**Bulletin of Electrical Engineering and Informatics**
Vol. 15, No. 6, December 2026, pp. 5636~5653
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.12164 

# Development of a model for detecting diseases of crops using spectral and vegetative features

**Aislu Kassekeyeva[1], Moldir Yessenova[1], Talgatbek Ayazbaev[2], Gulzira Abdikerimova[1], Gulmira Azieva[3], Zhanylsyn Sultanova[4], Assyl Bissengaliyeva[5], Assem Serikbayeva[6]**
[1]Department of Information Systems, Institute of Digital Sciences and Artificial Intelligence, Eurasian National University named L.N.Gumilyov, Astana, Kazakhstan
[2]Department of Information and Communication Technologies, Faculty of Information Technology, International Taraz Innovative Institute, Taraz, Kazakhstan
[3]Department of Cryptology, Faculty of Mechanics and Mathematics, Eurasian National University named L.N.Gumilyov, Astana, Kazakhstan
[4]Department of Biostatistics, Bioinformatics and Information Technologies, Astana Medical University, Astana, Kazakhstan
[5]Zhangir Khan University, Uralsk, Kazakhstan
[6]Department of Higher Mathematics, Kazakh Agrotechnical Research University named after S. Seifullin, Astana, Kazakhstan

**Article Info**

***Article history:***

Received Dec 18, 2025
Revised Sep 8, 2026
Accepted Sep 12, 2026

***Keywords:***

Agricultural land classification
Machine learning
Precision agriculture
Spectral analysis
Vegetation indices

**ABSTRACT**

This study presents a leakage-aware framework for feature engineering and spatial data quality assessment of multitemporal Sentinel-2 observations for agricultural land classification. After removing 29 duplicate coordinates, 5,563 samples formed 1,983 connected components under a 5 km rule and were assigned to spatially independent training, validation, and held-out test subsets. Twelve spectral bands and eight vegetation and moisture indices generated 648 spectral, 432 index-derived, and 1,080 combined features. On validation data, the combined representation achieved a macro-F1 of 0.7261 with HistGradientBoosting, while the proposed out-of-fold (OOF)-stacked multispectral machine learning fusion (OS-MMF) achieved 0.7408. Across ten repeated whole-component partitions, the combined representation obtained a mean macro-F1 of 0.7362. Removing the redundant McFeeters normalized difference water index (NDWI) features produced comparable performance. Final held-out evaluation of OS-MMF yielded an accuracy of 0.7803 and a macro-F1 of 0.7080, supporting reproducible spatially controlled classification without establishing universal model superiority.

*This is an open access article under the CC BY-SA license.*

***Corresponding Author:***

Moldir Yessenova
Department of Information Systems, Institute of Digital Sciences and Artificial Intelligence
Eurasian National University named L.N.Gumilyov
Astana, Kazakhstan
Email: moldirrespect@gmail.com

## 1. INTRODUCTION

Agricultural land classification is a fundamental task in Earth observation, providing critical information for precision agriculture, crop inventory, food security assessment, environmental monitoring, and sustainable land management. Accurate identification of crop types supports agricultural planning, resource allocation, yield assessment, and environmental policy development. The increasing availability of publicly available satellite observations has accelerated the development of large-scale agricultural monitoring systems, among which the Sentinel-2 constellation has become one of the most widely used optical missions due to its 10–20 m spatial resolution, high flyby rate, and rich multispectral information [1]–[3]. The combination of visible, red-edge, near-infrared, and shortwave infrared bands allows Sentinel-2 imagery to record crop

phenology, vegetation viability, chlorophyll content, canopy structure, and moisture dynamics throughout the growing season. Consequently, multi-temporal Sentinel-2 observations have become a primary data source for crop classification, agricultural land use mapping, and vegetation monitoring. Recent studies show that multi-temporal observations consistently outperform single-date imagery, as seasonal spectral trajectories provide significantly greater resolution between crop types [4]–[6]. However, systematic reviews show that, although benchmark classification accuracy often exceeds 95%, operational performance remains significantly lower due to limited spatial transferability, data quality issues, and inappropriate validation strategies. Recent advances in machine learning (ML) and deep learning (DL) have significantly improved agricultural land classification. Convolutional neural networks, recurrent neural networks, transformer-based architectures, and hybrid learning structures have demonstrated remarkable performance when applied to multi-temporal Sentinel-2 imagery [5]–[8]. However, most published studies primarily focus on classifier optimization, while significantly less attention has been paid to constructing and validating the input feature space. Consequently, improvements in forecasting accuracy are often achieved through increasingly complex models rather than through systematic improvement of the underlying analytical dataset.

Feature engineering has thus become a key research area in agricultural remote sensing. In addition to the original Sentinel-2 spectral bands, vegetation indices, seasonal descriptors, temporal statistics, and aggregated spectral characteristics provide additional information describing vegetation growth dynamics and physiological properties. Recent studies have shown that carefully designed feature engineering strategies significantly improve class separability while reducing computational complexity and spectral redundancy [9], [10]. Furthermore, spatially-aware feature selection methods have demonstrated that removing redundant variables can preserve predictive information while improving model robustness and interpretability. Another rapidly developing area of research concerns the quality of the training data itself. Existing agricultural datasets often contain duplicate observations, inaccurate reference labels, incomplete time series, cloud-contaminated data, and radiometric inconsistencies, which directly impact feature engineering and subsequent classification performance. Recent studies have shown that systematic quality control procedures and ground truth validation significantly improve classification reliability, emphasizing that data quality is as important as classifier selection [11]–[13].

At the same time, spatial validation is receiving increasing attention. Traditional random partitioning often distributes geographically adjacent observations into both the training and test sets, leading to information leakage due to spatial autocorrelation. This methodological issue often leads to overly optimistic estimates of predictive performance that are impossible to replicate in operational settings. Therefore, recent methodological reviews recommend geographically independent partitioning strategies and spatially explicit validation protocols as necessary requirements for realistic evaluation of remote sensing models. Recent applications also demonstrate that informative feature representations are essential in various remote sensing domains. Hybrid spatiotemporal modelling successfully improved probabilistic soil salinity mapping by integrating multi-source environmental variables while maintaining spatial consistency [14]. Crop segmentation based on pixels obtained from ultra-high-resolution unmanned aerial vehicle (UAV) imagery has shown that carefully designed spectral representations significantly improve crop discrimination in heterogeneous agricultural settings [15]. Similarly, the integration of remote sensing data with climate variables has significantly improved soil salinity mapping, highlighting the importance of constructing informative feature spaces prior to predictive modelling [16].

Despite recent advances in agricultural land classification, a number of methodological limitations remain. Most existing studies focus primarily on classifier performance, while significantly less attention is paid to the quality of the analytical dataset and the construction of robust feature representations. In particular, spatial integrity, temporal completeness, radiometric consistency, and feature redundancy are rarely systematically assessed before predictive modelling, even though these factors directly impact model robustness and reproducibility [17]–[19]. Another important challenge is the high dimensionality of multi-temporal Sentinel-2 datasets. The combination of spectral bands, vegetation indices, and temporal descriptors generates large feature spaces containing significant redundancy and nonlinear dependencies. Consequently, statistical feature diagnostics is increasingly important for identifying informative variables while reducing unnecessary complexity.

An additional limitation of existing feature engineering studies is that the statistical significance of engineered variables is not always assessed in terms of their incremental contribution to downstream classification. High univariate associations with crop labels do not necessarily imply additional predictive value when highly correlated spectral variables and deterministic index transformations are used jointly. Therefore, a rigorous evaluation of feature engineering should combine data quality diagnostics with controlled downstream evaluation, in which alternative feature representations are compared under the same spatial partitioning, pre-processing protocol, and classification model. This design allows the contribution of the engineered representation to be distinguished from improvements resulting from changes in the learning algorithm or validation design. To address these limitations, this study proposes an integrated framework for feature

engineering and spatial data quality assessment of multi-temporal Sentinel-2 observations. The proposed methodology integrates duplicate coordinate detection, graph-based spatial connectivity analysis, geographically independent partitioning, temporal completeness assessment, radiometric quality control, vegetation index construction, correlation analysis, and statistical feature diagnostics within a single reproducible workflow.

The novelty of this study lies in its approach to feature engineering as an independent analytical step rather than a traditional preprocessing procedure. The proposed framework jointly assesses spatial integrity, temporal completeness, spectral consistency, feature redundancy, and nonlinear information content using mutual information (MI), distance correlation, the Kruskal-Wallis test, effect size estimation ($\eta^2$ and $\varepsilon^2$), and multiple testing correction, providing a statistically validated and leakage-safe feature representation prior to model development [20]–[23]. The aim of this study is to develop a replicable framework for feature engineering and spatial data quality assessment using multi-temporal Sentinel-2 observations for agricultural land classification [24], [25]. The proposed approach aims to create a geographically independent, information-rich analytical dataset that improves the reliability, reproducibility, and interpretability of subsequent ML applications.

The main contributions of this study are summarized as follows:

- A spatial data quality assessment framework integrating duplicate detection, graph-based spatial connectivity analysis, and geographically independent partitioning.
- A reproducible feature engineering workflow combining Sentinel-2 spectral bands, vegetation indices, temporal descriptors, and aggregated seasonal statistics.
- A statistical feature diagnostic methodology integrating correlation analysis, MI, distance correlation, the Kruskal–Wallis test, effect-size estimation, and multiple-testing correction to construct a leakage-safe analytical feature space.

