FramePests: A Comprehensive Framework for Crop Pests Modeling and Forecasting

Emmanuel Lasso, Natacha Motisi, Jacques Avelino, Juan Carlos Corrales

Received June 28, 2021, accepted July 27, 2021, date of publication August 13, 2021, date of current version August 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104537

FramePests: A Comprehensive Framework for Crop Pests Modeling and Forecasting

EMMANUEL LASSO[1], NATACHA MOTISI[2,3,4], JACQUES AVELINO[2,3,4], AND JUAN CARLOS CORRALES[1]

[1]Telematics Engineering Group, University of Cauca at Tulcán, Popayán 190003, Colombia
[2]Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR PHIM, Turrialba 30501, Costa Rica
[3]PHIM Plant Health Institute, University of Montpellier, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), INRAE, Institut Agro, IRD, 34394 Montpellier, France
[4]Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), DID, Turrialba 30501, Costa Rica

Corresponding author: Emmanuel Lasso (eglasso@unicauca.edu.co)

This work was supported in part by the Doctoral Program of Telematics Engineering and the Telematics Engineering Group (GIT) of the University of Cauca, in part by the "Programa Centroamericano de Gestión Integral de la Roya del Café (PROCAGICA)" through the EU under Grant DCI-ALA/2015/365-17, and in part by the Tropical Agricultural Research and Higher Education Center (CATIE). The work of Emmanuel Lasso was supported by the InnovAccion Cauca Project of the Colombian Science, Technology and Innovation Fund (SGR-CTI) through the Ph.D. Scholarship.

ABSTRACT Crop pests are among the greatest threats to food security, generating broad economic, social, and environmental impacts. These pests interact with their hosts and the environment through complex pathways, and it is increasingly common to find professionals from different areas gathering into projects that attempt to deal with this complexity. We propose a framework called *FramePests* guiding steps and activities for crop pest modeling and forecasting. From theoretical references about carrying out mappings and systematic reviews of the literature, the framework proposes a series of steps leading to a state of science as a knowledge base for modeling tasks. Then, two modeling solutions, based on data and knowledge are used. Finally, the model outputs and performances are compared. The application of the proposed framework was demonstrated for coffee leaf rust modeling, for which we obtained a data-based model built using a gradient boosting algorithm (*XGBoost*) with a mean absolute error of 7.19% and a knowledge-based model represented by a hierarchical multi-criteria decision structure with an accuracy of 56.03%. A complementary study for our case study allowed us to explore how elements of a data-based model can improve a knowledge-based model, improving its accuracy by 7.07%. and showed that knowledge-based modeling can be an alternative to data-based modeling when the available dataset has approximately 60 instances. Data-based models tend to have better performance, but their replicability is conditioned by the diversity in the dataset used. Knowledge-based models may be simpler but allow expert supervision, and these models are not usually tied to specific sites.

INDEX TERMS Crop pest forecasting, data-based model, knowledge-based model, smart farming.

I. INTRODUCTION

According to the Food and Agriculture Organization (FAO), pests are among the greatest threats to food security, generating broad economic, social, and environmental impacts [1]. For integrated pest management, the term *Pest* refers to any living being (diseases caused by pathogens, fungi, viruses, insects, nematodes, etc.) that cause damage to crop plants [2]. These pests interact with their hosts and the environment at different spatial and temporal scales through complex pathways, and it is increasingly common to find professionals from different areas (farmers, technicians, plant pathologists, computer scientists, economists, etc.) gathering into projects that attempt to deal with this complexity. This complexity increases when multiple pests are simultaneously analyzed at the same time. If professionals' profiles are diverse, the challenge is to achieve a mutual understanding of the agroecosystem and coordination of activities within the work team.

Forecasting pest development is required for three reasons: economic impact, food safety, and justification of control

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

methods [3]. Pest forecasting can be addressed by constructing models that describe the conditions for pest propagation and the damage caused to host crops. Pest development modeling can provide explanatory models of pathosystems and/or predictive models that can be used in early warning systems [4]. The models can be built either from data or expert knowledge (expertise, literature, technical reports, expert interview) of the mechanisms involved in the interactions between the hosts, the pests, and the environment.

In the model generation, three contrasting situations were highlighted. In the first situation, few data exist on the pathosystem, but knowledge is available, allowing the development of knowledge-based models (e.g., mechanistic and qualitative) without the possibility of using data for model evaluation and validation. In the second situation, a large amount of data is available, but exhaustive knowledge of the pathosystem is lacking, which can be handled by exhaustive data processing through the induction of data-based models. In the ideal third situation, both sufficient knowledge and data are available, which allows evaluation of the knowledge-based models using the data, as well as the improvement of the data-based models with available knowledge. Knowledge and models are often found in academic publications, as well as in gray literature, but not directly available for the vast majority of farmers, the first actor implementing strategies for pest management.

Stenberg [5] highlighted the need for a conceptual framework that takes advantage of modern science to approach integrated pest management and thus optimize plant protection solutions. According to Jabareen [6], a conceptual framework (CF) is a set of concepts related to each other, explaining a phenomenon to achieve an understanding of it, which has had different applications in agriculture. Robert *et al.* [7] proposed a conceptual framework for data acquisition and analysis, integrated with expert knowledge and decision-making. The framework consists of four steps: definition of the problem, case selection, data collection and analysis, and model formalization.

There are methodologies for the construction of a knowledge base from scientific literature. Petersen *et al.* [8] presented a guide for carrying out an engineering systematic map. Mapping is a method of building a classification scheme and categorizing research reports and published literature. This methodology can be complemented by a systematic review proposed by Kitchenham [9]. This review is conducted to identify gaps in current research and appropriately position new research activities. Several methodologies are available for modeling. For knowledge-based modeling, Aubertot and Robin [10] proposed a multi-criteria qualitative modeling approach, called the injury profile simulator framework (IPSIM). IPSIM allows the building of knowledge-based models to predict injury profiles in crops as a function of cropping practices and the environment. In the case of data-based modeling, Chapman *et al.* [11] proposed the cross-industry standard process for data mining (CRISP-DM), a methodology to carry out data mining, as the induction of models from data, which is a topic of interest in the application of big data in smart farming [12].

For a group of researchers who want to start modeling work, there is no guide that considers all the elements to take into account to generate models from data or knowledge, depending on the conditions of the research to be carried out (presence or absence of data and formalized knowledge about pests) and taking into account the multiple entities and interactions that can affect the development of pests. Furthermore, a comparative and complementary study of models generated from knowledge versus those generated from data is necessary to understand the scope of each model and how these could complement each other.

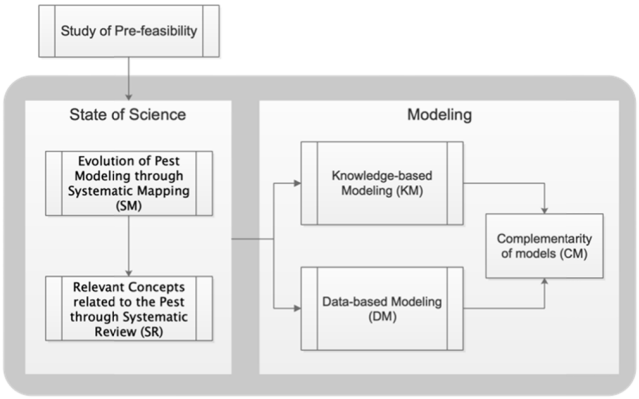

FIGURE 1. FramePests framework overview composed of three macroprocesses.

In this paper, we propose a comprehensive framework, *FramePests*, for crop pest development modeling and forecasting. FramePest is intended for interdisciplinary groups that are composed of both users without modeling experience (as a guided and ordered process), as well as for those with experience in that domain (as a formalization of all the tasks carried out in modeling). FramePest can be individually executed for various crop pests. This framework groups various methodologies developed in the field of conceptual framework construction, literature mappings and reviews, and modeling solutions [6], [8]–[11]. *FramePests* is described by three macroprocesses shown in Figure 1. The macroprocess study of pre-feasibility starts the *FramePests* execution. This is followed by the macroprocess state of science, which is an in-depth search in literature and expert knowledge described by two sub-processes: (i) the systematic mapping (module SM) describing the evolution of Pest Modeling and (ii) the Systematic Review (module SR) describing the relevant concepts related to the Pest. Then, based on the results obtained, we pass to the modeling macroprocess composed of three sub-processes: (i) knowledge-based modeling (module KM), (ii) Data-based modeling (module DM), and (iii) the complementarity of models (module CM) to compare the results of the two modeling approaches and explore how they can complement each other, both in the integration of knowledge in a data-based modeling process, as well as the use of

data analysis in the definition of a knowledge-based model. Finally, we present the functioning of *FramePests* through a coffee leaf rust case study.

The remainder of this paper is organized as follows. In Section 2, we describe the proposed framework. Section 3 discusses the application of the framework to a case study. Section 4 concludes the paper with some remarks.

II. THE FRAMEPESTS FRAMEWORK

This section presents the components of the *FramePests* framework.

A. STUDY OF PRE-FEASIBILITY

The study of pre-feasibility provides the necessary elements to start the *FramePests* execution (Figure 2).

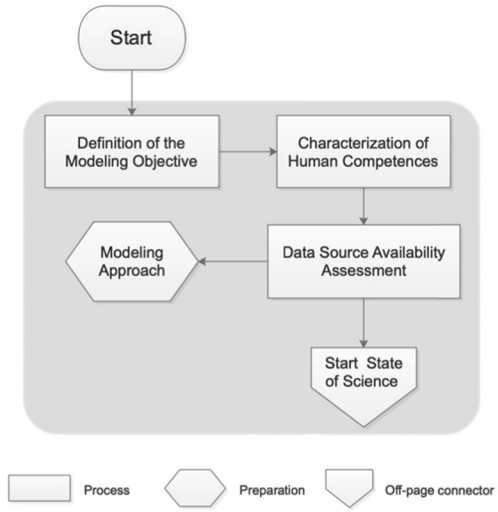

FIGURE 2. Study of pre-feasibility.

The components and activities are:

i. Definition of the modeling objective: This activity defines the scope of modeling in terms of the crop pest to be addressed, the scale of the analysis, the response variable, for what and for whom the modeling is carried out, among others.
ii. Characterization of human competences: This activity aims to identify the available human talent to execute the modeling macroprocess.
iii. Data Source Availability Assessment: This activity aims to identify the knowledge and data monitored in the crops required for pest modeling. The available data must be described, and it must be determined whether they are sufficient to carry out data-based modeling.
iv. Modeling approach: This preparation parameter sets the type of modeling that will be carried out: data-driven, knowledge-based, or both. If there are no databases describing the effect of variables on the pathosystem, data-based modeling cannot be performed, and only knowledge-based modeling can be carried out.
v. Start State of Science: this connector represents the beginning of processes in *FramePests*.

In the following sections, the macroprocesses are framed in one or more phases according to the Jabareen's methodology for conceptual framework building [6] described in seven steps: (i) the mapping of the selected sources, (ii) the extensive reading and categorization of the selected data, (iii) the identification and naming of concepts, (iv) the deconstruction and categorization of the concepts, (v) the integration of concepts, (vi) synthesis and resynthesis, and (vii) validation of the framework.

B. THE STATE OF SCIENCE

1) SYSTEMATIC MAPPING (SM) TO DESCRIBE THE EVOLUTION OF PEST MODELING

This sub-process aims to explore the studies that have addressed the pest's development (dynamics, patterns, temporal evolution) modeling in crops, based on the systematic mapping proposed by Petersen *et al.* [8] (Figure 3).

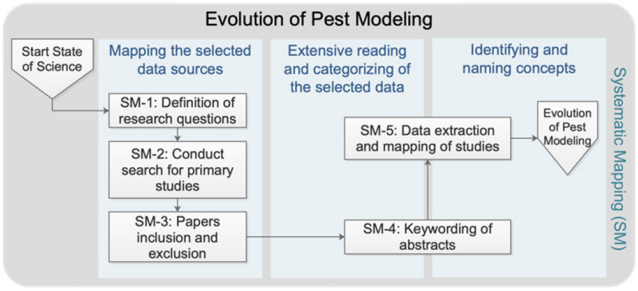

FIGURE 3. Systematic Mapping (SM) describing the evolution of Pest Modeling.

