AN ATTENTION-BASED FUSION ARCHITECTURE FOR MULTIMODAL CROP DISEASE DETECTION AND ITS DEPLOYMENT ON EDGE DEVICES

Sindarov O.M.

Samarkand State University named after Sharof Rashidov

Abstract. The thesis presents a conceptual multimodal architecture that fuses image and sensor data through an attention mechanism for the early detection of crop diseases, together with its mathematical formulation, dataset construction procedure, evaluation criteria, and the prospects of deploying the model on resource-constrained edge devices.

Keywords: *multimodal fusion, attention mechanism, deep learning, edge computing, crop diseases, IoT sensors, model compression.*

Introduction

A review of the existing literature shows that models relying solely on image data face serious limitations in terms of robustness and early diagnosis under field conditions [1; 3; 7]. Overcoming these limitations requires a multimodal architecture capable of jointly analyzing image and sensor data. The purpose of this thesis is to present, in detail, the structure, mathematical formulation, and practical deployment aspects of an attention-based fusion architecture.

The scientific novelty of the proposed approach lies in the fact that it relies not only on the visual features present in an image but also on environmental factors — temperature, humidity, and light — that indirectly affect the plant's physiological state, thereby extending the ability to detect disease at its earliest, not yet visually apparent, stage [2; 4].

Proposed multimodal architecture

The architecture has a two-branch (dual-branch) structure, in which each branch specializes in processing a specific modality. The first branch — the image encoder —

processes the digital or spectral image acquired under field conditions through a convolutional neural network (for example, an architecture from the ResNet or EfficientNet family) or a Vision Transformer, producing the feature vector . The second branch — the sensor encoder — processes the time series of temperature, humidity, soil, and light indicators obtained from IoT sensors through a recurrent network (LSTM) or a Transformer adapted for time series, forming the feature vector .

The feature vectors obtained from both modalities are then passed to an attention-based (cross-attention) fusion layer. In this layer, the relative importance of each modality is dynamically determined through learnable attention weights — for example, in situations where the sensor data indicate conditions favorable to disease but the image does not yet show a clear sign, this mechanism automatically increases the weight assigned to the sensor branch. The fused vector is then used by the final classifier to predict the disease class (or the healthy state).

Mathematical formulation

The operation of the architecture is expressed through the following stages. First, each modality is transformed into a separate feature vector by its own encoder:

(1)

(2)

Here, is the input image, is the time series of sensor data, and are the learnable parameters of the respective encoders. The two modalities are then fused by means of the attention mechanism:

(3)

(4)

Here, denotes the fusion of the two vectors, and are the learnable parameters of the attention layer, denotes the weight coefficients assigned to each modality. In the final stage, the fused vector is transformed by the classifier into a probability distribution over the disease classes:

(5)

Here, and — are the parameters of the final fully connected layer, and denotes the vector of predicted probabilities for each disease class (including the "healthy" class). During training, the cross-entropy loss function is minimized, which serves to reduce the discrepancy between the predicted and the true class distributions.

Training strategy

To optimize the model parameters, the Adam optimizer together with a step-wise learning-rate schedule is planned to be used. To prevent overfitting, dropout is applied in the encoder and classifier layers, and L2 regularization is applied to the weight parameters. For cases in which the disease classes are unevenly distributed in the dataset (some diseases occurring less frequently than others), the use of weighted cross-entropy or focal loss is envisaged — this prevents the model from neglecting the less frequent classes.

The weighting coefficients (α) of the attention layer are formed automatically during training; at the initial stage, both modalities are assigned equal importance, and over the course of training the model independently learns which modality is more informative under which circumstances. This property makes the model adaptive to varying conditions (for example, situations in which a sensor temporarily malfunctions), since in such cases the attention mechanism automatically relies more heavily on the image branch.

Dataset and preprocessing

The proposed architecture is trained on data obtained from two sources: digital/spectral images collected under field conditions using a dedicated camera, and temperature, air and soil humidity, and light-intensity readings obtained from IoT sensors placed around the plant. Images are first normalized by size, after which augmentation techniques such as rotation, mirroring, and variation of illumination

intensity are applied to increase the diversity of the dataset. Sensor data are grouped over a defined time window and normalized using the min–max or z-score method.

Both modalities are synchronized according to a common timestamp and geolocation marker, ensuring that each image sample is correctly paired with its corresponding sensor window. The dataset is divided into training, validation, and test subsets, with samples from different crop varieties and regions deliberately held out in the test subset — enabling the model's ability to generalize to new conditions to be assessed through cross-domain validation.

Evaluation criteria

Model performance is evaluated using several metrics: overall accuracy, the F1-score (which accounts for class imbalance), and the area under the precision–recall curve. In addition, the model's robustness to a new crop variety or geographic region is assessed through cross-domain validation, while its suitability for real production conditions is assessed through inference latency and operating-memory consumption. Together, these two groups of criteria — accuracy and efficiency — enable a comprehensive evaluation of the model's applicability both in the laboratory and under practical field conditions.

