Received June 4, 2021, accepted June 12, 2021, date of publication June 15, 2021, date of current version June 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3089665

Multi-Label Data Fusion to Support Agricultural Vulnerability Assessments

IVÁN DARÍO LÓPEZ[1], APOLINAR FIGUEROA[2], AND JUAN CARLOS CORRALES[1]
[1]Telematics Engineering Group, University of Cauca at Tulcán, Popayán 190003, Colombia
[2]Environmental Studies Group, University of Cauca at Tulcán, Popayán 190003, Colombia

Corresponding author: Iván Darío López (navis@unicauca.edu.co)

This work was supported in part by the University of Cauca through the Doctoral Program of Telematics Engineering, in part by the Water Security and Sustainable Development Hub through the U.K. Research and Innovation's Global Challenges Research Fund (GCRF) under Grant ES/S008179/1, and in part by the Project "Alternativas Innovadoras de Agricultura Inteligente para sistemas productivos agrícolas del departamento del Cauca soportado en entornos de IoT" under Grant VRI ID4633. The work of Iván Darío López was supported by a scholarship granted by the Colombian Ministry of Science, Technology, and Innovation (Minciencias).

ABSTRACT Identifying crop species and varieties adaptable to climate change impacts is one of the main aspects of climate vulnerability assessments. This estimation involves processing, integrating, and analyzing many information sources to provide accurate and timely responses. However, designing this evaluation, examine the information gathered, and reaching agreements among all stakeholders and experts, often requires considerable effort in time, money, and people. In this study, we propose a data fusion strategy to support climate vulnerability assessments by identifying the adaptability of crops in a territory in the short term. This strategy follows the Joint Directors of Laboratories' data fusion model guidelines. It was evaluated and validated through a case study in Colombia's upper Cauca river basin. For this purpose, we identified Climate, Soil, Water Quality, Productive Alliances, and Production as the most relevant data sources to be integrated, and using metrics such as Mean IR, SCUMBLE, TCS, among others, we evaluated the combined datasets according to their theoretical complexity. The adaptability of crops in a territory was addressed as a multi-label learning problem, assessing the performance of different multi-label classification and multi-view multi-label classification models with both test and actual data. Comparing the predicted crops with the actual ones, we obtained a 98% similarity without considering crop ranking using the Binary Relevance approach and the Random Forest and XGBoost algorithms. Using a more exhaustive test involving order, we obtained a maximum similarity of 67% applying Binary Relevance and Random Forest.

INDEX TERMS Climate vulnerability assessment, climate change, crop production, data processing, data fusion, machine learning, multi-label classification, multi-label dataset, sustainable agriculture.

I. INTRODUCTION

Identify crop species and varieties adaptable to climate change impacts is one of the most economical and environmentally friendly strategies for food security [1]. This concept promotes the interaction between four essential elements: food availability, food access, food utilization, and vulnerability [2]. In the agricultural context, the latter aspect refers to the degree of a system's susceptibility to climate change's adverse effects and measuring it is essential for executing sustainable actions and making decisions to develop food security scenarios [3]. In this sense, different areas and disciplines have involved experts such as scientists, decision-makers, farmers, among others stakeholders, to propose a large number of Climate Vulnerability Assessments (CVA). These stakeholders are responsible for designing a CVA to understand three important questions: who or what is vulnerable to climate variability and change, why and how they are vulnerable, and what opportunities exist to reduce these vulnerabilities. CVAs are designed to meet the specific needs of a strategy (globally or at a country level), project (at regional or sectoral level), or activity (specific organizations or sites) [4].

There exist several methods to conduct a CVA with different levels of complexity. Among the most widely used methods are desk reviews, stakeholder and expert workshops, community-based approaches, and additional specialized analysis (vulnerability indexes, simulations, modeling,

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

impact analysis, map generation, among others) [3]. These methods comprise stages such as conduct literature reviews; identify stakeholders; evaluate the information needs of stakeholders; evaluate the roles and capacities of stakeholders; select data, methods, and tools adjusted to the spatial and temporal scales; and design evaluations on different Climate-Smart Agriculture (CSA) objectives [5].

The aforementioned methods have shown significant contributions such as identifying entry points to address climate stress factors, highlighting opportunities to take advantage of in a changing climate, and determining adaptation measures [4]. However, designing these assessments often requires enormous efforts in time, money, and people. For instance, the interaction between experts in different areas makes it difficult to reach a consensus. Likewise, more specific problems have emerged, such as the variety of data sources [5]. In agriculture, as in many other domains, data inputs correspond to numerous sources such as sensors, structured and unstructured databases, plain text files, multimedia files, and reports. Additionally, many of these data sources have restricted access and are not freely available [6]. Given these points, data fusion represents a non-trivial activity in agricultural vulnerability assessments considering aspects like quantity, diversity, and restrictions of access to data sources.

As can be seen, the use of simple and robust scientific tools to guide stakeholder decision-making on a seasonal and long-term basis, are essential for planning climate-smart strategies, projects, and activities [4], [5]. In this sense, this research work strengthens the additional specialized analyzes carried out in a CVA. It is fundamental to note that this work does not replace an agricultural vulnerability assessment; instead, this is a tool to support and automate data-driven processes that are inherent in such evaluations. In this study, we propose a data fusion strategy to determine, in the short term, the adaptability of crops in a territory, which is determined from the information available on different agricultural vulnerability dimensions. Furthermore, the data fusion strategy was evaluated and validated through a case study in the upper Cauca river basin in Colombia, and we addressed the following contributions: i) a formal multi-dimensional process for preparing the gathered data sources, ii) a method for combining and labeling data sources of different dimensions, and iii) a scheme for training multi-label classification models and determining crop adaptability.

The remainder of this paper continues as follows. Section 2 presents the related works around data fusion and multi-label classification applied to CVAs and crop prediction. Section 3 exposes the components of the proposed data fusion strategy. Section 4 shows the results obtained from the data fusion strategy applied to a case study. Finally, section 5 provides the conclusions of this research.

II. RELATED WORKS

This section exposes the related works around Data fusion (DF) [7] and Multi-Label Classification (MLC) [8]. We also identify interrelations among different approaches and shortcomings regarding this study.

A. DATA FUSION

Agricultural Data Fusion groups different approaches depending on the type of data sources, the crops involved, and the integration aims. Most methods combine different types of satellite imagery (including Landsat-8, Sentinel-2, STRM, MODIS-EVI, and MODIS-NDVI) considering integration objectives such as possible planting areas for rice, soybeans, and corn [9]; estimation of cultivated areas for corn [10]; estimation of yield for corn, soybeans, and cotton [11]; and classification of large crop areas [12]. Satellite images are also combined with an in-situ, survey, or multi-sensor data to identify areas of climate vulnerability [13], determine variations in wheat production [14], detect areas suitable for tomato cultivation [15], and estimate the high spatio-temporal resolution land subsidence [16]. Other approaches integrate multi-sensor and in-situ data for predicting wheat and other crop yields [17], planning and monitoring of oil palm and barley plantations [18], detecting crop diseases for fungicide applications [19], and estimating climate variables from soil and air data [20].

Likewise, some works are focused on integrating historical data around crops such as production, yield, diseases, among others, to solve problems such as detection of genomic regions of pathogens in crops [21], management of viticulture and winemaking processes [22], production and yield estimation of sugarcane crops [23], and identification of crop management areas for application of agricultural inputs [24]. Other approaches integrate multispectral images from Unmanned Aerial Vehicles (UAVs) to determine soybean harvested area and improve crop production monitoring [25]. On the other hand, de Lange *et al.* [26] propose the integration of socio-economic and biophysical data for overcoming spatial incompatibilities. Finally, we highlight two theoretical works closely related to our research. These studies propose to integrate climate, environmental, social, economic, cultural, political, and institutional data for decision making in smart farming contexts using Big Data technologies [27], [28].

B. MULTI-LABEL CLASSIFICATION

Although MLC has addressed problems such as air pollution [29] and flood retention [30], research in agriculture is focused on issues such as the classification of land uses. Conventional classification models assign a single land use label to each spatial unit. Therefore, several approaches classify this coverage unit with several labels simultaneously (mixed use of land). Shendryk *et al.* [31] propose combining deep learning models with MLC to classify atmospheric conditions and land use coverage from satellite images of the Amazon forest. In the same research line, Omrani *et al.* [32] developed an integrated modeling framework (multi-label learning, cellular automata, and land transformation models) to classify land uses in Luxembourg. Using data from this

country, the same authors [33] proposed to solve the same classification problem but using the K-Nearest Neighbors (KNN) technique in the MLC paradigm.

On the other hand, the classification of diseases in crops and leaves is another widely addressed problem. Convolutional Neural Networks (CNN) combined with MLC algorithms are used to detect simultaneous diseases in crops [34]. Likewise, Abd El-Aziz *et al.* [35] detected diseases in apple fruit using the Multi-Label KNN (ML-KNN). Finally, we highlight the approach of Doshi *et al.* [36], which is the most related to our proposal. This approach presents *AgroConsultant*, an intelligent system to assist Indian farmers on which crop to plant depending on the planting season, geographic location, soil characteristics, and environmental factors such as temperature and rainfall. Algorithms such as Decision Trees (DT), KNN, Random Forest (RF), and Neural Networks (NN) were used for the prediction task. NN was selected for obtaining 91% accuracy.

New approaches have emerged to enforce multi-label classification such as multi-view multi-label classification. This approach, in addition to relating an object to multiple class labels simultaneously, assumes its representation across multiple data views [37]. Although this approach has not been applied to agriculture, it is important to mention some of the most relevant works. Some studies propose feature selection methods in Binary Relevance to learn specific features for missing [38], non-missing [39] and class-dependent [40] labels. Other studies exploit the complementarity between different views through multi-view approaches with latent semantic awareness [41] and view-specific information extraction [42]. Finally, Huang *et al.* [43] propose a new framework for multi-view multi-label learning with view-label specific features to identify contributions of different features to each label.