The definition of the research questions (SM-1) establishes the research scope. The questions should be oriented to what has been the evolution of the studies that addressed the pest and the most used research topics (multidisciplinary). Some recommended questions are as follows:

- What has been the evolution of pest modeling approaches?
- Which modeling techniques have been used for pest development forecasting?

The search for primary studies (SM-2) uses the defined scope to create search strings and submit them to bibliographic source systems. Since many results of experiments related to a pest have been published in technical bulletins, gray literature should be considered and can be characterized as basic knowledge.

The screening of papers for inclusion and exclusion (SM-3) applies exclusion and inclusion criteria to the studies obtained in the previous step. These criteria should be based on the relevance of the studies and ensure that their context is aligned with objective modeling. A reading of the abstracts can help define whether the studies provide valuable knowledge for pest modeling.

Next, the keywording of the abstracts (SM-4) aims to find keywords and concepts that reflect the contribution of each study, in addition to the keywords specified by the authors. The concepts are then analyzed to develop a high-level understanding of the research and generate a classification scheme according to categories of elements related to the pest, such as meteorology, cropping practices, and crop and pest properties.

Finally, the data from the studies were extracted, and the mapping was generated (SM-5) from the categories of concepts found. The visualization of the mapping, representing the *evolution of pest modeling*, corresponds to a comparison of elements such as publication frequencies, affiliations, years of publications, and main concepts.

2) SYSTEMATIC REVIEW (SR) TO DESCRIBE THE RELEVANT CONCEPTS RELATED TO THE PEST

The systematic review corresponds to a refinement of the results obtained in the systematic mapping, reaching a deeper level of analysis and identifying current trends in crop pest modeling. This sub-process deals with understanding the fundamental concepts around the development cycle of the Pest in crops, based on the systematic review proposed by Kitchenham [9] (Figure 4).

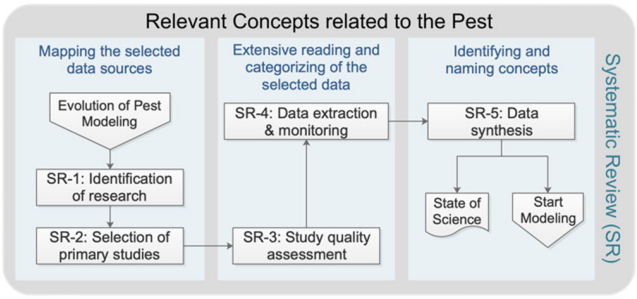

FIGURE 4. Relevant concepts related to the Pest through Systematic Review (SR).

The identification of research (SR-1) is based on the definition of research questions and search documentation. This process can take the elements of the previous work done in SM-1 and SM-2. At this point, the research questions seek more specific information than those in SM. Following the recommended questions for SM-1, the new questions are as follows:

- What are the variables most related to the most studied aspect of pest?
- How were the techniques used for pest development forecasting implemented?

The selection of primary studies (SR-2) takes into account the studies that provide direct evidence about the work around the pest specified in the research questions. Additionally, mapping visualization (SM-5) can help define the time window used to identify trends in crop pest development. The inclusion and exclusion criteria were similar to those used in SM-3, obtaining a set of relevant studies.

A quality assessment of the relevant studies (SR-3) ensures a more reliable filter for better contributions to pest studies. The highlights in each study must be interpreted according to the metrics and procedures used to compare them. Additionally, the future works proposed in these studies can guide current pest modeling.

Next, the information is extracted and monitored (SR-4) through forms according to the elements analyzed in the studies. Additionally, the databases used in the studies, their properties, and access conditions should be identified. Public datasets related to pests are potential resources for validating the current research results.

Finally, the results of the previous process were collected and summarized in a Data synthesis (SR-5). The most used modeling techniques, how they are implemented, the predictors used, the type of validation, and the principal authors about the pest development modeling must be identified. A document called the *state of science* is generated and must contain all the relevant findings as the materials, methods, and techniques used, principal authors, performance metrics, and highlights. The on-page connector *start modeling* corresponds to a state that gives way to start the modeling approaches (knowledge-based or data-based).

The application of systematic review (SR) after systematic mapping (SM) allows starting from a base of previously filtered and analyzed studies. Although these two methodologies have some similarities, we decided to use them in a complementary manner, rather than merging them. In our case study, human talent consisted of an interdisciplinary group. However, this situation is not always present, and our approach allows groups of pest/crop experts or groups of data scientists to carry out a successful modeling process with a crop pest knowledge base supported in the literature. While the findings consigned in scientific production (books, journal papers, etc.) show solid bases of knowledge on a pest, gray literature is still very important, since many resources in this category correspond to knowledge that is being applied by partner institutions to the crop production in each country or region.

C. MODELING

The modeling macroprocess describes the processes of building models based on the state of the science macroprocess. It is composed of three sub-processes: knowledge-based modeling, data-based modeling, and complementarity of models.

1) KNOWLEDGE-BASED MODELING (KM) THROUGH IPSIM

This sub-process aims to build a decision structure that allows characterization of the pest, based on the agronomic and environmental conditions of the crop (Figure 5), based on IPSIM [10].

The modeling is based on expert knowledge expressed as a tree-based structure composed of attributes, aggregated attributes, and the output variable representing the model. KM-1 collects the state of science available on the pest to

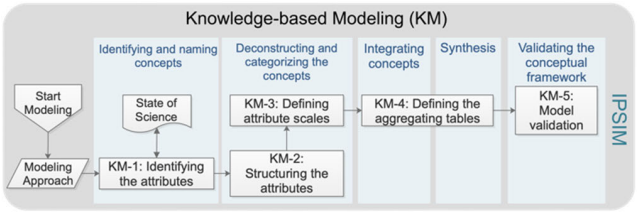

FIGURE 5. Knowledge-based modeling (KM) through injury profile simulator.

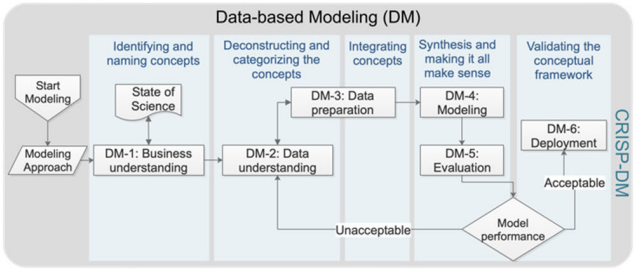

FIGURE 6. Data-based modeling (DM) through CRISP-DM.

be modeled and identifies the entry variables and concepts, called *basic attributes*, related to the pest from the state of science and their properties. KM-2 identifies the relationships between the basic attributes, which are *aggregated attributes*. The basic attributes correspond to the main concepts, and the aggregated attributes correspond to the mapping categories identified in SM-4. Each attribute is scaled by defining the possible values it can take (KM-3). The scales are defined by a threshold and should express the properties as a qualitative variable (nominal or ordinal). The aggregating tables (KM-4) represent the hierarchical multi-criteria decision structure. Each table represents a mapping of all combinations of attribute categories based on ''if-then'' rules. The rules are defined according to the effects and importance of an attribute over another (variable weight). Both scales and rules can be defined from the knowledge obtained in the state of science or expert knowledge. The main output is the response of the model. This corresponds to the successive aggregation of tables in a hierarchical order.

The last step was the validation of the model, in which the model performance was estimated from simulated events or historical data. This process was added to the end of the macroprocess (KM-5). The validation comprises the following activities:

- Define validation criteria: the standard performance metrics for classification suggested are accuracy, precision, recall, F1-score [13], and Cohen's weighted kappa [14].
- Prepare simulation cases from information: This can be done from historical data or hypothetical cases defined by an expert.
- Apply the model to the simulation cases and compare the output of the model with the actual expected output.
- Collect validation results according the validation criteria.

After carrying out the model validation, if the results are not acceptable, an iteration to the IPSIM-4 process is required to calibrate the category thresholds of the variables and the combinations of the attributes in the aggregation tables.

2) DATA-BASED MODELING (DM) THROUGH CRISP-DM

This sub-process is a process of induction of machine learning models from a dataset that represents the conditions (agronomic practices and environmental) of the crops, based on CRISP-DM [11] (Figure 6).

Business understanding (DM-1) takes the produced *state of science*. The business corresponds to the problem to be solved, in this case, pest modeling. Pest knowledge corresponds to the main concepts and categories in the classification schema identified in SM-4 and SR-4. Knowledge is expressed in technical terms as a data mining objective.

The data understanding (DM-2) process begins collecting all available data sources for the pest, relevant for the business, first the variable to explain (target), and second, the explanatory variables (predictors). This process can have iterations with data preparation (DM-3) for each modification of the dataset or generation of a new dataset from the original datasets. One of the most common ways of describing variables is based on descriptive statistics according to the type of variable (quantitative or qualitative).

Data preparation (DM-3) addresses the transformation of the original datasets, and it begins with manual feature inclusion or exclusion. The criteria must correspond to variables that affect the pest and its development from SM-4 and SR-5 from the *state of science*. The success of the following processes depends mainly on the quality of the data used, anomalies in the dataset must be detected and resolved, either by discarding faulty instances or processing them to correct their value. A framework for the data cleaning process can be consulted for regression [15] and classification [16] tasks. Finally, if there is more than one dataset, they must be merged, taking care of the dimensions that each represents and its temporality.

With a clean and structured dataset, the modeling (DM-4) process can be executed. The final dataset was used to train the machine learning model. Different algorithms can be used depending on the learning task. Unsupervised learning algorithms train a model with no target variable specifications and are focused on recognizing patterns. Supervised learning algorithms train a model according to labeled examples (data with meaningful and informative labels from which a model can learn). Semi-supervised learning is a technique that uses labeled unlabeled data to train a model [17]. The recommended procedure is to apply several of these algorithms to the final dataset, calibrating its parameters to obtain optimal results. Cross-validation [12] is needed to determine each algorithm's performance metrics, which also depends on the modeling task. The typical performance metrics used were accuracy, precision, recall, F1-score, receiver

operating characteristic curve (ROC) for classification, mean absolute error, mean squared error for regression, and correlation for statistical methods. A guide for choosing the algorithms (phase 4 in [11]) to be tested for pest modeling is presented in [19].

The evaluation (DM-5) process compares the performance metrics of the applied algorithms to identify the best result and determines whether the business and modeling objectives are achieved. If the results are not acceptable, a new iteration from the DM-2 process is suggested.

The deployment (DM-6) plan must be structured following the modeling objectives and must respond to the case study and the end-user who will benefit. Whether the result is a prediction model or knowledge induced from the data, it must be organized and presented so that a user can use it.

3) COMPLEMENTARITY OF MODELS (CM)

This sub-process aims to analyze and extract the benefits and challenges of the two modeling approaches (knowledge-based and data-based modeling) and how they can complement each other.

Complementarity can be approached in two ways. The first is training a data-based model (when the data are available) with variables similar to those of the knowledge-based model. The data-based modeling process can provide elements to improve the knowledge-based model through the definition of scales (KM-3) and the relationship structure of the variables (KM-4). These elements can be association rules, the importance of variables, and the impact of the range of variables on predictions, among others. The other way is integrating knowledge obtained in the state of science within the data-based modeling in the data preparation (DM-3) and modeling (DM-4) phases.

The comparison of prediction models is generally in terms of their performance metrics when they are applied to a test dataset or simulation cases. For quantitative models, the desired performance is a low bias and low prediction error. In the case of qualitative models, the desired performance is high accuracy, recall, sensitivity, specificity [13], and Cohen's weighted kappa [14].

