`

Data Fusion

Group Assignment

Athanasios Athanasopoulos i6275313

Chrysanthi Foti i6332640

Eleftherios Tetteris i6295677

Athanasopoulos Nikolaos i6310104

Vasileios Chaidas i6299447

30/03/2023

`

Table of contents

`

Introduction

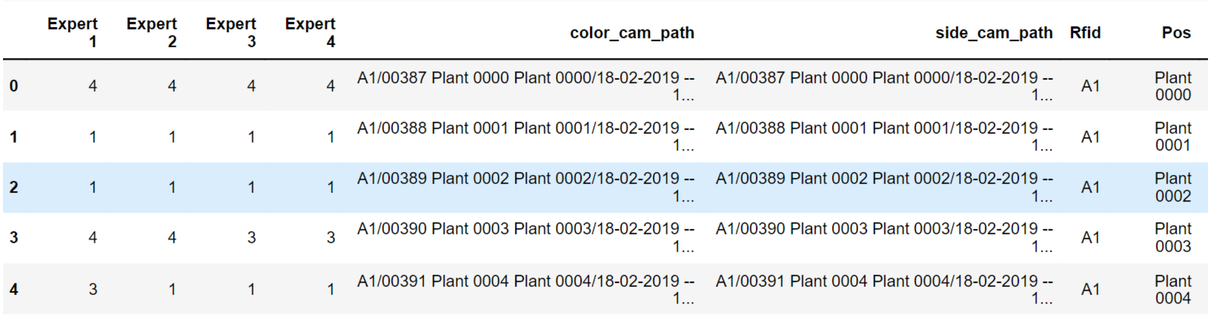
The classification of plants as normal or abnormal is an essential task in the field of plant biology. The classification process is usually done by experts (biologists) who possess a vast knowledge about various plants and their characteristics. However, relying on individual expert decisions may not always provide the most accurate results due to the possibility of human error or personal bias. Thus, it is very often necessary and crucial to fuse expert decisions in order to achieve more accurate and generalised plant classification results.

Motivation

The use of fusion methods in the field of plant classification has gained increasing attention in recent years due to the need for accurate and reliable results. The classification of plants is vital in various fields such as agriculture, botany and environmental science. The ability to accurately classify plants as normal or abnormal can aid in disease diagnosis, pest control, and ecosystem monitoring. However, relying on a single expert's decision to classify plants may result in inaccuracies due to the subjective nature of individual experts. Therefore, combining the decisions of multiple experts through fusion methods can lead to more accurate and reliable results. This study aims to use Low, Mid and High-level fusion methods along with the subcombination decisions (discussed later) to combine the decisions of four experts to achieve a more robust model for plant classification. Additionally, this study utilises various feature extraction techniques from plantCV from both RGB and Grayscale images respectfully to aid in the fusion process. Ultimately, the goal of this study is to show the preference of fusion techniques in achieving more accurate and reliable outcomes compared to relying on individual expert decisions. All in all, by demonstrating the superiority of fusion techniques, the client can make better-informed decisions based on the business value that is being delivered.

Methods and Approach

In order to see if fusion techniques can improve the overall prediction of the plant classification problem (healthy or unhealthy), a wide range of fusion methods was used. The dataset provided for this assignment has a specific structure which can be seen in Fig1.



	Expert 1	Expert 2	Expert 3	Expert 4	color_cam_path	side_cam_path	Rfid	Pos
0	4	4	4	4	A1/00387 Plant 0000 Plant 0000/18-02-2019 -- 1...	A1/00387 Plant 0000 Plant 0000/18-02-2019 -- 1...	A1	Plant 0000
1	1	1	1	1	A1/00388 Plant 0001 Plant 0001/18-02-2019 -- 1...	A1/00388 Plant 0001 Plant 0001/18-02-2019 -- 1...	A1	Plant 0001
2	1	1	1	1	A1/00389 Plant 0002 Plant 0002/18-02-2019 -- 1...	A1/00389 Plant 0002 Plant 0002/18-02-2019 -- 1...	A1	Plant 0002
3	4	4	3	3	A1/00390 Plant 0003 Plant 0003/18-02-2019 -- 1...	A1/00390 Plant 0003 Plant 0003/18-02-2019 -- 1...	A1	Plant 0003
4	3	1	1	1	A1/00391 Plant 0004 Plant 0004/18-02-2019 -- 1...	A1/00391 Plant 0004 Plant 0004/18-02-2019 -- 1...	A1	Plant 0004

Fig1: The structure of the dataset

`

The dataset consists of 994 different tomato seedlings with each one having both a top view coloured image and a side view grayscale image. Four experts gave their opinion based on the images on whether each tomato seedling is normal with a score of 1 or 2 and abnormal with a score of 3 or 4. Moreover, each plant also belongs to a certain location (“RFID” column, 8 in total). The “pos” column simply keeps track of the index of the plant. This structure can be manipulated for data fusion in various ways. The axis of the fusion could be the data (or the features that can be extracted from the data), the experts, the RFID and of course every possible combination of these aforementioned axes.