III. DATA FUSION STRATEGY

This section describes our data fusion approach and its main components. This strategy is based on the Joint Directors of Laboratories (JDL) data fusion model [7], one of the most widely used models for DF tasks. Unlike other models, JDL is a functional model rather than a process model, the reason why we select this one. JDL facilitates understanding data fusion techniques and communication between stakeholders to achieve common objectives. This premise implies that the data fusion strategy's actions do not always follow a strict or canonical order to achieve the final goal. In this sense, JDL categorizes the data fusion functions by levels according to different types of problems. Based on these assessments, our data fusion strategy was adapted as follows: Level 0 - Data Assessment, Level 1 - Relationship Analysis, Level 2 - Data Integration, Level 3 - Data Analysis, and Level 4 - Process Refinement. Fig. 1 presents an overview of the data fusion strategy, and its levels are described below.

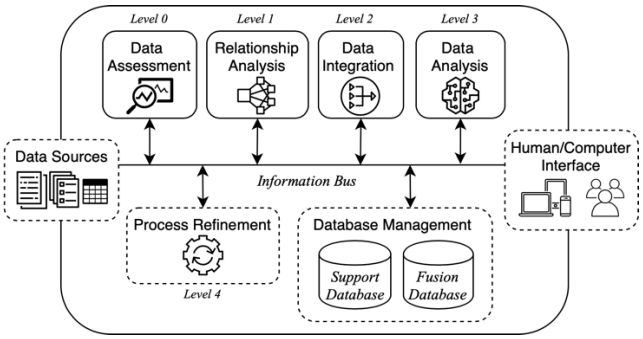


FIGURE 1. General architecture of the data fusion strategy, adapted from JDL data fusion model [7].

A. DATA SOURCES

The data fusion strategy uses open data as the primary input for analysis tasks. Through the data collection process, we identify different official public organizations related to the CVA study area. These organizations usually allow consulting freely accessible data through web portals according to specific information requirements. However, the search for these sources is not only limited to web portals, but it also includes the reuse of CVA results. On the other hand, private entities can provide access to supplementary data by authorizing the corresponding permissions.

B. DATA ASSESSMENT (LEVEL 0)

This level evaluates the gathered data around the agricultural vulnerability dimensions and defines a formal preparation process to improve their quality. The components present a modular scheme, where the results obtained in a module allow feedback to the next one. It is composed of three phases: data sources evaluation, data sources pre-processing, and variables prioritization. The complete data preparation process is detailed in [44].

C. RELATIONSHIP ANALYSIS (LEVEL 1)

In level 1, we identify the implicit relationships between data sources by analyzing the temporal and spatial scales (sampling intervals, producing municipalities, crops, and data coverage area). For this purpose, we establish a spatio-temporal characterization process where all possible relationships are consolidated and analyzed to verify their relevance in the data fusion strategy. Also, we build a relationship scheme to guide the data sources integration at level 2. The components of level 1 are presented in detail below.

1) SPATIO-TEMPORAL CHARACTERIZATION OF DATA SOURCES

In this component, we identify possible spatio-temporal relationships between data sources. Meta-features determine the temporal relationships in the data sources such as time window, temporal scale, and sampling intervals. On the other hand, spatial relationships are focused on the area covered by a territorial division. The spatial scales handle these divisions,

including farms, villages, municipalities, states, regions, and countries. Similarly, we also identify other types of relationships inherent to the data fusion objective. In our case, crops in a specific area could be an indication of these possible additional relationships. Finally, all the Spatio-Temporal Meta-Features (STMF) should be consolidated in a summary table for further analysis.

2) DATA SOURCE RELATIONSHIP SCHEME

The data source relationship scheme corresponds to a matrix that establishes the strength of the possible relationships. This matrix is generated by comparing the STMF consolidated in the previous component. The data source relationship scheme is described below and presented in Fig. 2.

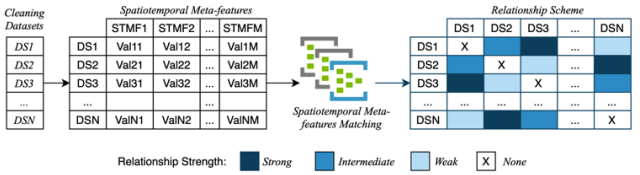


FIGURE 2. Relationship scheme. DS: Data Source, STMF: Spatial-Temporal Meta-Feature, Val: Attribute Value.

(i) *Creating the Relationship Scheme*. The consolidated table of STMF is shown on the left side of Fig. 2. The rows represent the pre-processed data sources of size N. The columns correspond to the STMF of size M. We transform the table into a matrix of N x N dimensions, where the color of each cell represents the relationship strength between the data sources (right side of Fig. 2). To obtain the strength of a relationship, we compare the STMF of one data source with the others using a similarity function. This function returns a value between 0 and 1. The values near 1 represent a high similarity, and the values near 0 a low similarity. We define the following ranges and colors to assign the strengths of the relationships: 0 to 0.25 (light blue, weak relationship), 0.25 to 0.75 (blue, intermediate relationship), and 0.75 to 1 (dark blue, strong relationship).

(ii) *Relationship Scheme Analysis.* After obtaining the relationship scheme, we can integrate a table with guidelines about the possible relationships identified and how the data sources are involved. These guidelines indicate which STMFs we should consider at level 2 (data integration), for example, the resulting time windows, the regularity of sampling intervals, and specific attributes that would facilitate integration.

D. DATA INTEGRATION (LEVEL 2)

Level 2 corresponds to data integration, a reduction process to generate new data sets (combined data sources) with a more synthesized and reliable added value. At this level, we select a data integration approach and apply a method to combine the data sources based on Entity Matching [45]. Finally, we label the resulting datasets according to the final objective of the data fusion strategy. In our case, identify the crops produced per municipality to establish the set of labels for each instance. Data integration steps are described in detail below.

1) SELECTING THE INTEGRATION STATE

Depending on the nature of the datasets and the statistical problem to be solved, data from different sources can be integrated into three different states: early, intermediate, or late [46]. These states of data integration are described below.

(i) *Early Integration.* In this state, a single feature space groups the data sources' attributes without changing their format and nature. However, a disadvantage lies in the increased dimensionality of the combined datasets.

(ii) *Intermediate Integration.* Before being combined, intermediate integration transforms the attributes of all data sources into a common feature space. Then, a model learns a joint representation of several data sets and merges them during a later stage of analysis.

(iii) *Late Integration.* Each dataset trains one or more models separately, and an assembly method combines the final results. This state has advantages such as the free choice of the best algorithm and the parallel analysis of each dataset.

2) INTEGRATING DATA SOURCES

The basic idea is to ensure that the matching attributes have the same structure and their content follows the same formats. Next, we apply indexing or blocking to reduce the computational effort when comparing record pairs, i.e., using highly tolerant similarity measures to filter out those record pairs that are "obvious" non-matches. After that, only those not pruned by the indexing or blocking step are inspected in record pair comparison. Finally, we apply a function to combine the records that have been matched.

3) LABELING COMBINED DATA SOURCES

Each record in the combined datasets is finally labeled with one (single label) or more (multi-label) target variables or classes. This process depends on the specific problem addressed in the data fusion strategy and must be guided by expert knowledge. In this sense, we can apply labeling techniques such as manual or automatic labeling. Manual labeling requires the supervision of one or more experts in the field, who assign the labels to the respective samples. This process can also be supported by reviewing the literature in the knowledge areas around the data sources. On the other hand, in automatic labeling, different clustering algorithms can be applied to find groups representing common labels or classes.

4) EXPLORATORY ANALYSIS IN MULTI-LABEL DATASETS

After combining and labeling, data sources require exploratory analysis to determine the effectiveness of data

integration before to the last level of the data fusion strategy (level 3 - data analysis). In the case of Multi-Label Datasets (MLD), the exploratory analysis is based on [47] and [8], which define a set of measures to describe the combined data sources. These measures provide information about the data distribution and the possible behavior of a classification algorithm or a pre-processing technique. We used Discarded Attributes, Number of Attributes, Number of Instances, Number of Inputs, Number of Labels, Number of Labelsets, Number of Single Labelsets, Maximum Frequency, Cardinality, Density, Imbalance Ratio of a Label (IRLbl), Mean Imbalance Ratio, Score of ConcUrrence among iMBalanced LabEls (SCUMBLE), SCUMBLE Variation Coefficient, and Theoretical Complexity Score (TCS).

E. DATA ANALYSIS (LEVEL 3)

In data analysis or Level 3, we apply several techniques or machine learning algorithms to train a set of models and estimate one or more target variables (predicted crops in a municipality). In this sense, we use different metrics to evaluate the trained models' performance and select the best ones. We apply an Analysis of Variance (ANOVA) [48], which identifies statistically significant differences among model performances. Finally, we validate the models' results with actual data, i.e., data from a real scenario (crop production and yield trends in subsequent years).

1) MODEL TRAINING SCHEME

In this component, we generate several predictive models and train them from both the Combined Data Sources (CDS) at level 2 and a set of variations of those sources. These variations correspond to modified versions of a specific combined DS, and these are mentioned below.

(i) *Original CDS.* An initial version of the combined data source without modifications.

(ii) *Decoupled CDS.* A majority label frequently appears in instances, while a minority label appears rarely. When the majority and minority labels coincide in the same instance, the minority labels are more difficult to classify due to the majority's bias. To separate the labels, we apply the REMEDIAL (REsampling MultilabEl datasets by Decoupling highly ImbAlanced Labels) [49] algorithm to the CDS obtaining a new version of this source (Decoupled CDS). The number of instances increases according to the proportion of instances containing both majority and minority labels through this technique. REMEDIAL is recommended for datasets with a high SCUMBLE value.

