MACHINE LEARNING-DRIVEN CROP CLASSIFICATION AND YIELD PREDICTION USING COMBINED MULTISPECTRAL, HYPERSPECTRAL, AND ENVIRONMENTAL DATA

Summary of PhD Dissertation

<u>FARMONOV NIZOM</u>

Supervisor:

DR. MUCSI LÁSZLÓ

associate professor

co-supervisor:

DR. SZATMÁRI JÓZSEF

associate professor

Doctoral School of Geosciences

Faculty of Science and Informatics

University of Szeged

SZEGED 2024

1. Problem Statement

In the agricultural landscape of Mezőhegyes, Hungary, accurately detecting different crop types which has almost similar spectral signatures and predicting crop yield at the field scale (e.g., wheat) is a significant challenge. As the world's leading producer and exporter of soybeans, the United States produces 115 million tons of soybeans in 2020, representing 39.9% of the world's soybean exports (USDA/NASS, 2021). Even minor shifts in the production of soybeans in the United States can result in significant volatility in the international soybean market. As a result, accurate and timely forecasts to produce soybeans in the United States are essential to global food trade and security (Fritz et al., 2019). However, due to complex processes that influence yield formation and the complex effects of weather, soil conditions, vegetation and management practices, it is difficult to predict the soybean yield quickly at the county level. Traditional crop mapping and yield estimation methods are often lacking the precision needed for effective decision-making by farmers and agricultural stakeholders. The integration of ML techniques with remote sensing data, including multispectral and hyperspectral imaging, presents a promising approach to enhance the precision of crop yield predictions. Hyperspectral data, with its rich spectral information, can provide detailed insights into crop health and environmental conditions, which are crucial for fine-tuning ML models. The complex relationships between environmental factors, crop health indicators, and historical performance demand advanced ML models capable of leveraging hyperspectral data to its fullest potential. The Mezőhegyes study area presents unique challenges, such as variable weather patterns, soil properties, and pesticide use, which complicate yield predictions. Thus, a robust ML-driven approach that effectively utilizes hyperspectral data is essential for improving agricultural productivity and supporting informed decision-making. Accurate yield forecasts can enhance agricultural risk management and assist farmers in planning optimal planting strategies.

2. Research objectives and questions.

The aim of this dissertation is to advance crop classification and yield prediction at the field and county level, methodologies through the integration of hyperspectral imaging data and ML techniques. Specifically, the study seeks to develop and evaluate sophisticated ML and DL models that utilize hyperspectral data to extract detailed spectral and spatial information about crops and their characteristics. By incorporating wavelet transforms, spectral and spatial attention mechanisms, and morphological operations and integration of multi-source satellite images with environmental data, the research aims to enhance the precision and robustness of crop type mapping and yield prediction models.

The research focuses on leveraging spaceborne hyperspectral imaging, which provides rich spectral information, in conjunction with Multispectral images and environmental variables to improve the accuracy of crop classification and yield forecasts. This approach aims to address the limitations of traditional methods by incorporating advanced feature extraction techniques and integrating diverse data sources. The goal is to offer a more comprehensive understanding of crop health and environmental influences, ultimately supporting better decision-making for farmers and agricultural stakeholders.

The study explores how combining these advanced methodologies can lead to improved predictions and insights into crop yield dynamics, thereby contributing to more effective agricultural management and risk mitigation strategies. Considering the theoretical framework and objectives, following scientific questions were developed:

1. How do the integration of wavelet transform and spectral attention mechanisms in a 2-D CNN framework and variations in spatial patch sizes and factor analysis dimensions affect the accuracy and efficiency of HSI classification for crop-type mapping compared to traditional ML and 3-D CNN approaches?

2. How do the spatial and temporal resolutions of PlanetScope and Sentinel-2 imagery, along with the introduction of vegetation indices and environmental data, influence the accuracy of wheat yield estimation and prediction, and which vegetation indices and phenological stages are most effective in enhancing this accuracy?

3. How effective are ML and DL models in forecasting soybean yields in the U.S. Corn Belt, considering geospatial and climatic data?

4. How does the integration of 3-D–2-D CNNs with morphological operators and attention mechanisms improve the accuracy and robustness of HSI classification compared to traditional deep learning methods?

5. What impact does the fusion of hyperspectral and LiDAR data, incorporating environmental and morphological features, have on the precision of crop classification and yield prediction in the Mezőhegyes agricultural landscape?

