Maastricht University

Department of Advanced Computer Sciences

Master's Program in Artificial Intelligence
and Data Science for Decision Making

Data Fusion Report

Theodor Antoniou,
Fivos Tzavellos,
Spyridon Giagtzoglou,
Ioannis Grepsios,
Stavros Nikoletopoulos

March 29, 2023

1 Introduction & Motivation

The increasing global population and the corresponding rise in food demand necessitate a more sustainable and efficient approach to agriculture. Climate change, soil degradation, and water scarcity further amplify the challenges faced by farmers in meeting this growing demand. Plant monitoring, as an integral aspect of precision agriculture, offers a promising solution to address these challenges.

By closely observing and analyzing individual plants or specific zones within a field, plant monitoring enables farmers to detect and address issues such as nutrient deficiencies, pests, diseases, and water stress in a targeted manner. This targeted approach allows for the efficient allocation of resources, reducing waste and minimizing the environmental impact of agricultural practices.

Moreover, plant monitoring facilitates the early identification of under-performing or abnormal plants, allowing farmers to take corrective measures or allocate resources to more productive plants, thereby optimizing overall crop yields. The adoption of plant monitoring technologies also contributes to reducing labor costs, as farmers can automate various time consuming tasks such as scouting, data collection, and analysis. We can take inspiration from the work of Barbedo et al [1], who used data fusion to bring together data from sensors taken for agricultural purposes. Feng et al [2] used raw data fusion, feature fusion and decision fusion to fuse spectral features for the categorization of rice diseases.

2 Problem Statement

The goal of this assignment is to develop a model for the binary classification of tomato seedlings using images as input. The classification task involves determining if the seedlings are normal (class 1,2) or abnormal (class 3,4). In order to enhance the model's performance and optimize the overall classification accuracy, we will utilize data fusion techniques. Specifically, we will employ High-Level Fusion and Mid-Level Fusion strategies in our approach.

By integrating High-Level Fusion and Mid-Level Fusion techniques, our model is expected to yield a more accurate and robust binary classification of tomato seedlings, distinguishing between normal and abnormal growth effectively. This will facilitate early detection of potential issues, enabling timely interventions to ensure the successful growth and development of the tomato plants.

A successful classification of tomato plants as normal and abnormal (in other words being able to tell if it will be fruitful or not) would benefit the farmer in the sense that a more efficient crop is yielded, less resources are wasted and pest infestation is regulated. The ideal scenario would include less involvement from manual laborers and more automated procedures that cut the cost of production by a large margin.

3 Data Preprocessing

The overall idea and approach for the preprocessing part was to take the images provided to us by Itility and extract some insightful features using the PlantCV library, firstly individually on one grayscale and one coloured image (just to see how the filters alter the image) and then as a pipeline to extract the features for all the images. After the extraction comes the data-cleaning, which we will explain in detail in the following sections.

3.1 Dataset

The dataset that was provided to us contains 994 gray-scale images taken from the side of the plant and 994 colored images taken from the top, along with the labels from 4 experts evaluating the plants with a scale from 1 - 4 (good to bad). This is our starting point and our main source of information.

3.2 Grayscale Images

For the black and white images we used some functions taken from the PlantCV library for removing the background, thresholding the image, using filters such as Laplace and Sobel, lowpass, inverting, removing noise, removing the edges, identifying objects, defining the Region Of Interest, performing distance transformation for object segmentation and lastly performing NIR Signal and shape analysis. We tune our parameters using this test for a single image, and then create the pipeline to scale it up to the size of the dataset and run through every grayscale image. The result is a dataframe that has attributes (277 in total) such as size of area, width, height, nir frequencies, center of mass etc.

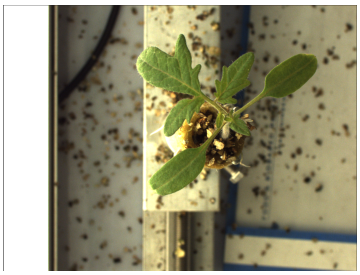
3.3 Colored Images

For the colored images we again use some functions derived from the PlantCV library, like for converting RGB to HSV and extracting the saturation channel, binary thresholding the plant from the background, cleaning noise with median blur, extracting colour frequencies, filtering out the dark and the bright noise, fill for small object removal and some of the same shape identifying analysis we did for the black and white case. We again formulate the pipeline to pass the filters through every image and are left with a dataset where each row (994 in total) represents one image and each column (1782 in total) is an attribute like area, perimeter, colour frequencies etc.