(iii) *Infrequent Positive Label Removal (IPLR).* Combined data sources excluding infrequent positive labels (labels with value 1), we must define the number of excluded labels according to the frequency distribution.

(iv) *Skewness Labels Removal (SLR).* Combined data source excluding skewness labels, i.e., majority and minority positive labels at a defined threshold according to the frequency distribution.

After obtaining the CDS variations, we apply a combination of a multi-label classification strategy plus a machine learning algorithm to each CDS. Through this combination, we generate a set of trained models to estimate the values of one (single label) or several target variables (multiple labels). MLC strategies transform a dataset to apply a base algorithm. These strategies include Binary Relevance (BR), Binary Relevance Plus (BRPLUS), Ensemble of Classifier Chains (ECC), Label Powerset (LP), Hierarchy Of Multi-label classifiER (HOMER), and Random k-labelsets (RAKEL). We also used two new multi-view multi-label classification strategies such as incomplete Multi-View Weak-label Learning (iMVWL) [37], Multi-View Weak-label Learning (McWL) [50]. On the other hand, Random Forest (RF), Support Vector Machines (SVM), K Nearest Neighbor (KNN), Sequential Minimal Optimization (SMO), C5.0 Decision Trees (C5.0), Naive Bayes (NB), eXtreme Gradient Boosting (XGB), Classification and Regression Trees (CART), Majority Class Prediction (MAJORITY), and Random Prediction (RANDOM) represent the base algorithms. Fig. 3 presents the generation and training of S models from the combined data sources (1 to N CDS), the CDS variations (1 to M variations), the MLC strategies (1 to P strategies), and the basic machine learning algorithms (1 to Q algorithms), where the number of models is represented by $S = N^*M^*P^*Q$.

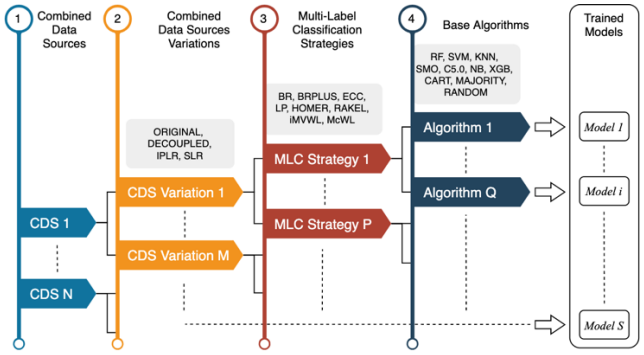


FIGURE 3. Model generation and training scheme for combined data source variations applying different multi-label classification strategies and base algorithms.

2) MODEL PERFORMANCE EVALUATION

To evaluate model performance with test data, we use different metrics for both traditional supervised learning and multi-label learning [51]. Metrics such as Accuracy, Recall, Precision, F1-score, among others, are included in supervised learning. While multi-label learning involves metrics such as Hamming-Loss, Ranking-Loss, One-Error, among others. On the other hand, the above metrics could eventually generate very approximate results between models. To address this issue, we apply an analysis of variance (ANOVA) [48], which allows us to determine possible significant differences in model evaluation results. We considered several aspects such as metrics distribution (normality test), homogeneity of variance across groups (homoscedasticity), paired tests, and post-hoc comparisons to apply the ANOVA model.

3) MODEL VALIDATION

To determine if the data fusion strategy is applicable in a real agricultural environment, we validate the models' best estimates with actual data. These data can be obtained in the first stage of the fusion strategy (data collection) or later, depending on the mentioned strategy's objectives. Finally, different techniques for comparing results must be defined in this validation component, again emphasizing the fusion strategy's objective.

F. PROCESS REFINEMENT (LEVEL 4)

Process refinement is a transversal level to monitor each of the components (or levels) in the data fusion strategy. Level 4 represents the planning, control, and allocation of resources to tasks. This refinement also includes experts in different knowledge domains around agricultural vulnerability, which provide constant feedback to validate the data fusion strategy. In this sense, we can also use CVAs' results in this validation process as a benchmark.

G. DATABASE MANAGEMENT

This component manages two databases, the support data-base and the fusion database. The first one stores the raw and processed data sources (data on different dimensions of agricultural vulnerability). The second one stores the combined and labeled data sources. The storage also supplies additional information about the datasets, dimensional characteristics, characterization of the agricultural area, and model calibration. Similarly, this component manages the data ingestion, i.e., the process of flowing data from its origin to one or more data stores such as a data lake, relational databases, search engines, among others. This study is oriented to the development of an innovative data fusion strategy using new technologies. Therefore, we managed the databases by implementing a Data Lake [52], which is described in detail in [53].

H. HUMAN/COMPUTER INTERFACE

This component presents the information obtained from the analysis of the different datasets. This software tool allows the user to consult the results according to the objective, which has been initially defined in the data fusion strategy. The information must be presented considering each user type as a farmer, technician, extensionist, researcher, or decision-maker. We must orient the tool towards how the user can take advantage of all the information visualized in the respective knowledge domain.

IV. RESULTS

In this section, we present the main findings from the instantiation of the proposed data fusion strategy. Through a case study, we show the results step by step at each level and component. Initially, we describe the study area, its characteristics, and its previous Climate Vulnerability Assessment (CVA) to contextualize the subsequent data analyses.

A. STUDY AREA

The Cauca River is the primary water source in the western region of Colombia. This river extends from the *Macizo Colombiano* area (approximately 3,200 meters above sea level - m.a.s.l.) to the Magdalena River in the north of the country, covering about 1,360 km through nine regions from south to north. The Upper Cauca River Basin (UCRB) has approximately 23,000 km^2, of which 32% corresponds to Cauca, 47% to Valle del Cauca, 13% to Risaralda, and 8% to Quindío. In this research, we focused specifically on the Cauca zone, where the altitude varies between 4,700 m.a.s.l. at the summit of the Puracé volcano, and 950 m.a.s.l. in the Cauca River's alluvial valley (approximate area of 7,368 km^2).

In this sub-basin, agriculture represents a significant percentage of the Colombian economy and benefits about 23 municipalities. Even the country's food security, in a certain percentage, is directly affected by agricultural production in this area. The food demand of urban centers is largely supplied by small-scale commercial farms of coffee, beans, corn, cassava, fruit trees, vegetables, medicinal plants, livestock, and fish farming. In social and economic terms, the rural economy is significant and a primary source of food security both in the region and in the country [54].

B. DATA SOURCES

We considered several data sources as the fundamental input of this study. In this case, we extracted 16 datasets from different web portals of official public organizations (details about these data sources can be consulted in [44]). We also used the results of previous climate vulnerability assessments developed at the UCRB as inputs.

C. DATA ASSESSMENT (LEVEL 0)

Level 0 developed a diagnosis of the initial data sources and tries to improve their quality through pre-processing. Initially, we grouped the data sources into the four dimensions defined in AVA (biophysical, economic-productive, socio-cultural, and political-institutional). Subsequently, their main meta-features were extracted and analyzed to build an overview of all datasets. Then, we identified the leading data quality problems through a statistical analysis of data distributions. Finally, we identify the most relevant attributes of each data source. The complete and detailed data sources evaluation (level 0) is referenced in [44].

D. RELATIONSHIP ANALYSIS (LEVEL 1)

At this level, we identify spatio-temporal relationships among data sources. First, we selected the meta-features at the temporal and spatial level, such as the time window, the temporal scale, the sampling interval, and the spatial scale. We also extract additional meta-features for guiding the data integration process, in this case, information about crops, such as the spatial units of sowing, production, or commercialization.

TABLE 1. Spatio-temporal Meta-features of data sources. A: Annual, BA: Biannual, M: Monthly, Mu: Municipality, De: Department. Group and subgroup represent hierarchical crop categories such as fruit trees, vegetables, among others.

Dimension	Data Source	Temporal Meta-features			Spatial Meta-features	Other Meta-features	Units
		Time Window	Temporal Scale	Sampling Intervals	Spatial Scale		
Biophysical	bp-sivicap	2015	A	Regular	De, Mu	-	-
	bp-corpoica	2013 - 2016	A	Irregular	De, Mu	Crop	Point
	bp-ideam	2012 - 2019	A, M	Regular	De, Mu	-	-
	bp-ava	2007 - 2011	-	-	De, Mu	Crop	Point
Economic-Productive	ep-finagro	2004 - 2019	A	Regular	De	-	-
	ep-dane-sipsa-p	2004 - 2019	A	Regular	De	Crop	Area (t/Ha)
	ep-agronet	2007 - 2016	A	Regular	De	Crop	Area (t/Ha)
	ep-minagricultura	2007 - 2015	A, BA	Regular	De	Group, Subgroup, Crop	Area (t/Ha)
	ep-agronet-p	2007 - 2016	A, M, D	Irregular	St	Crop	Point
	ep-dane-sipsa	2013 - 2017	A, M	Regular	De, Mu, Ma	Group, Crop	Point
Political-Institutional	pi-dnp-aib	2005 - 2013	A	Regular	De, Co	-	-
	pi-dnp-fi	1995 - 2014	A	Irregular	De	-	-
	pi-dnp-la	2010 - 2014	A	Irregular	De	-	-
	pi-dnp-pa	2002 - 2013	A	Regular	De, Mu	Crop	Area (Ha/Alliance)
Sociocultural	sc-dane-hh	2014	-	-	De, Mu	-	-
	sc-dane-h	2014	-	-	De, Mu	-	-

Table 1 summarizes the above meta-features for each of the 16 data sources.