The remainder of the paper is organized as follows: section 2 describes the study area, dataset characteristics, and the proposed methodology for spatial data quality assessment, temporal and spectral validation, vegetation index construction, feature engineering, and statistical feature diagnostics. Section 3 presents the experimental results, including dataset characterization, spatial and temporal quality assessment, spectral statistics, correlation analysis, feature diagnostics, and the final feature representation. Finally, section 4 discusses the implications, limitations, and potential applications of the proposed framework, and section 5 concludes and outlines future research directions.

## 2. METHOD

This study was conducted using the Canadian crop dataset (CCD), a publicly available benchmark dataset containing multi-temporal Sentinel-2 observations of agricultural land across Canada. Although the original dataset spans four growing seasons (2017–2020), to ensure temporal consistency and eliminate interannual variability, only observations acquired in 2020 were used in data quality assessment and feature development. The dataset contains parcel-level annotations derived from the annual crop inventory (ACI) and includes ten land-use classes. Each geographic sample can contain up to five Sentinel-2 observations acquired during the main growing season from June to October, corresponding to a maximum of one available observation for each monthly position. Because some monthly observations are unavailable, individual samples may contain fewer than five valid observations. Each available observation consists of twelve spectral bands (B1–B9, B8A, B11, and B12), with B10 excluded because it is primarily used for atmospheric sensing.

Prior to feature development, the dataset underwent a comprehensive quality control process. Duplicate geographic coordinates were identified and removed, resulting in an analytical set consisting of 5,563 unique georeferenced samples representing 22,904 monthly observations across eight Canadian provinces. Thus, the analytical cohort contained an average of approximately 4.12 available monthly observations per geographic sample, reflecting incomplete temporal coverage for a subset of samples. The remaining observations were further evaluated using spatial integrity checks, temporal completeness assessments, spectral quality analysis, and statistical diagnostics of characteristics. To ensure geographic independence of the analysis, adjacent samples located up to 5 km apart were grouped into related spatial components, which subsequently served as the basis for constructing spatially independent data partitions used throughout the study. To ensure seamless feature construction and subsequent model evaluation without leaking spatial or preprocessing data, the analytical workflow was organized as a sequence of clearly separated stages: data quality, feature engineering, and modeling. The pipeline begins with defining the 2020 Sentinel-2 analytical cohort and spatially independent partitioning, followed by coordinate and radiometric data quality control, construction of spectral and vegetation index representations, preservation of temporal availability information, and preprocessing of data obtained from the training set. All transformations requiring parameter estimation are performed exclusively on the training data and then applied to subsets of the data for validation and held-out testing without re-evaluation. The complete data-safe workflow, taking into account data leakage, is presented in Figure 1.

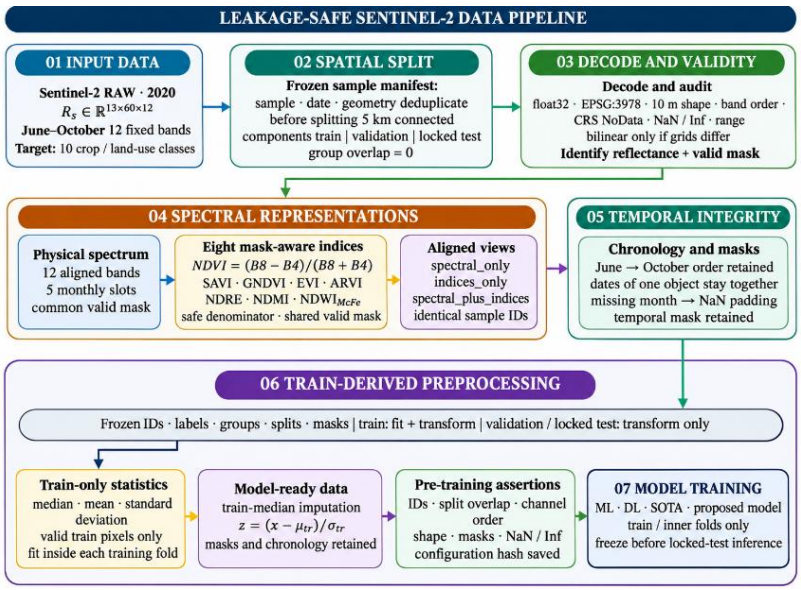



Figure 1. Methodological algorithm for assessing the quality of spatial data and designing features of multi-temporal Sentinel-2 data

As shown in Figure 1, the workflow consists of seven sequential stages. First, the analytical cohort is defined from the 2020 Sentinel-2 observations covering ten agricultural land-use classes. Second, geographic samples are partitioned using spatially connected components to prevent spatially related observations from being distributed across training, validation, and held-out test subsets. Third, coordinate validity, duplicate locations, CRS consistency, NoData/NaN/Inf values, reflectance ranges, and radiometric quality are verified before feature construction. Fourth, the quality-controlled observations are transformed into aligned spectral and vegetation-index representations while preserving identical sample identifiers and spatial splits. Fifth, the original June–October chronology is retained through temporal masks, so missing monthly observations remain explicitly represented rather than being reconstructed by temporal interpolation.

The final stages enforce separation between model development and evaluation. Preprocessing parameters, including imputation and scaling statistics, are estimated exclusively from the training subset and then applied unchanged to the validation and held-out test data. Pre-training assertions additionally verify the consistency of sample identifiers, split assignments, feature order, missing-value handling, and tensor dimensions before model fitting. Consequently, the pipeline preserves spatial independence and prevents information from the evaluation subsets from influencing feature preprocessing or model development, providing a reproducible basis for the subsequent feature-diagnostic and downstream classification experiments.

To enrich the spectral representation of agricultural surfaces, a set of widely used vegetation and moisture indices derived from the Sentinel-2 source spectral bands was developed. These indices highlight various biophysical properties of vegetation, including canopy viability, chlorophyll concentration, water content, biomass density, and canopy cover, thereby complementing the information contained in the source reflectance measurements. Eight vegetation indices were calculated for each available monthly observation: the normalized difference vegetation index (NDVI), the green normalized difference vegetation index (GNDVI), the atmospherically resilient vegetation index (ARVI), the normalized red edge difference index (NDRE), the normalized difference wetness index (NDMI), the soil-adjusted vegetation index (SAVI), the enhanced vegetation index (EVI), and the normalized difference water index (NDWI). The indices were calculated directly from radiometrically verified Sentinel-2 reflectivity values without additional normalization. The mathematical definitions of the vegetation indices used in this study are provided in Table 1.

Feature engineering was conducted to construct informative representations based on multi-temporal Sentinel-2 observations. The feature set included original spectral bands, derived vegetation indices, monthly (temporal) descriptors, and seasonal aggregated statistics calculated for the observation period from June to October. These variables were organized into groups of spectral, index, temporal, and aggregated features. To ensure consistent feature distributions, numerical variables were standardized after removing zero-variance features. All preprocessing operations, including imputation and scaling, were performed exclusively on the training data and then applied to the remaining spatial partitions without recalculation. This strategy, based solely on the training data, prevented information leakage while maintaining the spatial independence of the

analytical dataset. To assess the relevance of spectral and plant characteristics, diagnostic studies were conducted using additional statistical methods. MI and distance correlation were used to quantify nonlinear relationships with crop classes, and the Kruskal-Wallis test was used to assess differences between classes. To assess the magnitude of class separation, the corresponding effect sizes $\eta^2$ and $\varepsilon^2$ were calculated. To account for multiple comparisons, p values were adjusted using the Benjamini-Hochberg correction. The final feature matrix was constructed by combining the validated Sentinel-2 spectral bands, derived vegetation indices, monthly temporal characteristics, and seasonal aggregated statistics into a single analytical dataset. The resulting feature space included 648 spectral features, 432 vegetation index features, and 1,080 combined spectral and vegetation features. All features were organized into four groups: spectral, vegetation index, temporal, and aggregated features. This final representation served as the basis for subsequent statistical analysis and feature diagnostics.

Table 1. Vegetation indices derived from Sentinel-2 spectral bands

| Index | Formula | Sentinel-2 bands | Main application |
|---|---|---|---|
| NDVI | $\frac{B_8 - B_4}{B_8 + B_4}$ | NIR, red | Vegetation vigor |
| GNDVI | $\frac{B_8 - B_3}{B_8 + B_3}$ | NIR, green | Chlorophyll content |
| ARVI | $(\frac{B_8 - 2B_4 - B_2)}{B_8 + 2B_4 - B_2)}$ | NIR, red, blue | Atmospheric resistance |
| NDRE | $(B_{8A}–B_5)/(B_{8A}+B_5)$ | NIR, red-edge | Crop health |
| NDMI | $\frac{B_8 - B_{11}}{B_8 + B_{11}}$ | NIR, SWIR | Vegetation moisture |
| SAVI | $\frac{1.5(B_8 - B_4)}{B_8 + B_4 + 0.5}$ | NIR, red | Soil-adjusted vegetation |
| EVI | $2.5(B_8–B_4)/(B_8+6B_4–7.5B_2+1)$ | NIR, red, blue | Dense vegetation |
| NDWI | $\frac{B_3 - B_8}{B_3 + B_8}$ | Green, NIR | Surface water and moisture |