In this assignment, the axes considered for data fusion were the Experts and the data (their features). In particular, the following methods were used:

- Expert Fusion: The labels from the Experts can be either fused before the training of the model (Low Level Expert fusion) or after training a model for each Expert (High Level Expert fusion). Both of these methods were used and their results were compared to the respective individual models from each Expert (Fig2).

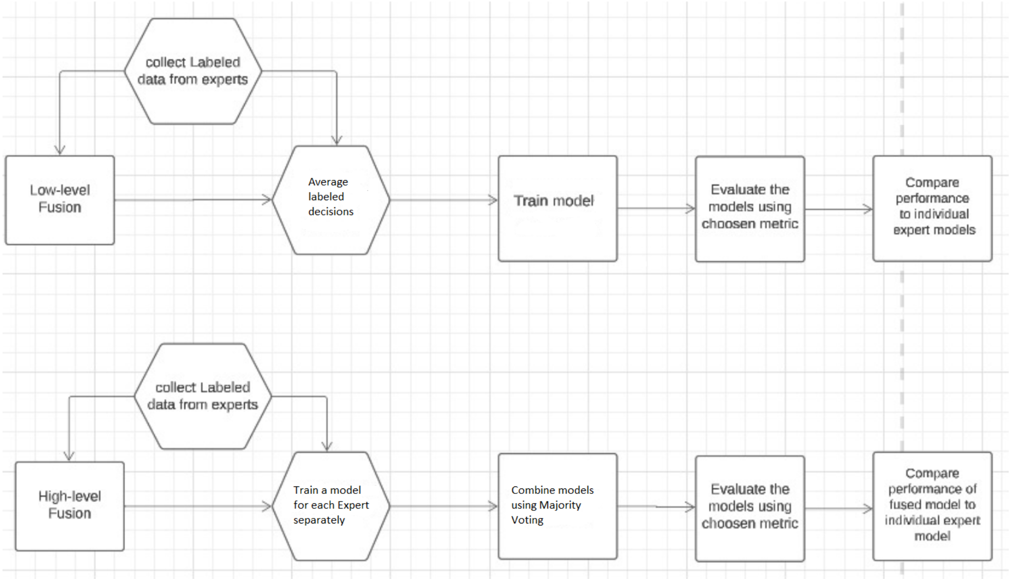

collect Labeled data from experts
Low-level Fusion
Average labeled decisions
Train model
Evaluate the models using choosen metric
Compare performance to individual expert models
collect Labeled data from experts
High-level Fusion
Train a model for each Expert separately
Combine models using Majority Voting
Evaluate the models using choosen metric
Compare performance of fused model to individual expert model

Fig2: Expert Fusion

- Image Fusion: The images provided can be either used “as is” to train the model (Low Level) or features can be extracted from them, fused for both coloured and grayscale images and be used for training the model (Mid Level fusion, Fig3).