The performance of data-driven models is limited by the quantity and quality of the data. The first comparative aspect is knowing whether the performance of the DM model exceeds that of the KM model, and if this is true, knowing the minimum amount of training data necessary for the DM model to be as good as the KM model. We propose a process to obtain an approximation to the minimum size that a dataset must have so that a DM model induced from it has a performance as good as that of the KM model, as shown in Figure 7. From a training dataset, subsets of different sizes were randomly generated incrementally. In each iteration of the cycle shown in Figure 7, the size of the subset increases from 1 until it reaches the size of the training dataset. Next, the DM model was trained with the subset, and its performance metrics were calculated using a test dataset. In this case, if the output of the DM model is different from that of the KM model (for example, qualitative and quantitative), this output must be transformed to match. If the performance of the DM model is less than that of the KM model, then a file with the information of the experiment (size of the subset and performance metrics) is updated; if the subset has the maximum size (equal to the training dataset), the process ends; otherwise, it increases by 1 the size of the subset to be generated for the next cycle. If the performance of the DM model reaches or exceeds that of the KM model, the size of the subset for which this happens is stored together with the performance of the DM model and the process continues.

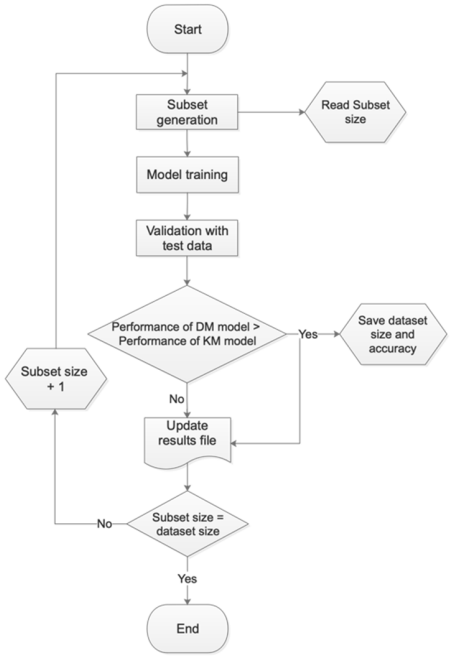

FIGURE 7. Train dataset minimum size estimation process.

Another comparison resource is to determine whether the difference between the outputs of the models is statistically significant [20]. To compare these models directly, the response of each can be transformed in terms of the other [21], [22], and test them using ANOVA and McNemar's metrics.

Assuming that the models are validated with the same dataset:

- For quantitative models, the variance (ANOVA) analysis was used to determine if there was a significant statistical difference between the means of two or more sets.
- For qualitative models, McNemar's test [23] can be used to determine whether the two methods (models) have the same accuracy. The test is based on the number of instances misclassified only by the first algorithm and the number of instances misclassified by the second algorithm.

However, if only one model is built, it can be compared with similar models identified in the relevant concepts related to the Pest through a systematic review (SR) macroprocess, from the application of models on the same validation dataset or the comparison of performance metrics.

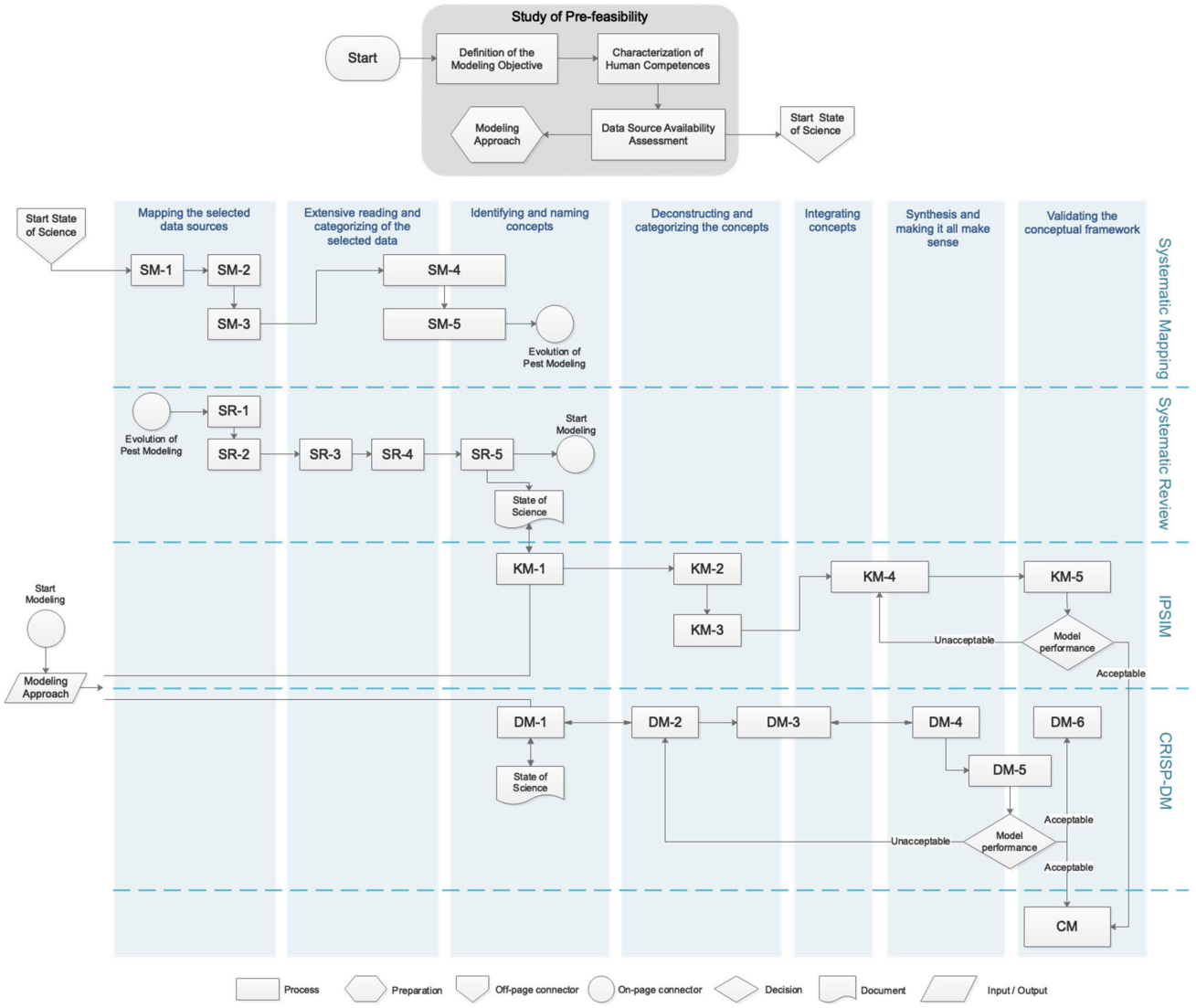

FIGURE 8. Execution flow of *FramePests* framework. after the Study of Pre-feasibility, the execution flow of activities in the framework instantiates each phase of the Jabareen's methodology for conceptual framework building [6].

The overall framework is shown in Fig. 8. After the Study of Pre-feasibility, the execution flow of activities in the framework instantiates each phase of Jabareen's methodology for conceptual framework building. The off-page connector ''Start State of Science'' corresponds to the execution of the evolution of pest modeling macroprocess through SM, followed by relevant concepts related to the Pest through SR. The on-page connector ''Start Modeling'' begins with knowledge-based and/or data-based modeling, depending on the modeling approach parameter. Finally, the CM module was used to compare the results of the two modeling approaches.

III. APPLICATION OF FRAMEPEST: THE CASE STUDY OF COFFEE LEAF RUST

We illustrate the *FramePests* framework and its execution flow with the coffee leaf rust (CLR), caused by the fungal pathogen *Hemileia vastatrix* Berk. & Broome (1869). In 2008-2011 Colombia suffered one of its most serious crises due to this pathogen, with a 30% decrease in production and income losses, notably for smallholder producers. In 2012-2013, a similar crisis occurred in Central America. The weather factors and production conditions conditioned their intensity [33]. The defoliation of the coffee tree and the death of branches in the worst cases are the main factors that cause production losses.

A. STUDY OF PRE-FEASIBILITY

Our modeling objective was to *forecast CLR disease dynamics at the field scale*. The human competences available for this study were a data scientist with experience in predictive modeling processes and two plant pathologists, experts in coffee Arabica-CLR pathosystem, all co-authors of this article. The study area corresponds to an experiment with several coffee-based agroforestry systems. These differ according to the tree species intercropped with coffee,

their number, and coffee management. The experiment was conducted at the Tropical Agricultural Research and Higher Education Center (CATIE) [24], [25], in the canton of Turrialba, province of Cartago, Costa Rica. The variety of coffee planted is Caturra of the species *Coffea arabica*. A CATIE weather station was located near the experiment and the data contained the following variables: maximum (*tMax*) and minimum (*tMin*) air temperature, average (*tAvg*) air temperature calculated over the day, average (*hAvg*) and minimum (*hMin*) relative humidity, and daily precipitation (*pre*). Additionally, the monitoring of pests in each plot was performed monthly. The available data sources are: monitoring of CLR incidence (CLRI), that is, the proportion of leaves that present the disease's symptoms or not [26]; vegetative growth of coffee trees; the properties of the plot (management and shade cover); and measurement date and weather data from the on-site station from 2002 to 2012. The number of instances in the dataset is 439.

We took the CLRI of the next month for each data sample as the future incidence to be predicted. The date of prediction (*DP*) corresponding to the day the previous incidence was measured, and the date of predicted incidence (*DPI*) corresponded to the day DP + 28 days later the predicted incidence was measured. The current incidence (*cCLRI*) corresponding to the CLRI measured each *DP*, and the predicted incidence (*pCLRI*) corresponded to the CLRI measured for each *DPI*. Meteorological attributes and host growth were calculated over 14 days preceding the date of incidence monitoring. This period was used to exclude part of the weather conditions that were already included in the monitored incidence, as this is the result of the weather conditions that occurred before. The period of 14 days is half of the time between the two monitoring periods. In addition, the monitored incidence provides a measurement of the inoculum stock available for new infections [27]. The host growth attribute was characterized by the difference between the number of leaves in coffee trees over a period of 14 days, as a proxy for characterizing growth dynamics in the following month. The attributes related to chemical control and nutrition refer to the two levels of conventional management in the CATIE experiment, as explained in [24].

The modeling objective was defined as *the model of CLRI development at the field scale*. The availability of experts enables the construction of KM, and the availability of an important dataset (number of data samples in the dataset) enables the building of DM. Thus, it was possible to carry out the modeling approach completely until the study of the complementarity of the models. After the Study of Prefeasibility, the flow of activities in the framework starts the sub-process evolution of pest modeling.

B. THE STATE OF SCIENCE

1) EVOLUTION OF CLR MODELING (SUB-PROCESS SM)

The research questions (SM-1) that establish the research scope were:

- What has been the evolution of coffee leaf rust modeling?
- Which modeling techniques have been used for coffee leaf rust forecasting?

We performed a search for primary studies (SM-2) in Web of Science to find studies published in scientific journals with impact factors and Google Scholar to find studies published in the gray literature. We created search strings according to the CLR name in different languages: English *coffee rust*, Spanish *roya del café*, and Portuguese *Ferrugem do cafeeiro*. Additionally, we use the following related modeling words: prediction, model, dynamics, and forecast. Table 1 lists the search strings for bibliographic source systems and the number of studies found. Owing to the large number of studies obtained in Google Scholar, these were ordered by relevance according to the tool that this search engine offers.

TABLE 1. Search strings and number of studies founded in bibliographic sources systems.

Search string	Source	Number
(TITLE-ABS-KEY (coffee AND rust) AND TITLE-ABS-KEY (prediction OR model OR dynamics OR forecast))	Web of Science	101
coffee AND rust AND (prediction OR model OR dynamics OR forecast)	Google Scholar	45200
roya AND café AND (predicción OR modelo OR dinámica)	Google Scholar	5570
Ferrugem AND cafeeiro AND (predição OR modelo OR dinâmica)	Google Scholar	6490

We considered the top of the most relevant documents, according to where they were published, who wrote them, as well as how often and how recently they were cited in other scholarly literature.[1] We selected 29 academic papers from the two bibliographic sources. The criteria for the screening of papers for the inclusion and exclusion process (SM-3) were: (i) studies directly related to the CLR modeling, not its detection on coffee leaves or studies of its impact on coffee production; (ii) studies corresponding to gray literature, such as technical manuals and bulletins of coffee institutions, providing a basis for knowledge of the principal drivers of CLR. We used *bibliometrix*, an R library for the science mapping analysis of the selected papers [28]. This library allowed us to perform an automatic analysis of academic papers about their references, authors, citations, affiliations, and keywords.

