Smart Agricultural Technology 3 (2023) 100070

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.elsevier.com/locate/atech

DSSApple: A hybrid expert system for the diagnosis of post-harvest diseases of apple

Gabriele Sottocornola[a,*], Sanja Baric[a], Maximilian Nocker[a], Fabio Stella[b], Markus Zanker[a,c]

[a] Free University of Bozen-Bolzano, Bolzano, Italy
[b] University of Milano-Bicocca, Milano, Italy
[c] University of Klagenfurt, Klagenfurt, Austria

ARTICLE INFO

Keywords:
Decision support system in agriculture
Bayesian network
Knowledge-based diagnosis
Knowledge elicitation
Explanation

ABSTRACT

Post-harvest diseases of apple can cause considerable economic losses. Thus, we developed *DSSApple*, an interactive web-based decision support system, that helps users to diagnose post-harvest diseases of domesticated apple based on observed macroscopic symptoms on fruit. Specifically, *DSSApple* is designed as a two-stream hybrid diagnostic tool, that can be effectively used by both expert and non-expert users to diagnose diseased instances of apple. The *image-based* stream allows the user to interact simply by selecting pictures, representing the variety of symptoms of diseases at different stages of the infection and on different cultivars. Instead, the *expert-based* stream of the system incrementally collects user feedback about the target disease by asking questions related to the macroscopic characteristics of the observed symptoms on a target apple. The *expert-based* reasoning mechanism of *DSSApple* is developed by leveraging the framework of Bayesian Networks (BNs). We detail the process of building this knowledge base with the support of a domain expert. We further exploit the BN to process incomplete or conflicting user feedback within the inference mechanism as well as to provide human-understandable explanations on the suggested diagnoses. The proposed hybrid approach has been thoroughly evaluated in two studies, involving simulated (by photos) as well as real infected apples. Thus, the proposed hybrid version of *DSSApple* is able to outperform both the single streams and the user intuition in terms of diagnostic accuracy.

1. Introduction

The apple (*Malus* x *domestica*) is cultivated on a global scale in temperate regions, reaching a world gross production value of 37.8 billion dollars in 2016 [29]. Apple fruits are a valuable contribution to human nutrition and are available throughout the year, as modern apple cultivars can be stored for a period of up to twelve months under controlled atmosphere conditions [40]. In the course of storage, however, apple fruit may deteriorate, due to physiological disorders or infectious post-harvest diseases. The most important post-harvest diseases of apple are caused by pathogenic fungi that can affect both the quantity and quality of the produce, not only during storage but also at the time of packing and shipment. For instance, in the USA, the annual losses caused by post-harvest diseases of apple were estimated at 4.4 million dollars [30], while in Northern Europe, storage losses due to pathogenic microorganisms were estimated to reach up to 10% in integrated production and up to 30% in organic production [25]. As fungal post-harvest pathogens differ in their biological characteristics, effective disease determination is crucial for containing damages, setting sales and marketing priorities as well as to implement a sanitation program, or to define pre-harvest plant protection measures for the following season. The most accurate diagnostic methods are based on microscopic, microbiological or molecular genetic examinations, which require dedicated laboratories and trained staff [3]. A method for disease determination that requires the lowest technical effort and can be directly applied in packing-houses and fields is based on the observation of decayed fruit for the presence of macroscopic symptoms or fungal signs. The former include the appearance, colour, texture and consistency of the rot induced on the peal and/or the pulp tissue, whereas the latter comprise mycelium, fruiting bodies or spore tufts. However, symptom-based disease diagnosis requires a good knowledge of the diseases involved and a trained eye of the user. Furthermore, symptoms can vary according to the cultivar, the stage of infection, as well as the cultivation and storage conditions. Therefore, a computer-guided decision support system able to support on-site practitioners to distinguish pathogens producing apparently similar symptoms is of crucial importance in such an agricultural sector [36,44].

To the best of our knowledge, few works in the literature addresses the problem of building a decision support system to diagnose post-harvest diseases of apple. One of these few was the seminal work by Roach et al. [35], introducing *POMME*, a knowledge-based diagnostic

[*] Corresponding author.
E-mail addresses: gsottocornola@unibz.it (G. Sottocornola), sanja.baric@unibz.it (S. Baric), fabio.stella@unimib.it (F. Stella), mzanker@unibz.it (M. Zanker).

https://doi.org/10.1016/j.atech.2022.100070
Received 21 March 2022; Received in revised form 14 May 2022; Accepted 15 May 2022

module for apple scab being part of a system supporting apple growers in managing their orchards. In the domain of agriculture, expert systems have been widely adopted to cope with the diagnosis and prevention of diseases and disorders that may appear at different developmental stages or parts of the crop plant, and largely affect the yield [2]. Knowledge-based expert systems for supporting diagnosis, founded on *if-then* or logical rules [5,24,42] have inspired our work, specifically in the context of knowledge elicitation and management. However, these works lack important features, which we consider fundamental for our decision support system. Specifically, we designed the system to be used by a different variety of practitioners, researchers, or interested users, with different expertise levels in the domain of apple diseases – from quality managers in packing-houses, to the farmers, from researchers in phytopathology, to the students of horticultural science. Thus, the application of strict expert-defined rules is shortcoming, given that it requires a good amount of domain knowledge by the final user to effectively use the diagnostic application. Namely, in order to correctly recognize some subtle symptoms on fruits associated to specific diseases, years of training and experience in the field are required. Therefore, we exploit multiple sources of information to facilitate the interaction with (non-expert) users. For instance, Kolhe et al. [18] reported a web-based intelligent diagnostic system for oilseed-crops. The system was designed to incorporate a dynamic knowledge base and to provide reasoning by means of a fuzzy logic approach. The interaction with the system was supported by means of an audio–visual–graphical user interface using text-to-speech conversion tools. Gonzalez-Andujar [12], Gonzalez-Diaz et al. [13], instead, built an expert system for the identification of weeds, insects, and diseases of olive trees and pepper plants. Knowledge was gathered by literature review and interviews with experts, and the system adopted a conventional *if-then* knowledge representation, but also employed digital images to assist users in the identification process. Finally, the *Identificator* [32] diagnosis system for strawberries inspired our contribution. Similarly, the framework lets users select macroscopic features (symptoms) associated to predefined images in order to diagnose the correct diagnosis.

However, a property missing in the applications referenced so far concerns the capability of dealing with uncertainty in the knowledge as well as in the user feedback, i.e., the presumed relations between symptoms and diseases are mediated by some degree of uncertainty as well as the feedback acquired from users might be wrong or misleading. Thus, for the development of the expert-defined stream of our model we relied on the *Bayesian Network (BN)* [19,31] framework, a probabilistic graphical model which allows reasoning under uncertainty about symptoms, signs, and diseases. The application of the BN technique for diagnosis tasks has its roots in the late 1980s and early 1990s with the first decision support systems for medical diagnosis, such as *MammoNet*, presented by Kahn et al. [15], for the diagnosis of breast cancer. In particular, we took inspiration from the methodology employed by Spiegelhalter et al. [39] on the construction of a probabilistic expert system diagnosing the “blue baby” disease. In contrast to the application of BN models in agriculture [7,33] that were fully bootstrapped from data, we applied the guidelines for the elicitation of expert knowledge for setting model parameters provided by Kuhnert et al. [20].

The objective of this study is to develop *DSSApple*, an interactive web-based decision support system, that helps users to diagnose post-harvest diseases of apple, based on observed macroscopic symptoms, in order to suggest possible counter-measures for future prevention. The system is designed to provide a practical interface to elicit information about an unknown disease on a target apple from both, expert and non-expert users. Specifically, our application allows for a two-stream hybrid interaction based on the similarity of images, depicting symptoms variety (*image-based* stream), as well as on multiple-choice questions related to the macroscopic characteristics of symptoms, that are exploited by the BN reasoning mechanism (*expert-based* stream). In order to compute the diagnosis (i.e., a ranked list of suggested diseases), both, the *image-based* and the *expert-based* stream are combined to boost the accuracy. The system has been thoroughly tested by means of two real-world evaluation studies involving semi-expert users. The users were challenged to use the *DSSApple* application in order to correctly diagnose two sets of infected apple, i.e., simulated through photos, and real infected apple provided by storage houses. In both cases, the ground-truth disease for each apple was assessed in laboratory through microbiological analysis of the pathogen. This real-user evaluation proved the effectiveness of the hybrid *DSSApple* system in the diagnostic task, which generally outperformed both the *image-based* and the *expert-based* approaches, as well as the users’ intuition based on their self-report.