From Table 1, we compare the meta-features of one data source with the other 15 to obtain the strength of each relationship. We use a function to compare, one by one, the temporal and spatial attributes among two data sources. The similarity intervals used were: 0 - 0.25 (weak relationship), 0.25 - 0.75 (intermediate relationship), 0.75 - 1 (strong relationship). Subsequently, we averaged the strength of all spatio-temporal attributes for each compared tuple of datasets. For example, if we compare the attribute "time scale" in the datasets bp_sivicap and bp_corpoica, the strength of this relationship will be 1 (100%) considering the same scale (annual) in both datasets. While the strength will be 0.5 (50%) if the compared datasets are bp_sivicap (annual) and bp_ideam (annual, monthly). Finally, we generated the Data Source Relationship Matrix shown in Fig. 4. According to these relationships' strengths, we established the following best combinations among the datasets: *bp-ideam + (bp-sivicap or bp-corpoica); ep-agronet + (bp-sivicap, bp-corpoica, or bp-ideam); ep-minagricultura + ep-agronet; ep-dane-sipsa + (ep-agronet or ep-minagricultura); pi-dnp-pa + (bp-corpoica or ep-agronet); sc-dane-hh + bp-ideam; sc-dane-h + (bp-ideam or sc-dane-hh).*

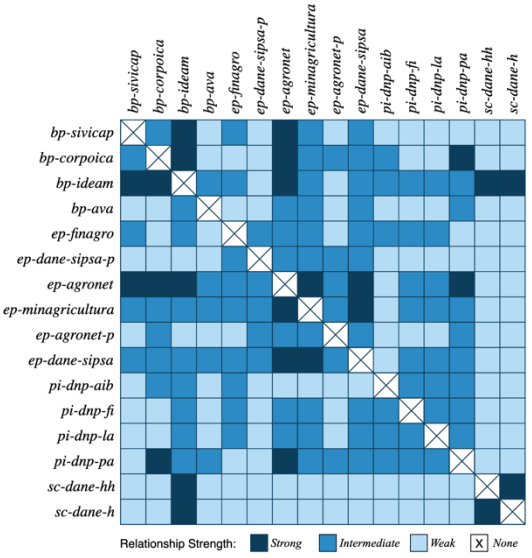


FIGURE 4. Data source relationship matrix.

E. DATA INTEGRATION (LEVEL 2)

In this level, we generate new data sets with a more synthesized and reliable added value through a reduction process. To develop the integration, we first identify the most related data to the data fusion strategy objective. In our case, considering attributes on production and crop yield per municipality, we identified the *ep-agronet* as the central data source. We applied entity matching to associate the selected data sources based on *ep-agronet* and the relationship matrix information (Fig. 4). We used the *Jaro-Winkler* similarity function [55] to identify the matches between municipality names in each tuple of data sources. We selected this similarity algorithm by considering two key aspects such as good performance comparing short strings and the name-comparison oriented design. Once Jaro-Winkler was applied, we obtained 4 combined data sources, *CDS1 (bp-ideam + ep-agronet), CDS2 (bp-corpoica + ep-agronet), CDS3 (bp-sivicap + ep-agronet), and CDS4 (pi-dnp-pa + ep-agronet)*. We discarded other combinations considering the low number of resulting instances (between 4 and 15). Furthermore, it was possible to combine five data sources (*bp-ideam + bp-corpoica + bp-sivicap + pi-dnp-pa + ep-agronet*) to obtain a Combined Global Dataset (CGD), which represents the combination of the maximum possible number of data sources. It is worth mentioning that a conventional multi-label approach can be transformed into a multi-view multi-label approach using two strategies such as independent and integrated views [43]. In our case, we consider the four combined datasets as independent views and the CGD dataset as their integration.

On the other hand, we labeled each of the five combined datasets according to the data fusion strategy's objective, i.e., to predict the crops that can best adapt in an area in the short term. This objective induces a multi-label classification problem, where a label represents a crop. Based on ep-agronet dataset yield data, we identify the crops in a municipality by assigning binary labels to each instance in the combined datasets. Therefore, if a municipality produces a crop in a specific year, the value one is assigned, and 0 otherwise. To determine the datasets integration quality, we applied a preliminary exploratory analysis. In Table 2, we considered the metrics previously mentioned in section 3.4.4 to identify the most appropriate datasets to apply the MLC algorithms used in level 3.

TABLE 2. Spatio-temporal meta-features.

Metric	CDS1	CDS2	CDS3	CDS4	CGD
Discarded attributes	3	3	3	4	3
Number of attributes	72	87	80	62	125
Number of instances	69	80	39	55	73
Number of inputs	6	25	14	7	47
Number of labels	66	62	66	55	78
Number of labelsets	48	44	39	43	68
Number of single labelsets	31	27	39	36	63
Maximum Frequency	3	8	1	5	2
Cardinality	12.92	10.21	12.12	9.9	11.21
Density	0.19	0.16	0.18	0.18	0.14
Mean IR	10.1	18.21	11.41	8.78	19.56
SCUMBLE	0.27	0.23	0.28	0.21	0.29
SCUMBLE.CV	0.39	0.68	0.49	0.59	0.58
TCS	9.85	11.13	10.49	9.71	12.42

The results presented in Table 2 indicate acceptable quality in all five combined data sources for the next level of the data fusion strategy (Level 3). This finding is based on three key metrics, such as TCS, SCUMBLE, and Mean IR. The first corresponds to a low value in Theoretical Complexity Score (TCS) compared to more complex MLDs used in different studies [56]. The TCS values for the combined data sources were between 9.71 and 12.42, indicating less complexity in learning a predictive model. The second refers to the global level of unbalanced labels (Mean IR), with values between 8.78 and 19.56, which indicates an acceptable average level of imbalance in all combined data sources compared to other well-known MLDs [57]. Finally, the third metric indicates the Score of ConcUrrence among iMBalanced LabEls (SCUMBLE) with values between 0.21 and 0.29, which indicates a low concurrence among minority and majority labels, considering that this measure is in the range [0,1] [58].

F. DATA ANALYSIS (LEVEL 3)

In Level 3, we explore different MLC strategies and machine learning algorithms to generate models that predict one or more target variables. We trained these models according to the data fusion strategy's objective, from the combined data sources in Level 2 for predicting the crops produced at a specific site in the short term. Following the model generation scheme, previously defined in Fig. 3, we trained 2,430 models ($S = N^*M^*P^*Q$, $S = 5^*9^*6^*9$) and evaluated them using the metrics presented in Table 3.

Tab. 3 summarizes the best models for each MLC Strategy (MLCS) applied to the combined data sources. We did not consider the results obtained from the combined data sources' variations since they did not exceed the original combined data sources' results. In the case of the accuracy, precision, recall, and F1-score metrics, we have highlighted the highest values in each combined data source. On the other hand, for hamming-loss, ranking-loss, and one-error metrics, we have highlighted the lowest values considering that these are loss functions, i.e., the best results correspond to the lowest values. These results showed two crucial findings in the performance of predictive models. From the point of view of conventional performance measurements (Accuracy, Precision, Recall, and F1-measure), the Label Powerset (LP) strategy obtained the best results when combined with the C5.0 and Naïve Bayes algorithms in the CDS1, CDS4, and CGD datasets, while the RAKEL strategy performed well with a broader set of algorithms such as RF, C5.0, NB, and SVM. On the other hand, from the perspective of MLC-oriented metrics (Hamming-Loss, Ranking-Loss, and One-Error), we only observed a similar behavior of Hamming-Loss concerning conventional metrics for the same combined datasets. However, for specialized label ranking evaluation metrics such as Ranking-Loss and One-Error, the best performances were obtained using the BR and BRPLUS strategies pre-dominantly in conjunction with the Random Forest algorithm. Regarding the multi-view multi-label classification strategies (iMVWL and McWL), these showed acceptable performances in all measures, however, these were below the best values.

Although previous results showed better performance in some MLC strategies, it is impossible to establish a significant difference among methods at first sight. Considering the above, we performed a test of statistical significance using an Analysis of Variance (ANOVA) [48]. We identified the total variance from the variance among sample groups (in this case, the groups correspond to each applied MLC strategy). We evaluated the Hamming-Loss metric for the classification task and the Ranking-Loss metric for the ranking task. We selected Hamming-Loss considering relevant aspects such as its behavior similar to conventional metrics and also because it is a metric oriented to MLC approaches. Also, Ranking-Loss was selected because it is the main MLC-oriented metric for evaluating label rankings. Furthermore, we also applied several a priori and a posteriori tests of the ANOVA to determine those groups with significant differences. The first of these analyses was the normality test, where we checked an adequate metrics distribution in each group. We used the *Shapiro-Wilk* normality test [59] because it is the most used, efficient, and useful when the samples are small. This test showed high levels of normality in the group distributions for both Hamming-Loss and Ranking-Loss metrics. After checking the normality, we verified the homogeneity of variances by applying a homoscedasticity test. One of

TABLE 3. Performance metrics for the best predictive models in the MLC approach. CDS: Combined Data Source, MLCS: Multi-Label Classification Strategy, Alg: Machine Learning Algorithm.