The keywording of the abstracts (SM-4) allowed us to find the following concepts: *Hemileia vastatrix, machine learning, decision trees, rust resistance, incidence, severity, climate change, temperature, humidity, precipitation, shade, and data mining*. The main categories found were *weather, agricultural activities, crop properties, and diseases*.

The incidence has been the most studied response variable in the most recent studies, while from the year 2000, the

[1] https://scholar.google.com/intl/en/scholar/about.html

emergence of works based on machine learning (ML) techniques is visible in the mapping results (SM5) (Figure 9).

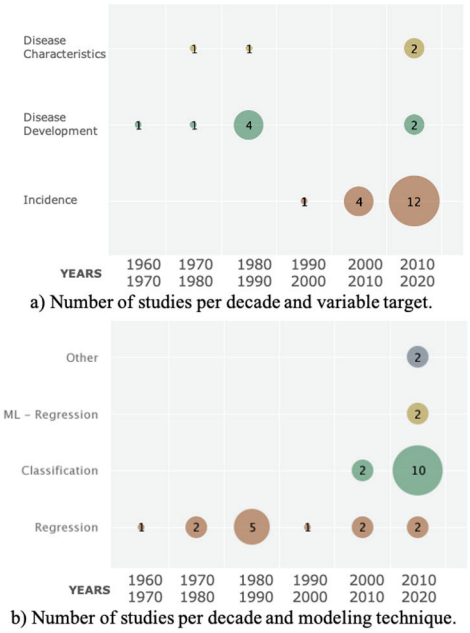

FIGURE 9. Mapping of studies in CLR modeling techniques used and elements around the CLR like its characteristics (genetics, resistance), development (Dynamics of CLR, spread of disease, etc.) and incidence studies.

2) RELEVANT CONCEPTS RELATED TO THE CLR (SUB-PROCESS SR)

We used the primary studies obtained in the evolution of CLR modeling as a basis for research identification (SR-1). Additionally, the research questions are updated as follows:

- What are the variables most related to rust incidence (CLRI)?
- How were the techniques used for CLRI modeling implemented?

The selection of primary studies (SR-2) took into account the same criteria as in the screening of papers for inclusion and exclusion process (SM-3) and the selection of those studies that were directly related to disease development and modeling. Thus, studies on modeling of disease resistance from its genetics, identification of severity in leaves from computer vision, socio-economic studies, and descriptive analyses were ruled out. From the density of studies shown in SM-5, we considered those published since 2000, as they represent the trends in CLR modeling. The results of the quality assessment (SR-3) and data extraction (SR-4) are synthesized (SR-5) in Table 2. This table presents the final relevant studies. The columns show the publication year, times cited (TC), target variable addressed, predictors of the target variable, modeling technique (MT), metric of the modeling validation, best metric value, and highlights of each study (main contributions, findings, approaches, and/or future works). Among the predictors considered, weather was the most used category, and its analysis can be improved using various time periods of different sizes (time windows) [29]. The other most used predictors were shade as a quantitative or qualitative variable, previous inoculum, use of fungicide, and fruit load. Regression-based models showed significant results, and the use of machine learning algorithms improved the modeling processes, providing feature selection methods, optimization of learning functions, and use of ensembles such as boosting. From the analysis of the most cited references in the articles, the following studies were identified on a theoretical basis: [27], [30]–[47]. These documents are cited in most primary studies and provide important insights for modeling CLRI and discussing the results. Lastly, the findings found in SM and SR, theoretical basis, concepts, categories, and synthesis tables constituted the *state of science* of the framework.

C. MODELING

1) KNOWLEDGE-BASED MODELING (KM) OF CLR

We built a model based on knowledge acquired in the state of science to predict a CLRI in the next month. The basic and aggregated attributes and their relationships (KM-1 and KM-2 phases) were defined based on, but not necessarily equal to, the categories found in the keywords of the abstracts (SM-4), and elements of the synthesis of systematic review (SR-5). The tree structure of the model is shown in Fig. 10. The basic attributes, that is, the model input, are shown in green, and aggregated in gray. The output variable (incidence category) is indicated in red. We considered aggregate attributes as processes in pathogen–host–environment interactions. The processes can represent the relationship between two or more basic attributes, as well as two or more aggregated attributes or a combination of them. Ordinal scales were used for all attributes (KM-3).

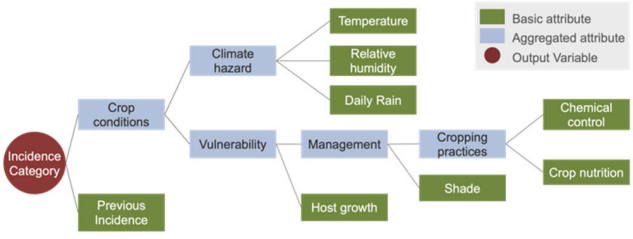

FIGURE 10. Tree-based representation of knowledge-based model for coffee leaf rust incidence.

The scales of the basic and aggregated attributes (KM-3) were *favorable to the disease* and *unfavorable for the disease*. The fruit load attribute has three ordered categories. Both the *previous incidence* basic attributes and the final output *incidence category* have a different scale (Table 3). The colors in scales represent whether the value of the scale is favorable to the disease (red), unfavorable to the disease (green), or a medium effect (black).

The disease incidence, a continuous variable ranging from 0 to 100%, was discretized into four categories according to the literature and expert knowledge:

TABLE 2. Synthesis of systematic review for incidence of coffee leaf rust forecasting.

Study	Year	TC	Target variable	Predictors	MT	Metric	BMV	Highlights
[48]	2020	1	The onset of coffee leaf rust symptoms and signs	Microclimatic variables in time windows, fruit load, lesion data, sporulation	Statistical analysis	RMSE	0.012	CLRI monitoring data is an important predictor. The analysis of weather variables in different time windows improve the modeling.
[49]	2020	1	Rust life stages, inoculum	Host leaf renewal, fruit load, shade, fungicide	Structural equation modeling	p-value	$p < 0.0001$	Importance of host growth, disease monitoring and fungicide application as predictors. Antagonist effect of shade.
[51]	2019	0	Coffee Rust Level	Maximum and minimum temperature, rainfall, relative humidity, altitude	Rule-based expert system (classification)	Accuracy	66.67%	Use of expert knowledge and technical reports. Future works: consider flowering date and knowledge representation for reasoning.
[52]	2018	7	CLR infection rate from incidence data	Temperature, rainfall, rainy days, relative humidity, leaf wetness	Multiple linear regression	R squared	0.785	The Gompertz growth model was the best to describe CLR epidemics accurately. Monthly minimum air temperature and relative humidity were the main weather variables to estimate CLR apparent infection rate.
[53]	2018	1	Incidence value and expected growth	Weather variables (temperature, relative humidity, rainfall) during the day and in hours of leaf wetness. Coffee variety, crop age, shade, crop management	Ensemble method, decision tree	MAE, Precision	MAE 1.2. Precision 92.2%	The modeling based on ensemble methods gets better performance. Considering expert knowledge to generate the features of datasets improves the modeling task.
[54]	2020	8	Incidence	Temperature, rainfall, rainy days, relative humidity, leaf wetness	Multiple linear regression, k-neighbors regression, random forest, neural networks	RMSE and R squared	RMSE 5.87, R squared 0.866	Training of various machine learning algorithms. Variables analyzed in three periods of time. Models calibrated for low and high yields years.
[50]	2008	12	Infection rate from incidence data	Temperature, rainfall, relative humidity, leaf wetness, temperature in leaf wetness condition.	Decision Trees	Accuracy	88%	The weather variables were characterized according to incubation and infection periods. Temperature in conditions of leaf wetness is the most important predictor.

TC: times cited. MT: modeling technique. BMV: best metric value

TABLE 3. Basic attributes scale (KM-3) for coffee leaf rust incidence.

Basic attribute	Scales	Values
Temperature [43],[35], [42]	Favorable	Between 21°C and 25°C
	Unfavorable	Other values
Relative humidity [44]	Favorable	>= 95%
	Unfavorable	Other values
Daily rain [45],[43], [48], [46]	Favorable	Between 1mm and 15 mm
	Unfavorable	Other values
Chemical control [42], [49]	Favorable	Medium or Low
	Unfavorable	High
Crop nutrition [31]	Favorable	Not adequate or null
	Unfavorable	Adequate
Shade [32], [39], [45]	Favorable	Shaded crop
	Unfavorable	Full sun exposure
Host growth [27], [33], [49]	Favorable	Decrease
	Unfavorable	Growth
Fruit load [47] [31] [50]	Very favorable	High fruit load
	Favorable	Medium fruit load
	Unfavorable	Low fruit load
Previous Incidence [27], [33], [49], [52]	>50	CLRI greater than 50%
	25-50	25 to 50% of CLRI
	5-25	5 to 25% of CLRI
	0-5	0 to 5% of CLRI

- According to the recommendations for preventive application of fungicides from 5% incidence [42], we used this value to define the two lowest categories of incidence:
- According to expert knowledge, a peak of 50% of incidence is a value that already represents great negative impacts on crops (next year's maximum production around 14% due to branch death); therefore, we used this value to define the two highest categories [55].

As a result, *previous incidence* and the final output *incidence category* were classified into four categories: *0-5 (0%–5% of CLRI), 5-25 (5%–25% of CLRI), 25-50 (25%–50% of CLRI), and >50 (CLRI greater than 50%).*

We applied expert knowledge to compile the heterogeneous information found in the literature to determine the ranges used in the model. Although the presented KM model considers the fruit load, this attribute is not available in the dataset used. For illustration purposes of the application of all the *FramePest* phases, the KM model did not include the fruit load attribute. The rules represented in aggregation tables (KM-4) were built considering an equal weight for all basic attributes. All aggregation tables derived from KM-4 are presented in Appendix A.

We validated the model (KM-5) to obtain an accuracy of 56.03% and a Cohen's weighted kappa of 0.31, which can be interpreted as a fair strength of agreement [56] between the model predictions and the observed data. The data from the CATIE experiment and the meteorological

station located next to it were used to validate the model. We sought to improve accuracy from the complementarity of the model (CM) process.

2) DATA-BASED MODELING (DM) OF CLR

The detailed process of this modeling, as well as the analysis of the results, was published in [57]. Next, we present a summary of the results, framing them in the DM phases of *FramePests*.

We took the *state of science* obtained in the SM and SR macroprocesses to carry out business understanding (DM-1). The business objective was to generate a CLRI prediction model on a monthly scale from data on crop properties (shade, management, vegetative growth) and climatic variables available in the dataset characterized in time windows. The data mining objective was to process a dataset, select the features with the most significant impact on a target variable, generate a regression model, and analyze each feature's impact on model predictions.

Data understanding (DM-2) and preparation (DM-3) start by collecting the datasets of the experiment of coffee-based agroforestry and meteorological stations reported in the CATIE experiment. We used the same data as in the KM validation. The data in the files did not contain null data. The thermal amplitude (*tAmp*), which represents the difference between the maximum and minimum temperatures, and the characterization of each day as a rainy day (precipitation greater or equal to 1 mm) (*rDay*), were calculated and added to the dataset. We relied on weather 14 days before *DP*, similar to KM. We used the concept of time windows [29], [57] to generate consecutive subperiods of each climatic variable within the main period of 14 days before *DP*.

The target variable was *pCLRI*, and the predictors were the rest of the experimental variables (*cCLRI, shade, host growth (hGrowth), and management (mgmt)*, as well as climatic variables characterized in time windows of various lengths of consecutive days: 3D, 4D, 7D, and 14D. Each subset had 439 instances, and the dimension depended on window size: 14D had 13 variables (8 related to climate), 7D had 69 features (64 related to climate), 4D had 93 features (88 related to climate), and 3D had 101 features (96 related to climate). The datasets were downloaded from [58].