MI was estimated on the 3,879 training samples with mutual_info_classif (n_neighbors=3; random_state=1729) in ten one-versus-rest analyses. Distance correlation used double-centered absolute-distance matrices on a deterministic 100-sample-per-class subset. Reflectance was stored as float32 and was not divided by 10,000. Pixel validity required raster validity, finite values, reflectance within [-0.2, 2], a nonzero spectrum, denominator safeguards, and index bounds; no explicit Sentinel-2 scene classification layer (SCL) or cloud-probability mask was used. The spatial structure of the analytical corpus was formalized using an undirected graph (1):

$$G = (V, E) \tag{1}$$

where each vertex $v_i \in V$ corresponds to one unique geographic sample, and the spatial edge is defined by condition (2):

$$E = \{(i, j): \| s_i - s_j \|_2 \leq 5000 \text{ m}\} \tag{2}$$

where $s_i$ and $s_j$ represent the coordinates of the centers of two samples in a metric coordinate system. Transitively connected vertices form spatial connected components. For each component c and partition $p$, a binary variable is introduced (3):

$$x_{cp} \in \{0, 1\} \tag{3}$$

and constraint (4) is satisfied:

$$\sum_p x_{cp} = 1 \tag{4}$$

The 70/15/15 outer split was optimized over whole components by scipy.optimize.milp (HiGHS). The objective minimized normalized deviations in total, class, and province counts using weights 8, 20, and 1. Each component was assigned to exactly one partition, with at least five class-bearing components in training and one in each evaluation subset. Presolve was enabled, the mixed-integer programming (MIP) relative gap was 1e-4, the time limit was 180 s, and deterministic tie-breaking used seed 1729. Validation and held-out

performance were not used in partition construction. A component belongs entirely to only one subset and cannot be shared between training, validation, and held-out testing. This restriction prevents directly related spatial samples from appearing simultaneously in the training and evaluation sets. The feature space is also formed deterministically. For each physical spectral/index variable, five fixed monthly positions and one full-season aggregate are used. Nine statistical features are calculated in each temporal scope. Therefore, the number of engineered features is determined as (5):

$$d = q \times 6 \times 9 \quad (5)$$

where $q$ is the number of physical variables. The deterministic feature-engineering procedure generated a total of 648 spectral features from the twelve Sentinel-2 bands and 432 index-derived features from the eight vegetation and moisture indices. The integration of these two feature groups resulted in a combined representation comprising 1,080 engineered features. All transformation parameters that require evaluation on data are defined only within the corresponding training scope and then applied to the evaluation subset without re-evaluation.

Each physical variable produced nine statistics-mean, standard deviation (SD), median, P10, P25, P75, P90, interquartile range, and valid-pixel fraction-for five monthly positions and one full-season scope. Missing months kept their chronological positions; numerical summaries remained missing and valid-pixel fraction was zero. Where required, missing values were filled with medians fitted only on the corresponding training scope.

To evaluate whether the engineered feature representation provided measurable predictive information beyond the individual feature groups, a controlled downstream multiclass classification experiment was conducted. Four alternative input representations were evaluated: i) spectral-only features comprising 648 engineered variables derived from the twelve Sentinel-2 bands; ii) index-only features comprising 432 variables derived from the eight vegetation and moisture indices; iii) the complete spectral-plus-index representation comprising 1,080 features; and iv) a reduced spectral-plus-index representation containing 1,026 features after removal of the 54 McFeeters NDWI-derived variables identified as deterministically redundant.

To isolate the effect of feature representation from the effect of classifier selection, all four configurations were evaluated using the same HistGradientBoosting classifier and the same spatially independent validation subset. The classifier configuration and preprocessing protocol were kept fixed across the four experiments. All preprocessing operations requiring parameter estimation, including missing-value imputation and feature scaling where applicable, were fitted exclusively on the corresponding training data and subsequently applied to the validation subset without re-estimation. Classification performance was assessed using overall accuracy, balanced accuracy, macro-F1, and the Matthews correlation coefficient (MCC). Particular emphasis was placed on macro-F1 and balanced accuracy because the analytical cohort exhibits unequal class frequencies. The validation subset was not used for feature construction, preprocessing-parameter estimation, or feature selection. This controlled design therefore enables differences in predictive performance to be interpreted primarily in relation to changes in the input feature representation.

OS-MMF stacks class-probability vectors from extreme gradient boosting (XGBoost), HistGradientBoosting, multinomial logistic regression, and extra trees using a class-balanced multinomial logistic meta-learner. Meta-regularization and blending with the strongest component were selected only from five spatial training folds; validation and held-out data were not used for fitting. Results average seeds 17, 29, and 43. The post-2024 random forest (RF) backward-feature-reduction comparator [6] used training out-of-fold (OOF) selection only. Because the published method also uses Sentinel-1 SAR, it is reported as a same-data adaptation rather than an exact reproduction.

Two complementary experimental designs were used. First, HistGradientBoosting was employed as a fixed classifier to evaluate the contribution of alternative feature representations. Second, the proposed OOF-stacked multispectral machine learning fusion (OS-MMF) framework was compared with HistGradientBoosting, masked temporal attention, RF-BFR, and RF under the same spatially independent validation protocol. Thus, the feature removal experiment and the model comparison experiment address different research questions while employing a consistent leakage-aware spatial partitioning strategy.

## 3. RESULTS

Following the quality control procedures described in section 2, the analytical sample consisted of 5,563 unique georeferenced agricultural accessions represented by 22904 multi-temporal Sentinel-2 observations acquired during the 2020 growing season. The dataset covers ten agricultural land-use classes distributed across eight Canadian provinces, providing broad geographic coverage and diverse crop growing conditions. The main characteristics of the analytical dataset are summarized in Table 2.

Table 2. Characteristics of the analytical dataset used for spatial data quality assessment and feature engineering

| Characteristic | Value |
|---|---|
| Dataset | CCD |
| Satellite mission | Sentinel-2 MSI |
| Study year | 2020 |
| Study area | Agricultural regions of Canada |
| Number of crop/land-use classes | 10 |
| Number of unique samples | 5,563 |
| Monthly observations | 22904 |
| Observation period | June–October |
| Maximum observations per sample | 5 |
| Spectral bands | 12 (B1–B9, B8A, B11, B12) |
| Band excluded | B10 |
| Coordinate duplicates removed | 29 |
| Analytical cohort | 5,563 unique georeferenced samples |
| Coordinate reference system | EPSG:3978 |

Table 3 presents the distribution of the resulting analytical cohort across ten categories of agricultural crops and land-use types. The distribution demonstrates substantial heterogeneity in the class structure, with noticeable differences in the number of samples assigned to individual categories. Some crop classes are represented by considerably larger numbers of observations, whereas others contain relatively fewer samples. This class imbalance should therefore be considered when interpreting the subsequent classification results and evaluating model performance across individual categories.

Table 3. Distribution of the analytical cohort across agricultural land-use classes

| Class code | Number of samples | Share of samples, % | Number of monthly observations | Observations per sample |
|---|---|---|---|---|
| Barley | 487 | 8.75 | 2,083 | 4.28 |
| Canola | 763 | 13.72 | 3,263 | 4.28 |
| Corn | 1574 | 28.29 | 6,518 | 4.14 |
| Oat | 204 | 3.67 | 805 | 3.95 |
| Orchard | 471 | 8.47 | 1,479 | 3.14 |
| Pasture | 522 | 9.38 | 2,188 | 4.19 |
| Pea | 142 | 2.55 | 606 | 4.27 |
| Potato | 149 | 2.68 | 622 | 4.17 |
| Soybean | 358 | 6.44 | 1,509 | 4.22 |
| Spring_wheat | 893 | 16.05 | 3,831 | 4.29 |
| Total | 5,563 | 100.00 | 22,904 | 4.12 |

Table 3 presents the distribution of the analytical cohort across ten agricultural land-use classes. The dataset exhibits a moderately imbalanced class distribution, with corn representing the largest category (1,574 samples; 28.29%), followed by spring_wheat (893 samples; 16.05%) and canola (763 samples; 13.72%). Together, these three dominant classes account for 58.06% of the analytical cohort. In contrast, pea (142 samples; 2.55%), potato (149 samples; 2.68%), and oat (204 samples; 3.67%) are the least represented classes. The ratio between the largest and smallest classes is approximately 11:1. Despite this class imbalance, temporal coverage remains relatively stable across most categories, with approximately four available monthly observations per sample. Orchard represents the main exception, with an average of 3.14 observations per sample, indicating lower temporal availability than the other classes. Overall, the analytical cohort retains representation across all ten agricultural land-use classes while providing sufficient temporal coverage for subsequent feature engineering and statistical analysis.

Figure 2 shows the number of unique geographic centers after removing exact coordinate duplicates and before decoding the pixel arrays. The final analytical cohort includes 5,563 samples distributed across ten target categories. Plotting the diagram at the unique center level eliminates duplicate counts of the same spatial object and reflects the actual number of independent units of analysis used in the subsequent formation of the training, validation, and held-out test subsets.

Figure 2 shows the distribution of the spatially deduplicated analytical sample across ten agricultural land-use classes after removing duplicate geographic coordinates. The class distribution is noticeably unbalanced: the corn category is the largest (1,574 samples), followed by spring_wheat (893 samples) and canola (763 samples). In contrast, pea, potato, and oat constitute the smallest classes, each containing fewer than 250 samples. Despite this imbalance, all agricultural land-use categories are represented by a sufficient number of geographically unique observations to support subsequent statistical analysis and feature diagnostics. The observed class composition reflects the natural distribution of agricultural land in the CCD and provides a representative basis for assessing feature quality and spatial variability across crop types.