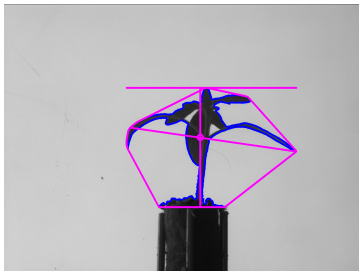
3.4 Data Cleaning

For cleaning up the data we remove the rows that had plants that were out of the bounds of the picture frame, remove the attributes that only contained values of zero and concatenate the dataset with the expert labels. To run them through our models we removed some attributes that were considered noisy (they were driving the accuracy down and we speculate it was duo to the different units of measurement) like the size and area of the plants. In figure 1 we can notice the grayscale, colored and shape filtered image of a plant considered normal (average expert label: 1). In figure 2 we notice the grayscale, colored and shape filtered image of a plant considered abnormal (average expert label: 4).

(a) Grayscale Image

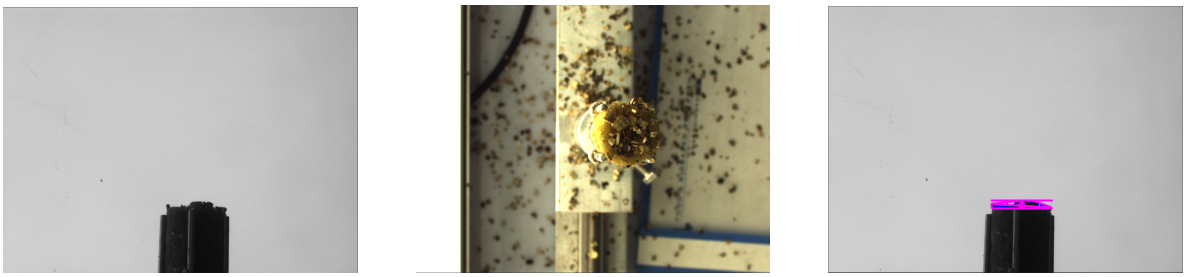


(b) Colored Image



(c) Shape Filter

Figure 1: Normal Plant



(a) Grayscale Image (b) Colored Image (c) Shape Filter

Figure 2: Abnormal Plant

4 Data Fusion

In this section, we present our approach to binary classification of tomato seedlings using High-Level Fusion and Mid-Level Fusion techniques. We detail the implementation of these techniques, their rationale, and the results obtained.

Implementation of High-Level Fusion and Mid-Level Fusion Techniques:
High-Level Fusion:

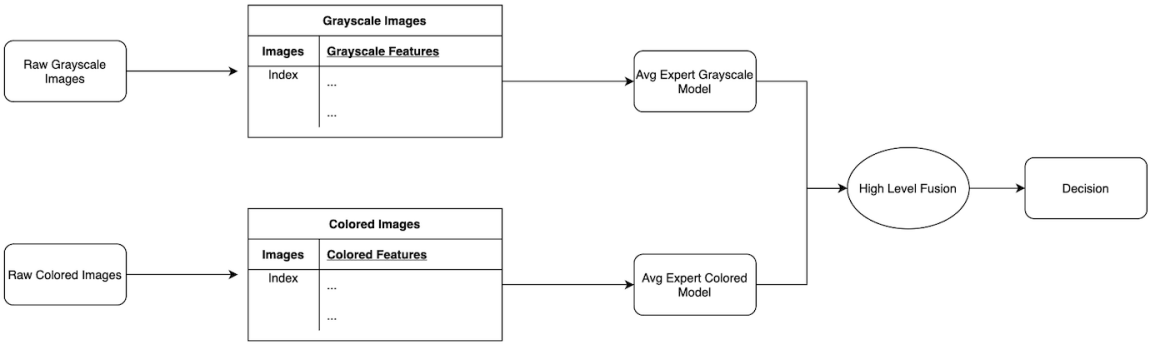


Figure 3: High Level Image Fusion

- We trained four different classifiers (Logistic Regression, Random Forest, MLPClassifier, AdaBoostClassifier) on the tomato seedling images.
- We used a validation set to determine the optimal weights for each classifier based on their individual performance.
- We combined the output decisions of the four classifiers using weighted majority voting to achieve a more accurate final decision.