Thus, the methodological contribution of this paper is manifold, namely:

a) We describe a novel application based on a hybrid expert system for the challenging task of diagnosing post-harvest diseases of apple (Section 2.1) and we discuss the design choices made for building our system (Section 2.1).
b) We illustrate the knowledge elicitation process for the construction of an ad-hoc knowledge base, responsible of the reasoning mechanism of the BN model (Section 2.4).
c) We formally define a practical and adaptive inference mechanism, based on the BN framework and able to deal with soft evidence (Section 2.5), as well as an hybridization technique to increase the diagnostic effectiveness of the system (Section 2.6).
d) We present a practical algorithm for explaining the suggested diagnosis, based on the BN processed evidence provided by the user (Section 2.7).

The rest of the paper is organized as follows. In Section 2, we report the methodological contributions of this work. Firstly, we present the *DSSApple* application, illustrating its features and design choices. Secondly, after introducing background theory on BN, we describe the process of expert knowledge elicitation for BN construction, and the mechanism of expert-based inference with BN. Then, the algorithm for the computation of hybrid diagnoses, as well as the technique for diagnosis explanation are detailed. In Section 3, we outline the experimental user studies conducted in order to test the effectiveness of *DSSApple* and comment on the results. Finally, in Section 4 we draw conclusions and depict future extensions of the presented research.

2. Materials and methods

2.1. Application overview

DSSApple [1] is designed to be an interactive easy-to-use web application allowing expert and non-expert users in the area of apple production and storage (e.g., scholars, researchers, and storage workers) to perform diagnosis of post-harvest diseases of apple fruit, based solely on the observed macroscopic symptoms on the fruit. We leave full flexibility to users to select either or both of the two streams of the system harnessing the *image-based* and *expert-based* sources of information. Of course, this choice depends on the degree of expertise of the user and the difficulty of the diagnosis under investigation. While the image-based stream offers a more intuitive interface, solely based on pictures, the expert-based stream requires a deeper knowledge and understanding of disease symptoms. After the user’s feedback collection, the system proposes a ranked list of diagnoses along with explanations. The interaction process with the *DSSApple* application is represented in Fig. 1.

In more detail, in the **image-based** stream, the user interaction with the system is conducted simply by clicking on pictures, representing the variety of symptoms of different diseases at different stages of infection and on different cultivars. We divided the interactive session into two rounds for both the outer part and the inner part of the apple. At each round of interaction, the user is requested immediate feedback on

[1] Accessible at http://dssapple.unibz.it.

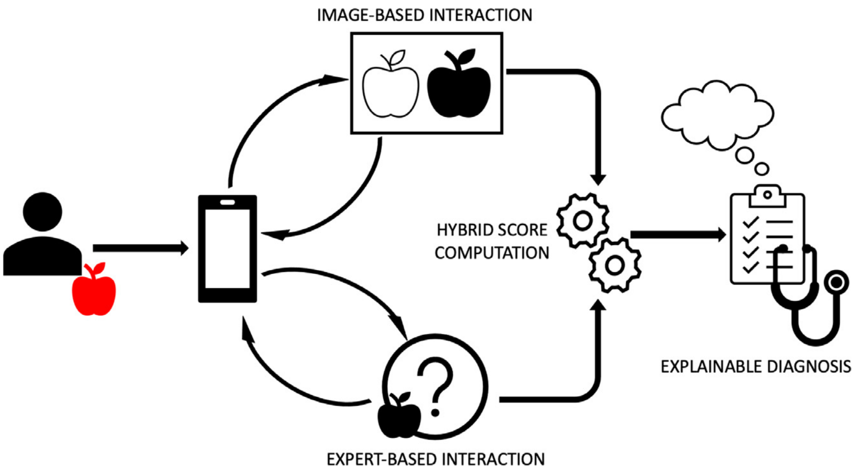


Fig. 1. A diagram representing the interaction of the user with the *DSSApple* hybrid system.

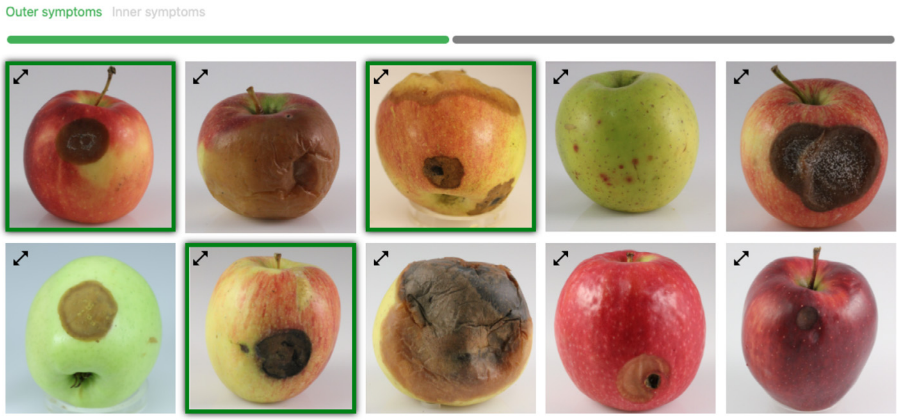


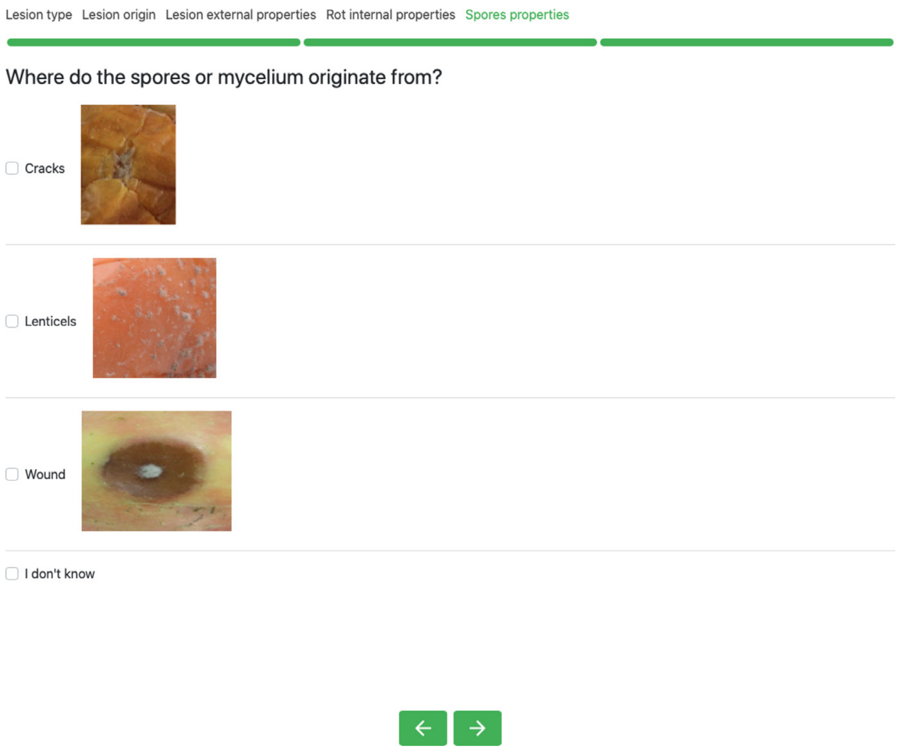
Fig. 2. The interface for a round of interaction with the image-based stream of *DSSApple*, namely referred to the outer symptoms of the apple fruit.

a small set of images, depicting reliably determined disease symptoms on apple fruit, based on the perceived similarity with the diseased target apple. A round of interaction with the *image-based* part of the system is represented in Fig. 2. The user can navigate back and forth the four rounds of interaction and revise her choices. At the current stage, the system included ten high-quality photos for each disease, which are randomly sampled at each round with stratification over the disease. Given the unavailability of a large dataset of reliable high-quality apple diseases images, we produced our own set of images. In particular, infected apples were collected from orchards and storage-houses in the Bolzano region and the ground-truth disease was reliably determined in laboratory by microbiological analysis of the spores. Hence, different photographs of the manifested inner and outer symptoms were taken for each sampled apple, to be included into the system. An extensive analysis on this part of the system is provided in our previous works [37,38].

In the **expert-based** stream, the system collects user's observations about the target disease by asking a set of dynamic multiple-choice questions related to the macroscopic features of the observed symptoms (e.g., the shape of the rot, the origin of the infection, etc.). Each question is illustrated with exemplary pictures, facilitating also non-expert users in their understanding. Each symptom-related question is mapped to a specific variable in the BN model as described in Section 2.4. This part of the system is dynamic, since the system incrementally adapts the questions based on the previous answers given by the user. For instance, when the system gets the information that fungal structures are visible on the infected apple, it will inquire the user about further features of those structures (i.e., their distribution, colour, or origin). Furthermore, the system again provides full flexibility to the user, i.e., it allows to navigate back and forth the questions path, to revise previous answers, and to skip those questions for which the user does not have enough confidence to answer. Fig. 3 reports a section of the expert-based part of *DSSApple*.