3. Data and methods

The Materials and Methods sections of Chapters 2, 3, 4, 5 and 6 contain specific details about the data and methodologies employed. The following summaries provide a concise overview.

a) **Hyperspectral Data**: Hyperspectral data were acquired from the DESIS (DLR Earth Sensing Imaging Spectrometer) aboard the International Space Station (ISS). Two Level-2A bottom-of-atmosphere (BOA) reflectance DESIS images from June were downloaded via the EOWEB Geoportal. DESIS images were not available publicly to the users. Therefore, scientific proposal was presented to the DLR by Dr. Mucsi László and limited amount of quota were granted. The dataset spans a 30 km x 30 km area and offers detailed spectral data across 235 bands.
b) **PlanetScope Data**: Next generation of commercial PlanetScope (PS) data were collected between November 2020 and July 2021. A total of 72 cloud-free images were ordered and downloaded, providing surface reflectance products across eight spectral bands at a spatial resolution of 3 meters. These images were harmonized with Sentinel-2 imagery to ensure radiometric consistency. It worth to be mentioned that PS imagery is not also free and open access as well. However, as result of research proposal leveraging “Education and

Research Program" of Planet Labs submitted by Mr. Nizom Farmonov, access was obtained.

c) **Sentinel-2 Data**: Twenty-five Sentinel-2 (S2) Level-2A images, covering the same period, were obtained. These images offer spectral information across 13 bands with spatial resolutions of 10, 20, and 60 meters.

d) **MODIS Data:** The Moderate Resolution Imaging Spectroradiometer (MODIS) provided key vegetation indices such as the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI), which were critical to capturing crop growth dynamics. These indices were derived from the MOD13Q1 dataset, available at a 250 m resolution and a 16-day interval, complementing the higher-resolution data.

In addition, high-resolution LiDAR digital terrain model (DTM) data were integrated into our analysis to provide detailed elevation and morphology information. Acquired via airborne radar on April 19, 2019, with a spatial resolution of 5 cm, this data helped assess terrain influences on crop fields. The LiDAR data resampled using the cubic curve method in ERDAS IMAGINE 2020 to align with the 30 m resolution of DESIS imagery, enhanced the accuracy of the study. Also, to validate our method, we tested it on the public Houston 2013 dataset, which includes hyperspectral and LiDAR data. This dataset, with 144 sub bands and 15 classification classes, ensured the generalizability of our approach across different data sources.

Crop yield data were collected from Mezőhegyes farm during the growing season using a John Deere W650i combine harvester with a yield-mapping system. Ground truth data was obtained from Mezőhegyes company. The data, recorded every 2 seconds, were filtered to remove outliers and inaccuracies, such as near-zero yields from overlapping rows. The cleaned data, corresponding to the harvester's header size (2 m × 6 m), were provided by the farming company. Using QGIS v.3.16, the data were interpolated into raster format via inverse distance weighting (IDW) to match satellite image resolution.

Data pre-processing and analysis were carried out using specialized software tools such as ESA SNAP 8.0 for Sentinel data, Python 3.9 and R program for ML model development, and QGIS 3.16 and ERDAS IMAGINE 2020 for geospatial data manipulation and visualization.

4. Dissertation outline

The chapters draw upon scientific articles published in peer-reviewed journals. Each chapter outlined below articulates its research objectives, with the overall conclusions and future outlook provided in the comprehensive conclusion and perspectives section. Furthermore, each individual chapter can be regarded as an exploration of an independent research query.

Chapter 1 presents a summary of the dissertation, featuring a concise literature review that addresses the primary topics of the study. It also includes a description of the research area, defines the problem statement, explains the research objective, formulates the hypotheses, and outlines the overall structure of the dissertation.

Chapter 2 outlines the methodology employed in the study conducted on wheat fields in Mezőhegyes, Hungary, utilizing remote sensing data from PS and S2 satellites alongside environmental data. The study involved advanced agricultural practices, including seeding, weed control, and harvesting, with data collected using yield-mapping systems. The RFRVI model was used to estimate wheat yield by integrating spectral bands, VIs, and environmental variables such as climate and topography. The model demonstrated high accuracy, achieving R^2 values up to 0.81 and root mean square errors (RMSE) as low as 0.287 t/ha. Notably, combining PS and S2 data with environmental inputs enhanced yield prediction accuracy, especially during critical phenological stages.

Chapter 3 details the study area, data sources, and methodology for crop classification in Mezőhegyes, Hungary. The region, characterized by fertile chernozem soil, hosts various crops and non-crop classes. Data were collected from DESIS and Sentinel-2 satellites, and field samples from 5080 pixels were used for model training and validation. The methodology involved RFR and SVM algorithms, with comparisons to advanced deep learning models like MSRN and MDBRSSN. While RFR and SVM were effective, the wavelet attention CNN model achieved the highest accuracy, with an overall accuracy of 97.89% and a kappa coefficient of 0.97, showcasing significant improvements in classification precision and efficiency.