Because a high-dimensionality dataset can generate multicollinearity problems, and a large number of variables may be irrelevant or generate noise in the modeling process, we applied feature selection methods [59] to obtain new reduced subsets. Five algorithms were used for the modeling (DM-4) applied to the reduced subsets: XGBoost, random forest regressor, support vector regression, sequential (neural network), and decision tree regressor, using Scikit-learn for Python. Each algorithm was configured multiple times from a pipeline with a randomized search to obtain the best hyperparameter configuration.

In the evaluation process (DM-5), we obtained the mean absolute error (MAE) for the model training process using cross-validation. The best result was obtained (MAE 7.19% and Bias 0.025%) in the model built with the subset of four consecutive days window reduced by the embedded method, which uses the feature selection as part of the training process of the learning algorithm, and *XGBoost* as the learning algorithm. As some windows for a climatic variable can have days in common, we filtered the highly correlated variables in the reduced subset. The final variables were *rDay14-11, pre11-8, tMax9-6, pre6-3, tMin4-1, hGrowth, cCLRI, shade*, and *mgmt*. Appendix B relates the variables of the KM and DM models.

The deployment (DM-6) was addressed as a functional prototype for the Central American Program for the Comprehensive Management of Coffee Rust (PROCAGICA), which is available as a web application.[2] The description of the prototype can be found in Appendix B. Additionally, the deployment was also addressed as an analysis of the impact of each feature on the model output (predicted incidence), through SHAP (SHapley Additive exPlanations) values [60]. The SHAP values allowed the expert in plant health to interpret the values of each feature that had the greatest impact on the predicted incidence according to a base value (average value of the model's predictions in its training). Figure 12 shows some examples of SHAP values representing the conditions in the features so that the predicted value differs from the base value (increases or decreases).

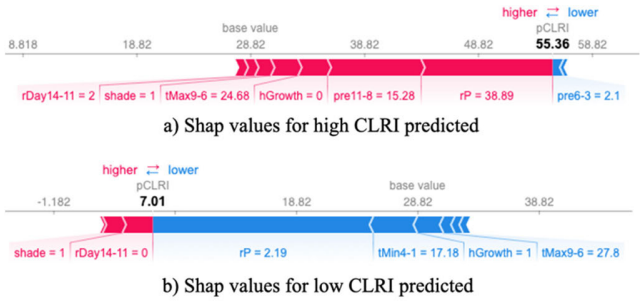

FIGURE 11. Examples of SHAP values for some predictions made by the model.

The features that cause an increase in the value of the predicted CLRI (*pCLRI*) are shown in red, and those that cause a decrease in pCLRI are shown in blue. The size of the segments of each feature represents the magnitude of its effect over the prediction, and the value of features with low importance for each specific prediction is not shown. The variables are maximum (tMax) and minimum (tMin) air temperature, daily precipitation (pre), rainy day (precipitation greater or equal to 1 mm) (rDay), and host growth (hGrowth). The index of the variables indicates the days covered by the window.

The DM model was derived from variables available in the experimental dataset. For this reason, this model is flexible due to the lack of information on fruit load, as the model seeks its best fit from the available data. The results of

[2]https://www.redpergamino.net/app-stadinc

DM allowed us to validate and compare findings from other studies, such as the importance of previous inoculum for future incidence [27], [49], the effect of wash-off given high rainfall [45], [48], favorable temperature for germination and penetration [43], and the effect of host growth [32], [39], [45].

3) COMPLEMENTARITY OF MODELS

We explored the complementarity between the two modeling approaches to improve the accuracy of the KM model. Because the variables of the obtained models are not equal, we carried out an additional DM process using a training dataset composed of KM variables (Table 3). In this case, the learning task was classified because the variable output used in the KM was categorical. We tested the following algorithms for classification: *XGBoost, decision tree, random forest, AdaBoost, and support vector classifiers*. The best accuracy and Cohen's weighted kappa values were obtained using XGBoost. The model allowed us to rank the importance of the variables, as shown in Table 4.

TABLE 4. Variable importance in a model trained with a dataset composed of the same variables of KM model.

Variable	Importance
Previous Incidence	0.5107
Host growth	0.1197
Daily rain	0.0924
Temperature	0.0853
Relative Humidity	0.0718
Shade	0.0713
Management (crop nutrition and chemical control)	0.0485

We modified the rules of the aggregation tables (KM-4) of the KM model to roughly represent the ranking of importance in Table 4 and carried out the validation process (DM-5) again. The model accuracy of the updated KM model was 63.1%, which is a 7.07% improvement over the first KM model built. The Cohen's weighted kappa obtained was 0.41, which can be interpreted as a moderate strength of agreement [56] between the model predictions and the observed data. The updated aggregation tables are presented in Appendix A.

Table 5 shows the distribution of the predicted and observed categories of incidence (CLRI) and the confusion matrix related to the results. The model tends to underestimate the CLRI category, that is, to predict lower categories than the original, for example, predict *the 5-25* category when *the 25-50* category was actually observed in the experimental data.

The highest precision for *the 5%–25%* category represents the ability to predict this category among all categories (Table 6). The low recall value for *the 0-5* category shows a higher proportion of false negatives for this category, most

TABLE 5. Confusion matrix between predicted and observed categories for knowledge-based CLRI model.

Predicted category of CLRI	Real category of CLRI			
	0-5	5-25	25-50	>50
0-5	4	14	2	0
5-25	2	131	70	2
25-50	0	34	110	12
>50	0	0	26	32

TABLE 6. Precision, recall and F1-score for each category of CLRI for the knowledge-based model.

Category	Precision	Recall	F1-score
0-5	0.67	0.20	0.31
5-25	0.73	0.64	0.68
25-50	0.53	0.71	0.60
>50	0.70	0.55	0.62

of which occur for predicting instances in *the 0-5* category when these instances corresponded to *the 5-25* category. The F1-score shows that, except for *0-5*, all categories have a good balance between precision and recall (VI).

For our case study, we estimated that the DM model required at least 59 instances to reach the accuracy of the KM model (Figure 12). Given that the data contained information from four different plots (given the management and shade combinations), this means that at least one year of monitoring data must be obtained to achieve the same accuracy in DM than in KM. The dataset of the CATIE experiment was used to incrementally generate subsets until reaching the size of the entire dataset (439 instances). Each subset was used to train a model from *XGBoost* with previously found hyperparameter settings (DM-4 process). Additionally, the accuracy was calculated using the data from Block 2 as the test dataset.

Finally, we compared the two models directly, transforming the responses into terms of the other. Although both models address the incidence, the output variables in the two models differ from each other, being a number for one and a range for the other. First, the KM model output, CLRI, was transformed into a quantitative variable, taking the center of the category value ranges, for example, *for 5-25*, the center is 15. The predicted CLRI (*pCLRI*) was compared to the observed incidences by calculating the MAE and bias. The results were MAE 11.29% and Bias −2.66 %, which show a lower performance than the DM model. Second, the DM model output was transformed into a qualitative variable corresponding to the categories used in KM. The accuracy obtained was 84.93%, which corresponds to a good value for the predictions. The precision, recall, and F1-score, as well as the summary of metrics obtained for each model and the associated transformation are presented in Appendix C. There was a significant difference between the predictions of the

two models. For qualitative models, after applying McNemar's test, the *p-value* obtained was 1.3×10^{-11}. For quantitative models, after applying the ANOVA test, the p-value obtained was 1.3×10^{-15}.

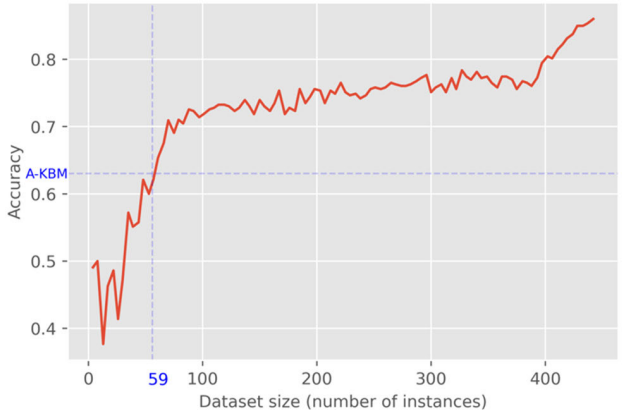

FIGURE 12. Accuracy according the training dataset size for data-based model. The blue marker corresponds to the accuracy of the knowledge-based model (A-KBM).

The KM model represents CLR mechanisms that occur in a general way in coffee crops, while the data-based model (DM) may be linked to the conditions present in the experiment site from where these data were monitored. We are aware that evaluations and comparative studies may be biased by the data from the case study. In this sense, the KM model could guide the process of generalization of the DM model, such as the approach carried out by Herr *et al.* [61]. The improvement of the KM model in the comparative study represents how a model that describes the general mechanisms of the disease can be adjusted to the characteristics of the study area.

IV. CONCLUSION

Our approach presents a comprehensive framework that guides a robust crop pest modeling process, from obtaining knowledge of the crop pest to be modeled, to the modeling alternatives according to the available resources necessary for modeling, such as data and knowledge. For example, a common problem is the amount of data with which the models are trained. If the data are insufficient, a modeling alternative that does not require data is required. Several approaches to crop pest modeling assume knowledge of the problem that is already present, without considering steps to obtain and refine it, and others carry out the modeling process empirically without following a methodology. Although this does not mean that the results are less reliable, the use of methodologies is recommended to achieve an orderly, reliable, and well-presented process.

The application of the proposed framework was demonstrated for coffee leaf rust modeling. All *FramePests* processes were applied to provide a better understanding of its use. This allowed the modeling tasks to be done with knowledge about pests and the investigations that have addressed

TABLE 7. Aggregation table for model output.

currentIncidence	CropConditions	Incidence category
>50	**Favorable**	**>50**
>50	>=Moderately favorable	25-50
<=25-50	Moderately favorable	25-50
25-50	<=Moderately favorable	25-50
25-50:5-25	**Favorable**	25-50
25-50:5-25	***Unfavorable***	5-25
5-25	>=Moderately favorable	5-25
>=5-25	Moderately favorable	5-25
0-5	<=Moderately favorable	5-25
0-5	***Unfavorable***	***0-5***

TABLE 8. Aggregation table for crop conditions.

climate.hazard	Vulnerability	CropConditions
<=Moderately favorable	<=Moderately favorable	**Favorable**
*	**Favorable**	**Favorable**
<=Moderately favorable	***Unfavorable***	Moderately favorable
Unfavorable	>=Moderately favorable	***Unfavorable***

it, which were acquired from formal processes that facilitate its assimilation. A possible improvement would be to obtain knowledge from experts who have studied the biological mechanisms of crop pests from the application of forms or interviews. For data-based modeling, the process suggested by the framework allowed us to obtain a model with a MAE of 7.19% for CLRI forecasting. For knowledge-based modeling, the multi-criteria and hierarchical structuring of the model made it possible to represent the pathogen × host–environment relationships that affect CLRI development, from associations that can be easily inspected and validated by experts. This model had an initial accuracy of 56%. Both models were validated using data from a real agroforestry experiment. The scientific bases of knowledge-based modeling were the same as those used by data-based modeling, allowing us to obtain a model that considers similar predictors. The study of the complementarity of the models allowed us to explore how elements of a data-based model can improve a knowledge-based model. From an estimate of the importance of the model variables in relation to the variable output obtained from the data, we were able to increase the accuracy of the KM model by 7.07%. For our case study, the results show that knowledge-based modeling can be an alternative to generate a prediction model when the available dataset has approximately 59 instances.

We are aware that modeling tasks can become very complex for a group of human talent without experience in it, so the framework is structured in such a way that its steps are easily followed. The results obtained when applying the framework in a case study show that it can become a valuable tool for different institutions and research groups that wish to start a crop pest modeling process. In addition, the execution of the proposed framework can be included in integrated pest management plans [5].

TABLE 9. Aggregation table for climate hazard.

Daily Rain	Temperature	RHumidity	climate.hazard
Favorable	Favorable	Favorable	Favorable
Favorable	*	*Unfavorable*	Moderately favorable
*	Favorable	*Unfavorable*	Moderately favorable
Favorable	*Unfavorable*	*	Moderately favorable
*	*Unfavorable*	Favorable	Moderately favorable
Unfavorable	Favorable	*	Moderately favorable
Unfavorable	*	Favorable	Moderately favorable
Unfavorable	*Unfavorable*	*Unfavorable*	*Unfavorable*

TABLE 10. Aggregation table for vulnerability.