CDS	MLCS	Accuracy		Precision		Recall		F1-Score		Hamming-Loss		Ranking-Loss		One-Error	
		Value	Alg.	Value	Alg.	Value	Alg.	Value	Alg.	Value	Alg.	Value	Alg.	Value	Alg.
CDS1	BR	0.6	RF	0.83	RF	0.73	NB	0.74	RF	0.09	RF	0.04	RF	0.04	SVM
	BRPLUS	0.56	RF	0.82	RF	0.67	NB	0.69	RF	0.11	RF	0.05	RF	0.07	C5.0
	ECC	0.51	RF	0.69	C5.0	0.7	RF	0.66	RF	0.13	RF	0.1	RF	0.05	RF
	HOMER	0.12	RF	0.49	RF	0.13	RF	0.2	RF	0.19	RF	0.49	SMO	0.44	RF
	LP	0.77	C5.0	0.88	C5.0	0.84	C5.0	0.85	C5.0	0.05	C5.0	0.09	C5.0	0.11	C5.0
	RAKEL	0.67	C5.0	0.82	RF	0.79	C5.0	0.79	C5.0	0.08	C5.0	0.07	C5.0	0.04	XGB
	iMVWL	0.52	LR	0.73	LR	0.65	LR	0.65	LR	0.11	LR	0.08	LR	0.07	LR
	McWL	0.52	KNN	0.75	KNN	0.62	KNN	0.67	KNN	0.11	KNN	0.08	KNN	0.08	KNN
CDS2	BR	0.63	SMO	0.84	RF	0.78	C5.0	0.74	SMO	0.08	RF	0.05	RF	0.03	RF
	BRPLUS	0.66	SMO	0.87	RF	0.76	C5.0	0.75	SMO	0.09	RF	0.05	RF	0.01	RF
	ECC	0.61	SMO	0.73	SMO	0.78	SMO	0.72	SMO	0.1	RF	0.07	C5.0	0.04	C5.0
	HOMER	0.17	RF	0.47	RF	0.19	RF	0.25	RF	0.17	RF	0.46	SMO	0.49	RF
	LP	0.69	SMO	0.79	C5.0	0.78	SMO	0.77	SMO	0.09	C5.0	0.14	SMO	0.23	C5.0
	RAKEL	0.68	C5.0	0.83	RF	0.81	C5.0	0.78	C5.0	0.08	RF	0.07	C5.0	0.09	C5.0
	iMVWL	0.62	LR	0.75	LR	0.77	LR	0.73	LR	0.09	LR	0.8	LR	0.6	LR
	McWL	0.61	KNN	0.73	KNN	0.75	KNN	0.73	KNN	0.1	KNN	0.7	KNN	0.6	KNN
CDS3	BR	0.38	RF	0.75	SVM	0.58	NB	0.54	RF	0.14	RF	0.17	RF	0.13	CART
	BRPLUS	0.36	RF	0.75	RF	0.56	NB	0.52	RF	0.14	RF	0.17	RF	0.13	C5.0
	ECC	0.38	XGB	0.57	CART	0.58	XGB	0.54	XGB	0.17	RF	0.19	C5.0	0.15	SVM
	HOMER	0.06	SMO	0.24	SMO	0.07	SMO	0.1	SMO	0.22	RF	0.56	SMO	0.75	C5.0
	LP	0.32	C5.0	0.75	XGB	0.5	RF	0.47	C5.0	0.16	CART	0.34	RF	0.2	CART
	RAKEL	0.39	RF	0.77	SVM	0.59	NB	0.55	RF	0.14	RF	0.23	C5.0	0.18	RF
	iMVWL	0.34	LR	0.73	LR	0.57	LR	0.52	LR	0.16	LR	0.25	LR	0.17	LR
	McWL	0.35	KNN	0.7	KNN	0.57	KNN	0.53	KNN	0.17	KNN	0.28	KNN	0.16	KNN
CDS4	BR	0.38	RF	0.66	SVM	0.51	RF	0.51	RF	0.14	RF	0.11	RF	0.18	RF
	BRPLUS	0.39	RF	0.67	RF	0.49	XGB	0.51	RF	0.14	RF	0.11	RF	0.16	RF
	ECC	0.36	CART	0.56	C5.0	0.54	XGB	0.49	CART	0.15	C5.0	0.13	NB	0.22	SVM
	HOMER	0.07	RF	0.26	C5.0	0.09	RF	0.11	RF	0.21	C5.0	0.45	SMO	0.67	C5.0
	LP	0.44	NB	0.57	NB	0.56	NB	0.54	NB	0.13	NB	0.21	NB	0.43	NB
	RAKEL	0.38	XGB	0.69	SVM	0.48	NB	0.5	XGB	0.13	RF	0.17	XGB	0.25	XGB
	iMVWL	0.38	LR	0.62	LR	0.51	LR	0.52	LR	0.15	LR	0.14	LR	0.19	LR
	McWL	0.37	KNN	0.63	KNN	0.5	KNN	0.51	KNN	0.15	KNN	0.14	KNN	0.2	KNN
CGD	BR	0.61	RF	0.83	RF	0.74	NB	0.74	RF	0.09	RF	0.04	RF	0.04	SVM
	BRPLUS	0.56	RF	0.83	RF	0.68	NB	0.7	RF	0.11	RF	0.05	RF	0.07	C5.0
	ECC	0.52	RF	0.7	C5.0	0.7	RF	0.67	RF	0.13	RF	0.1	RF	0.05	RF
	HOMER	0.12	RF	0.49	RF	0.14	RF	0.2	RF	0.19	RF	0.49	SMO	0.44	RF
	LP	0.78	C5.0	0.89	C5.0	0.85	C5.0	0.86	C5.0	0.05	C5.0	0.09	C5.0	0.11	C5.0
	RAKEL	0.68	C5.0	0.83	RF	0.79	C5.0	0.79	C5.0	0.08	C5.0	0.07	C5.0	0.04	XGB
	iMVWL	0.62	LR	0.84	LR	0.73	LR	0.75	LR	0.08	LR	0.06	LR	0.08	LR
	McWL	0.61	KNN	0.83	KNN	0.71	KNN	0.77	KNN	0.09	KNN	0.07	KNN	0.07	KNN

TABLE 4. Results of normality (percentage of groups with a normal distribution), homoscedasticity (p-value), and ANOVA (p-value) analyses.

CDS	Normality Test		Homoscedasticity Test		ANOVA	
	Hamming-Loss	Ranking-Loss	Hamming-Loss	Ranking-Loss	Hamming-Loss	Ranking-Loss
CDS1	94.4%	92.3%	$6.5 * 10^{-15}$	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$
CDS2	98.1%	90.1%	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$
CDS3	100%	99.6%	0.17	0.71	$1.5 * 10^{-11}$	$2.2 * 10^{-16}$
CDS4	100%	95.3%	0.9	0.05	$9.9 * 10^{-12}$	$2.2 * 10^{-16}$
CDS5	96.2%	95.3%	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$	$2.2 * 10^{-16}$

the most used is *Levene's* test [60]. We identified variations in the sample groups through this test, considering a p-value lower than 0.05 in the CDS1, CD2, and CGD datasets. After these two analyses, we applied linear ANOVA models for each combined data source. Table 4 summarizes the results of normality, homoscedasticity, and ANOVA analyses.

Although the homoscedasticity test showed variations in the CDS1, CD2, and CGD datasets, Tab. 4 indicates variations in all datasets, considering that the ANOVA p-values were less than 0.05. To accurately determine such variations, we analyzed each group or model independently in each combined data source. For this purpose, we used the ANOVA results and two additional analyses, such as the Pairwise t-test [61] and the Tukey's test [62]. From these tests, we observed those specific groups (MLC strategies) where variations occurred (p-value < 0.05 for both Hamming-Loss and Ranking-Loss metrics), and we identified two predominant patterns. The former was the combination of the MAJORITY and RANDOM algorithms with all the tested MLC strategies. The latter was the combination of the HOMER strategy with all the applied algorithms. The significant differences correspond mainly to the worst-performing MLC strategies but not to the best models. Therefore, there were no significant differences among the

best performing MLC strategies. At this point, we could not establish which models were the most suitable for crop prediction. For this reason, we validated the crop rankings predicted by all models except those with significant differences previously mentioned.

We validated the MLC models with actual data from the Colombian Agriculture Ministry through the Agronet web platform. Predictive models were trained with data from 2012 to 2015 and validated with data from 2016 to 2018. These data were obtained from the production trends and yield of different crops by the site in Colombia. To transform the validation data into actual rankings comparable to the predicted ones, we converted crop yield data into annual growth rates from 2016 to 2018. We then averaged the growth rates by crop and ranked these averages from highest to lowest in each municipality. To compare rankings, we initially focused on conventional correlation indices such as Pearson, Spearman, and Kendall. However, these indexes had some disadvantages for our work. For example, they only compare rankings of equal length, and they also work with a homogeneous weight distribution regardless of the position of the items.

Considering the above, we used two similarity measures for comparing indefinite ranked and unranked lists. For unranked lists, we used a simple similarity measure that indicates the percentage of items from the actual ranking contained in the predicted one. We refer to this measure as *Unranked Lists' Similarity (ULS)*. Furthermore, for ranked lists, we used the *Rank Biased Overlap (RBO)* [63], which is a similarity measure that assumes that the top rank is more important than the bottom rank. In other words, exchanges or perturbations in the top rank are more significant and more strongly penalized than those in the bottom rank. The RBO range varies from 0 to 1, where 1 corresponds to identical rankings, and 0 represents disjointed rankings. We can also adjust weight ranking positions through the p parameter (between 0 and 1). For RBO, a low p represents a high weighting in the top-ranking items (top-weighted). On the other hand, when p is equal to 0, only the top-k items are considered (k is the evaluation depth parameter). Finally, when p is close to 1, weights are arbitrarily flat, and the evaluation is arbitrarily deeper in the rankings. To determine the similarity between the predicted and actual crop rankings, we extracted the ULS and RBO values for each model applied in each municipality, and then averaged them to obtain an overall value. To display these values for a particular dataset, Fig. 5 shows the global ULSs and RBOs of each MLC model for the CDS1 dataset.