After the system finishes the data collection phase (i.e., the user completed the *expert-based*, or the *image-based*, or both paths), the user can access the diagnosis page. Here the application displays a ranked list of suggested diagnoses, based on the information provided by the user. Each suggested diagnosis is supplied with a score, representing the confidence of the system towards that diagnosis, and an explanation motivating that suggestion in the light of the collected user feedback. The score is computed based on the path followed by the user. Namely, if just the *expert-based* stream is followed, the score of each diagnosis is computed based on the reasoning on the underlying expert model; if just the *image-based* stream is followed, the score of each diagnosis is based on the frequency of the coherent symptoms selected for each candidate disease; if both streams are completed, *DSSApple* computes a hybrid score for each disease, which is a linear combination of the two scores. More details about the hybrid diagnosis are provided in Section 2.6.

Moreover, the explanation component is crucial for such a decision support system in order to increase the transparency of the suggested diagnosis and thus the trust by the user. For *DSSApple*, we designed it as a pop-up box for each suggested disease. Based on the path followed by the user for the diagnosis, the explanation box will show fingerprints of the clicked images which belong to that disease (for the image-based stream), or the most representative answers provided by the user describing that disease (for the expert-based stream). Furthermore, a brief description of the disease is provided, followed by a link to a wiki page where the user can find additional information about the disease (e.g., causal pathogen, disease cycle, recommendation for disease manage-



Lesion type Lesion origin Lesion external properties Rot internal properties Spores properties

Where do the spores or mycelium originate from?

□ Cracks

□ Lenticels

□ Wound

□ I don't know

Fig. 3. A section of the expert-based stream of *DSSApple*, in which the user is queried about the origin of fungal growth.

ment, etc.). Fig. 4 shows an example of an explanation box, which includes both the *expert-based* and the *image-based* explanation.

The reasoning system for the *DSSApple expert-based* stream is developed on the framework of *Bayesian Network (BN)* [19,31]. The choice of BN allows to define and reason about relationships between causes (e.g., the disease) and effects (i.e., the symptoms) under uncertainty, namely when these relations are not deterministic but should be mediated by probability. The procedure of eliciting these relations (i.e., the BN conditional probabilities) from a domain expert is described in Section 2.4. Furthermore, other tools, naturally derived from the BN framework, are perfectly suited for the task of *DSSApple*. For instance, the capability of including incomplete and stochastic information (i.e., soft evidence) in the inference mechanism is described in Section 2.5, while the possibility to explain the suggested diagnosis in the light of the collected evidence, is described in Section 2.7.

2.2. Design choice

We are aware of the fact that automated image analysis and classification have achieved important results for the task of disease identification in agriculture, in particular in recent years thanks to the deep learning paradigm [9]. Nevertheless, we considered such a fully-automated approach inadequate for our scope for both methodological and practical reasons.

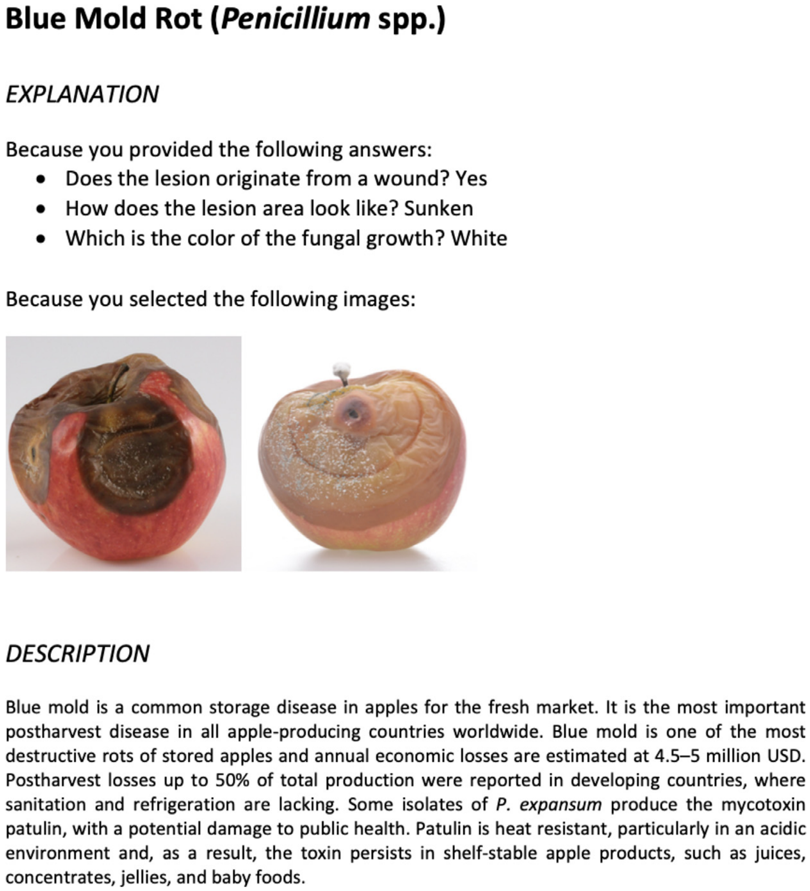
First of all, the majority of the proposed approaches presented so far in the literature [9,10,23] was tested solely on an "offline" batch evaluation. Namely, the reported results were optimized for the prediction on a previously collected dataset, disregarding the non-trivial effort of transposing such an approach in a real-world environment. Specifically, this would have required a user to produce some high-quality photographs in the field or to equip storage facilities with proper cameras. Secondly, deep learning is a black-box tool by construction: the automated classification is non-transparent and can hardly be explained to the user [41]. This might lead, in case of poor diagnostic performance, to a distrust towards the system. These shortcomings are both counteracted by our design choice of adopting a "human-in-the-loop" paradigm [43] in which the user is directly involved (with a minimum technical effort) into the diagnostic process, having the effect of increasing the usability, transparency, and trustability of the system.

Furthermore, the adoption of deep learning generally requires a large amount of images to cope with the high variance embedded in the classification problem, while our manually curated dataset of high-quality and trustable images counted just few hundreds of photographs, derived from around one hundred distinct instances of apple. Indeed, the task of classifying post-harvest diseases of apple shows a particularly high intra-disease variance. The same pathogen induces different symptoms on different species, also based on the progression of the diseases (i.e., days after an infection). At the same time, for a non-expert evaluation, and even for experts without a microscopic or microbiological analysis, it is very difficult to understand the subtle differences of symptom appearances just by observing images of macroscopic symptoms, particularly at early stages of an infection.

To corroborate this intuition with an illustrative example, in Fig. 5, we show three photos of external symptoms. When comparing these images the difficulty of the classification task clearly emerges. The two symptoms looking most similar, given also that they appear on the same apple cultivar, are in fact manifestations of the two different diseases (*Neofabrea* and *Alternaria*). On the other hand, two examples of *Alternaria* symptoms appear to be largely different, since they manifest themselves on different cultivars and at different stages of the infection.

2.3. Background on bayesian network

A *Bayesian Network (BN)* [16,19] is defined by its two main components: the qualitative part represented by its graphical structure and the quantitative part consisting of the conditional probabilities. More formally, a BN is graphically represented as a *directed acyclic graph (DAG)* $\mathcal{G} = (N, E)$, where $N = \{n_1, n_2, \dots, n_l\}$ denotes the set of l nodes and $E \subseteq N \times N$ the set of directed edges between pairs of nodes. Each node $n_i \in N$ in the DAG $\mathcal{G}$ is mapped one-to-one with a random variable $X_i \in \mathcal{X}$, where $\mathcal{X}$ denotes the set of random variables involved in the model. A random variable $X_i \in \mathcal{X}$ is represented by a set of exclusive values (or states) in which the variable might be observed $Val(X_i) = \{x_i^1, x_i^2, \dots, x_i^m\}$, where $x_i^j \in Val(X_i)$ denotes the jth value of



Blue Mold Rot (*Penicillium* spp.)

EXPLANATION

Because you provided the following answers:

- Does the lesion originate from a wound? Yes
- How does the lesion area look like? Sunken
- Which is the color of the fungal growth? White

Because you selected the following images:

DESCRIPTION

Blue mold is a common storage disease in apples for the fresh market. It is the most important postharvest disease in all apple-producing countries worldwide. Blue mold is one of the most destructive rots of stored apples and annual economic losses are estimated at 4.5–5 million USD. Postharvest losses up to 50% of total production were reported in developing countries, where sanitation and refrigeration are lacking. Some isolates of *P. expansum* produce the mycotoxin patulin, with a potential damage to public health. Patulin is heat resistant, particularly in an acidic environment and, as a result, the toxin persists in shelf-stable apple products, such as juices, concentrates, jellies, and baby foods.

