y. Sun, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects,” IEEE Access, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[7] P. Wu, R. Ma, and T. T. Toe, “Stacking-Enhanced Bagging Ensemble Learning for Breast Cancer Classification with CNN,” in 2023 3rd International Conference on Electronic Engineering (ICEEM), IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICEEM58740.2023.10319517.

[8] D. H. Wolpert, “Stacked generalization,” Neural Networks, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[9] [9] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” Mar. 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep Residual Learning for Image Recognition,” Jun. 2016.

[11] J. A. K. Suykens, “Nonlinear modelling and support vector machines,” in IMTC 2001. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No.01CH 37188), Budapest, Hungary: IEEE, May 2001, pp. 287–294. doi: 10.1109/IMTC.2001.928828.

[12] Tian Yu Liu and Baharan Mirzasoleiman, “Data-Efficient Augmentation for Training Neural,” vol. 35, pp. 5124–5136, 2022, Accessed: Jan. 02, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/2130b8a44e2e28e25dc7d0ee4eb6d9cf-Paper-Conference.pdf

[13] Alec Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” vol. 139, 2021, Accessed: Jan. 02, 2025. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[14] Z. M. Khaing, Y. Naung, and P. H. Htut, “Development of control system for fruit classification based on convolutional neural network,” in 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), IEEE, Jan. 2018, pp. 1805–1807. doi: 10.1109/EIConRus.2018.8317456.

[15] H. M. Zawbaa, M. Hazman, M. Abbass, and A. E. Hassanien, “Automatic fruit classification using random forest algorithm,” in 2014 14th International Conference on Hybrid Intelligent Systems, IEEE, Dec. 2014, pp. 164–168. doi: 10.1109/HIS.2014.7086191.

[16] H. W. Zhang et al., “Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases,” Sci. Rep., vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-80210-x.

[17] S. C. Hidayati, A. A. I. Rahardja, and N. Suciati, “Stacking-based ensemble learning for identifying artist signatures on paintings,” Indones. J. Electr. Eng. Comput. Sci., vol. 36, no. 3, pp. 1683–1693, 2024, doi: 10.11591/ijeecs.v36.i3.pp1683-1693.

[18] N. K. Sarkar, M. M. Singh, and U. Nandi, “Ensemble Transfer Learning for Image Classification,” pp. 25–36, doi: 10.37936/ecti-cit.2025191.257836.

[19] V. Sharma, A. K. Tripathi, H. Mittal, and L. Nkenyereye, “SoyaTrans: A novel transformer model for fine-grained visual classification of soybean leaf disease diagnosis,” Expert Syst. Appl., vol. 260, no. August 2024, p. 125385, 2025, doi: 10.1016/j.eswa.2024.125385.

[20] T. Ji, H. Fang, R. Zhang, J. Yang, Z. Wang, and X. Wang, “Plastic waste identification based on multimodal feature selection and cross-modal Swin Transformer,” Waste Manag., vol. 192, no. November 2024, pp. 58–68, 2025, doi: 10.1016/j.wasman.2024.11.027.