Mid-Level Fusion:

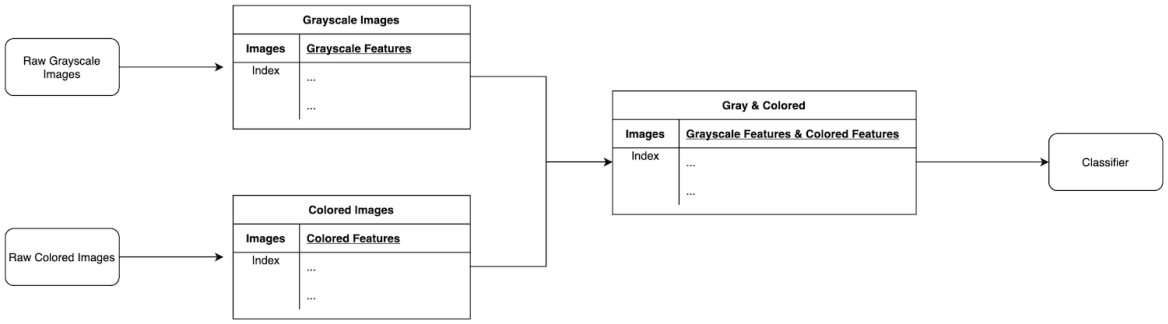


Figure 4: Mid Level Image Fusion

- We trained again four different classifiers (Logistic Regression, Random Forest, MLPClassifier, AdaBoostClassifier) on the tomato seedling images.
- We used a validation set to determine the optimal weights for each classifier based on their individual performance.
- We combined the output decisions of the four classifiers using the average of the experts.

Mid-Level & High-Level Fusion:

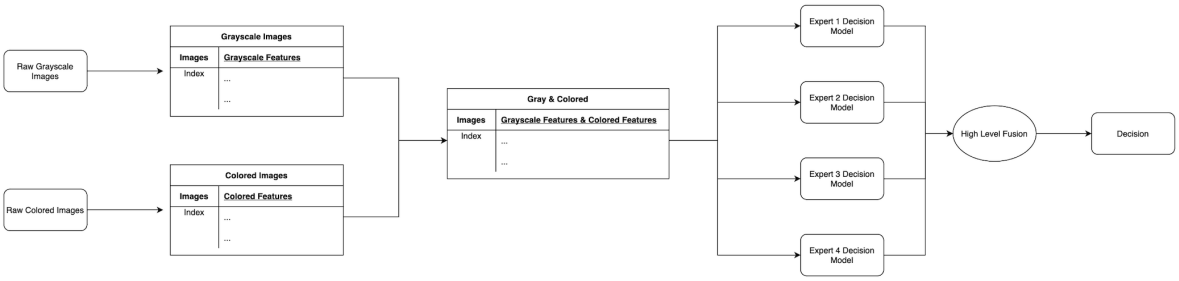


Figure 5: Mid-High Level Expert Fusion

- Feature extraction: First, we extracted various features from the input images. In the case of tomato seedlings, relevant features could include color frequencies, texture features, and shape features.
- Feature combination: Combined the extracted features into a unified feature set. We implemented this by using different techniques such as feature concatenation, feature selection,and feature fusion. Feature concatenation is a simple method that involves concatenating feature vectors from different feature types to form a single, longer feature vector. Feature selection or fusion methods may involve selecting or combining the most relevant features to create a more compact and effective feature set.
- Classifier training: Trained classifiers, such as Logistic Regression, RandomForest Classifier, MLPClassifier and AdaBoost, on the combined feature set.
- Model evaluation: We evaluated the performance of the trained classifier using appropriate performance metrics such as accuracy, F1 score, recall, and AUC.

Results: The results of our classification model, which incorporated both High-Level and Mid-Level Fusion techniques, demonstrated improved accuracy compared to using a single classifier or a single feature type.

	Random Forest		MLP Classifier	
	Accuracy	F1	Accuracy	F1
Individual	0.9643	0.9548	0.9286	0.9009
Weighted Majority Voting	0.9762	0.9762	0.9603	0.9482
Bayesian	0.9603	0.9510	0.9603	0.9492

	Logistic Regression		AdaBoost	
	Accuracy	F1	Accuracy	F1
Individual	0.9246	0.9039	0.9722	0.9653
Weighted Majority Voting	0.9444	0.9274	0.9841	0.9799
Bayesian	0.9286	0.9146	0.9683	0.9611

Table 1: Classification results for High Level Image Fusion

	Random Forest		MLP Classifier	
	Accuracy	F1	Accuracy	F1
Individual	0.9643	0.9570	0.9425	0.9281
Majority Voting	0.9623	0.9550	0.9425	0.9278
Weighted Majority Voting	0.9544	0.9451	0.9425	0.9278
Bayesian	0.9623	0.9550	0.9464	0.9340