Fig. 4. Example of an explanation box interface for the Blue mold rot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

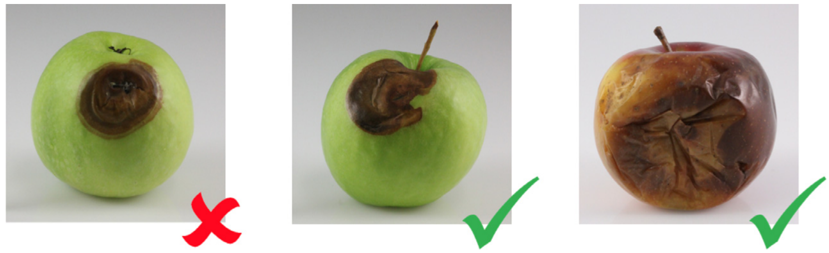


Fig. 5. An example of how difficult could be to automatically classify apple diseases based on photos - The left-most apple is infected by *Neofabrea*, while the others are infected by *Alternaria*.

variable X_i. We use the notation $X_i = x_i^j$ for an observed event, to express that variable $X_i \in \mathcal{X}$ is observed (or instantiated) in the state $x_i^j \in Val(X_i)$. Quantitatively, a *conditional probability table (CPT)* is associated to each random variable $X_i \in \mathcal{X}$. The CPT specifies the conditional probability distribution $P(X_i|pa(X_i)) \in \mathcal{P}$ over the states of X_i. Where, $\mathcal{P}$ represents the set of conditional probabilities in the model, and $pa(X_i) \subset \mathcal{X}$ denotes the set of the so-called *parents* of the variable X_i associated to the node n_i in the DAG $\mathcal{G}$. Specifically, the parent set of X_i is composed by every variable $X_j \in \mathcal{X}$ associated to the node n_j in the DAG $\mathcal{G}$, connected with a directed edge to n_i (the so-called *child* node). More formally, $pa(X_i) = \{X_j \in \mathcal{X} : (n_j, n_i) \in E\}$. We can further define an *ancestor* variable $an(X_i)$ of the variable X_i, and a *descendant* variable $de(X_i)$ of variable X_i, if exists a directed path (i.e., a set of directed edges) connecting node n_a (associated with variable $an(X_i)$) to n_i (associated with variable X_i), and n_i to n_d (associated with variable $de(X_i)$); namely $\{(n_a, n_j), (n_j, n_i), (n_i, n_h), \dots, (n_g, n_d)\} \subset E$. It is important to mention that the DAG $\mathcal{G}$ of the BN typically specifies a set of probabilistic or causal relationships among variables in the model. Namely, if an edge $(n_j, n_i) \in E$ exists in the graph, this usually implies that a causal relation holds between the variables X_j and X_i, associated to nodes n_j and n_i. Specifically, the parent X_j represents the cause and child X_i represents the effect in the modeled domain. Thus, a fundamental assumption of conditional (in)dependence between variables could be derived. Specifically, this assumption is referred as *Local Markov Assumption* (or *Local Independence Assumption*), and it states that: given its parents $pa(X_i) \subset \mathcal{X}$, defined in the DAG $\mathcal{G}$, a variable X_i is conditionally independent of all its non-descendent variables. More formally, for each variable X_i: $(X_i \perp X_j | pa(X_i))$, where $X_j \notin de(X_i)$, set of descendants of X_i. This property allows to specify the joint distribution over the space of the variables $\mathcal{X}$ in the BN model through the following probability factorization:

$$P(\mathcal{X}) = \prod_{i=1}^{l} P(X_i|pa(X_i)) \tag{1}$$

The equation is usually referred to as the *chain rule for Bayesian networks*. This joint distribution implicitly allows the BN to compute other probabilities of interest. For instance, one may be interested in reasoning about the impact of any set of observations (or *evidence*) $\mathbf{E}$ on any assignment $\mathbf{X}_q = \mathbf{x}_q^g$, where $\mathbf{X}_q \subset \mathcal{X}$. We define an evidence $\mathbf{E}$ as an observation of any proper subset of variables in the BN model $\mathbf{E} = \{X_1 = x_1^i, X_2 = x_2^j, \dots, X_d = x_d^t\}$, where $d < l$. The probability distribution we want to compute, hence, becomes the posterior probability

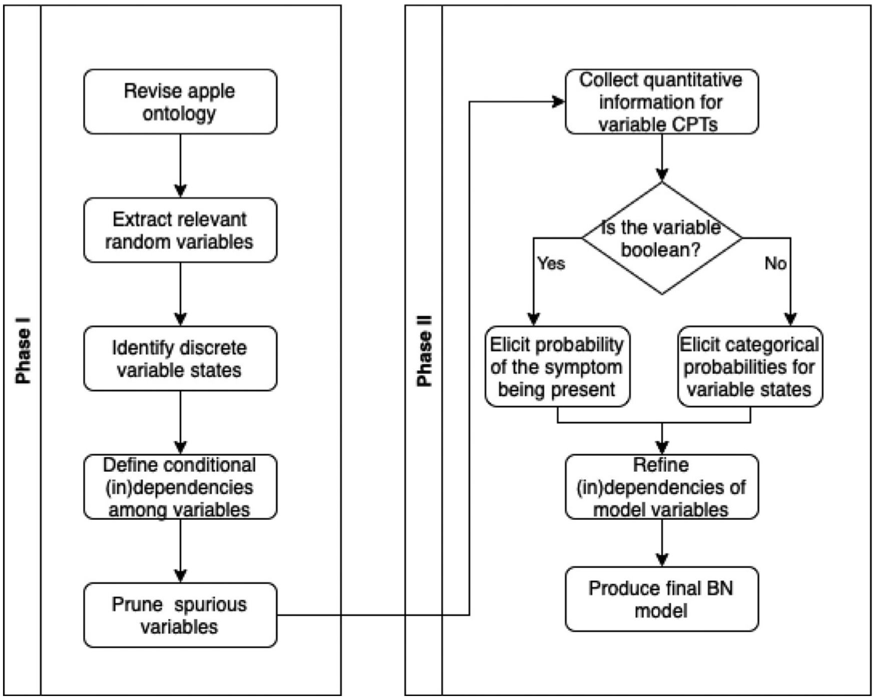

Phase I
Revise apple ontology
Extract relevant random variables
Identify discrete variable states
Define conditional (in)dependencies among variables
Prune spurious variables
Phase II
Collect quantitative information for variable CPTs
Is the variable boolean?
Yes
No
Elicit probability of the symptom being present
Elicit categorical probabilities for variable states
Refine (in)dependencies of model variables
Produce final BN model

Fig. 6. Flow chart of the knowledge elicitation process followed for the BN construction.

$P(X_q = x_q^g|\mathbf{E})$, of the query (unobserved) event x_q^g given the evidence $\mathbf{E}$. This posterior probability can be computed directly from the joint distribution $P(\mathcal{X})$, by conditioning it on the observation $\mathbf{E}$ eliminating the entries in the joint inconsistent with the observation and re-normalizing the results such that they sum up to 1; we compute the probability of the event $X_q = x_q^g$ by summing the probabilities of the entries in the resulting posterior distribution which are consistent with x_q^g.

2.4. Expert knowledge elicitation for Bayesian network construction

The elicitation of expert knowledge for constructing the Bayesian network (i.e., both the network structure and the CPTs) is a crucial task. To accomplish this challenging operation two options are available: learn it from data or elicit knowledge from domain experts. At the best of our knowledge, no datasets are publicly available allowing to learn significant relationships among apple diseases and macroscopic symptoms. Thus, we focused on interviewing a domain expert for the construction of the BN knowledge base. Following the common scheme presented in the literature [8,11,34], we divided the task into two distinct phases: during the first phase, we identified the random variables (i.e., the macroscopic symptoms) relevant for the diagnostic process; during the second phase, we determined the probability values (i.e., the CPTs) quantitatively linking the diseases to the symptoms, and revised the conditional dependencies among random variables in the BN model. In Fig. 6, we schematize the whole knowledge elicitation process through a flow chart.

With the help of a domain expert we limited a general apple ontology [27] to the relevant parts for relating diseases with visible macroscopic symptoms. Thus, we ended up with a stable configuration of around 30 observed random variables, grouped into 8 categories, together with two hidden random variables, namely *Disease* and *Stage*. One basic assumption underlying our model is that there are no multiple infections with two or more diseases and that the list of diseases is complete, namely, a target apple is always infected by one and only one disease. In addition, we also assumed that all the symptom variables are conditionally independent given the *Disease* variable, as commonly done for Naive Bayes [16]. Thus, the *Disease* variable encodes the whole set of fungal diseases of our study, namely *Val(Disease) = {alternaria_rot, alternaria_spot, black_rot, blue_mold, bitter_rot, bulls_eye, fusarium_rot, grey_mold, mucor_rot, side_rot}*. The *Stage* variable was introduced to facilitate the probability elicitation task. It represents three discrete and symbolic stages of advancement of the post-harvest infection, namely *Val(Stage) = {early, medium, late}*. This workaround allows experts to visualize a specific condition of the disease and thus specify a more reliable likelihood of the symptoms. At the end of this phase, we removed variables for which it was too hard to assess differences among different manifestations of the symptom, and thus to define an exhaustive set of mutually exclusive states. We mention for this issue the variables *Lesion_colour* and *Rot_colour*. In Tables 1 and 2, we report all the variables and respective states included in the final model, defined after this first phase.