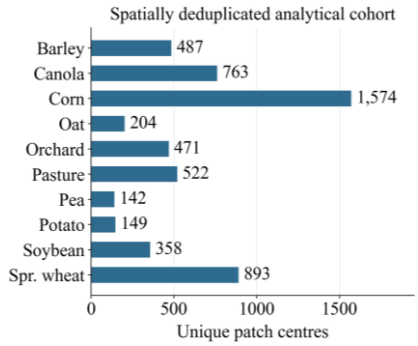



Figure 2. Distribution of spatially deduplicated analytical cohort by crop categories and land use types

To assess the spatial quality of the analytical dataset, the geographic distribution of the samples was examined after duplicate removal and spatial-integrity verification. Application of the 5 km connectivity criterion to the 5,563 unique geographic samples resulted in 1,983 connected spatial components. These components were treated as indivisible spatial units during partitioning and were assigned entirely to the training, validation, or held-out test subset. Specifically, 1,145 components containing 3,879 samples were assigned to the training subset, 475 components containing 842 samples to the validation subset, and 363 components containing 842 samples to the held-out test subset. No connected component was divided across analytical partitions. The characteristics of the resulting spatial partitions are summarized in Table 4.

Table 4. Spatial statistics of partitioning of analytical dataset

| Spatial partition | 5-km connected components | Samples (n) | Samples (%) | Monthly acquisitions | Acquisitions (%) |
|---|---|---|---|---|---|
| Training | 1,145 | 3,879 | 69.73 | 15,836 | 69.14 |
| Validation | 475 | 842 | 15.14 | 3,415 | 14.91 |
| Held-out test | 363 | 842 | 15.14 | 3,653 | 15.95 |
| Total | 1,983 | 5,563 | 100.00 | 22,904 | 100.00 |

The component-based partitioning resulted in approximately 70%, 15%, and 15% of the geographic samples being assigned to the training, validation, and held-out test subsets, respectively. Because each connected component was retained as an indivisible spatial unit, no component was shared across analytical partitions. To further assess the consistency of the spatial partitioning strategy, the relative distribution of agricultural land-use classes across the three spatially independent subsets was examined. The analysis was conducted to determine whether the component-based partitioning preserved the overall class composition while maintaining geographic independence between partitions. The resulting class proportions are presented in Table 5. Direct coordinate verification gave minimum centre-to-centre distances of 5007.28 m between training and validation, 5002.77 m between training and held-out test, and 5089.93 m between validation and held-out test; no cross-partition pair fell below the 5 km threshold.

Table 5. Class-wise proportions of the spatially independent partitions

| Agricultural land-use class | Total samples | Training (%) | Validation (%) | Test (%) |
|---|---|---|---|---|
| Barley | 487 | 70.02 | 14.99 | 14.99 |
| Canola | 763 | 70.12 | 14.94 | 14.94 |
| Corn | 1.574 | 70.01 | 14.99 | 14.99 |
| Oat | 204 | 69.61 | 15.20 | 15.20 |
| Orchard | 471 | 66.45 | 16.77 | 16.77 |
| Pasture | 522 | 70.11 | 14.94 | 14.94 |
| Pea | 142 | 70.42 | 14.79 | 14.79 |
| Potato | 149 | 70.47 | 14.77 | 14.77 |
| Soybean | 358 | 69.83 | 15.08 | 15.08 |
| Spring wheat | 893 | 69.99 | 15.01 | 15.01 |
| Overall | 5.563 | 69.73 | 15.14 | 15.14 |

The component-based spatial partitioning maintained virtually identical class proportions across all three partitions. Minor deviations, particularly for the orchard class, arose from assigning complete, connected spatial components to a single partition rather than splitting adjacent observations. Overall, the partitioning strategy achieved a balanced class distribution while maintaining strict geographic independence between the analytical subsets. To visually assess the effectiveness of the spatial partitioning strategy, the geographic

distribution of the analytical dataset was examined after assigning each related spatial component to a single spatially independent partition. The resulting partitioning is presented in Figure 3.

Figure 3 shows the spatial distribution of three geographically independent partitions formed from related spatial components. The partitioning preserves the natural spatial structure of the analytical dataset while ensuring that each related component is entirely assigned to a single partition. No spatial overlap is observed between adjacent components, indicating that the adopted partitioning strategy effectively minimizes spatial dependency while maintaining representative geographic coverage across the study area. Figure 4 shows the distribution of 1,983 connected spatial components by the number of geographic samples they comprise. Components were formed using a 5 km connectivity criterion: two samples were considered connected if the distance between their centers did not exceed a set threshold, including transitive connections through intermediate points. The vertical axis is plotted on a logarithmic scale, as the frequencies of components of different sizes differ by several orders of magnitude.

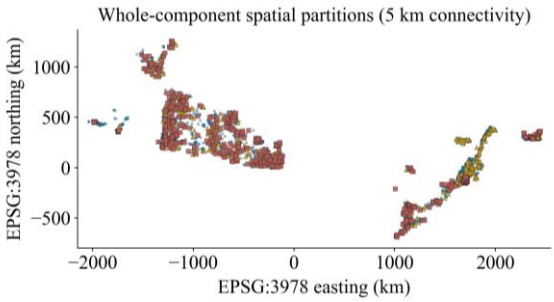



Figure 3. Spatial distribution of geographically independent training, validation, and test samples

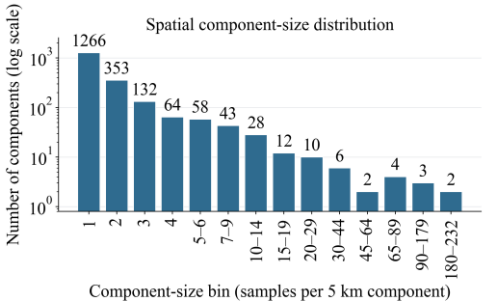



Figure 4. Distribution of sizes of 5 km spatial components of the analytical cohort

Figure 4 shows a highly right-skewed distribution of spatial component sizes. Most components are small: 1,266 (63.84%) consist of a single geographic sample, and nearly 90% contain no more than three samples. Only a small number of components contain several dozen or more than one hundred samples. This distribution demonstrates the heterogeneous spatial organization of the analytical cohort and supports the use of component-based spatial partitioning to preserve geographic independence while reducing the risk of information leakage caused by spatial autocorrelation. Following the spatial assessment, temporal completeness was evaluated according to the number of available monthly Sentinel-2 observations for each geographic sample during the June-October growing season. Because each sample could contain up to five monthly observations, the analysis quantified the frequency of samples with one to five available observations without assuming complete temporal coverage. Missing monthly observations were retained as unavailable temporal positions rather than reconstructed through temporal interpolation. The resulting distribution of temporal completeness is presented in Table 6.

Table 6. Temporal completeness of multitemporal Sentinel-2 observations

| Number of available monthly observations | Samples (n) | Samples (%) | Cumulative (%) |
|---|---|---|---|
| 1 | 34 | 0.61 | 0.61 |
| 2 | 199 | 3.58 | 4.19 |
| 3 | 1,051 | 18.89 | 23.08 |
| 4 | 2,076 | 37.32 | 60.40 |
| 5 | 2,203 | 39.60 | 100.00 |
| Total | 5,563 | 100.00 | — |

The temporal completeness analysis indicates that most geographic samples retained substantial coverage of the June-October growing season. Of the 5,563 samples, 2203 (39.60%) contained all five monthly observations, while 2076 (37.32%) contained four observations. Thus, 76.92% of the analytical cohort had four or five available monthly observations. A further 1051 samples (18.89%) contained three observations, meaning that 95.81% of the samples retained at least three monthly observations. Only 233 samples (4.19%) contained one or two observations. Overall, the analytical cohort comprised 22904 available monthly observations, corresponding to an average of approximately 4.12 observations per geographic sample. These results indicate generally high temporal coverage while also confirming the presence of incomplete temporal sequences, supporting the use of temporal availability masks rather than assuming that all samples contain five observations.

Following the temporal completeness assessment, the numerical distributions of the twelve Sentinel-2 spectral bands were examined as part of the radiometric quality-control procedure. Table 7 summarizes the distributional characteristics of the analyzed spectral values using robust descriptive statistics, including the minimum, first percentile, median, 99th percentile, and maximum. Median reflectance values generally increase from the visible wavelengths toward the red-edge and near-infrared regions, with the highest median values observed for B8, B8A, and B9. In the shortwave-infrared region, the median values decrease to approximately 0.2249 for B11 and 0.1399 for B12.

Table 7. Sample distribution characteristics of the values of twelve spectral channels of Sentinel-2 (analyzed values=63744)

| Bands | Median | P99 | Max |
|---|---|---|---|
| B1 | 0.0359 | 0.1361 | 0.58800 |
| B2 | 0.0521 | 0.1988 | 0.61320 |
| B3 | 0.0938 | 0.2685 | 0.61880 |
| B4 | 0.0812 | 0.3750 | 0.68580 |
| B5 | 0.1527 | 0.4319 | 0.72115 |
| B6 | 0.2800 | 0.5074 | 0.72700 |
| B7 | 0.3304 | 0.6086 | 0.74340 |
| B8 | 0.3484 | 0.6107 | 0.77640 |
| B8A | 0.3678 | 0.6211 | 0.73700 |
| B9 | 0.3697 | 0.6256 | 1.23960 |
| B11 | 0.2249 | 0.4703 | 0.54930 |
| B12 | 0.1399 | 0.3658 | 0.52505 |

Zero values are present in both the minimum and first percentile of all bands. Therefore, zero was not automatically interpreted as a physical measure of reflectance: its validity was determined in conjunction with NoData, finiteness of values, and the overall pixel mask. The most prominent extreme value was found in B9-1.23960, while its 99th percentile is 0.62560. This indicates the rarity of such observations and confirms the need to rely on robust quantiles, not just absolute minimums and maximums. The resulting distributions were used for quality control and mask tuning, but not for normalization across the entire analytical cohort. To assess the relationships between Sentinel-2 spectral bands and vegetation indices and to identify potentially redundant variables, a correlation analysis was conducted. Pairwise associations were quantified using Pearson and Spearman correlation coefficients, the latter providing a robust estimate of monotonic relationships between multispectral characteristics. The resulting correlation matrix of Sentinel-2 spectral bands is presented in Figure 5.