In the same line, Table 5 summarizes the best MLC models validated with actual crop rankings. We obtained ULS values above 90% for most of the combined data sources. These results indicate a good performance of the predictive models without considering the crop ranking. However, we prioritized the RBO measure for being more exhaustive in evaluating the position of each element within the ranking. The maximum average RBO value was 0.67 for the CDS1 dataset with the BR-RF model. On the other hand, CDS3 obtained a maximum average of 0.56 using the BR-CART model. At first sight, we could consider these values at a low level of similarity concerning to the actual rankings; however,

TABLE 5. Five highest overall RBO (Rank Biased Overlap) and its respective ULS (Unranked Lists' Similarity) values for each MLC model across all combined data sources.

CDS	MLC Model	RBO	ULS
CDS1	BR-RF	0.67	0.87
	BRPLUS-RF	0.64	0.88
	BR-SVM	0.64	0.9
	BR-C5.0	0.64	0.9
	BR-XGB	0.63	0.9
CDS2	BR-RF	0.61	0.95
	BR-XGB	0.61	0.96
	BRPLUS-XGB	0.61	0.96
	BRPLUS-RF	0.61	0.96
	BRPLUS-C5.0	0.6	0.96
CDS3	BR-CART	0.56	0.94
	BR-RF	0.56	0.96
	BRPLUS-CART	0.56	0.93
	BR-XGB	0.55	0.96
	BRPLUS-C5.0	0.54	0.96
CDS4	BR-RF	0.61	0.94
	BR-XGB	0.59	0.94
	BRPLUS-RF	0.58	0.94
	BR-NB	0.58	0.91
	BRPLUS-XGB	0.58	0.94
CGD	BRPLUS-RF	0.65	0.97
	BR-RF	0.64	0.98
	BR-XGB	0.61	0.98
	BRPLUS-XGB	0.61	0.98
	ECC-RF	0.57	0.81

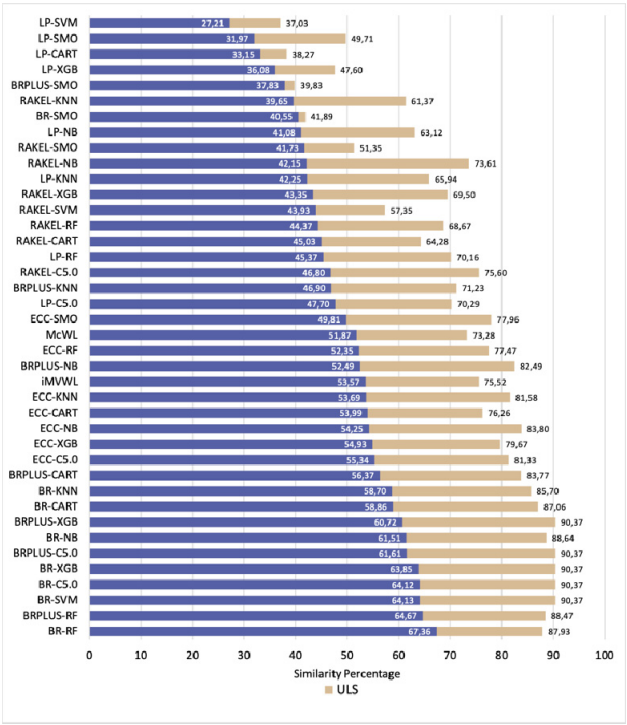
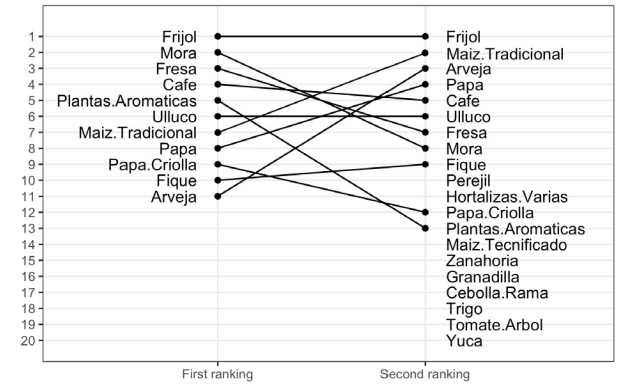
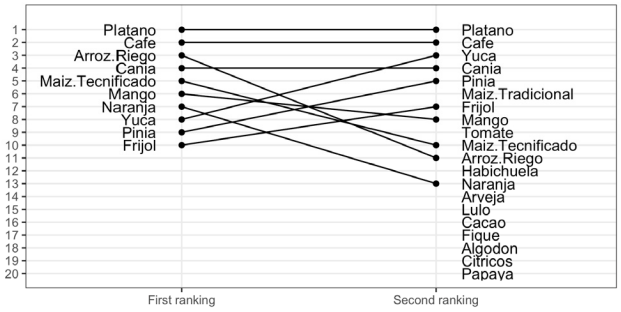


FIGURE 5. Global similarities (percentage values) for each MLC model in the CDS1 dataset using RBO (Rank Biased Overlap) and ULS (Unranked Lists' Similarity) metrics.



(a)



(b)

FIGURE 6. **Comparison of crop rankings by municipality in the (a) CDS1 and (b) CGD datasets applying the BR-RF model. The first ranking corresponds to the actual ranking, while the second is the predicted ranking.**

these values could be acceptable considering that the test data correspond to data unknown for the MLC models. Furthermore, although the difference was not significant between these models, BR-RF obtained a high and more generalized performance in the 5 datasets. This finding allows considering the application of this unique model in the crop prediction contemplated in our proposal. As an example of the rankings predicted by the BR-RF model, Fig. 6 presents a comparison of the actual and predicted rankings for the municipalities of Totoro (Fig. 6(a)) and Santander de Quilichao (Fig. 6(b)) in Cauca, Colombia.

As shown in the examples in Fig. 6, the selected model is correct for most of the crops in the top-rank. Although these crops' order was not the same as the actual order (which is difficult to obtain in practical terms), the model matched in the first or second position. Furthermore, the predicted ranking gets additional crops in the bottom-rank, which could be considered by experts for further analysis of new crops' adaptation in a territory. We obtained similar results with the other municipalities considered in this study. On the other hand, we performed a final test comparing RBO values with both training and actual data. We identify similarities in the predicted crops using known and unknown data for the model through this test. Fig. 7 presents the RBO values obtained in all municipalities using training and actual data in the CGD dataset. The other datasets showed similar behavior, with most municipalities retaining the same trend according

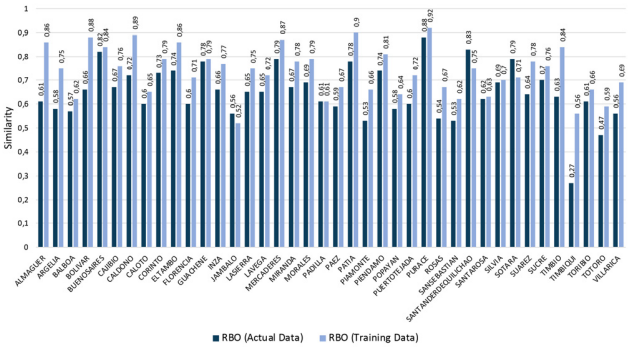


FIGURE 7. **Global RBO values for all municipalities using the BR-RF model in the CDS1 dataset.**

TABLE 6. **Variation and correlation between RBOs obtained with training and actual data in all combined data sources.**

CDS	Average RBO Variation	Pearson's Correlation Coefficient	p-value
CDS1	6%	0.71	$1.2 * 10^{-4}$
CDS2	9%	0.67	$2.8 * 10^{-3}$
CDS3	7%	0.6	$3.7 * 10^{-5}$
CDS4	9%	0.72	$7.1 * 10^{-5}$
CDS5	9%	0.73	$7.9 * 10^{-8}$

to the results presented in Table 6. These results indicate a low difference between the RBO obtained with training data and actual data (Training Data RBO - Actual Data RBO). Likewise, Pearson's correlation coefficient reaffirms a high level of agreement between these two trends (correlation coefficient close to 1 and p-value < 0.05).

G. HUMAN/COMPUTER INTERFACE

We developed IoT-Agro (https://www.iot-agro.com/ servicios/cropmodel), a software prototype to deploy the predicted crops in a municipality from all the previous results. This platform allows the user to select the municipality, the combined data source, and its related variables to consult the crop's planting probability in the short term. We used a combination of three programming languages to implement this tool. The R language was used to generate the predictive models, Java for the deployment of web services, and php to build the web site.

V. CONCLUSION

In this study, we proposed a data fusion strategy to support Climate Vulnerability Assessments (CVA), specifically in the study case of predicting the adaptability of crops in a territory in the short term, where three main contributions are highlighted: a multi-dimensional data preparation process specifically oriented towards CVAs, an adaptation of the JDL data fusion model to define a strategy for merging data from different agricultural vulnerability dimensions, and the modeling and implementation of a multi-label classification approach for crop prediction. Through our proposal, the policy-makers can establish public policy decisions on the risks and vulnerabilities in a region with an acceptable

certainty level. However, replicating these analyses usually requires complex and organized processes, where all stakeholders make their decisions through coordinated efforts. Our approach can guide CVAs in both land use and crops located in non-optimal areas for growth and production.

From a data-driven perspective, we analyzed the data source meta-features before and after preparation. It representing a reduction of time and effort in the training and application of many predictive models, which, in many cases, will not be applied in a real environment. By consolidating meta-features, we established an overview of data quality and strategies to improve it. These strategies can be defined by analyzing each meta-feature with expert knowledge. Similarly, we identify the most relevant attributes in a data source through a mixed approach (algorithms and expert knowledge). However, prioritization methods are not the same for all data sources. If they are labeled, we can apply logistic regression or random forest techniques, while for unlabeled data sources, we use correlations among variables.