In the second phase, we interviewed the domain expert in order to define the quantitative probabilistic dependencies (i.e., the CPTs) among variables. For simplicity, we decided to start from a situation where all the symptom variables are conditionally independent upon one another and also conditionally dependent upon the two hidden variables (i.e., *Disease* and *Stage*). We indicate the *Disease* variable as $D \in \mathcal{D}$, where $\mathcal{D}$ defines the set of hidden variables for the model. $Val(D) = \{d^1, d^2, \dots, d^n\}$ represents the set of states of the variable D, where d^i is the ith state of the *Disease* variable (i.e., the ith disease in our pool). The *Stage* variable is referred as $T \in \mathcal{D}$ and $Val(T) = \{t^1, t^2, \dots, t^m\}$ represents the set of states of variable T, where t^i is the ith state of the *Stage* variable. All other (observed) variables in the model are generally referred to as symptom variables and they belong to the set $\mathcal{S}$. A generic symptom variable $S_i \in \mathcal{S}$ is represented by a set of states $Val(S_i) = \{s_i^1, s_i^2, \dots, s_i^q\}$, where s_i^j is the jth state of the symptom variable S_i. Moreover, we adopted a mixed-questionnaire approach inspired by Gaag et al. [11], for facilitating the expert knowledge elicitation process and thus to define the model CPTs. In more details, two techniques were applied depending on the support of the variable. For Boolean variables (for each symptom variable $S_i \in \mathcal{S}$ such that $Val(S_i) = \{true, false\}$), the expert was requested to answer the question: *"How frequently do you observe symptom $S_i = true$, given that you have an apple infected by disease $D = d_i$ at stage $T = t_j$?"*. We provided her the choice among a pre-defined 6-point scale, including *Always (A)*, *Very often (V)*, *Often (O)*, *Sometimes (S)*, *Rarely (R)*, and *Never*

Table 1
Summary of all the variables and states involved in the model, grouped by categories (first part).

Category	Variable	States
Diagnosis	Disease	{alternaria_rot, alternaria_spot, black_rot, blue_mold, bitter_rot, bulls_eye, fusarium_rot, grey_mold, mucor_rot, side_rot}
Diagnosis	Stage	{early, medium, late}
Lesion type	Rot	{true, false}
Lesion type	Spot	{true, false}
Lesion type	Scab	{true, false}
Other traits	Halo	{true, false}
Other traits	Mycelium_spores	{true, false}
Other traits	Sclerotia	{true, false}
Other traits	Odour	{true, false}
Lesion origin	Lenticel	{true, false}
Lesion origin	Wound	{true, false}
Lesion origin	Calyx	{true, false}
Lesion origin	Stalk	{true, false}
Lesion origin	Core	{true, false}
Lesion properties	Number_lesions	{single, few, multiple}
Lesion properties	Lesion_form	{circular, irregular}
Lesion properties	Lesion_margin	{sharp, indistinct}
Lesion properties	Lesion_area	{plane, flat, sunken, collapsed}
Lesion properties	Lesion_appearance	{dry, watery, baked}
Lesion properties	Lesion_surface	{unwrinkled, slightly_wrinkled, wrinkled, corky}
Lesion properties	Lesion_intactness	{uncracked, cracked, parchment}
Lesion properties	Lesion_size	{xs, s, m, l, xl}

Table 2
Summary of all the variables and states involved in the model, grouped by categories (second part).

Category	Variable	States
Odour properties	Odour_type	{sweet_cider, earthy_musty, bandage}
Halo properties	Halo_colour	{brown, red, yellow, light_green}
Fungal properties	Fungal_colour	{white, grey, dark_grey, pink, yellow, brown, green_blue, peppered}
Fungal properties	Fungal_distribution	{random, concentric}
Fungal properties	Fungal_origin	{wound, lenticels, cracks}
Rot properties	Rot_shape	{conical, rounded, irregular}
Rot properties	Rot_margin	{sharp, indistinct}
Rot properties	Rot_moisture	{dry, moist, juicy}
Rot properties	Rot_transparency	{opaque, glassy}
Rot properties	Rot_consistency	{firm, spongy, soft}

Table 3
Scale to convert expert knowledge into probability distributions.

Question	Answer	$P(S_i = true\|d,t)$
How frequently do you observe symptom $S_i = true$, given that you have apples infected by disease d at stage t?	Always (A)	0.999
	Very often (V)	0.8
	Often (O)	0.6
	Sometimes (S)	0.3
	Rarely (R)	0.01
	Never (N)	0.001

(N). The expert had to fill a form, answering for each combination of $d_i \in Dxt_j \in T$. The symbolic scale was converted into an actual probability value $P(S_i = true|D = d_i, T = t_j)$ according to the scheme reported in Table 3. The complementary probability was consequentially defined as $P(S_i = false|D = d_i, T = t_j) = 1 - P(S_i = true|D = d_i, T = t_j)$.

For categorical variables (for each symptom variable $S_l \in \mathcal{S}$ such that $Val(S_l) = \{s_l^1, s_l^2, \dots, s_l^m\}$, where $m \geq 2$), we adopted a lighter, yet effective, approach. For each categorical symptom variable $S_l \in \mathcal{S}$, given a specific disease $D = d_i$ at stage $T = t_j$, the expert was invited to simply indicate which values of $Val(S_l)$ are likely to be observed. Furthermore, we agreed on a 3-point symbolic annotation to denote the likelihood of each reported state, namely, *common* (no parenthesis), *less common* (one parenthesis), and *rare* (two parentheses). The assumption underneath this choice was that many symptom states are never observed under specific conditions (i.e., resulting CPTs are sparse) and could be ignored to speed up the elicitation process. In order to convert likelihood annotations into actual probability distributions, we designed the following heuristic. Consider a random variable R with $Val(R) = \{a, b, c, d\}$, which is annotated as follows by the expert: *a: common*, *b: less common*, *c: rare*, and d is ignored; then $P(a) = 2P(b) = 4P(c) = 1.0$ and $P(d) = 0.0$. Furthermore, a small value $\epsilon = 0.001$ is added to each probability value in order to avoid null probabilities, then values are normalized such that $\sum_{r \in Val(R)} P(r) = 1.0$. This process completely defines a probability distribution for the categorical random variable R.

After the complete definition of the quantitative components of the model (i.e., the CPTs), we further refined the model to converge to the final specification of the BN model for *DSSApple*. We leveraged the expert-defined CPTs to identify conditional independence in the model and thus prune the graph from superfluous edges. Consider a set of three random variables X, Y, and Z. X and Y are defined as conditionally independent given Z if and only if $P(X|Y, Z) = P(X|Z)$, and this is written $(X \perp Y|Z)$. In our context, we decided to start from a graphical representation where all the symptom variables are dependent upon both the disease variable D and the stage variable T. After probability elicitation, we were able to identify conditional independence between a symptom variable S and stage variable T, given disease variable D, namely when $P(S|D, T) = P(S|D)$. We exploit this property to identify conditional independence for all the variables belonging to the *lesion origin* category and for the variables *Number_lesions*, *Odour_type*, and *Fungal_distribution*, with respect to *Stage*. Finally, we identified some “second-order” variables to be further connected to other symptom nodes. This is the case of