`

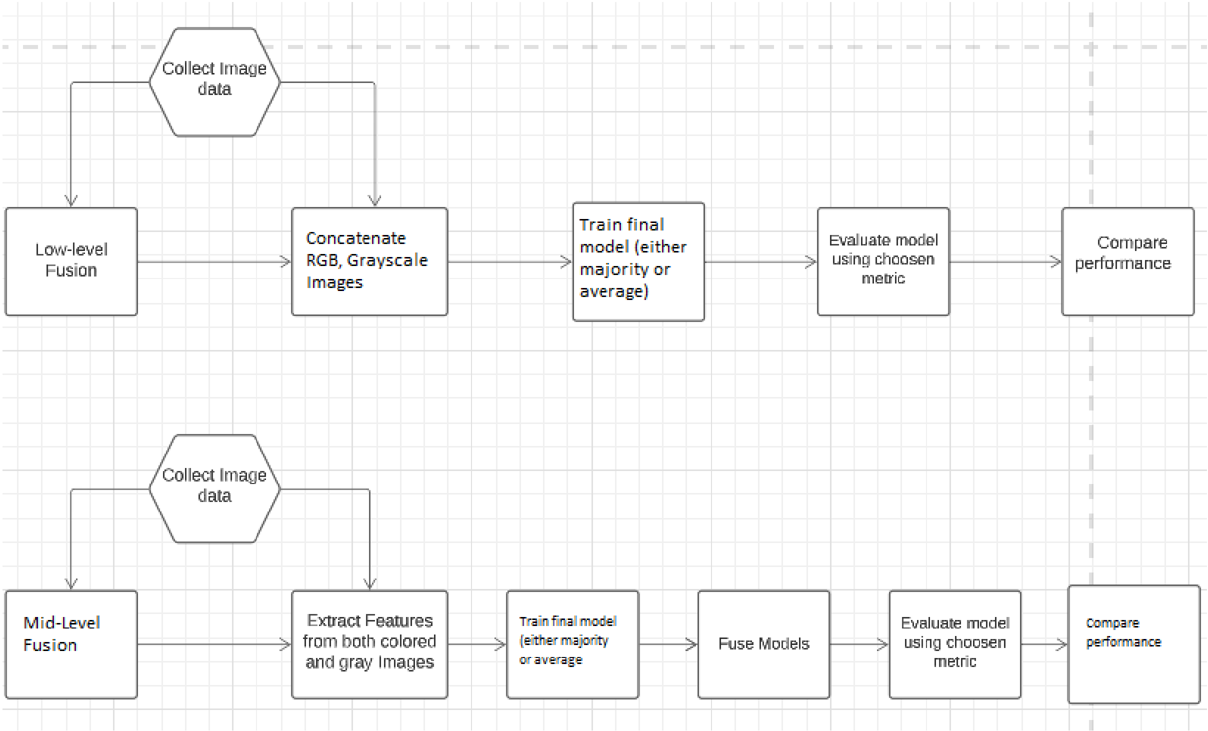

Collect Image data
Low-level Fusion
Concatenate RGB, Grayscale Images
Train final model (either majority or average)
Evaluate model using choosen metric
Compare performance
Collect Image data
Mid-Level Fusion
Extract Features from both colored and gray Images
Train final model (either majority or average
Fuse Models
Evaluate model using choosen metric
Compare performance

Fig3: Image Fusion

It can be easily seen that for this assignment, all low level, mid level and high level fusion techniques were used. In all the cases, Logistic Regression, Decision Tree and Random Forests classifiers were tested to obtain the final predictions and will be discussed further in following sections. The next chapter discusses the data preprocessing and feature extraction techniques used, while the observed results can be seen in the Results section.

Data Preprocessing-Feature Extraction

In order to perform the final classification, the data has to be preprocessed. The first thing is to convert the scores from 1,2,3,4 to 0 and 1 in order to perform binary classification. The normal (1,2) and the abnormal (3,4) classes were converted to 1 and 0 respectively. The next thing is to investigate whether there is class imbalance by summing all the 1s in each expert and then dividing by the total number of photos (994). The percentage of 1s for each Expert is around 70%, which means that the class 0 is under-represented. As a result, random oversampling was used to achieve a class representation of 50%-50%. This random oversampling was performed when the final features were defined for each method.

Features of both grayscale and RGB images were extracted to perform the mid-level fusion as mentioned before. To do so, the approach was based on the PlantCV tutorials' general concept. The tutorials followed were the NIR tutorial (feature extraction for grayscale images) and the VIS tutorial (feature extraction for the colour RGB images). The feature extraction is carried out for each plant in the data (994 plants * 2 images per plant). The optimal values for the various parameters for each step of the tutorials (for example threshold for binary images etc) were found experimentally via trial and error. The final data

`

is added to the main dataframe, and the image arrays are flattened, concatenated and saved for easier use. In the end, each image has 2,999,040 features. Of course, classification on these features would be impossible, so dimensionality reduction was performed.

Due to the slow performance of regular PCA and Truncated SVD (which took 20 minutes to find 100 principal components), an alternative method called Kernel PCA was employed. This technique uses kernels and randomization methods to find 100 principal components that produce the higher variance over a sampling procedure, and ultimately runs in 4 minutes instead of 20 for all images. The final variance of the dataset that is explained by the 100 principal components is 50%, which is usually low for machine learning purposes (typically 80-90% is desired), but as it can be seen in the Results chapter the models trained on these 100 components perform very well. It should be noted that the models using mid-level data fusion were trained both on the regular PCA components and the KernelPCA components, and it was found that the KernelPCA models perform just as well as the models trained on the regular PCA components. After the components are found, the data is again saved for easier access (so there is no need to re-run the whole notebook to find the components again). Also, each process containing randomness is defined with a specific random state in order for the results to be reproducible.