On the other hand, a key finding was to identify a central dataset (Agronet) for labeling all Combined Data Sources (CDS). It allowed us to adjust the CDS to our crop prediction objective, considering that the central dataset was related to information about production and crop yield per municipality. Therefore, the labeling process should be done in parallel with the integration, considering the target variables. In our case, such target variables corresponded to the crops associated with a municipality. The multi-label approach was selected because the agricultural production of a territory focuses on a wide variety of crops, and the relationship between them is relevant for this type of analysis. The exploratory analysis in MLDs provided key metrics such as Theoretical Complexity Score (TCS) to identify those CDS that might be most appropriate to train subsequent predictive models. We identified Climate, Soil, Water Quality, Productive Alliances, and Production as the most relevant data sources to be integrated.

We used the BR-RF (Binary Relevance - Random Forest) model to perform the crop prediction considering two important findings. The first is related to prediction and classification tasks. Although we found models with better prediction performance, such as Label Powerset (LP) and Random k-label sets (RAKEL), the results were poor in crop probability ranking. Statistical significance tests proved no significant differences in classification task applying the best models for ranking. These tests and the validation with actual crop production data allowed us to select the BR-RF model as the best performance for our final prediction objective. The performance of the applied multi-label multi-view multi-label classification models was not superior to the performance of the conventional multi-label models in any of the combined datasets. Furthermore, the performance of the applied multi-view multi-label classification models was not better than the performance of the conventional multi-label models in any of the combined datasets. These results can be explained by considering issues such as the low number of instances in such datasets.

The ULS (Unranked Lists' Similarity) exceeded 90% regardless of the order of elements in both the predicted and actual rankings, RBO (Rank Biased Overlap) similarity reached a maximum of 67% strictly considering the order. Nevertheless, these results indicates that for more exhaustive ranking comparisons, the last similarity percentage is acceptable, considering the difficulty in comparing the position of each element within the ranking. In this sense, we can use the predicted rankings to provide crop recommendations at the same level of relevance, i.e., which crops could be produced in the short term without considering the probability (ranking positions). On the other hand, if we provide a ranking of crops, we require a strategy to improve the RBO similarity in the predictive models.

REFERENCES

[1] H. C. J. Godfray, J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, and C. Toulmin, "Food security: The challenge of feeding 9 billion people," *Science*, vol. 327, no. 5967, pp. 812–818, Feb. 2010, doi: 10.1126/science.1185383.

[2] R. B Singh, S. Shastun, S. Chibisov, A. Itharat, F. De Meester, D. W Wilson, G. Halabi, R. Horiuchi, and T. Takahashi, "Functional food security and the heart," *J. Cardiology Therapy*, vol. 4, no. 1, pp. 599–607, 2017, doi: 10.17554/j.issn.2309-6861.2017.04.125.

[3] USAID. (2014). *Climate-Resilient Development: A Framework for Understanding and Addressing Climate Change, U.S. Agency for International Development*. Accessed: Nov. 7, 2020. [Online]. Available: http://www.usaid.gov/sites/default/files/documents/1865/climate-resilient-developmentframework.pdf

[4] USAID. (2018). *Designing Climate Vulnerability Assessments, U.S. Agency for International Development*. Accessed: Nov. 7, 2020. [Online]. Available: https://www.climatelinks.org/sites/default/files/asset/document/2018_USAID-ATLAS-Project_Designing-Climate-Vulnerability-Assessments.pdf

[5] FAO. *Climate Smart Agriculture Sourcebook, Climate-Smart Crop Production Practices and Technologies*. Accessed: Oct. 7, 2020. [Online]. Available: http://www.fao.org/climate-smart-agriculture-sourcebook/production-resources/module-b1-crops/chapter-b1-2/en/

[6] TheWorldBank. (2013). *Open Data + Agriculture Can Transform How Farmers Respond to Looming Crises*. Accessed: Jul. 13, 2020. [Online]. Available: https://www.worldbank.org/en/news/feature/2013/04/26/open-data-can-transform-farmers-response-to-crisis

[7] A. N. Steinberg, C. L. Bowman, and F. E. White, "Revisions to the JDL data fusion model," in *Proc. Sensor Fusion, Archit., Algorithms, Appl. III*, Mar. 1999, pp. 430–441, doi: 10.1117/12.341367.

[8] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus, *Multilabel Classification—Problem Analysis, Metrics and Techniques*. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-41111-8.

[9] N. You and J. Dong, "Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth engine," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 109–123, Mar. 2020, doi: 10.1016/j.isprsjprs.2020.01.001.

[10] J. Zhang, L. Feng, and F. Yao, "Improved maize cultivated area estimation over a large scale combining MODIS–EVI time series data and crop phenological information," *ISPRS J. Photogramm. Remote Sens.*, vol. 94, pp. 102–113, Aug. 2014, doi: 10.1016/j.isprsjprs.2014.04.023.

[11] C. Liao, J. Wang, T. Dong, J. Shang, J. Liu, and Y. Song, "Using spatio-temporal fusion of Landsat-8 and MODIS data to derive phenology, biomass and yield estimates for corn and soybean," *Sci. Total Environ.*, vol. 650, pp. 1707–1721, Feb. 2019, doi: 10.1016/j.scitotenv.2018.09.308.

[12] L. Piedelobo, D. Hernández-López, R. Ballesteros, A. Chakhar, S. Del Pozo, D. González-Aguilera, and M. A. Moreno, "Scalable pixel-based crop classification combining Sentinel-2 and Landsat-8 data time series: Case study of the Duero river basin," *Agricult. Syst.*, vol. 171, pp. 36–50, May 2019, doi: 10.1016/j.agsy.2019.01.005.

[13] A. de Sherbinin, T. Chai-Onn, M. Jaiteh, V. Mara, L. Pistolesi, E. Schnarr, and S. Trzaska, "Data integration for climate vulnerability mapping in West Africa," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 4, pp. 2561–2582, Nov. 2015, doi: 10.3390/ijgi4042561.

[14] M. L. Mann and J. M. Warner, "Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach," *Field Crops Res.*, vol. 201, pp. 60–74, Feb. 2017, doi: 10.1016/j.fcr.2016.10.014.

[15] D. De Benedetto, A. Castrignano, M. Diacono, M. Rinaldi, S. Ruggieri, and R. Tamborrino, "Field partition by proximal and remote sensing data fusion," *Biosyst. Eng.*, vol. 114, no. 4, pp. 372–383, Apr. 2013, doi: 10.1016/j.biosystemseng.2012.12.001.

[16] H.-J. Chu, M. Z. Ali, and T. J. Burbey, "Spatio-temporal data fusion for fine-resolution subsidence estimation," *Environ. Model. Softw.*, vol. 137, Mar. 2021, Art. no. 104975, doi: 10.1016/j.envsoft.2021.104975.

[17] W. Jing and L. Xin, "Progresses on data fusion technology of crop growth model and multi-source observation information," *Remote Sens. Technol. Appl.*, vol. 30, no. 2, pp. 209–219, 2015. [Online]. Available: http://www.rsta.ac.cn/EN/10.11873/j.issn.1004-0323.2015.2.0209 and http://www.rsta.ac.cn/EN/Y2015/V30/I2/209, doi: 10.11873/j.issn.1004-0323.2015.2.0209.

[18] C. Pohl, K. D. Kanniah, and C. K. Loong, "Monitoring oil palm plantations in Malaysia," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 2556–2559, doi: 10.1109/IGARSS.2016.7729660.

[19] D. Moshou, I. Gravalos, D. K. C. Bravo, R. Oberti, J. S. West, and H. Ramon, "Multisensor fusion of remote sensing data for crop disease detection," in *Geospatial Techniques for Managing Environmental Resources*, J. K. Thakur, S. K. Singh, A. Ramanathan, M. B. K. Prasad, and W. Gossel, Eds. Amsterdam, The Netherlands Springer, 2011, pp. 201–219, doi: 10.1007/978-94-007-1858-6_13.

[20] L. Johansson, V. Epitropou, K. Karatzas, A. Karppinen, L. Wanner, S. Vrochidis, A. Bassoukos, J. Kukkonen, and I. Kompatsiaris, "Fusion of meteorological and air quality data extracted from the Web for personalized environmental information services," *Environ. Model. Softw.*, vol. 64, pp. 143–155, Feb. 2015, doi: 10.1016/j.envsoft.2014.11.021.

[21] P. Schweizer and N. Stein, "Large-scale data integration reveals colocalization of gene functional groups with meta-QTL for multiple disease resistance in barley," *Mol. Plant-Microbe Interact.*, vol. 24, no. 12, pp. 1492–1501, Dec. 2011, doi: 10.1094/MPMI-05-11-0107.

[22] A. R. Muljarto, J.-M. Salmon, B. Charnomordic, P. Buche, A. Tireau, and P. Neveu, "A generic ontological network for Agri-food experiment integration—Application to viticulture and winemaking," *Comput. Electron. Agricult.*, vol. 140, pp. 433–442, Aug. 2017, doi: 10.1016/j.compag.2017.06.020.

[23] P. Grover and R. Johari, "PAID: Predictive agriculture analysis of data integration in India," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2016, pp. 184–188.

[24] X. E. Pantazi, D. Moshou, A. M. Mouazen, T. Alexandridis, and B. Kuang, "Data fusion of proximal soil sensing and remote crop sensing for the delineation of management zones in arable crop precision farming," in *Proc. HAICTA*, 2015, pp. 765–776.

[25] S. Ehrmann, R. Seppelt, and C. Meyer, "Harmonise and integrate heterogeneous areal data with the R package arealDB," *Environ. Model. Softw.*, vol. 133, Nov. 2020, Art. no. 104799, doi: 10.1016/j.envsoft.2020.104799.

[26] W. J. de Lange, R. M. Wise, G. G. Forsyth, and A. Nahman, "Integrating socio-economic and biophysical data to support water allocations within river basins: An example from the Inkomati water management area in South Africa," *Environ. Model. Softw.*, vol. 25, no. 1, pp. 43–50, Jan. 2010, doi: 10.1016/j.envsoft.2009.06.011.