	Logistic Regression		AdaBoost	
	Accuracy	F1	Accuracy	F1
Individual	0.9345	0.9225	1.0000	1.0000
Majority Voting	0.9306	0.9169	0.9742	0.9692
Weighted Majority Voting	0.9504	0.9394	0.9742	0.9692
Bayesian	0.9306	0.9169	0.9702	0.9638

Table 2: Classification results for Mid-High Level Expert Fusion

Classifier	Accuracy	F1	Recall	AUC
LogisticRegression	0.949495	0.897959	0.916667	0.938333
RandomForestClassifier	0.959596	0.916667	0.956522	0.958524
MLPClassifier	0.954545	0.901099	1.000000	0.971338
AdaBoostClassifier	0.964646	0.927835	0.957447	0.962167

Table 3: Classification results for Mid-Level Image Fusion

Analysis of Techniques: High-Level Fusion worked better than using a single classifier because it leveraged the strengths of different classifiers to enhance the overall performance. By combining the output decisions of multiple classifiers, our model was able to compensate for the weaknesses of individual classifiers and reduce the chances of overfitting. Especially on the combination with mid-Level image fusion, **AdaBoostClassifier** manages **F1 Score: 1.0000** perfectly classifying all instances. Keep in mid that this is a 7.7% increase when compared to the F1 Score of Adaboost for the mid level fusion.

Mid-Level Fusion was effective because it utilized complementary information from different features, leading to a more robust representation of the input data. By combining color, texture, and shape features, the model could capture various aspects of the tomato seedlings, making it more sensitive to the differences between normal and abnormal growth patterns.

In conclusion, the Mid-Level Fusion approach effectively improves the classification performance for tomato seedlings by combining different features. The ensemble methods, such as **AdaBoostClassifier** and **RandomForestClassifier**, appear to be particularly well-suited for this task, as they achieve high accuracy, F1 scores, and AUC values. Last but not least,The MLPClassifier demonstrates the best recall (1.000000), which means it correctly identifies all abnormal seedlings. This is a crucial aspect in a classification task where it is important to avoid missing any abnormal seedlings.

5 Future Work

Our study aimed to develop a binary classification model for tomato seedlings based on images using two data fusion techniques, Mid-level fusion and High-level fusion. Despite achieving promising results, several areas of improvement could be explored in future work to enhance the accuracy and robustness of the classification system. Firstly, advanced feature extraction techniques such as CNN Autoencoder or Transformers like BERT could be incorporated to create a more accurate semantic representation of the raw images. Secondly, more complex deep learning classifiers, such as CNN and LSTMs, could be utilized to improve the overall accuracy of the classification system. Thirdly, Federated Learning could be examined to train the model on data distributed across multiple sources while preserving data privacy. Finally, multiple axis data fusion architectures that combine both image and expert information simultaneously could be applied for more accurate classification, overcoming the limitation of one-axis fusion methods.

In addition to the aforementioned enhancements, further improvement of our tomato seedling classification system could involve the implementation of more advanced and sophisticated models. These models could incorporate techniques such as transfer learning or ensembling to further improve the accuracy of the classification system. Furthermore, it may be beneficial to explore methods that perform more than one axis fusion, such as combining image and expert information at the same time, to improve the overall performance of the classification system. By exploring these avenues of research, we can continue to refine and improve our tomato seedling classification system, leading to potential benefits for the agricultural industry.

One last thing we want to note, is that after gaining the full view of the results, we came up with one insight that could be beneficial.The characteristic that maybe can help from a business point of view (lowering the costs and computational requirements) is to run our models by using resized lower resolution images, to compare the impact. If the results are good enough we could proceed with lower resolution cameras minimising the costs.

References

[1] Jayme Barbedo. "Data Fusion in Agriculture: Resolving Ambiguities and Closing Data Gaps". In: *Sensors* 22 (Mar. 2022), p. 2285. DOI: 10.3390/s22062285.

[2] Lei Feng et al. "Investigation on Data Fusion of Multisource Spectral Data for Rice Leaf Diseases Identification Using Machine Learning Methods". In: *Frontiers in Plant Science* 11 (2020). ISSN: 1664-462X. DOI: 10.3389/fpls.2020.577063. URL: https://www.frontiersin.org/articles/10.3389/fpls.2020.577063.