Chapter 4 represented applied machine learning models to forecast soybean yields across major U.S. states using satellite-derived data. MODIS provided indices such as NDVI, EVI, surface reflectance, and temperature data, while yield records from USDA (2012-2021) were used as ground truth. Among the five machine learning models tested, Random Forest outperformed others, achieving an R^2 of 0.75 and RMSE of

0.342 t/ha. NDVI, EVI, and surface reflectance were identified as the most influential features for accurate yield predictions.

Chapter 5 outlines the study area and data for the proposed method. Data sources include DESIS hyperspectral imagery and high-resolution LiDAR data. The DESIS data provides detailed spectral information, while LiDAR offers topographic precision. The proposed method integrates these data sources using a deep learning model with 2-D and 3-D convolutional layers and spatial-spectral morphological attention. Key techniques include PCA for dimensionality reduction, 3-D CNNs for spatial-spectral analysis, and dropout for generalization. The Adam optimizer with a learning rate of 0.001 and a decay function of 1e-06 is used for training, involving 50 epochs and a batch size of 256. The model features ReLU activation for all layers except the output, which uses SoftMax. Convolutional windows of various sizes are tested, and PCA components range from one to ten. The model demonstrates superior classification performance, with PCA=3 and a 11×11 window size offering the best results. Morphological operators enhance LiDAR data, and increasing the structuring element size generally decreases accuracy. Evaluation shows the proposed method outperforms other techniques in Overall Accuracy (OA), Kappa, and F1-Score, and is robust against noise. Ablation tests reveal the impact of different model components on performance, highlighting their significance in achieving high accuracy.

Chapter 6 describes the dissertation's main conclusions, implications, limitations, and suggestions.

5. Key findings

By addressing the research hypotheses to achieve the research objective, my study has achieved the following key findings:

Thesis 1. My results show that integrating wavelet transforms and spectral attention mechanisms into a 2-D CNN framework enhances hyperspectral image classification accuracy for crop mapping by efficiently focusing on both spatial and spectral features. This method achieves a high accuracy of 97.89% OA and $\kappa = 0.97$, while reducing noise, computational costs, and overfitting compared to classical ML and 3-D CNNs. Additionally, using medium-sized spatial patches optimizes classification accuracy by balancing detail and computational efficiency, and factor analysis further improves model performance by reducing spectral dimensions and minimizing data redundancy. These results in this thesis corresponds to the first publication.

Thesis 2. Based on my investigation, combining PS imagery with 3-meter resolution and S2 imagery with 10-meter resolution both deliver accurate wheat yield estimations, with PS showing slightly lower RMSE (0.336 t/ha) compared to S2 (0.325 t/ha) due to its finer spatial detail. Integrating Vegetation Indices (VIs) further enhances prediction accuracy, with MTCI being the most effective for PS and MSAVI2 for S2. The milk and dough stages (210–225 days) are most critical for accurate predictions, and including environmental data further improves the model's accuracy. These findings were reported from the second publication.

Thesis 3. My study found that both ML and DL models effectively forecast soybean yields, with the RF model outperforming classical ML models such as LASSO, XGBoost, and decision tree regression, achieving an R^2 of 0.77 and an RMSE of 0.334 t/ha. The 1D-CNN deep learning model surpassed all, with an R^2 of 0.86 and an RMSE of 0.276 t/ha in 2021. These findings highlight the value of integrating geospatial and climatic data for improving yield forecasts, especially in the U.S. Corn Belt. These thesis findings correspond to the third publication.

Thesis 4. According to my research, Integration of 3D–2D CNNs with morphological operators and attention mechanisms enhances HSI classification by outperforming current DL (e.g., HLDC or DECL) models with 1%–3%. The 3D CNN captures spatial context, the 2D CNN refines spectral details, morphological operators highlight geometric features, and attention mechanisms focus on critical areas, leading to improved accuracy and robustness. These results were detailed in the fourth publication.

Thesis 5. Following my findings, HypsLiDNet model demonstrated significant improvement in crop type detection by integrating HSI and LiDAR data with OA of 0.98% compared to all other competing methods. This comprehensive approach allows for a more precise analysis of crop characteristics and conditions, leading to better prediction accuracy in the Mezőhegyes agricultural landscape. These thesis outcomes were found in the fourth publication.