[27] H. Mousannif and J. Zahir, "AgriFuture: A new theory of change approach to building climate-resilient agriculture," in *Advanced Intelligent Systems for Sustainable Development (AI2SD)*, vol. 911, M. Ezziyyani, Ed. Cham, Switzerland: Springer, 2019, pp. 88–97, doi: 10.1007/978-3-030-11878-5_10.

[28] I. D. López and J. C. Corrales, "A smart farming approach in automatic detection of favorable conditions for planting and crop production in the upper basin of Cauca River," in *Advances in Information and Communication Technologies for Adapting Agriculture to Climate Change*, vol. 687, P. Angelov, J. A. Iglesias, and J. C. Corrales, Eds. Cham, Switzerland: Springer, 2018, pp. 223–233, doi: 10.1007/978-3-319-70187-5_17.

[29] G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," *Environ. Model. Softw.*, vol. 80, pp. 259–264, Jun. 2016, doi: 10.1016/j.envsoft.2016.02.030.

[30] Q. Yang, J. Shao, M. Scholz, C. Boehm, and C. Plant, "Multi-label classification models for sustainable flood retention basins," *Environ. Model. Softw.*, vol. 32, pp. 27–36, Jun. 2012, doi: 10.1016/j.envsoft.2012.01.001.

[31] I. Shendryk, Y. Rist, R. Lucas, P. Thorburn, and C. Ticehurst, "Deep learning—A new approach for multi-label scene classification in planetscope and Sentinel-2 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 1116–1119, doi: 10.1109/IGARSS.2018.8517499.

[32] H. Omrani, A. Tayyebi, and B. Pijanowski, "Integrating the multi-label land-use concept and cellular automata with the artificial neural network-based land transformation model: An integrated ML-CA-LTM modeling framework," *GISci. Remote Sens.*, vol. 54, no. 3, pp. 283–304, May 2017, doi: 10.1080/15481603.2016.1265706.

[33] H. Omrani, F. Abdallah, O. Charif, and N. T. Longford, "Multi-label class assignment in land-use modelling," *Int. J. Geographical Inf. Sci.*, vol. 29, no. 6, pp. 1023–1041, Jun. 2015, doi: 10.1080/13658816.2015.1008004.

[34] Y. Zhong and M. Zhao, "Research on deep learning in apple leaf disease recognition," *Comput. Electron. Agricult.*, vol. 168, Jan. 2020, Art. no. 105146, doi: 10.1016/j.compag.2019.105146.

[35] A. A. Abd El-aziz, A. Darwish, D. Oliva, and A. E. Hassanien, "Machine learning for apple fruit diseases classification system," in *Advances in Intelligent Systems and Computing*, vol. 1153. Cham, Switzerland: Springer, 2020, pp. 16–25, doi: 10.1007/978-3-030-44289-7_2.

[36] Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "AgroConsultant: Intelligent crop recommendation system using machine learning algorithms," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–6, doi: 10.1109/ICCUBEA.2018.8697349.

[37] Q. Tan, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Incomplete multi-view weak-label learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2703–2709, doi: 10.24963/ijcai.2018/375.

[38] J. Huang, F. Qin, X. Zheng, Z. Cheng, Z. Yuan, and W. Zhang, "Learning label-specific features for multi-label classification with missing labels," in *Proc. IEEE 4th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–5, doi: 10.1109/BigMM.2018.8499080.

[39] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label specific features for multi-label classification," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 181–190, doi: 10.1109/ICDM.2015.67.

[40] J. Huang, G. Li, Q. Huang, and X. Wu, "Learning label-specific features and class-dependent labels for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3309–3323, Dec. 2016, doi: 10.1109/TKDE.2016.2608339.

[41] C. Zhang, Z. Yu, Q. Hu, P. Zhu, X. Liu, and X. Wang, "Latent semantic aware multi-view multi-label classification," in *Proc. AAAI*, 2018, pp. 4414–4421.

[42] X. Wu, Q.-G. Chen, Y. Hu, D. Wang, X. Chang, X. Wang, and M.-L. Zhang, "Multi-view multi-label learning with view-specific information extraction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3884–3890, doi: 10.24963/ijcai.2019/539.

[43] J. Huang, X. Qu, G. Li, F. Qin, X. Zheng, and Q. Huang, "Multi-view multi-label learning with view-label-specific features," *IEEE Access*, vol. 7, pp. 100979–100992, 2019, doi: 10.1109/ACCESS.2019.2930468.

[44] I. D. Lopez, A. Figueroa, and J. C. Corrales, "Multi-dimensional data preparation: A process to support vulnerability analysis and climate change adaptation," *IEEE Access*, vol. 8, pp. 87228–87242, 2020, doi: 10.1109/ACCESS.2020.2992255.

[45] P. Christen, "Concepts and techniques for record linkage, entity resolution, and duplicate detection," in *Data Matching*, 1st ed. 2012, p. 272, doi: 10.1007/978-3-642-31164-2.

[46] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," in *Information Fusion*, vol. 50. Amsterdam, The Netherlands: Elsevier, Oct. 2019, pp. 71–91, doi: 10.1016/j.inffus.2018.09.012.

[47] D. Charte and F. Charte, "mldr: Paquete R para Exploración de Datos Multietiqueta," in *Proc. 16th Conferencia de la Asociación Española Para la Inteligencia Artificial (CAEPIA)*, Albacete, Spain, 2015, pp. 695–704.

[48] H. Scheffé, *The Analysis of Variance*. Hoboken, NJ, USA: Wiley, 1999.

[49] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "Resampling multilabel datasets by decoupling highly imbalanced labels," in *Hybrid Artificial Intelligent Systems*. Cham, Switzerland: Springer, 2015, pp. 489–501, doi: 10.1007/978-3-319-19644-2_41.

[50] Q. Tan, G. Yu, C. Domeniconi, J. Wang, and Z. Zhang, "Multi-view weak-label Learning based on matrix completion," in *Proc. SIAM Int. Conf. Data Mining, (SDM)*, San Diego, CA, USA: San Diego Marriott Mission Valley, May 2018, pp. 450–458, doi: 10.1137/1.9781611975321.51.

[51] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014, doi: 10.1109/TKDE.2013.39.
[52] B. Inmon, *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Osaka, Japan: Technics Publications, 2016.
[53] I. D. López, J. F. Grass, A. Figueroa, and J. C. Corrales, "A proposal for a multi-domain data fusion strategy in a climate-smart agriculture context," *Int. Trans. Oper. Res.*, pp. 1–22, Oct. 2020, doi: 10.1111/itor.12899.
[54] C. Peterson, A. Nowak, A. Jarvis, C. Navarrete, A. Figueroa, N. M. Riano, and J. Vargas, "Analysing vulnerability: A multi-dimensional approach from Colombia's upper Cauca river basin," in *Policy Brief. Cali, Colombia, Climate and Development Knowledge Network (CDKN)*. Cali, Colombia: Climate and Development Knowledge Network, 2012.
[55] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *J. Amer. Stat. Assoc.*, vol. 84, no. 406, pp. 414–420, Jun. 1989.
[56] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "On the impact of dataset complexity and sampling strategy in multilabel classifiers performance," in *Hybrid Artificial Intelligent Systems*. Cham, Switzerland: Springer, 2016, pp. 500–511, doi: 10.1007/978-3-319-32034-2_42.
[57] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, Sep. 2015, doi: 10.1016/j.neucom.2014.08.091.
[58] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Dealing with difficult minority labels in imbalanced mutilabel data sets," *Neurocomputing*, vols. 326–327, pp. 39–53, Jan. 2019, doi: 10.1016/j.neucom.2016.08.158.
[59] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, nos. 3–4, pp. 591–611, Dec. 1965, doi: 10.1093/biomet/52.3-4.591.
[60] J. L. Gastwirth, Y. R. Gel, and W. Miao, "The impact of Levene's test of equality of variances on statistical theory and practice," *Stat. Sci.*, vol. 24, no. 3, pp. 343–360, 2009.
[61] J. F. Box, "Guinness, Gosset, Fisher, and small samples," *Stat. Sci.*, vol. 2, no. 1, pp. 45–52, Feb. 1987, doi: 10.1214/ss/1177013437.
[62] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949.
[63] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Trans. Inf. Syst.*, vol. 28, no. 4, pp. 1–38, Nov. 2010, doi: 10.1145/1852102.1852106.

IVÁN DARÍO LÓPEZ received the degree in information systems engineering and the master's degree in telematics engineering from the University of Cauca, Colombia, in 2011 and 2016, respectively, where he is currently pursuing the Ph.D. degree in telematics engineering. He is also a Researcher with the Telematics Engineering Group (GIT), University of Cauca. His current research interests include big data analysis, data mining, and machine learning techniques for modeling and developing applications in the agricultural sector, specifically oriented to climate-smart agriculture (CSA).

APOLINAR FIGUEROA received the degree in biology from the University of Cauca, Colombia, in 1982, the master's degree in ecology from the University of Barcelona, Spain, in 1986, and the Ph.D. degree in biological sciences from the University of Valencia, Spain, in 1999. He is currently a Full Professor and leads the Environmental Studies Group, University of Cauca. His research interests include environmental impact assessment and biodiversity management.

JUAN CARLOS CORRALES received the Dipl.-Ing. and master's degrees in telematics engineering from the University of Cauca, Colombia, in 1999 and 2004, respectively, and the Ph.D. degree in sciences from the University of Versailles Saint-Quentin-en-Yvelines, France, in 2008, with a focus on computer science. He is currently a Full Professor and leads the Telematics Engineering Group (GIT), University of Cauca. His research interests include machine learning and data analytics.