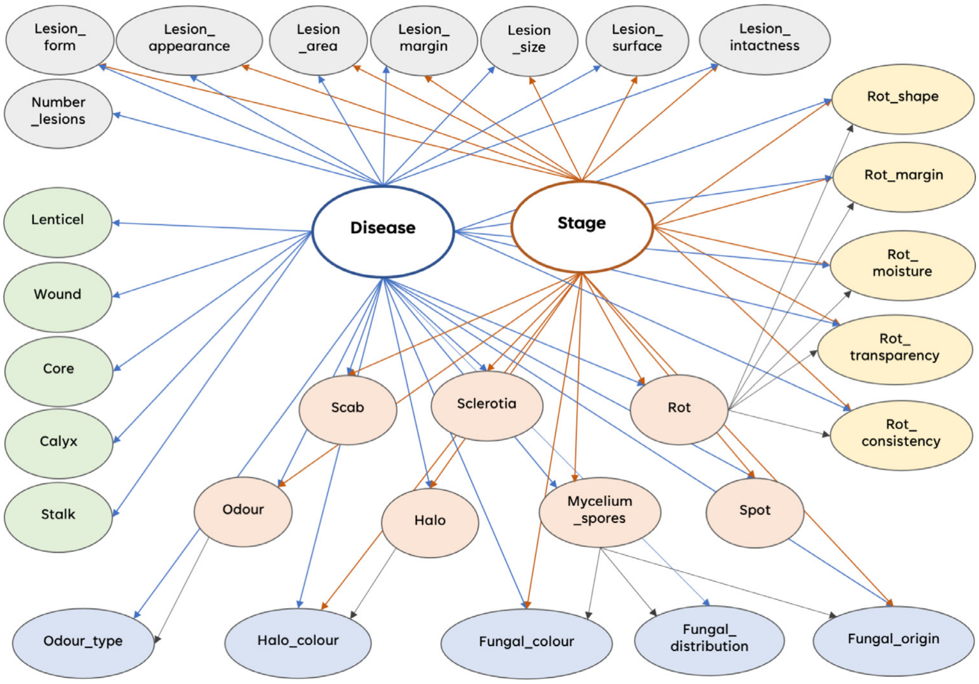


Fig. 7. The graph of the Bayesian network for *DSSApple*.

the variables in the category *rot properties* connected with node *Rot*, *halo properties* connected with node *Halo*, *fungal properties* connected with node *Mycelium_spores*, and *odour properties* connected with node *Odour*. These edges will increase the expressiveness of the model by conditioning the property of the symptom with the actual symptom. For instance, in a situation where our system will receive the evidence $Rot = false$, it will immediately set the values of the variables in the *rot properties* category to na (not applicable), or more formally $P(R_p = na) = 1.0$ for every variable $R_p \in \mathcal{R}$, rot properties category.

The final BN graph used as a reasoning mechanism for *DSSApple* is reported in Fig. 7. The central nodes in the network, bolded and empty, represent the two unobserved diagnosis variables, namely *Disease* and *Stage*. On the top part of the network, coloured in grey, are the nodes related to the *lesion properties*. On the right-most part, coloured in yellow, are the *rot properties*, while on the left-most part, coloured in green, are the *lesion origin* nodes. Finally, in the central-bottom part, coloured in orange, are the nodes related to the *lesion type* and *other traits* represented, under those, coloured in cyan, the nodes representing the properties of the other traits.

2.5. Bayesian network inference mechanism

The reasoning system of the constructed BN allows to perform the inference, namely, to estimate the posterior probability distribution on a target unobserved variable (i.e., the *Disease* variable D), given any set $\mathbf{S} \in \mathcal{S}$ of observed variables as provided by the user (i.e., the *evidence* $\mathbf{E}$). The evidence set $\mathbf{E}$ is constructed incrementally by the *DSSApple* application. At each step, the application requests the user to answer a multiple-choice question, related to a symptom variable $S_i \in \mathcal{S}$. When the user submits the observed state $s_i^j \in Val(S_i)$, *DSSApple* includes the new information into the evidence set, $\mathbf{E} \cup S_i = s_i^j$. At the end, of this feedback collection process, the application will have access to the complete information provided by the user on the target disease, she want to diagnose. It is important to mention that the BN inference mechanism is robust to missing values, hence, the user is not forced to provide an observation for every symptom variable $S_i \in \mathcal{S}$ in the model. Thus, if the user skips the question related to variable $S_m \in \mathcal{S}$, the evidence set $\mathbf{E}$ will not include an observation for that variable, $S_m \notin \mathbf{E}$, and the inference will be computed anyway. Thus, the goal of the reasoning system is to provide a probability over the candidate diseases (i.e., the diagnosis). We estimate the posterior probability distribution $P(D|\mathbf{E})$ through an algorithm called *loopy belief propagation* [14]. The loopy belief propagation is an approximate message-passing method to perform inference on graphical models. In few words, the algorithm iteratively updates the marginal distribution $P(N)$ of a node $N \in \mathcal{G}$, by updating the outgoing message, at the current iteration, from the node N to each of its neighbors $\mathbf{V} \in \mathcal{G}$ in terms of the previous iteration's incoming messages from $\mathbf{V}$.

Furthermore, we decided to provide additional flexibility to the application, by allowing the user to submit multiple answers to the same categorical symptom variable. This may be useful in situation where the observed symptom state on target apple is ambiguous (e.g., it is difficult to distinguish whether the shape of the internal rot is rounded or conical), or the target apple presents different states of the same symptom (e.g., the spore originate both from a mechanical wound and from lenticels). Thus, a categorical variable $S_d \in \mathcal{S}$ which is instantiated to more than one value $S_d = \{s_d^1, s_d^2, \ldots, s_d^l\}$, such that $l \leq |Val(S_d)|$, it is converted by the system into a uniform distribution $P(S_d)$ on the set of observed values. In more details, $P(S_d = s_d^i) = 1/l$, where $s_d^i \in S_d$ is an observed value for the variable S_d. Obviously, if a value $s_d^j \in Val(S_d)$ is not observed, i.e., $s_d^j \notin S_d$, its probability is set to $P(S_d = s_d^j) = 0$. This type of evidence is referred in the literature as *soft evidence* and it allows to define and reason about uncertain evidence [26]. The soft evidence can be included into the evidence set, $\mathbf{E} \cup P(S_d)$, and the inference mechanism works as usual.

2.6. Hybrid diagnosis computation

In this section, we clarify how a ranked list of suggested diseases (i.e., a diagnosis) is computed when user completed just the image-based stream, just the expert-based stream, or completed both streams of *DSSApple*.

For the image-based path alone, at the end of the image selection rounds, an *image-based score* $score(d_i)_{img}$ for each disease $d_i \in D$ is computed. The score should be proportional to the number of coherent symptoms for disease d_i depicted in the pictures clicked by the user. More formally, given the set of m clicked images $C = c_1, c_2, \ldots, c_m$, during the image-based selection rounds, the image-based score for disease

$d_i \in D$ is computed as:

$$score(d_i)_{img} = \sum_{c_j \in C} \mathbb{1}_{I_{d_i}}(c_j)/m \tag{2}$$

where $\mathbb{1}_{I_{d_i}}(c_j) : I \rightarrow \{0,1\}$ is an indicator function, equals to 1 when clicked image c_j belongs to the set I_{d_i}, of the images depicting symptoms of disease d_i, 0 otherwise. The score is normalized by the total number of clicks m such to constraint the score in the interval $[0,1]$. Thus, the ranked list of k suggested diseases $R^k_{img} = \{d^1, d^2, \ldots, d^k\}$ for the image-based path is then based on the score for each disease, such that $score(d^i)_{img} \geq score(d^{i+1})_{img}$.

For the expert-based path alone, after completing the evidence collection phase, the system computes an *expert-based score* $score(d_i)_{\exp}$ for each disease $d_i \in D$ as the posterior probability for the BN, as described in Section 2.5. More formally, given the provided evidence set $\mathbf{E} = S_1 = s^o_1, S_2 = s^p_2, \ldots S_l = s^q_l$, defined as the set of instantiated state $s^j_i \in Val(S_i)$ for each random variable $S_i \in \mathcal{S}$, the expert-based score for disease $d_i \in D$ is computed as:

$$score(d_i)_{\exp} = P(D = d_i|\mathbf{E}) \tag{3}$$

Again, the ranked list of the k suggested diseases $R^k_{\exp} = \{d^1, d^2, \ldots, d^k\}$ for the expert-based path is then based on the score for each disease, such that $score(d^i)_{\exp} \geq score(d^{i+1})_{\exp}$.

When the user completes the path on both *DSSApple* components, the system computed an aggregated *hybrid score* $score(d_i)_{hyb}$ for each disease $d_i \in D$, which is a linear combination of $score(d_i)_{img}$ and $score(d_i)_{\exp}$. Specifically:

$$score(d_i)_{hyb} = (1-\alpha) \cdot score(d_i)_{img} + \alpha \cdot score(d_i)_{exp} \tag{4}$$

where $\alpha \in [0,1]$ is an hyperparameter of the system which allows to control the contribution of the expert-based stream with respect to the image-based stream of *DSSApple*. The α hyperparameter can be adapted to leverage the target user group characteristics (i.e., degree of expertise) for the deployed application, in order to optimize the diagnostic performance. A standard technique in machine learning is the one of using a limited set of controlled interactions with the system as a fine-tuning set from which to learn the α maximizing the diagnostic accuracy. Of course, α could be also manually set by the experimenter. For instance, with $\alpha = 0.5$, we impose an equal contribution of the two components in the computation of the aggregated score. Finally, the ranked list of the k suggested diseases $R^k_{hyb} = \{d^1, d^2, \ldots, d^k\}$ for the hybrid path is then based on the hybrid score for each disease, such that $score(d^i)_{hyb} \geq score(d^{i+1})_{hyb}$.