CropPract	hostGrowth	Vulnerability
Favorable	Decrease	Favorable
Favorable	*Growth*	Moderately favorable
Moderately favorable	Decrease	Moderately favorable
>=Moderately favorable	*Growth*	*Unfavorable*
Unfavorable	*	*Unfavorable*

TABLE 11. Aggregation table for crop practices.

Management	Shade	CropPract
Favorable	Shaded	Favorable
Favorable	*Full sun*	Moderately favorable
Unfavorable	Shaded	Moderately favorable
Unfavorable	*Full sun*	*Unfavorable*

TABLE 12. Aggregation table for management.

ChemicalC	Nutrition	Management
Medium or low	*	Favorable
*	Not adequate or null	Favorable
High	*Adequate*	*Unfavorable*

TABLE 13. Aggregation table for model output.

currentIncidence	CropConditions	Incidence category
>50	Favorable	>50
>50	>=Moderately favorable	25-50
<=25-50	Moderately favorable	25-50
25-50	<=Moderately favorable	25-50
25-50:5-25	Favorable	25-50
25-50:5-25	*Unfavorable*	5-25
5-25	>=Moderately favorable	5-25
>=5-25	Moderately favorable	5-25
0-5	<=Moderately favorable	5-25
0-5	*Unfavorable*	*0-5*

APPENDIX A
KNOWLEDGE-BASED MODEL OF CLR

A. AGGREGATION TABLES FOR THE FIRST KM MODEL

IPSIM-based modeling is addressed through a software called Dexi[3] for multi-attribute decision making. Below are the aggregation tables (Table 7 to 12) that describe the relationships between the base and aggregated attributes for CLR model. The colors in scales represent whether the value of the scale is favorable to the disease (red), unfavorable to the disease (green), or a medium effect (black). The symbol * indicates that the value of the attribute does not influence the rule, and the logical operators "<" means less than, ">" means greater than, "=" equals to, and ":" indicates a range of values.

[3]https://kt.ijs.si/MarkoBohanec/dexi.html

TABLE 14. Aggregation table for crop conditions.

climate.hazard	Vulnerability	CropConditions
Favorable	<=Moderately favorable	Favorable
<=Moderately favorable	Favorable	Favorable
Favorable	*Unfavorable*	Moderately favorable
Moderately favorable	Moderately favorable	Moderately favorable
Unfavorable	Favorable	Moderately favorable
>=Moderately favorable	*Unfavorable*	*Unfavorable*
Unfavorable	>=Moderately favorable	*Unfavorable*

TABLE 15. Aggregation table for climate hazard.

Daily Rain	Temperature	RHumidity	climate.hazard
*	Favorable	Favorable	Favorable
*	Favorable	*Unfavorable*	Moderately favorable
Favorable	*Unfavorable*	*	Moderately favorable
Unfavorable	*Unfavorable*	*Unfavorable*	*Unfavorable*
*	*Unfavorable*	*Unfavorable*	*Unfavorable*

TABLE 16. Aggregation table for vulnerability.

CropPract	hostGrowth	Vulnerability
Favorable	Decrease	Favorable
Favorable	*Growth*	Moderately favorable
>=Moderately favorable	Decrease	Moderately favorable
>=Moderately favorable	*Growth*	*Unfavorable*

TABLE 17. Aggregation table for crop practices.

Management	Shade	CropPract
Favorable	Shaded	Favorable
Favorable	*Full sun*	Moderately favorable
Unfavorable	Shaded	Moderately favorable
Unfavorable	*Full sun*	*Unfavorable*

TABLE 18. Aggregation table for management.

ChemicalC	Nutrition	Management
Medium or low	*	Favorable
*	Not adequate or null	Favorable
High	*Adequate*	*Unfavorable*

B. AGGREGATION TABLES FOR THE UPDATED KM MODEL

The following aggreggation tables (Tables 13 to 18) correspond to the updated model presented in the Complementary of models (CM) process.

TABLE 19. Comparison of variables between KM and DM models.

Variable	Knowledge-based model		Data-based model	
	Description (favorability to the disease)	Type	Description	Type and unit
Temperature	Favorable if it is between 21°C and 25°C, unfavorable for other values	Categorical	Average maximum temperature between days 9 and 6 before DP	Numerical (°C)
			Average minimum temperature between days 4 and 1 before DP	Numerical (°C)
Relative humidity	Favorable if it is greater than or equal to>= 95%, unfavorable for other values	Categorical	-	
Precipitation	Favorable if it is between 1mm and 15 mm per day, unfavorable for other values	Categorical	Accumulated rainy days between days 14 and 11 before DP	Numerical (mm)
			Average daily precipitation between days 11 and 8 before DP	Numerical (mm)
			Average daily precipitation between days 6 and 3 before DP	Numerical (mm)
Chemical control	Favorable if it is medium or low, unfavorable for other values	Categorical	Type of management (1 if it was highly conventional, 0 if it was medium conventional)	Binary
Crop nutrition	Favorable if it is not adequate or null, unfavorable for other values	Categorical		
Shade	Favorable if the crop is under shade, unfavorable for other values	Categorical	Type of shade (1 if the crop was under the dense shade, 0 if it was in full sun)	Binary
Host growth	Favorable if the host is in a growing stage, unfavorable for other values	Categorical	Increase of leaves in the host (1 if the number of leaves increases between 14 days before the day of prediction (DP) and DP, and 0 otherwise)	Binary
Previous incidence	According to categories defined in KM	Categorical	The CLRI measured in the day of prediction (DP)	Numerical (%)

APPENDIX B
CLRI DATA-BASED MODEL

A. COMPARISON BETWEEN VARIABLES OF KM AND DM MODELS

Table 19 shows the variables of the KM and DM models. In this table you can see how the variables of climate, properties of crops and disease were characterized.

B. DEPLOYMENT OF CLRI DATA-BASED MODEL

The Deployment of CLRI model obtained in DM was addressed as a functional prototype for PROCAGICA (available at https://www.redpergamino.net/app-stadinc). The module that allows the model to be used is called STADINC (Statistical Development of Incidence prediction), available in the *Tools* section of the PERGAMINO platform.

PROCAGICA is the Central American Program for Comprehensive Management of Coffee Rust, whose objective is: Increase the capacity of the region to design and implement policies, programs and measures for a better adaptation, response capacity and resilience of the most vulnerable population, living in the coffee production areas of Central America and the Dominican Republic, and that it is exposed to the adverse effects of climate change and variability.

1) SYSTEM FUNCTIONALITIES

The objective of STADINC is to provide a tool to obtain a CLRI prediction 28 days after the consultation date. The system is presented in Figure 13 and is composed of the following modules.

- **Data from climate model retrieval**: Reusable module offered by the PERGAMINO platform, which allows obtaining the maximum and minimum temperature and precipitation data for the coffee areas covered by PROCAGICA using a climate model.

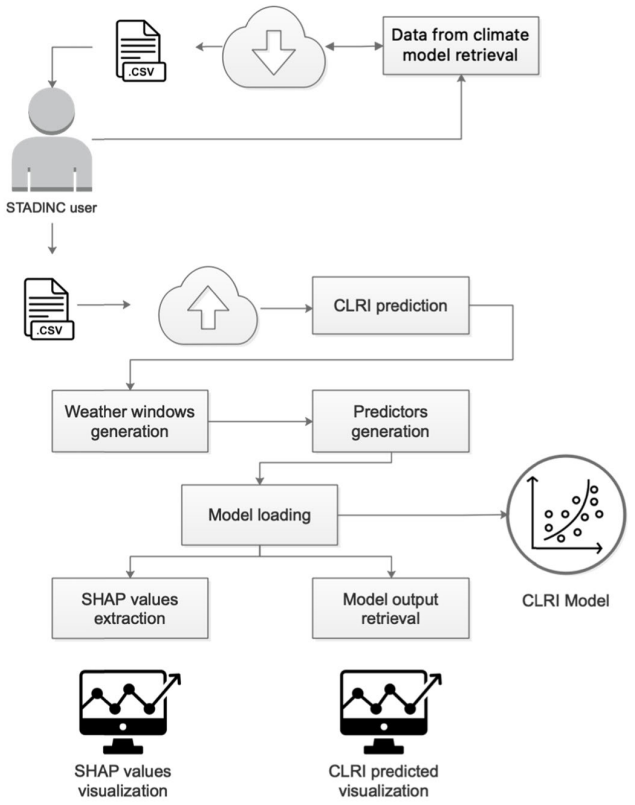

FIGURE 13. STADINC modules.

- **CLRI prediction:** Module that allows setting the predictor values associated with crop and CLR properties, as well as loading the CSV file with the weather data to be used.
- **Weather windows generation**: To avoid the user having to generate the weather windows that the model requires, this module is in charge of calculating them from daily values of temperature and precipitation.
- **Predictors generation:** Construction of the instance required by the model composed of its predictors.
- **Model loading:** Deserialization of the model stored on the server.
- **SHAP values extraction:** Calculation of the impact of each predictor on the output of the model.
- **Model output retrieval:** Extraction of the CLRI value predicted by the model.

2) SYSTEM FUNCTIONALITIES

The STADINC architecture represented by the logical view shown in the Figure 14. This view organizes the software classes into packages and three layers: Application, Mediation and Foundation.

a: APPLICATION LAYER

Provides the functionalities to a STADINC user. It is composed by the following package:

- **Graphical user interface:** contains the software classes and forms to provide a visual representation for data submission and response deployment. This allows user interaction with STADINC, with graphic elements such as plots, icons, text boxes, among others. The graphical user interface was developed in the R package Shiny.[4]

TABLE 20. Precision, recall and F1-score for each class of CLRI for transformed output of data-based model.

Class	Precision	Recall	F1-score
0-5	1.00	0.60	0.75
5-25	0.86	0.87	0.87
25-50	0.81	0.83	0.82
>50	0.88	0.91	0.90

b: MEDIATION LAYER

Contains the software classes named controllers. In our case, its structure corresponds to the one suggested for creating R-Shiny apps. It is composed by the following packages:

- **Global**: It contains the methods of loading the model and obtaining its output, as well as the SHAP values of the predictors for a specific prediction. This allows the implementation of functions in Server package.
- **Server**: Implements the mechanism for information, prediction and SHAP values retrieval. This class controller gets the user input and processes it to validate the input data and generate the response elements in the graphical interface.

c: FOUNDATION LAYER

This layer is composed by the software used in the STADINC:

- **R Engine**[5]: programming language and environment for statistical computing. The PERGAMINO platform is based on this language. We used R 3.6 and different from its core functions are used for data manipulation.
- **Shiny**[6]: R package that allows the creation of interactive web apps encoded in R.
- **R Shiny Server**[7]: Web server for Shiny applications that provides its hosting and access through the internet. It allows host an app in a controlled environment.
- **XGBoost**[8]: Gradient boosting library that implements machine learning algorithms based on the gradient boosting framework. Since the model based on CLRI data was generated with this technique, this library allows to load said model and make predictions.
- **SHAPforxgboost**[9]: Library that implements the calculation of SHAP values specifically for models built from XGBoost.