Results

This chapter contains the results for each model used in this assignment. It should be noted that for each method, 3 types of classifiers were used, namely Random Forest, Logistic Regression and Decision Trees in order to compare results as mentioned earlier. It can be easily seen in the code that the Random Forest model outperforms the 2 other models in every case.

The first model uses high level Expert fusion with majority voting and the features being the original RGB images (resized to 1/9 of their original size for easier training because this resizing was found to not have any negative effect on the final accuracy scores). The final model combining the 4 Expert models' decisions is compared to each individual model in order to see if this type of fusion is indeed performing better than the individual models. The results for the Random Forest (which is always the best performing classifier) are:

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Expert 1	0.95	0.95	0.95
Expert 2	0.95	0.95	0.95
Expert 3	0.95	0.95	0.95
Expert 4	0.94	0.94	0.94
Majority Voting Fusion Model	**0.97**	**0.97**	**0.97**

Fig4: Results for Majority Voting (High Level Expert Fusion) with resized RGB images as features

`

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Expert 1	0.96	0.96	0.96
Expert 2	0.95	0.95	0.95
Expert 3	**0.97**	**0.97**	**0.97**
Expert 4	0.94	0.94	0.94
Majority Voting Fusion Model	0.95	0.95	0.95

Fig5: Results for Majority Voting (High Level Expert Fusion) with resized RGB and Grayscale Images concatenated as features (Low Level Image Fusion)

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Expert 1	0.95	0.95	0.95
Expert 2	0.96	0.96	0.96
Expert 3	**0.98**	**0.98**	**0.98**
Expert 4	0.96	0.96	0.96
Majority Voting Fusion Model	0.96	0.96	0.96

Fig6: Results for Majority Voting (High Level Expert Fusion) with KernelPCA components for features (Mid Level Image Fusion)

It can be seen that when using KernelPCA components, the final majority voting model does not outperform Expert's 3 model, which implies that the model from Expert 3 should be given more weight on the final decision, because it also outperforms all the other 3 experts' models (Experts 1, 2 and 4).

The next results which are presented concern the models trained when the labels of the Experts are averaged before training any model (Low Level expert fusion). This means that only a single "average" model is trained on the data, which is then compared to the respective individual models trained on the same features. The results are:

`

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Expert 1	0.95	0.95	0.95
Expert 2	0.95	0.95	0.95
Expert 3	0.95	0.95	0.95
Expert 4	0.94	0.94	0.94
Average Expert Label Model	**0.97**	**0.97**	**0.97**

Fig7: Results for Average Label (Low Level Expert Fusion) with resized RGB images for features

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Expert 1	0.95	0.95	0.95
Expert 2	0.95	0.95	0.95
Expert 3	0.95	0.95	0.95
Expert 4	0.94	0.94	0.94
Average Expert Label Model	**0.98**	**0.98**	**0.98**

Fig8: Results for Average Label (Low Level Expert Fusion) with resized RGB and Grayscale Images concatenated for features (Low Level Image Fusion)

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Expert 1	0.95	0.95	0.95
Expert 2	0.96	0.96	0.96
Expert 3	0.98	0.98	0.98
Expert 4	0.96	0.96	0.96
Average Expert Label Model	**0.99**	**0.99**	**0.99**

Fig9: Results for Average Label (Low Level Expert Fusion) with KernelPCA components for features (Mid Level Image Fusion)

`

We can see that the Average Label model outperforms the individual models both for normal image features and the KernelPCA components-features which means that this type of fusion is better. The Average Label model trained on KernelPCA components has a 99% accuracy, recall precision, F1 and AUC scores, indicating that it is indeed a very good model (almost impossible to be outperformed).

Taking all of this into account, it is worth noting that in cases where the individual expert model outperformed the fused model (Fig 5 & Fig 6), the observed results differ by only 1%.On the whole, all fused models resulted to a 95% or higher score indicating that a fused model is worthwhile to implement.

The reasons behind why some techniques work better than others could be many. First of all, all models achieve a final AUC score of ≥96%, which means that all fusion techniques worked great. But averaging the labels seems to achieve the best overall result because the final model is trained on "correct" data. When the High level Expert fusion is used, each model learns from the individual labels, which could be wrong. This means that maybe the final model is worse compared to when 1 model is trained with the average labels, which are more "correct", since we are using the average expert's opinions.

Deliver our experience to a meaningful business value

Investing in the right plants can help avoid waste of resources, prevent the spread of diseases and pests, and save manual labour. It is essential to understand the market demand and consumer preferences when selecting crops to invest in. Researching and analysing the soil quality, climate, and growing conditions of the region can also help determine the most suitable plants to grow. Leveraging technology and implementing models can assist in predicting plant health, yield, and quality, thereby helping farmers make informed decisions. Additionally, it is crucial to invest in crops that are resistant to diseases and pests to minimise the use of harmful chemicals. By prioritising these factors, food production can be scaled up, and it will not be limited by the availability of domain experts.