2.7. Explanation of diagnosis

Explanation is a key component in modern machine learning approach. It allows to increase trustability towards the model by letting its decision to be understood by human being in real-world application [4]. Even more so, in a costly domain, such as the one of decision support system for diagnosis, where decisions can have huge economical impact, if not on human well-being (e.g., in medical area). Some previous work has been presented in the field of Bayesian Network explanation, which is reviewed by Lacave and Díez [21]. The authors classified the explanation methods into three categories: explanation of *reasoning*, explanation of the *model*, and explanation of *evidence*, according on the BN component interested by the explanation.

Based on the proposed classification, we present a novel BN *reasoning*-related explanation technique, which is inspired by the *forward feature selection* [17] in supervised classification. The goal is to identify, among all the pieces of evidence provided by the user, which subset of it better justifies a given diagnosis. More formally, given the complete set of evidence $\mathbf{E}$ provided by a user performing the diagnosis task, we want to find the subset $\mathbf{B}^d_n \subseteq \mathbf{E}$, with cardinality n, which represent the best explanation (i.e., the most representative evidence set) towards the diagnosed disease $D = d$. The approach we propose receives as input the evidence provided by the user $\mathbf{E} = \{e_1, e_2, \ldots, e_k\}$, a target diagnosis $D = d$, and an integer number $n \leq k$, namely, the cardinality of the subset of evidence that should explain the diagnosis d. Please notice, that a piece of evidence $e_i \in \mathbf{E}$ corresponds to the observation of a symptom random variable $S_i = s_j$, as provided by the user to the system. The algorithm starts from an empty set $\mathbf{B}^d_0 = \emptyset$ and computes a likelihood metric $L(\mathbf{B}^d_0 \cup e_i, d)$ for each piece of evidence $e_i \in \mathbf{E}$ with respect to the target diagnosis d. The evidence $e^*_i \in \mathbf{E}$ with the highest likelihood $L^*(\mathbf{B}^d_0 \cup e_i, d)$ is added to $\mathbf{B}^d_0$ and removed from $\mathbf{E}$. Thus, we construct the best set of evidence $\mathbf{B}^d_1 = \mathbf{B}^d_0 \cup e^*_i$ of cardinality 1, with respect to target disease d. Then, all the remaining pieces of evidence in $\mathbf{E} \setminus e^*_i$ are tested in conjunction with the best evidence set constructed so far. For each piece of evidence $e_j \in \mathbf{E} \setminus e^*_i$, we compute the likelihood metric $L(\mathbf{B}^d_1 \cup e_j, d)$ on the set $\mathbf{B}^d_1 \cup e_j$ with respect to d, and select the $e^*_j \in \mathbf{E} \setminus e^*_i$ that achieves the highest score $L^*(\mathbf{B}^d_1 \cup e_j, d)$, which is added to $\mathbf{B}^d_1$ and removed from $\mathbf{E}$. Thus, we obtain a best subset of evidence explaining d, $\mathbf{B}^d_2 = \{e^*_i, e^*_j\}$, of cardinality 2. This process iterates until the best subset of evidence $\mathbf{B}^d_n$ of cardinality n is built, or until step t, if $L^*(\mathbf{B}^d_t, d) > L^*(\mathbf{B}^d_{t+1}, d)$ for $0 \leq t < n$.

We define the likelihood metric $L(.)$ for a subset of evidence $\mathbf{E}' \subseteq \mathbf{E}$ with respect to a target disease $D = d$, according to the measure of *normalized likelihood (NL)* described by Kjaerulff and Madsen [16]. Thus, we formulate $NL(\mathbf{E}', d)$ as following:

$$NL(\mathbf{E}', d) = \frac{P(\mathbf{E}'|d)}{P(\mathbf{E}')} = \frac{P(\mathbf{E}', d)/P(d)}{P(\mathbf{E}')} = \frac{P(d|\mathbf{E}')P(\mathbf{E}')/P(d)}{P(\mathbf{E}')} = \frac{P(d|\mathbf{E}')}{P(d)} \tag{5}$$

$NL(\mathbf{E}', d)$ is a measure of the impact of a subset of evidence $\mathbf{E}' \subseteq \mathbf{E}$ on the target disease d. By comparing the normalized likelihoods of different subsets of the evidence, we compare the impacts of the subsets of evidence on the target variable D. Investigating the impact of different subsets $\mathbf{E}'$ of the evidence on states $d \in D$ helps to determine subsets of the evidence acting in favor of or against each possible hypothesis state. The higher the measure of $NL(\mathbf{E}', d)$, the more the subset of evidence $\mathbf{E}'$ acts in favour (and hence explain) the target disease $D = d$.

3. Experiments and results

3.1. User study evaluation

In order to evaluate the performance of the *DSSApple* application, we analyzed data on the usage of the system derived from a large user study. Specifically, we involved students from the 2021 Phytopathology class of the Bachelor in Agricultural, Food and Mountain Environmental Sciences at the Free University of Bozen-Bolzano. The students, at the end of the Phytopathology course, were instructed on how to use *DSSApple* application and were asked to interact with the system to diagnose the actual disease of a set of infected target apples. The user study was divided into two distinct phases. In the first phase, called **simulated challenge**, when the participant started a new diagnosis, the application sampled a random target apple (i.e., an apple infected by a ground truth disease, the user had to diagnose) and showed to her two photos of the target apple, namely, one view of the external part of the apple, and one internal view of the apple. The task of the user was to carefully analyze the target apple and interact with the application (as described in Section 2.1) in order to get a suitable diagnosis. The number of available target apples within the system was 50 (i.e., five apples for each candidate disease), thus, each user had the possibility to interact with the simulated part of the challenge up to 50 times. In the second phase, called **real-world challenge**, we distributed to each participant a set of four real apples, infected by a ground truth disease, inferred by laboratory analysis. Again, the task of the user was to carefully analyze the macroscopic symptoms on the apple and interact with the system in order to get to the correct diagnosis. In both phases, a single session of diagnosis is considered completed, once the user interacted with both the expert-based and the image-based stream of *DSSApple*, and she provided her blind guess on the responsible disease(s) of target apple infection. Namely, after the user provided all the evidence on a target apple to the system but before knowing the system's suggestions, she was asked to provide her guess on the actual diagnosis. In this case, the user could select up to three diseases, among the ten candidates, she considers responsible for the target apple decay. This information is used to compare the "a-priori" diagnostic capability of the user with the one of our decision support system.

The number of users participating in the simulated challenge is 21 (10 females and 11 males), performing a total number of 146 diagnoses. The average age of the participants is 24.2. The average number of diagnoses performed by each user is around 7, with a maximum of 46, a minimum of 1, and a median value of 2. The users provided an average number of 1.5 diagnostic guesses for each target apple. The number of participants involved in the real-world challenge is 16 (9 females and 7 males, of which 12 overlap with the ones involved in the simulated challenge). The average age of the participants is 24.8. The total number of performed diagnoses is 68. Each user received a balanced set of 4 infected apples, but in 4 cases the users tried to diagnose the same apple twice. These repeated trials were considered as two distinct diagnoses. The users provided an average number of 1.2 diagnostic guesses for each target apple. In both challenges, the original set of participants (i.e., the students of the Phytopathology class) were integrated with few external users to enrich the set of experiments with more data and variability. It is important to highlight that the guest participants were selected based on a similar or higher degree of knowledge in Pythopathology (i.e., PhD students, former graduate students, or colleagues). We stress the fact that all the participants of the challenge had a moderate level of expertise in the domain of post-harvest diseases of apple, in order to effectively use the full version of the hybrid *DSSApple* application. This ensured that the performed user study was a simulated yet realistic test of the system in the wild, namely, employed by storage workers, researchers, or practitioners in the field of Pythopathology.