[4]https://shiny.rstudio.com
[5]https://www.r-project.org
[6]https://shiny.rstudio.com
[7]https://rstudio.com/products/shiny/shiny-server/
[8]https://xgboost.readthedocs.io/en/latest/
[9] https://cran.r-project.org/web/packages/SHAPforxgboost/index.html

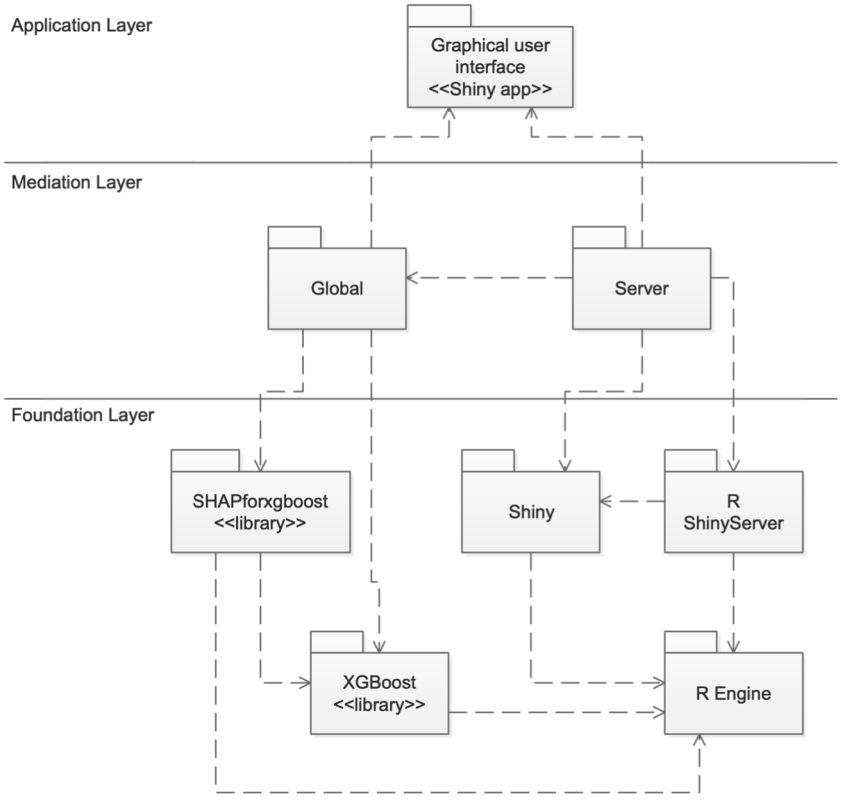

FIGURE 14. Logical view of STADINC.

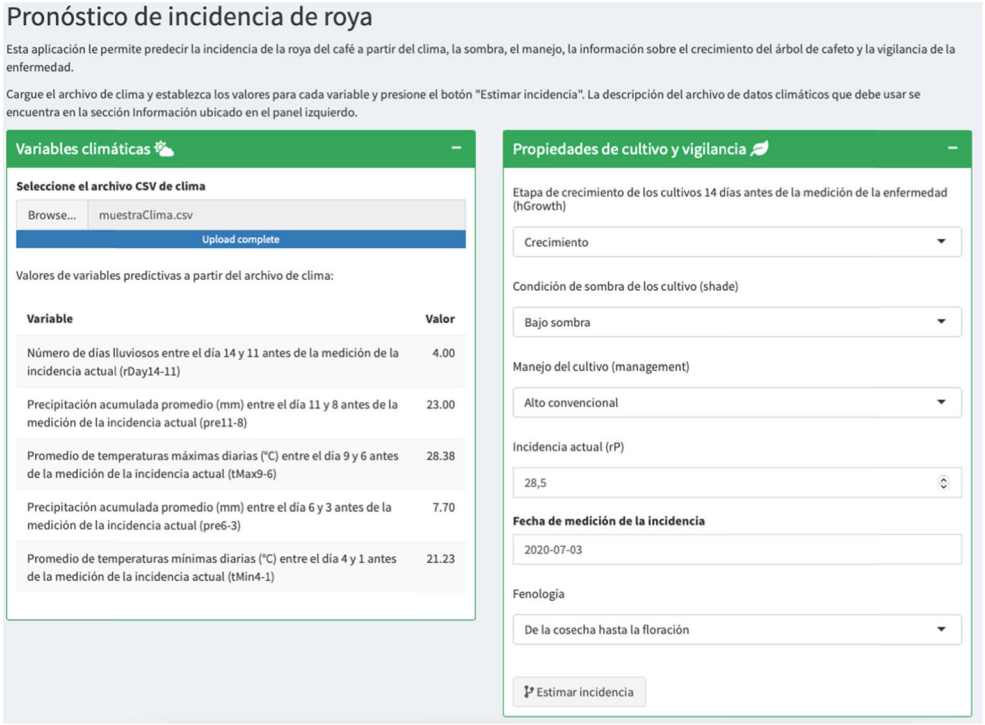

FIGURE 15. STADINC data entry forms.

3) USER INTERFACES

The main interface for the use of STADINC is composed of a form that allows to load the CSV data file for the calculation of the predictors related to weather and another one for the user to enter the data of the crop properties and the previous incidence (Figure 15). After the data submission, the response

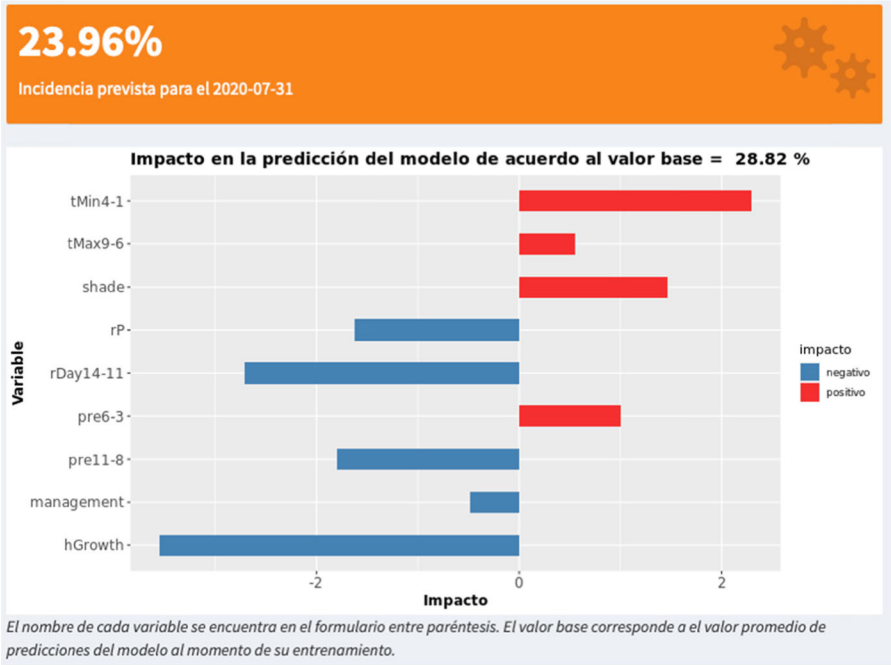

FIGURE 16. CLRI prediction visualization and impact of model variables in STADINC.

TABLE 21. Comparison of models for CLRI.

Metric		Knowledge-based model	Data-based model
Quantitative output variable	MAE (%)	10.93	7.19
	Bias (%)	2.9	0.03
Qualitative output variable	Accuracy (%)	64.45	84.93
	F1-score	0.57	0.83

of the model and the impacts of the variables are shown as presented in the Figure 16.

APPENDIX C
COMPLEMENTARITY OF MODELS

Results: First, the KM model output was transformed into a quantitative variable (numerical), taking the center of the class value range, e.g., *for 5-25*, the center is 15. This value was used together with the real incidences to calculate the MAE and Bias. The results were MAE 10.93% and Bias 2.9%. On the other hand, the DM model output was transformed into a qualitative variable (category). The incidence percentage predicted by the model (numerical), for which a CLRI class was assigned following the same ranges of values for the data-based model, e.g., a predicted 21.3% of CLRI corresponds to the *5-25* class. The accuracy obtained was 84.93%. The precision, recall, and F1-score are shown in Table 20.

The summary of metrics obtained for each model and the associated transformation is shown in Table 21.

ACKNOWLEDGMENT

The authors are grateful to Dr. Elias de Melo Virginio Filho for supplying the experimental data.

REFERENCES

[1] *The Future of Food and agriculture—Trends and Challenges*, FAO, Rome, Italy, 2017.
[2] M. Kogan, "Integrated pest management: Historical perspectives and contemporary developments," *Annu. Rev. Entomol.*, vol. 43, no. 1, pp. 243–270, Jan. 1998.
[3] N. Hardwick, "Disease forecasting," in *The Epidemiology of Plant Diseases*, B. Cooke, D. Jones, and B. Kaye, Eds. Dordrecht, The Netherlands: Springer, 1998, doi: 10.1007/1-4020-4581-6_9.
[4] Y. Prasad and M. Prabhakar, "Pest monitoring and forecasting," in *Integrated Pest Management: Principles and Practice*. Oxfordshire, U.K.: Cabi, 2012, pp. 41–57.
[5] J. A. Stenberg, "A conceptual framework for integrated pest management," *Trends Plant Sci.*, vol. 22, no. 9, pp. 759–769, Sep. 2017.
[6] Y. Jabareen, "Building a conceptual framework: Philosophy, definitions, and procedure," *Int. J. Qualitative Methods*, vol. 8, no. 4, pp. 49–62, Dec. 2009.
[7] M. Robert, J. Dury, A. Thomas, O. Therond, M. Sekhar, S. Badiger, L. Ruiz, and J.-E. Bergez, "CMFDM: A methodology to guide the design of a conceptual model of farmers' decision-making processes," *Agricult. Syst.*, vol. 148, pp. 86–94, Oct. 2016.
[8] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. 12th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, Jun. 2008, pp. 1–10.
[9] B. Kitchenham, "Procedures for performing systematic reviews," Dept. Comput. Sci., Keele Univ., Keele, U.K., Tech. Rep. TR/SE-0401, 2004, pp. 1–26, vol. 33.
[10] J.-N. Aubertot and M.-H. Robin, "Injury profile SIMulator, a qualitative aggregative modelling framework to predict crop injury profile as a function of cropping practices, and the abiotic and biotic environment. I. Conceptual bases," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e73202.

[11] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. (2000). *CRISP-DM 1.0: Step-by-Step Data Mining Guide*. [Online]. Available: http://www.citeulike.org/group/1598/article/1025172

[12] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming—A review," *Agricult. Syst.*, vol. 153, pp. 69–80, May 2017.

[13] D. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *ArXiv*, vol. abs/2010.16061, 2020.

[14] J. L. Fleiss and J. Cohen, "The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, vol. 33, no. 3, pp. 613–619, 1973.

[15] D. Corrales, J. Corrales, and A. Ledezma, "How to address the data quality issues in regression models: A guided process for data cleaning," *Symmetry*, vol. 10, no. 4, p. 99, Apr. 2018.

[16] D. Corrales, A. Ledezma, and J. Corrales, "From theory to practice: A data quality framework for classification tasks," *Symmetry*, vol. 10, no. 7, p. 248, Jul. 2018.

[17] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 3, no. 1, pp. 1–130, 2009.

[18] M. W. Browne, "Cross-validation methods," *J. Math. Psychol.*, vol. 44, no. 1, pp. 108–132, 2000.

[19] D. C. C. Muñoz, J. C. C. Muñoz, and A. Figueroa, "Towards detecting crop diseases and pest by supervised learning," *Rev. Ing. Univ.*, vol. 19, no. 1, pp. 207–228, 2015.

[20] Y. Roggo, L. Duponchel, C. Ruckebusch, and J.-P. Huvenne, "Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data," *J. Mol. Struct.*, vol. 654, nos. 1–3, pp. 253–262, Jun. 2003.

[21] D. Naidu and A. Patel, "A comparison of qualitative and quantitative methods of detecting earnings management: Evidence from two Fijian private and two Fijian state-owned entities," *Australas. Accounting, Bus. Finance J.*, vol. 7, no. 1, pp. 79–98, 2013.

[22] B. G. J. S. Sonneveld, M. A. Keyzer, and L. Stroosnijder, "Evaluating quantitative and qualitative models: An application for nationwide water erosion assessment in Ethiopia," *Environ. Model. Softw.*, vol. 26, no. 10, pp. 1161–1170, Oct. 2011.

[23] B. S. Everitt, *The Analysis of Contingency Tables*. Boca Raton, FL, USA: CRC Press, 1992.

[24] J. Haggar, M. Barrios, M. Bolaños, M. Merlo, P. Moraga, R. Munguia, A. Ponce, S. Romero, G. Soto, C. Staver, and E. D. M. F. Virginio, "Coffee agroecosystem performance under full sun, shade, conventional and organic management regimes in Central America," *Agroforestry Syst.*, vol. 82, no. 3, pp. 285–301, Jul. 2011.