6. Implications

The research presented in this dissertation offers significant implications for both the scientific community and practical applications in agriculture. First, by combining hyperspectral imaging (HSI) with LiDAR data and advanced deep learning techniques, the study demonstrates the potential for improving crop classification accuracy and yield prediction. This has direct implications for precision agriculture, offering farmers and agricultural stakeholders more reliable tools for monitoring crop health, optimizing resource use, and predicting yields. The findings also highlight the importance of spectral-spatial attention mechanisms and wavelet transforms, which could be applied to other areas of remote sensing, such as environmental monitoring and land-use mapping.

Moreover, the integration of 2D and 3D CNNs in analyzing hyperspectral and LiDAR data opens up new possibilities for handling complex datasets. The method's ability to achieve high accuracy with limited training data reduces the dependency on large, labeled datasets, which is a significant hurdle in machine learning. This could lead to more accessible and scalable applications of remote sensing technology in agriculture, especially in regions with limited resources for data collection. Furthermore, the successful fusion of multiple data sources (HSI, LiDAR, and environmental data) showcases the potential of multimodal approaches for enhancing decision-making in agriculture and beyond.

6. Limitations, Recommendations and Future Research

While this study presents an innovative approach to crop classification and yield prediction, several limitations need to be addressed. First, the generalizability of the HypsLiDNet model is somewhat constrained by the datasets used. The study focuses on specific crops in a particular geographic region (Mezőhegyes, Hungary), and while the model performed well in this context, its performance across different agricultural landscapes and diverse crop types remains uncertain. Testing on more varied datasets, especially in regions with different environmental conditions, is necessary to confirm the model's robustness.

Another limitation lies in the computational requirements of the proposed method. Although the model is effective in terms of accuracy, it involves complex architectures that demand substantial computational resources, which may limit its practical application for users with limited access to high-performance computing facilities. Additionally, the study's reliance on hyperspectral and LiDAR data, which can be costly and difficult to acquire in some regions, may reduce the scalability of the method, particularly in developing countries.

Lastly, the model's effectiveness was tested on relatively small datasets. While the use of dimensionality reduction techniques like PCA helped improve performance, larger-scale datasets may present new challenges related to computational efficiency and model optimization.

For future research, several areas should be explored to enhance the applicability and effectiveness of the proposed method. First, expanding the model's testing to include various crop types and agricultural regions globally would help assess its robustness and adaptability. This would involve acquiring additional datasets from different environments, possibly leveraging other available hyperspectral and LiDAR sources, to validate the model's generalizability.

Additionally, addressing the computational demands of the HypsLiDNet model is crucial for wider adoption. Future work could focus on optimizing the model's architecture to reduce computational complexity without sacrificing accuracy. Exploring more efficient training techniques, such as transfer learning, could help make the model accessible to users with limited computing resources.

Another important area for future research involves refining the fusion of hyperspectral and LiDAR data with other environmental variables. Incorporating real-time climate data or integrating additional sensor modalities could further enhance the predictive accuracy of crop yields. Moreover, testing the model on temporal datasets to track changes in crop growth over time could provide more dynamic predictions, helping farmers make more informed decisions throughout the growing season.

Finally, future research should investigate the model's application beyond agriculture. Given its ability to classify complex spatial-spectral data, HypsLiDNet could be applied to other domains, such as forestry, environmental monitoring, and urban planning.

7. List of publications used in the dissertation.

Farmonov, N., Amankulova, K., Szatmári, J., Sharifi, A., Abbasi-Moghadam, D., Mirhoseini Nejad, S.M., Mucsi, L., 2023. Crop Type Classification by DESIS Hyperspectral Imagery and Machine Learning Algorithms. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 16, 1576–1588. https://doi.org/10.1109/JSTARS.2023.3239756

Farmonov, N., Amankulova, K., Szatmári, J., Urinov, J., Narmanov, Z., Nosirov, J., Mucsi, L., 2023. Combining PlanetScope and Sentinel-2 images with environmental data for improved wheat yield estimation. International Journal of Digital Earth 16, 847–867. https://doi.org/10.1080/17538947.2023.2186505

Farmonov, N., Amankulova, K., Khan, S.N., Abdurakhimova, M., Szatmári, J., Khabiba, T., Makhliyo, R., Khodicha, M., Mucsi, L., 2024a. Effectiveness of machine learning and deep learning models at county-level soybean yield forecasting. HunGeoBull 72, 383–398. https://doi.org/10.15201/hungeobull.72.4.4

Farmonov, N., Esmaeili, M., Abbasi-Moghadam, D., Sharifi, A., Amankulova, K., Mucsi, L., 2024b. HypsLiDNet: 3-D–2-D CNN Model and Spatial–Spectral Morphological Attention for Crop Classification With DESIS and LiDAR Data. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 17, 11969–11996. https://doi.org/10.1109/JSTARS.2024.3418854