Deployment prospects on edge devices

The practical value of the proposed multimodal architecture is determined by the feasibility of running it on resource-constrained (edge) devices — for example, compact computing platforms operating under field conditions. Such deployment requires compressing the model, for which three principal methods are applied: pruning — removing connections of low importance; quantization — representing parameters with lower-precision numbers; and knowledge distillation — transferring the knowledge of a large “teacher” model to a smaller “student” model.

Applying these methods reduces the model's size and computational complexity, though this process may be accompanied by a certain degree of accuracy loss — for

this reason, finding the balance between the degree of compression and the loss of accuracy is treated as a separate research task. Operating in real time on the edge device, not requiring a constant internet connection, and increasing response speed through local data processing are among the practical advantages of the proposed solution [6].

The comparative characteristics of the three compression methods are summarized in Table 1.

Method	**Mechanism**	**Size reduction**	**Effect on accuracy**
Pruning	Removal of low-importance connections/neurons	Moderate	Small, controllable
Quantization	Representing parameters with lower-precision (e.g., 8-bit) numbers	High	Usually minimal
Knowledge distillation	“Training” a small model from a larger model	High	Depends on teacher quality

Table 1. Comparative characteristics of model compression methods

In practice, these methods achieve the greatest effect not in isolation but when applied jointly — for example, pruning followed by quantization. In the proposed research, all three methods will be tested together, and the optimal combination will be selected based on the accuracy–size trade-off.

Expected scientific and practical results

The experimental verification of the proposed architecture is expected to yield the following results: first, quantitative confirmation that the multimodal model outperforms image-only baseline models in terms of accuracy and F1-score; second, a

significant increase in the detection rate for the early, visually imperceptible stage of disease, since sensor data provide an advance signal of physiological changes not yet reflected in the image; and third, greater robustness of the model across different crop varieties and regions in cross-domain validation results compared to unimodal models.

As a practical outcome, the compressed version of the model (with pruning and quantization applied) is expected to operate with satisfactory inference latency and memory consumption on a resource-constrained device, turning the architecture into a prototype applicable under real field conditions, even without a constant internet connection. These results form the theoretical and practical basis for building a decision-support system for farms.

Further stages of the research

To develop the conceptual architecture presented in this thesis into a full experimental study, the following stages are planned: in the first stage, an initial dataset of image and sensor data will be constructed for selected crop types; in the second stage, the proposed dual-branch architecture will be implemented and trained under the same conditions as the unimodal baseline models; in the third stage, the results will be compared according to the defined criteria, and the architecture's hyperparameters (attention-layer dimensionality, learning rate, regularization coefficient) will be optimized; in the fourth stage, the best model will be compressed using the methods described above and tested on an edge device.

The results of each stage are planned to be presented in separate scientific publications, which constitutes an integral part of the further research plan.

Conclusion

The attention-based multimodal fusion architecture presented in this thesis enables the detection of crop diseases at an early stage through the joint analysis of image and sensor data. The architecture's mathematical formulation, dataset construction procedure, and evaluation criteria serve as a theoretical foundation for

subsequent experimental research. Investigating the feasibility of deploying the model on edge devices further enhances the practical significance of the work and marks it as an important step toward real-world application on farms.

References

1. Dolatabadian A. [et al.]. Image-based crop disease detection using machine learning // Plant Pathology. – 2025. – DOI: 10.1111/ppa.14006.
2. Hamid [et al.]. Enhancing Agricultural Disease Detection: A Multi-Model Deep Learning Novel Approach // Engineering Reports. – 2025. – DOI: 10.1002/eng2.13113.
3. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture // Artificial Intelligence Review. – 2024. – DOI: 10.1007/s10462-024-11100-x.
4. Multimodal Deep Learning and Foundation Models for Early Detection and Forecasting of Plant Diseases // Plants. – 2026. – Vol. 15, No. 17. – Art. 2564. – DOI: 10.3390/plants15172564.
5. Sable N.P., Shukla V., Khedkar V., Ujalambkar D.M. Hybrid multimodal learning framework for crop disease detection, adaptive treatment, and price forecasting // Frontiers in Computer Science. – 2026. – DOI: 10.3389/fcomp.2026.1855851.
6. Artificial Intelligence-Driven Plant Disease Detection and Diagnosis: A Comprehensive Review of Deep Learning Approaches, Multimodal Sensing Technologies, and Future Perspectives in Precision Agriculture // The Plant Pathology Journal. – 2026. – DOI: 10.5423/PPJ.RW.01.2026.0004.
7. A Review on the Detection of Plant Disease Using Machine Learning and Deep Learning Approaches // Journal of Imaging. – 2025. – DOI: 10.3390/jimaging11100326.