[25] E. Rossi, F. Montagnini, and E. D. M. V. Filho, "Effects of management practices on coffee productivity and herbaceous species diversity in agroforestry systems in Costa Rica," in *Agroforestry as a Tool for Landscape Restoration*. New York, NY, USA: Nova Science Publishers, 2011, pp. 115–132.

[26] L. V. Madden, G. Hughes, and F. Van Den Bosch, *The Study of Plant Disease Epidemics*. Saint Paul, MN, USA: APS Journal, 2007.

[27] A. C. Kushalappa, M. Akutsu, and A. Ludwig, "Application of survival ratio for monocyclic process of *Hemileia vastatrix* in predicting coffee rust infection rates," *Phytopathology*, vol. 73, no. 1, pp. 96–103, 1983.

[28] M. Aria and C. Cuccurullo, "Bibliometrix : An R-tool for comprehensive science mapping analysis," *J. Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017.

[29] S. M. Coakley, R. F. Line, and L. R. McDaniel, "Predicting stripe rust severity on winter wheat using an improved method for analyzing meteorological and rust data," *Phytopathology*, vol. 78, no. 5, pp. 543–550, 1988.

[30] J. Avelino, M. Cristancho, S. Georgiou, P. Imbach, L. Aguilar, G. Bornemann, P. Läderach, F. Anzueto, A. J. Hruska, and C. Morales, "The coffee rust crises in Colombia and Central America (2008–2013): Impacts, plausible causes and proposed solutions," *Food Secur.*, vol. 7, no. 2, pp. 303–321, Apr. 2015, doi: 10.1007/s12571-015-0446-9.

[31] J. Avelino, H. Zelaya, A. Merlo, A. Pineda, M. Ordoñez, and S. Savary, "The intensity of a coffee rust epidemic is dependent on production situations," *Ecol. Model.*, vol. 197, nos. 3–4, pp. 431–447, Aug. 2006.

[32] D. F. López-Bravo, E. D. M. Virginio-Filho, and J. Avelino, "Shade is conducive to coffee rust as compared to full sun exposure under standardized fruit load conditions," *Crop Protection*, vol. 38, pp. 21–29, Aug. 2012.

[33] J. Avelino, L. Willocquet, and S. Savary, "Effects of crop management patterns on coffee rust epidemics," *Plant Pathol.*, vol. 53, no. 5, pp. 541–547, Oct. 2004.

[34] L. Zambolim, "Current status and management of coffee leaf rust in Brazil," *Tropical Plant Pathol.*, vol. 41, no. 1, pp. 1–8, Feb. 2016.

[35] A. C. Kushalappa and A. B. Eskes, "Advances in coffee rust research," *Annu. Rev. Phytopathol.*, vol. 27, no. 1, pp. 503–531, Sep. 1989.

[36] D. Cressey, "Coffee rust regains foothold: Researchers marshal technology in bid to thwart fungal outbreak in Central America," *Nature*, vol. 493, no. 7434, pp. 587–588, 2013.

[37] S. McCook, "Global rust belt: *Hemileia vastatrix* and the ecological integration of world coffee production since 1850," *J. Global Hist.*, vol. 1, no. 2, pp. 177–195, Jul. 2006.

[38] R. W. Rayner, "Germination and penetration studies on coffee rust *(Hemileia vastatrix B. & Br.)*," *Ann. Appl. Biol.*, vol. 49, no. 3, pp. 497–505, Oct. 1961.

[39] A. Boudrot, J. Pico, I. Merle, E. Granados, S. Vílchez, P. Tixier, E. de Melo Virginio Filho, F. Casanoves, A. Tapia, C. Allinne, R. A. Rice, and J. Avelino, "Shade effects on the dispersal of airborne *Hemileia vastatrix* uredospores," *Phytopathology*, vol. 106, no. 6, pp. 572–580, 2016.

[40] J. Vandermeer, D. Jackson, and I. Perfecto, "Qualitative dynamics of the coffee rust epidemic: Educating intuition with theoretical ecology," *BioScience*, vol. 64, no. 3, pp. 210–218, Mar. 2014.

[41] J. M. Waller, "Coffee rust—Epidemiology and control," *Crop Protection*, vol. 1, no. 4, pp. 385–404, 1982.

[42] C. Rivillas, C. Serna, M. Cristancho, and A. Gaitán, "Roya del cafeto en Colombia: Impacto, manejo y costos del control, Chinchiná bol," *Téc*, vol. 82, no. 36, pp. 285–301, 2011.

[43] F. J. Nutman, F. M. Roberts, and R. T. Clarke, "Studies on the biology of *Hemileia vastatrix* Berk. & Br," *Trans. Brit. Mycol. Soc.*, vol. 46, no. 1, pp. 27–44, Mar. 1963.

[44] J. C. Sutton, T. J. Gillespie, and P. D. Hildebrand, "Monitoring weather factors in relation to plant disease," *Plant Disease*, vol. 68, no. 1, pp. 78–84, 1984. Accessed: May 20, 2015. [Online]. Available: http://agris.fao.org/agris-search/search.do?recordID=US19850001594

[45] J. Avelino, S. Vílchez, M. B. Segura-Escobar, M. A. Brenes-Loaiza, E. M. de Virginio Filho, and F. Casanoves, "Shade tree *Chloroleucon eurycyclum* promotes coffee leaf rust by reducing uredospore wash-off by rain," *Crop Protection*, vol. 129, Mar. 2020, Art. no. 105038, doi: 10.1016/j.cropro.2019.105038.

[46] J. M. Waller, M. Bigger, and R. J. Hillocks, *Coffee Pests, Diseases and Their Management*. Wallingford, U.K.: CABI, 2007. Accessed: May 20, 2015. [Online]. Available: https://books.google.com/books?hl=es&lr=&id=qm54fhoV1U4C&oi=fnd&pg=PR5&dq=Coffee+pests,+diseases+and+their+management.&ots=weOQn4miqr&sig=6sIdmdp3am_x7X7Heis9f26IxGc

[47] A. C. Kushalappa, M. Akutsu, S. H. Oseguera, G. M. Chaves, C. A. Melles, J. M. Miranda, and G. F. Bartolo, "Equations for predicting the rate of coffee rust development based on net survival ratio for monocyclic process of *Hemileia vastatrix [Coffea arabica]*," *Fitopatologia Brasileira*, vol. 9, no. 2, pp. 255–271, 1985. Accessed: May 20, 2015. [Online]. Available: http://agris.fao.org/agris-search/search.do?recordID=BR8500210

[48] I. Merle, P. Tixier, E. D. M. V. Filho, C. Cilas, and J. Avelino, "Forecast models of coffee leaf rust symptoms and signs based on identified microclimatic combinations in coffee-based agroforestry systems in Costa Rica," *Crop Protection*, vol. 130, Apr. 2020, Art. no. 105046.

[49] I. Merle, J. Pico, E. Granados, A. Boudrot, P. Tixier, E. D. M. V. Filho, C. Cilas, and J. Avelino, "Unraveling the complexity of coffee leaf rust behavior and development in different *Coffea arabica* agroecosystems," *Phytopathology*, vol. 110, no. 2, pp. 418–427, Feb. 2020.

[50] C. A. A. Meira, L. H. A. Rodrigues, and S. A. Moraes, "Análise da epidemia da ferrugem do cafeeiro com árvore de decisão," *Tropical Plant Pathol.*, vol. 33, no. 2, pp. 114–124, Apr. 2008, doi: 10.1590/S1982-56762008000200005.

[51] E. J. G. Buitrón, D. C. Corrales, J. Avelino, J. A. Iglesias, and J. C. Corrales, "Rule-based expert system for detection of coffee rust warnings in Colombian crops," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4765–4775, May 2019.

[52] F. D. Hinnah, P. C. Sentelhas, C. A. A. Meira, and R. N. Paiva, "Weather-based coffee leaf rust apparent infection rate modeling," *Int. J. Biometeorol.*, vol. 62, no. 10, pp. 1847–1860, Oct. 2018.

[53] D. C. Corrales, E. Lasso, A. F. Casas, A. Ledezma, and J. C. Corrales, "Estimation of coffee rust infection and growth through two-level classifier ensembles based on expert knowledge," *Int. J. Bus. Intell. Data Mining*, vol. 13, no. 4, pp. 369–387, 2018.

[54] L. E. de Oliveira Aparecido, G. de Souza Rolim, J. R. da Silva Cabral De Moraes, C. T. S. Costa, and P. S. de Souza, "Machine learning algorithms for forecasting the incidence of *Coffea arabica* pests and diseases," *Int. J. Biometeorol.*, vol. 64, no. 4, pp. 671–688, Apr. 2020, doi: 10.1007/s00484-019-01856-1.
[55] S. M. Chalfoun, C. M. de Silva, A. A. Pereira, and F. A. de Paiva, "Relaçao entre diferentes índices de infecçao de ferrugem (*Hemileia vastatrix* Berk et Br.) e produçao de cafereiros (*Coffea arabica* L.) em algunas localidades de Minas Gerais," in *Proc. Congresso Brasileiro de Pesquisas Cafeeiras*, Ribeirao Preto, Brazil, Oct. 1978, pp. 392–394.
[56] D. G. Altman, *Practical Statistics for Medical Research*. Boca Raton, FL, USA: CRC Press, 1990.
[57] E. Lasso, D. C. Corrales, J. Avelino, E. de Melo Virginio Filho, and J. C. Corrales, "Discovering weather periods and crop properties favorable for coffee rust incidence from feature selection approaches," *Comput. Electron. Agricult.*, vol. 176, Sep. 2020, Art. no. 105640, doi: 10.1016/j.compag.2020.105640.
[58] E. Lasso, E. de Melo Virginio Filho, and J. C. Corrales, "Subsets according time windows for coffee leaf rust incidence modeling. Paper: Discovering weather periods and crop properties favorable for coffee rust incidence from feature selection approaches," Mendeley Data, V1, 2021, doi: 10.17632/wpy54dw6t7.1.
[59] D. C. Corrales, E. Lasso, A. Ledezma, and J. C. Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?" *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, pp. 2825–2835, May 2018.
[60] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
[61] A. Herr, J. M. Dambacher, E. Pinkard, M. Glen, C. Mohammed, and T. Wardlaw, "The uncertain impact of climate change on forest ecosystems—How qualitative modelling can guide future research for quantitative model development," *Environ. Model. Softw.*, vol. 76, pp. 95–107, Feb. 2016, doi: 10.1016/j.envsoft.2015.10.023.

EMMANUEL LASSO received the degree in electronics and telecommunications engineering, the master's degree in telematics engineering, and the Ph.D. degree in telematics engineering from the University of Cauca, Popayán, Colombia, in 2013, 2016, and 2021, respectively. He is currently a Researcher of technology transfer at the University of Cauca and the Telematics Engineering Group (GIT). His current research interests include modeling solutions for smart farming, data science, and machine learning applications.

NATACHA MOTISI received the Ph.D. degree in epidemiology from Agrocampus Rennes, France. She is currently a Researcher with the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), France, and posted at the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Costa Rica. She has ten years of experience in epidemiology of coffee arabica diseases, and mechanistic modeling of the dynamics of the diseases at the tree, plot, and territorial scales. Her objective is to define the levers of action, particularly agro-ecological practices, making it possible to reduce the disease durably.

JACQUES AVELINO received the Ph.D. degree in plant pathology from the University of Paris—XI, Orsay France. His Ph.D. thesis research was carried out in Honduras, from 1994 to 1997, on the epidemiology of coffee leaf rust. He currently works with the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), France. He is posted at the Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Costa Rica. He has 35 years of experience in coffee pests and diseases, particularly coffee leaf rust. During his career, he established research and development operations in Mexico, Central America, the Dominican Republic, Venezuela, Indonesia, Laos, and Papua New Guinea.

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, specialty in computer science from the University of Versailles Saint-Quentin-en-Yvelines, France, in 2008. He is currently a Full Professor and leads the Telematics Engineering Group (GIT) at the University of Cauca. His research interests include machine learning and data analytics.