Figure 5 shows strong positive correlations between adjacent Sentinel-2 spectral bands. The highest correlation coefficients are observed in the red and near-infrared regions (B6–B9; ρ=0.95–0.99), indicating significant spectral redundancy. Conversely, weaker correlations between the visible and near-infrared bands suggest that these spectral regions provide additional information for subsequent feature engineering. This analysis evaluates monotonic associations between variables and provides insight into the complementarity and redundancy of spectral and index characteristics. The resulting correlation matrix is presented in Figure 6.

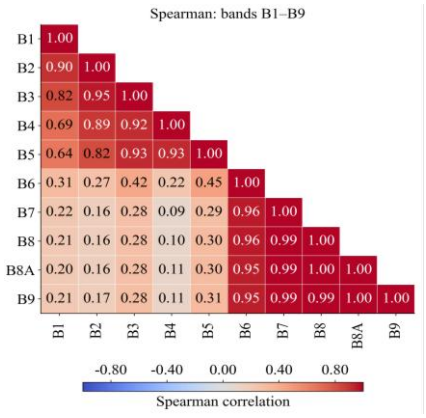



Figure 5. Spearman correlation matrix of Sentinel-2 spectral bands (B1–B9)

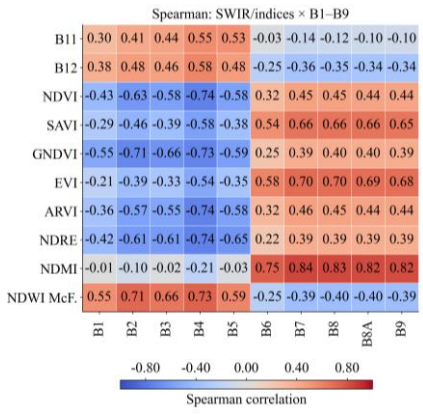



Figure 6. Spearman correlation matrix between Sentinel-2 SWIR bands, vegetation indices, and spectral bands (B1–B9)

Figure 6 shows clear correlation patterns between the vegetation indices and the Sentinel-2 spectral bands. Strong positive correlations are observed between most vegetation indices and the red edge and near infrared bands (B6–B9), reaching ρ=0.84 for NDMI and ρ=0.70 for EVI, while moderate to strong negative correlations are found with the visible bands (up to ρ=–0.74 for NDRE). The SWIR bands (B11 and B12) show moderate positive correlations with the visible bands (ρ=0.30–0.58) and generally weak to moderate relationships with the vegetation indices. These results indicate that the vegetation indices contain additional information beyond the original spectral bands, supporting their inclusion in the final feature presentation. To assess the nonlinear relationship between individual Sentinel-2 spectral bands and agricultural land use, MI was calculated using the one-versus-all (OVR) strategy for each crop class. Unlike correlation indices, MI takes into account both linear and nonlinear relationships without assuming a specific functional relationship between variables. The resulting MI values are presented in Figure 7.

Figure 7 shows that the information content of spectral bands varies significantly across agricultural land-use classes. The highest MI values are observed for rapeseed, reaching 0.12 for B5 and 0.11 for B3, followed by corn, where MI values for B5 and B2 reach 0.08. Most of the remaining classes exhibit relatively low MI values (typically <0.05), indicating that individual spectral bands themselves provide limited discriminatory information. To further evaluate the nonlinear discriminatory power of the derived vegetation indices and SWIR ranges, MI was calculated using a OVR strategy for each agricultural land-use class. This analysis quantifies the amount of information provided by each derived feature for individual crop categories. The resulting MI values are presented in Figure 8.

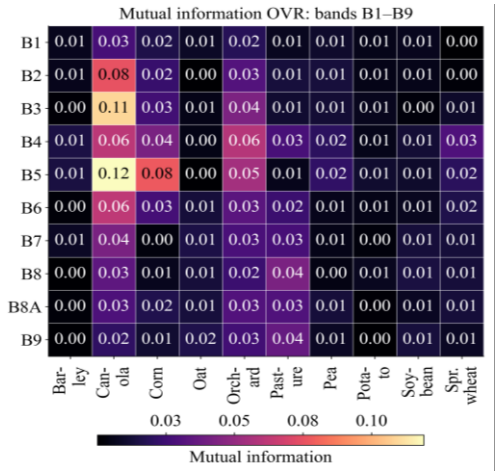



Figure 7. MI between Sentinel-2 spectral bands and agricultural land use classes

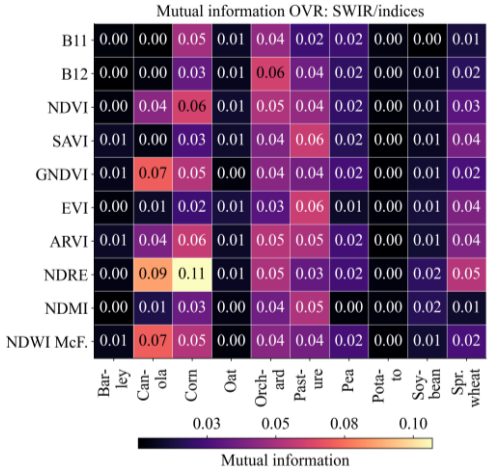



Figure 8. MI indices between vegetation indices, SWIR ranges, and agricultural land use classes

Figure 8 shows that the resulting vegetation indices generally provide higher nonlinear information content than the original spectral bands. The highest MI values are obtained for NDRE, reaching 0.11 for maize and 0.09 for rapeseed, indicating high discriminatory power for these crop types. Moderate MI values are also observed for GNDVI and McFeeters' NDWI, with maximum values of 0.07, while NDVI, SAVI, EVI, and ARVI reach peak MI values of approximately 0.06. In contrast, the SWIR bands (B11 and B12) demonstrate relatively low information content, with maximum MI values not exceeding 0.06. In addition to the MI analysis, the distance correlation coefficient (dCor) was calculated to quantify both linear and nonlinear relationships between Sentinel-2 spectral bands and agricultural land-use classes. Unlike standard correlation coefficients, distance correlation reveals any statistical relationship between variables and is zero only when they are statistically independent. The resulting distance correlation coefficients are presented in Figure 9.

Figure 9 shows that the strength of the relationship between Sentinel-2 spectral bands and agricultural land-use classes varies significantly depending on crop type. The highest distance correlation values are observed for pastures, reaching 0.36 for B4, 0.35 for B5, and 0.32 for B3, indicating a strong nonlinear relationship. Higher values are also found for rapeseed, where the maximum coefficient reaches 0.34, followed by orchards with values up to 0.29. In contrast, barley, peas, potatoes, soybeans, and spring wheat exhibit comparatively weaker relationships, with most coefficients remaining below 0.15. To further evaluate the nonlinear relationships between the obtained vegetation indices, SWIR ranges, and agricultural land-use classes, the distance correlation coefficient (dCor) was calculated using a OVR strategy. This analysis

complements the results of the MI analysis by quantifying the overall statistical relationship between the obtained traits and individual crop categories. The resulting distance correlation coefficients are presented in Figure 10.

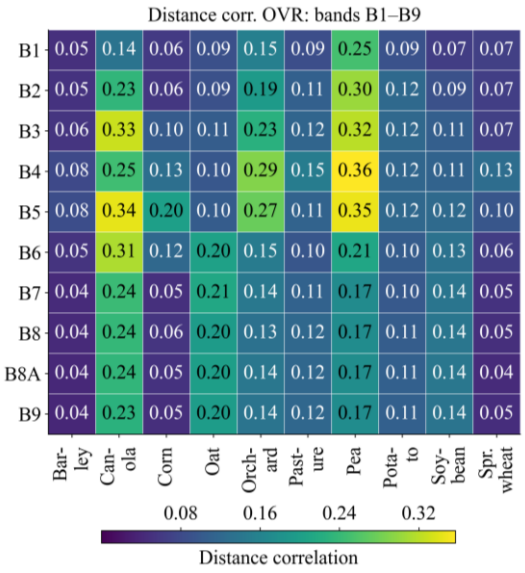



Figure 9. Correlation of distances between Sentinel-2 spectral bands and agricultural land use classes