Delivering a business value to our collaborator company (Itility B.V.) means providing a solution that addresses a particular problem or a specific need that the company has, while also creating economic benefits or improving their overall operations. In this case, the client is likely interested in plant health monitoring by considering the robust model that scores an accuracy of 99% that almost perfectly distinguishes between healthy and unhealthy plants which provides significant benefits to the company. One of the benefits of having a robust model is that there may be no need for expert labelling of the images. This is because the model is accurately predicting 99% of the cases based on the patterns it has learned from the training data. By using the model to provide a final prediction(the business value), the client can save time and money that would have been spent on expert labelling. Achieving a robust and generalizable model requires a careful selection and tuning of algorithms and techniques for feature extraction, as well as appropriate training and validation, which are addressed by the model architecture. These aspects are stored and can be accessed at any time, providing relatively quick results with respect to the techniques being used.The robustness of the model depends primarily on the quality and consistency of the input images. In our use case, high-quality and consistent images are provided to us so as the model can easily be trained to accurately deliver potential healthy and unhealthy plants with a high degree of reliability without making any resistant parameter(image denoising). Thus, in our case, image capturing from a farmers perspective plays an important role in order to ensure that the model remains robust and reliable. Overall, our team delivered a reliable and robust model plan and architecture with appropriate fuse techniques that rely on feature extraction, training and validation. The business value of the model is its ability to predict plant health status with high accuracy, benefiting both the company and their respective clients.

Bonus

In order to find which labeler is the least “trustworthy”, the Cohen kappa measure between all Expert pairs is calculated and presented below. The Cohen kappa is calculated both for when the labels are 4 and again when the labels are binary. Fig10 contains the Cohen kappa.

Expert i	**Expert j**	**Cohen Kappa for 4 labels**	**Cohen Kappa for 2 labels**
1	2	0.7580	0.8514
1	3	**0.7514**	**0.8397**
1	4	0.7894	0.8446
2	3	0.7636	0.8852
2	4	0.7976	0.8842
3	4	0.7982	0.8632

Fig10: Cohen Kappa for all possible Expert Pairs

It can be seen that Expert 1 and 3 disagree more with each other than anybody else. Expert 1 also disagrees with Expert 2. This is an indication that one of them is the least “trustworthy” with regards to their prediction power. In order to find out who, the accuracy of all individual models of Experts have to be compared. The Expert with the highest number of models that have the lowest accuracy scores out of the individual models should be the least “trustworthy” which means that he needs to be excluded from all the models. This means that all the model structures need to be investigated (so not only Random Forest Models, but also Logistic Regression and Decision Trees). Looking at the Jupyter Notebook, it can be derived that the Expert with the highest number of models that perform the worst out of all 4 models is Expert 1, so this Expert should be excluded from the models, and these models should be retrained only based on Experts 2, 3 and 4. Because of time limitations, only one model was retrained with only the 3 labels, namely the Logistic Regression trained with the KernelPCA features and majority voting for the experts (because this model had less predictive power than all the others, so if an Expert exemption made a difference, this model would show it easier). When this model is trained with all 4 Experts, it achieves the metrics described in the first row of Fig11, while when it is trained only on Experts 2,3,4 it achieves the metrics described in the second row of Fig11.

Model	Weighted Average Precision	Weighted Average Recall	Weighted Average F1 score
Logistic, KPCA, Majority, 4 Experts	0.89	0.87	0.87
Logistic, KPCA, Majority, 3 Experts	**0.91**	**0.91**	**0.91**

Fig11: Comparison of models when Expert 1 is excluded from final model (KernelPCA features, majority voting experts

It can be easily concluded that Expert 1 indeed lowers the final prediction accuracy when majority voting is trained with them included. With their labels excluded, the final model is re-trained and its final prediction accuracy is increased. This means that indeed Expert 1 is somewhat more “unreliable” when it comes to their labels for tomato seedlings, compared to the other 3 Experts (2,3,4). It should be noted that to generalise the conclusion better, more models will have to be re-trained using only Experts 2, 3 and 4, in order to see if Expert 1 is the least “reliable” in every case.

`

References

Deep Learning for Plant Identification in Natural Environment:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5458433/

The feature extraction methods used follow the tutorials from PlantCV :
https://plantcv.readthedocs.io/en/stable/analysis_approach/