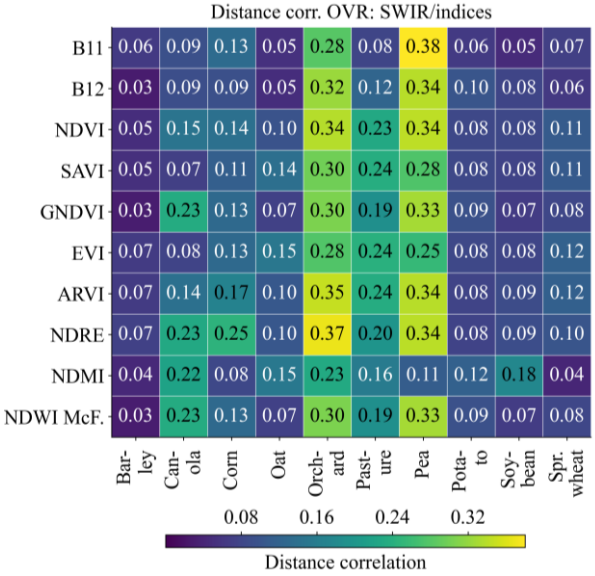



Figure 10. Correlation of distances between vegetation indices, SWIR ranges, and agricultural land use classes

Figure 10 shows that the derived vegetation indices exhibit stronger nonlinear relationships with agricultural land-use classes than the original spectral bands. The highest distance correlation is observed for B11 with pasture (dCor=0.38), followed by NDRE (0.37) and ARVI (0.35) for the orchard. Several vegetation indices, including NDVI, GNDVI, McFeeters' NDWI, and B12, also exhibit consistently high correlations with pasture and pea, with coefficients ranging from 0.30 to 0.34. In contrast, barley, potato, soybean, and spring wheat generally exhibit weaker relationships, with most coefficients remaining below 0.15. Table 8 summarizes the train-only univariate diagnostics. NDRE showed the strongest overall association, with MI of 0.3178, eta squared of 0.3277, maximum one-versus-rest distance correlation of 0.3691, Kruskal-Wallis H of 1426.25, and epsilon squared of 0.3663; its Benjamini-Hochberg adjusted p-value was below 0.0001. B5 ranked next on MI and effect size, further supporting the diagnostic value of the red-edge region.

Table 8. The most pronounced train-only associations of spectral bands and indices with agricultural cover categories

| Variable | MI | $\eta^2$ | Max one-vs-rest distance correlation | Kruskal–Wallis H | $\varepsilon^2$ |
|---|---|---|---|---|---|
| NDRE | 0.3178 | 0.3277 | 0.3691 | 1 426.2457 | 0.3663 |
| B5 | 0.2661 | 0.2952 | 0.3464 | 1 398.3839 | 0.3591 |
| ARVI | 0.2526 | 0.2480 | 0.3475 | 1 065.0436 | 0.2730 |
| B4 | 0.2331 | 0.2237 | 0.3568 | 1 078.1431 | 0.2763 |
| NDVI | 0.2293 | 0.2317 | 0.3419 | 988.2847 | 0.2531 |
| GNDVI | 0.2282 | 0.2352 | 0.3292 | 999.4582 | 0.2560 |
| NDWI по McFeeters | 0.2282 | 0.2352 | 0.3292 | 999.4582 | 0.2560 |
| B3 | 0.1919 | 0.1895 | 0.3286 | 1 034.6007 | 0.2651 |
| EVI | 0.1751 | 0.1587 | 0.2771 | 669.8895 | 0.1708 |
| SAVI | 0.1726 | 0.1614 | 0.2978 | 669.8511 | 0.1708 |
| B6 | 0.1713 | 0.1659 | 0.3130 | 578.6094 | 0.1472 |
| B12 | 0.1543 | 0.1118 | 0.3395 | 490.7070 | 0.1245 |
| B2 | 0.1507 | 0.1102 | 0.3035 | 714.4121 | 0.1823 |
| NDMI | 0.1456 | 0.0903 | 0.2303 | 395.7294 | 0.1000 |
| B11 | 0.1415 | 0.1454 | 0.3803 | 527.8582 | 0.1341 |

Among the derived traits, ARVI, NDVI, and GNDVI also demonstrate relatively strong associations with several agricultural land-use classes. The identical diagnostic results obtained for GNDVI and McFeeters NDWI are attributable to their deterministic mathematical relationship rather than to independent information content. Although B11 achieves the highest one-versus-rest distance correlation (0.3803), it exhibits lower MI and effect sizes, indicating class-specific rather than general discriminatory power. All traits remained statistically significant after the Benjamini–Hochberg correction (adjusted $p<0.001$); however, statistical significance alone does not imply independent predictive value, as the analysis evaluates each trait individually without considering inter-feature correlations or joint effects.

The statistical feature diagnostics presented above quantify the association between individual variables and agricultural land-use classes but do not establish whether these variables provide incremental predictive information when considered jointly. Therefore, the practical contribution of the engineered feature space was further evaluated through controlled downstream multiclass classification. The same classifier, pre-processing constraints, and spatially independent validation subset were used for all feature configurations, with only the input representation varied Table 9.

Table 9. Downstream classification performance of alternative feature representations on the spatially independent validation subset

| Feature representation | Number of features | Accuracy | Balanced accuracy | Macro-F1 | MCC |
|---|---|---|---|---|---|
| Spectral only | 648 | 0.8005 | 0.7079 | 0.7181 | 0.7638 |
| Indices only | 432 | 0.7910 | 0.6662 | 0.6819 | 0.7518 |
| Spectral+indices | 1,080 | 0.8052 | 0.7088 | 0.7261 | 0.7691 |
| Spectral+indices without McFeeters NDWI | 1,026 | 0.8052 | 0.7104 | 0.7255 | 0.7692 |

The spectral-only representation yielded an accuracy of 0.8005, a balanced accuracy of 0.7079, and a macro-F1 score of 0.7181. The index-only representation performed worse: a macro-F1 score of 0.6819 and a balanced accuracy of 0.6662. Combining spectral and index features increased the accuracy to 0.8052 and the macro-F1 score to 0.7261. The absolute improvement in macro-F1 over the spectral-only representation was 0.0080, indicating a moderate additional contribution to prediction from the derived indices when combined with the original spectral information. Removing 54 features derived from the McFeeters NDWI reduced the dimensionality of the combined representation from 1,080 to 1,026 features, with little change in predictive performance. The macro-F1 value changed from 0.7261 to 0.7255, accuracy remained at 0.8052, and balanced accuracy increased slightly from 0.7088 to 0.7104. This result is consistent with the deterministic redundancy identified during the statistical feature diagnostics stage and indicates that increasing feature dimensionality does not necessarily provide additional predictive information.

Overall, the subsequent experiment provides quantitative evidence that the generated representations retain useful information for multi-class agricultural land classification within the spatially independent evaluation protocol. At the same time, the limited difference between the full and reduced representations demonstrates the importance of distinguishing the statistical significance of features from their independent predictive contribution. To assess the effectiveness of the proposed OS-MMF model, a head-to-head comparison was performed with four reference approaches using the same spatially independent validation data and evaluation metrics. The comparative results are presented in Table 10.

Table 10. Same-data validation comparison of the proposed method and reference approaches

| Model | Number of features | Accuracy | Balanced accuracy | Macro-F1 | MCC | Cohen's κ |
|---|---|---|---|---|---|---|
| OS-MMF—proposed model | 1,080 | 0.8100 | 0.7570 | 0.7408 | 0.7766 | 0.7764 |
| HistGradientBoosting | 1,080 | 0.8052 | 0.7088 | 0.7261 | 0.7691 | 0.7687 |
| Masked temporal attention | 1,080 | 0.7922 | 0.7692 | 0.7269 | 0.7591 | 0.7577 |
| RF-BFR post-2024 same-data adaptation | 270 | 0.7708 | 0.6122 | 0.6381 | 0.7268 | 0.7244 |
| RF—complete representation | 1,080 | 0.7660 | 0.6224 | 0.6456 | 0.7213 | 0.7190 |

OS-MMF achieved the highest accuracy (0.8100), macro-F1 (0.7408), MCC (0.7766), and Cohen's κ (0.7764). Masked temporal attention achieved the highest balanced accuracy (0.7692), compared with 0.7570 for OS-MMF. HistGradientBoosting showed competitive performance but lower macro-F1 (0.7261), while both RF configurations produced lower results across the reported metrics. Overall, OS-MMF demonstrated the strongest performance in four of the five evaluation metrics under the same-data validation protocol. To further examine class-specific predictive performance, precision, recall, and F1-score were

calculated for each of the ten agricultural land-use classes using the complete spectral-plus-index representation. The results are presented in Table 11.

Table 11. Class-wise validation performance for the complete spectral-plus-index representation

| Class | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| Barley | 0.470 | 0.425 | 0.446 | 73 |
| Canola | 0.929 | 0.921 | 0.925 | 114 |
| Corn | 0.925 | 0.941 | 0.933 | 236 |
| Oat | 0.450 | 0.290 | 0.353 | 31 |
| Orchard | 0.951 | 0.987 | 0.969 | 79 |
| Pasture | 0.763 | 0.782 | 0.772 | 78 |
| Pea | 0.929 | 0.619 | 0.743 | 21 |
| Potato | 0.722 | 0.591 | 0.650 | 22 |
| Soybean | 0.727 | 0.741 | 0.734 | 54 |
| Spring_wheat | 0.688 | 0.791 | 0.736 | 134 |

The class-wise results reveal substantial variation in classification performance. Orchard, corn, and canola achieved the highest F1-scores of 0.969, 0.933, and 0.925, respectively. In contrast, oat and barley were the most difficult classes to distinguish, with F1-scores of 0.353 and 0.446. Pea showed high precision (0.929) but lower recall (0.619), indicating that correctly predicted pea samples were generally reliable, although a considerable proportion of actual pea samples were missed. Overall, these results demonstrate that aggregate performance metrics should be interpreted together with class-specific results, particularly for the less represented and more difficult-to-discriminate classes. To assess the stability of classification performance with respect to spatial partitioning, the four feature representations were additionally evaluated across ten repeated whole-component spatial partitions. Connected spatial components were kept intact in every repetition to preserve geographic independence between the training and evaluation subsets. Performance variability was quantified using the mean and SD of accuracy, balanced accuracy, macro-F1, and MCC across the ten repetitions. The results are presented in Table 12.

Table 12. Performance across ten repeated whole-component spatial partitions

| Representation | Accuracy, mean±SD | Balanced accuracy, mean±SD | Macro-F1, mean±SD | MCC, mean±SD |
|---|---|---|---|---|
| Spectral only | 0.8018±0.0223 | 0.7225±0.0297 | 0.7319±0.0295 | 0.7665±0.0252 |
| Indices only | 0.8004±0.0210 | 0.7086±0.0214 | 0.7178±0.0262 | 0.7645±0.0238 |
| Spectral+indices | 0.8060±0.0192 | 0.7252±0.0283 | 0.7362±0.0279 | 0.7711±0.0219 |
| Spectral+indices without McFeeters NDWI | 0.8081±0.0175 | 0.7269±0.0288 | 0.7372±0.0292 | 0.7736±0.0198 |

The repeated spatial evaluation showed generally stable performance across the ten partitions. The complete spectral-plus-index representation achieved a mean macro-F1 of 0.7362±0.0279, compared with 0.7319±0.0295 for spectral-only features and 0.7178±0.0262 for index-only features. The reduced representation without McFeeters NDWI achieved the highest mean performance, with an accuracy of 0.8081±0.0175, balanced accuracy of 0.7269±0.0288, macro-F1 of 0.7372±0.0292, and MCC of 0.7736±0.0198. The similar performance of the complete and reduced representations further indicates that removal of the redundant McFeeters NDWI-derived features did not adversely affect predictive performance. Overall, the relatively small SDs indicate that the observed results were reasonably stable across alternative spatial partitions.

## 4. DISCUSSION

The revised analysis shows that the main methodological value of the framework lies in controlling spatial dependence before feature construction and model evaluation. After removal of 29 duplicate coordinates, 5,563 samples formed 1,983 connected components under the 5 km rule, with no component shared across training, validation, or held-out test subsets. Direct coordinate checks yielded minimum cross-partition distances of 5,007.28 m, 5,002.77 m, and 5,089.93 m for the three partition pairs. Temporal coverage was high but incomplete: 76.92% of samples contained four or five observations, whereas 39.60% contained all five monthly positions. Retaining unavailable months in their original chronological positions therefore preserves the observed seasonal structure without introducing interpolated measurements.

The feature diagnostics also clarify the distinction between statistical association and independent predictive contribution. NDRE showed the strongest overall train-only univariate association, with MI of

0.3178, eta squared of 0.3277, maximum one-versus-rest distance correlation of 0.3691, and epsilon squared of 0.3663. At the same time, GNDVI and McFeeters NDWI are deterministically sign-reversed formulations of the same B8 and B3 information. Removing the 54 McFeeters NDWI-derived variables reduced the combined representation from 1.080 to 1.026 features with essentially unchanged validation performance: macro-F1 changed from 0.7261 to 0.7255, while balanced accuracy changed from 0.7088 to 0.7104. This result supports using feature diagnostics to identify redundancy rather than interpreting statistical significance as evidence of incremental predictive value.

The downstream experiments provide a direct test of the engineered representations under a common classifier and spatial split. HistGradientBoosting achieved macro-F1 values of 0.7181 with spectral-only features, 0.6819 with index-only features, and 0.7261 with the complete representation. Across ten repeated whole-component development partitions, the complete representation achieved a mean macro-F1 of 0.7362 with a SD of 0.0279, while the reduced representation without McFeeters NDWI achieved 0.7372 with a SD of 0.0292. These results indicate that combining spectral and index information is useful, but they do not support the stronger claim that every engineered index or additional feature dimension improves classification.

OS-MMF achieved the highest validation accuracy, macro-F1, MCC, and Cohen's kappa in the same-data comparison, with values of 0.8100, 0.7408, 0.7766, and 0.7764, respectively, whereas masked temporal attention produced the highest balanced accuracy of 0.7692. The validation macro-F1 advantage of OS-MMF over its strongest component was 0.0127, but the paired spatial-cluster interval ranged from -0.0127 to 0.0392 and the cluster-permutation test gave p=0.316. On the final held-out test, OS-MMF achieved an accuracy of 0.7803, balanced accuracy of 0.7027, macro-F1 of 0.7080, and MCC of 0.7414; the macro-F1 spatial-component bootstrap interval was 0.5958 to 0.7694. The evidence therefore supports competitive, spatially controlled performance, but not statistically established universal superiority. The conclusions remain limited to the ten classes and the 2020 Canadian cohort.

## 5. CONCLUSION

In this study, a reproducible workflow for spatially supervised feature engineering and classification of agricultural lands using multi-temporal Sentinel-2 observations was developed. The final sample included 5,563 unique samples, 22,904 measurements, and 1,983 associated spatial components. The engineered feature space contained 648 spectral features, 432 features derived from the vegetation index, and 1,080 combined features. In supervised validation, the full representation achieved a macro-F1 score of 0.7261 using HistGradientBoosting, while ten repeated spatial partitions into integer components yielded an average macro-F1 score of 0.7362. Removing deterministically redundant features derived from the McPheeters NDWI index had little effect on performance, demonstrating that increasing the dimensionality of the feature space does not necessarily provide additional predictive information.

The proposed OS-MMF model achieved a validation accuracy of 0.8100 and a macro-F1 score of 0.7408. On the final test set, it achieved an accuracy of 0.7803, a balanced accuracy of 0.7027, a macro-F1 score of 0.7080, and a macro-filter correlation coefficient (MCC) of 0.7414. A pairwise spatial uncertainty analysis did not demonstrate a statistically significant advantage over all strong baseline models; therefore, the proposed model should be interpreted as providing competitive and reproducible performance, rather than as universally superior. More broadly, the results indicate that reliable agricultural land classification depends not only on classifier complexity but also on spatially independent estimation, robust preprocessing, and data-driven feature construction.

The study is limited to one growing season, ten agricultural land-use classes, and the 2020 Canadian cohort. Furthermore, the current model, based on Sentinel-2 data, does not explicitly account for short-term meteorological variability, soil properties, crop management information, or hyperspectral observations. Future research should evaluate the proposed leakage-inclusive model over multiple years, in different geographic regions, and using independent datasets, and explore the integration of UAV imagery, hyperspectral observations, meteorological variables, and soil and agronomic factors. Such extensions will help determine the generalizability of the identified feature relationships and classification performance in more heterogeneous agricultural settings.

**FUNDING INFORMATION**

Authors state no funding involved.

## AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

| Name of Author | C | M | So | Va | Fo | I | R | D | O | E | Vi | Su | P | Fu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aislu Kassekeyeva | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | | ✓ | |
| Moldir Yessenova | | ✓ | | | | ✓ | | ✓ | ✓ | | ✓ | ✓ | | |
| Talgatbek Ayazbaev | ✓ | | ✓ | ✓ | | | ✓ | | | ✓ | ✓ | | ✓ | ✓ |
| Gulzira Abdikerimova | | ✓ | | | | ✓ | | ✓ | | ✓ | ✓ | ✓ | | |
| Gulmira Azieva | | | | ✓ | ✓ | | ✓ | | | ✓ | | ✓ | | ✓ |
| Zhanylsyn Sultanova | | ✓ | | | | ✓ | | | ✓ | ✓ | ✓ | ✓ | | |
| Assyl Bissengaliyeva | | | | | ✓ | | ✓ | | | ✓ | | ✓ | | ✓ |
| Assem Serikbayeva | ✓ | | ✓ | | ✓ | | ✓ | | | ✓ | | ✓ | | ✓ |

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

## REFERENCES


[1] T. Qu *et al*., "A fine crop classification model based on multitemporal Sentinel-2 images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 134, Nov. 2024, doi: 10.1016/j.jag.2024.104172.

[2] Z. M. Mobarakeh, S. Pourmanafi, and M. Ahmadi, "Employing sentinel-2 time-series and noisy data quality control enhance crop classification in arid environments: A comparison of machine learning and deep learning methods," *International Journal of Applied Earth Observation and Geoinformation*, vol. 142, Aug. 2025, doi: 10.1016/j.jag.2025.104678.

[3] T. Mahmood, M. Usman, and C. Conrad, "Selecting relevant features for random forest-based crop type classifications by spatial assessments of backward feature reduction," *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 93, no. 2, pp. 173–196, Apr. 2025, doi: 10.1007/s41064-024-00329-4.

[4] X. Huang, H. Wang, and X. Li, "A multi-scale semantic feature fusion method for remote sensing crop classification," *Computers and Electronics in Agriculture*, vol. 224, Sep. 2024, doi: 10.1016/j.compag.2024.109185.

[5] W. Li *et al*., "Fine-grained hierarchical crop type classification from integrated hyperspectral EnMAP data and multispectral sentinel-2 time series: A large-scale dataset and dual-stream transformer method," *Remote Sensing of Environment*, vol. 344, Oct. 2026, doi: 10.1016/j.rse.2026.115525.

[6] C. Wang *et al*., "Understanding the potentials of early-season crop type mapping by using Landsat-8, Sentinel-1/2, and GF-1/6 data," *Computers and Electronics in Agriculture*, vol. 224, Sep. 2024, doi: 10.1016/j.compag.2024.109239.

[7] R. Tufail, P. Tassinari, and D. Torreggiani, "Assessing feature extraction, selection, and classification combinations for crop mapping using Sentinel-2 time series: A case study in northern Italy," *Remote Sensing Applications: Society and Environment*, vol. 38, Apr. 2025, doi: 10.1016/j.rsase.2025.101525.

[8] X. Liu *et al*., "Comparisons between temporal statistical metrics, time series stacks and phenological features derived from NASA Harmonized Landsat Sentinel-2 data for crop type mapping," *Computers and Electronics in Agriculture*, vol. 211, Aug. 2023, doi: 10.1016/j.compag.2023.108015.

[9] A. Orynbaikyzy, U. Gessner, and C. Conrad, "Spatial transferability of random forest models for crop type classification using Sentinel-1 and Sentinel-2," *Remote Sensing*, vol. 14, no. 6, Mar. 2022, doi: 10.3390/rs14061493.

[10] K. Maolan, Y. Rusuli, Z. XuHui, and Y. Kuluwan, "Sentinel-2 image based smallholder crops classification and accuracy assessment by UAV data," *Geocarto International*, vol. 39, no. 1, Jan. 2024, doi: 10.1080/10106049.2024.2361733.

[11] A. Moumni, M. Oujaoura, J. Ezzahar, and A. Lahrouni, "A new synergistic approach for crop discrimination in a semi-arid region using Sentinel-2 time series and the multiple combination of machine learning classifiers," *Journal of Physics: Conference Series*, vol. 1743, no. 1, Jan. 2021, doi: 10.1088/1742-6596/1743/1/012026.

[12] A. F. Jocea, L. Porumb, L. Necula, and D. Raducanu, "Sentinel-2 land cover classification: State-of-the-art methods and the reality of operational deployment," *Preprint*, Oct. 15, 2025, doi: 10.20944/preprints202510.1174.v1.

[13] M. Narimani, A. Pourreza, and P. Farajpoor, "Mapping tomato cropping systems in california using AlphaEarth geospatial embeddings and deep learning analysis," *2026 ASABE Annual International Meeting*, May 2026, doi: 10.13031/aim.202600545.

[14] G. Abdikerimova *et al*., "A hybrid spatio-temporal model for probabilistic mapping of soil salinity in arid regions," *IEEE Access*, vol. 14, pp. 42465–42484, 2026, doi: 10.1109/ACCESS.2026.3659988.

[15] A. Ismailova, M. Yessenova, G. Murzabekova, J. Tussupov, and G. Abdikerimova, "A system-level approach to pixel-based crop segmentation from ultra-high-resolution UAV imagery," *Applied System Innovation*, vol. 9, no. 1, Dec. 2025, doi: 10.3390/asi9010003.

[16] G. Abdikerimova *et al.*, "Development of a model for soil salinity segmentation based on remote sensing data and climate parameters," *Algorithms*, vol. 18, no. 5, May 2025, doi: 10.3390/a18050285.
[17] S. Sharma, R. Sedona, M. Riedel, G. Cavallaro, and C. Paris, "Sen4Map: Advancing mapping with Sentinel-2 by providing detailed semantic descriptions and customizable land-use and land-cover data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 13893–13907, 2024, doi: 10.1109/JSTARS.2024.3435081.
[18] C. Sanchez, F. Mena, M. Charfuelan, M. Nuske, and A. Dengel, "Assessment of Sentinel-2 spatial and temporal coverage based on the scene classification layer," in *GARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, Jun. 2024, pp. 4099–4103, doi: 10.1109/IGARSS53475.2024.10642213.
[19] X.-Y. Tong and S. Wang, "Invariant features for global crop type classification," *arXiv Preprint*, Apr. 2026, doi: 10.48550/arXiv.2509.03497.
[20] M. Pasternak and K. Pawluszek-Filipiak, "The evaluation of spectral vegetation indexes and redundancy reduction on the accuracy of crop type detection," *Applied Sciences*, vol. 12, no. 10, May 2022, doi: 10.3390/app12105067.
[21] T. Lu, L. Wan, and L. Wang, "Fine crop classification in high resolution remote sensing based on deep learning," *Frontiers in Environmental Science*, vol. 10, Oct. 2022, doi: 10.3389/fenvs.2022.991173.
[22] H. Zhang *et al.*, "A rapid field crop data collection method for complexity cropping patterns using UAV and YOLOv3," *Frontiers of Earth Science*, vol. 18, no. 1, pp. 242–255, Mar. 2024, doi: 10.1007/s11707-024-1109-y.
[23] A. Htitiou, A. Boudhar, Y. Lebrini, R. Hadria, H. Lionboui, and T. Benabdelouahab, "A comparative analysis of different phenological information retrieved from Sentinel-2 time series images to improve crop classification: a machine learning approach," *Geocarto International*, vol. 37, no. 5, pp. 1426–1449, Mar. 2022, doi: 10.1080/10106049.2020.1768593.
[24] A. Rahmati, M. J. V. Zoej, and A. T. Dehkordi, "Early identification of crop types using Sentinel-2 satellite images and an incremental multi-feature ensemble method (Case study: Shahriar, Iran)," *Advances in Space Research*, vol. 70, no. 4, pp. 907–922, Aug. 2022, doi: 10.1016/j.asr.2022.05.038.
[25] R. Chen *et al.*, "Fine-scale classification of horticultural crops using Sentinel-2 time-series images in Linyi country, China," *Computers and Electronics in Agriculture*, vol. 236, Sep. 2025, doi: 10.1016/j.compag.2025.110425.

## BIOGRAPHIES OF AUTHORS



**Aislu Kassekeyeva** received her Ph.D. degree in 2021 in the specialty "Information Systems" named after. L.N. Gumilyov, Kazakhstan. Currently she is a senior lecturer at the Department of Information Systems named after. L.N. Gumilyov. Her research interests include computational linguistics, data analytics, artificial intelligence (AI), and data mining. She can be contacted at email: aibike_7474@mail.ru.



**Moldir Yessenova** received a bachelor's degree in Information Systems in 2014 and a master's degree in 2017 in Information Technology from L.N. Gumilyov Eurasian National University, Kazakhstan, Nursultan. Currently, she is a doctoral student at the Department of Information Systems at the L.N. Gumilyov Eurasian National University. Her research interests include image processing, computer vision, satellite imagery, AI, and ML. She can be contacted at email: moldirrespect@gmail.com.



**Talgatbek Ayazbaev** is, at the moment, a candidate in Physical and Mathematical Sciences and an associate professor at the Taraz Innovation and Humanitarian University, Taraz, Kazakhstan. His research interests include image processing, computer vision, satellite imagery, AI, and ML. He can be contacted at email: ayazbaev.talgat@mail.ru.



**Gulzira Abdikerimova** received her Ph.D. in 2020 in Information Systems from L.N. Gumilyov Eurasian National University, Kazakhstan. Currently, she is an associate professor at the Department of Information Systems at the same university. Her research interests include image processing, computer vision, satellite imagery, AI, and ML. She can be contacted at email: gulzira1981@mail.ru.



**Gulmira Azieva** completed her Ph.D. in Information Systems at L. Gumilyov Eurasian National University in 2021. In her dissertation, she investigated the application of a simulation model of the crude oil and petroleum products supply chain aimed at optimizing logistics processes and improving the efficiency of resource management in this area. Her research interests cover a wide range of issues in the field of information systems, including the development and implementation of models for analyzing and predicting complex systems. She began her education at the Faculty of Physics and Mathematics of the same university, where she completed her bachelor's degree in 2004. During her studies and research at Aktobe University, she focused on the application of mathematical and simulation models to solve practical problems, which became the basis for her subsequent research in the field of information systems. She can be contacted at email: gulmira_azieva@mail.ru.



**Zhanylsyn Sultanova** received her master's degree in 2012 in Information Systems from Turan University, Kazakhstan. Currently, she is a senior lecturer of the Department of Biostatistics, Bioinformatics, and ICT in Astana Medical University. Her research interests include mathematical modeling, ML, AI, and statistics. She can be contacted at email: zhanna.s.d@mail.ru.



**Assyl Bissengaliyeva** is master of Technical Sciences, Zhangir Khan West Kazakhstan Agrarian-Technical University. The main scientific interests of senior lecturer A. Bissengaliyeva - physics, pedagogical sphere. Over the past 5 years, she has published 12 scientific articles. She currently working as a senior lecturer at the Polytechnic Institute. She can be contacted at email: B.a.m69@mail.ru.



**Assem Serikbayeva** is a senior lecturer at the Department of Higher Mathematics at S. Seifullin Kazakh Agrotechnical University (KazATU), Kazakhstan. She obtained her degree in Mathematics from E.A. Buketov Karaganda State University (1986–1991) and later completed her master's degree at Bolashak University, Karaganda (2013–2015). She has extensive academic and teaching experience spanning over three decades. Her professional career includes positions as an engineer and lecturer at Karaganda State University, followed by a senior lecturer role at the Department of Applied Mathematics and Informatics. From 2021 to 2023, she worked as a senior lecturer at the Department of Information Systems at KazATU, and currently continues her academic activities at the Department of Higher Mathematics. Her research interests include informatics, teaching methodology in computer science, algorithmization and programming, inclusive education, and numerical methods. She teaches courses related to information and communication technologies, discrete mathematics applications, and numerical methods. She can be contacted at email: ser_assem@mail.ru.