3.2. Evaluation metrics

The way in which to evaluate the performance of *DSSApple* is not straightforward, hence, in this section, we illustrate and justify the evaluation metrics used in the results. The output of the system at each diagnosis on target apple a is a ranked list of suggested diseases $R^k_a = \{d^1_a, d^2_a, \dots, d^k_a\}$ of length k. The ith disease d^i_a is ranked based on the score $score(d^i_a)$ computed by the diagnostic model based on the information provided by the user on apple a, such to ensure that $score(d^i_a) \geq score(d^{i+1}_a)$. The desired property of the ranked list R^k_a is that it includes the ground truth target disease t_a for target apple a, within the smallest possible k. Specifically, we would like our diagnostic model to assign the highest score $s * (d^i_a)$ to the ground truth disease, namely $d^i == t_a$. In order to evaluate this property, we borrow two metrics from the information retrieval domain, namely *recall* and *precision* [1]. In our domain, recall measures the share of diagnosis where the ground truth disease t_a is correctly retrieved within the ranked list R^k_a, for apple a diagnosis. Precision measures the share of guesses (or suggested diseases in R^k_a) that correctly identify the ground truth disease t_a, on the total number of guesses in every diagnosis. To better formalize these two metrics, consider a situation in which a set N of n diagnosis is performed by *DSSApple*. The set N is composed by n lists of suggested diagnosis, namely $N = \{R^k_{a_1}, R^k_{a_2}, \dots R^k_{a_n}\}$, where a_i represents the ith apple processed by the system. Also consider that, in the presented scenario, the length k is static and fixed for all $R^k_{a_i}$, but it could be easily extended to the case in which it is dynamically adapted to each diagnosis. Thus, we formally define *recall* and *precision* as:

$$Recall(N,k) = \frac{\sum_{R^k_{a_i} \in N} \mathbb{1}_{R^k_{a_i}}(t_a)}{n} \tag{6}$$

$$Precision(N,k) = \frac{\sum_{R^k_{a_i} \in N} \mathbb{1}_{R^k_{a_i}}(t_a)}{n * k} \tag{7}$$

The function $\mathbb{1}_{R^k_{a_i}}(t_a) : D \rightarrow \{0,1\}$ is an indicator function equal to 1 if $t_a \in R^k_{a_i}$ and 0 otherwise. It is easy to understand how these two metrics are strongly influenced by the choice of k. Specifically, *recall* is directly correlated with k (i.e., it monotonically increases as k increases), while *precision* is inversely correlated with k (i.e., it monotonically decreases as k increases). Thus, we define a metric which is more robust on different k and it achieves a trade-off between recall and precision. This is the *Fβ-score*, formally computed as:

$$F\beta(N,k) = (1+\beta^2)\frac{Recall(N,k) \cdot Precision(N,k)}{\beta^2 \cdot Precision(N,k) + Recall(N,k)} \tag{8}$$

For the diagnostic application presented in this work, we value recall to be a more relevant metric than precision, since a type I error (i.e., identifying a wrong disease as diagnosis) is less harmful than a type II error (i.e., failing to identify the correct disease as diagnosis). Hence, in the presented results, we set $\beta = 2$ and use the *F2-score*, which weights recall twice as precision.

Based on a qualitative pre-study on usability and on application interface constraints, as well as on performance optimization, we decided to set $k = 2$, i.e., the length of the list of suggested disease for every diagnosis is 2. Thus, the results presented in next sections are intended to be evaluated on the top-2 ranked diseases for each diagnosed target apple a_i. For the benchmark method, namely the user-made diagnosis, we could not force a fixed length k, hence, in this case, the results are computed on the full list of (up to 3) diseases selected by each user for the target apple a_i.

3.3. Overall results

Tables 4 and 5 summarizes the overall results of recall, precision and F2-score achieved by the two diagnostic components of *DSSApple* alone (namely, *image* and *expert*), the full hybrid method (*hybrid*), and the user-made diagnosis (*user*), for the simulated and real-world challenge respectively. We would like to stress that we always considered the top-2 diseases, namely, the two highest scoring diseases for each DSS-related method as a diagnosis. Vice versa, for the user-made diagnosis, we evaluated all the disease selections made by the user at the end of the diagnosis, which is on average 1.5 (std. 0.75) for the simulated challenge and 1.2 (std. 0.62) for the real-world challenge. For all these experiments, we fixed the hyperparameter for the hybrid diagnosis score computation to $\alpha = 0.5$.

In Table 4, we notice that the hybrid system appears to be the best performing method in the diagnostic task. Specifically, hybrid outscores

Table 4
Results achieved by each diagnostic model (i.e., *image*, *expert*, *hybrid*, and *user*-made diagnosis) during the simulated challenge.

	image	*expert*	*hybrid*	*user*
Recall	0.678	0.527	**0.699**	0.555
Precision	0.339	0.264	0.349	**0.371**
F2-score	0.565	0.439	**0.582**	0.504

Table 5
Results achieved by each diagnostic model (i.e., *image*, *expert*, *hybrid*, and *user*-made diagnosis) during the real-world challenge.

	image	*expert*	*hybrid*	*user*
Recall	0.441	0.397	**0.544**	0.308
Precision	0.221	0.199	**0.272**	0.263
F2-score	0.368	0.331	**0.453**	0.298

the user-made selection for +14% in recall (70% against 56% of correctly retrieved diseases) and +8% in F2-score. Nevertheless, user-made diagnosis is slightly more precise (+2%, 37% against 35%) than the one made by the hybrid system. This fact might be attributed to the more conservative decisions taken by the user, which, in most circumstances, were motivated to select a single disease for which they were very confident being the correct one. Note, how the image-based method appears to be the second best performing method for both recall and F2-score metrics, separated from the hybrid method of around 2%. The expert-based method, instead, despite being the worst-performing method, it is fundamental in regularizing the image-based selection, as testified by the improved performances achieved by the hybrid. It is important to mention how the poor performances of the expert-based system are influenced by the prior expertise of the users involved in the study. Specifically, even a group of semi-expert users might find difficulties in correctly identifying the requested symptom characteristics. Such a limitation, namely, how to bridge the gap between expert and users knowledge and perception is currently under investigation by means of a transfer learning approach [22].

Other considerations can be drawn from the results of the real-world challenge, as reported in Table 5. Firstly, it is evident how difficult it was for the user to switch from a simulated environment, where the proposed target diseases are controlled and somehow paradigmatic of the disease, to an in-field environment, where the real apple could present much more variability inducing an increased challenge for the user in perceiving the correct macroscopic symptoms. Thus, the results are generally worse than the one registered during the simulated challenge. Nevertheless, hybrid appears once more as the most effective method in all the measured metrics. Specifically, it achieves the largest increase in recall with respect to the user-made diagnosis (+23%, 54% against 31%), and it appears to be also slightly more precise (27% against 26%), despite having a larger average k (2 against 1.2), number of suggested diseases. Finally, we could notice how both components (*image* and *expert*) outscore the user diagnosis in terms of recall and F2-score. In particular, the expert model is the one which gets more benefits from the real-world environment, getting closer results to the ones of the image-based method (around 4% difference in recall and F2-score). This is probably due to the fact that a user is facilitated in using the expert-based stream of the system by having a real apple in her hands. In fact, she can fulfill a more careful inspection of the macroscopic symptoms on the fruit and, hence, improve her capability of distinguish the subtlety of the symptom characteristics requested by the system.

This is, at the best of our knowledge, the first quantitative evaluation, in the form of a user study, assessing the effectiveness of a decision support system in the field of post-harvest disease of apple diagnosis. A similar procedure was followed by Kolhe et al. [18] for the evaluation of an expert system for crops disease identification. A group of 20 agriculture under-graduate students were involved, to diagnose 8 test cases each, and producing a recall of 65%, which is in line with our findings. Other relevant applications of expert systems in agriculture domain [12,13,24,32], presented an evaluation based on user-reported effectiveness, usefulness, and usability, assessed by means of questionnaires, such as the well-known System Usability Scale (SUS) [6]. This kind of qualitative evaluation is outside the scope of this work, since it has already been investigated in previous publications [28,37].

Table 6
F2-score achieved by the four diagnostic methods, i.e., *image*, *expert*, *hybrid*, and *user* diagnosis, conditioned on the target ground truth disease, for the simulated challenge. The number of diagnoses taken for each disease (*#diag*) is also reported.

	#diag	*image*	*expert*	*hybrid*	*user*
alternaria_rot	14	**0.833**	0.357	0.417	0.667
alternaria_spot	15	0.833	0.278	0.833	**0.933**
black_rot	15	0.333	**0.556**	**0.556**	0.366
blue_mold	15	**0.389**	0.111	**0.389**	0.341
bitter_rot	14	**0.655**	**0.655**	**0.655**	0.608
bulls_eye	16	0.313	0.313	0.417	**0.506**
fusarium_rot	13	0.192	**0.385**	0.321	0.201
grey_mold	15	0.722	0.778	**0.833**	0.536
mucor_rot	15	0.611	0.556	**0.667**	0.542
side_rot	14	**0.774**	0.417	0.714	0.389

Table 7
F2-score achieved by the four diagnostic methods, i.e., *image*, *expert*, *hybrid*, and *user* diagnosis, conditioned on the target ground truth disease, for the real-world challenge. The number of diagnoses taken for each disease (*#diag*) is also reported.

	#diag	*image*	*expert*	*hybrid*	*user*
black_rot	7	0.357	**0.595**	**0.595**	0.286
blue_mold	16	**0.729**	0.156	0.677	0.366
bitter_rot	9	0.093	0.093	0.093	**0.392**
bulls_eye	5	0.333	0.333	0.333	**0.741**
fusarium_rot	3	0.0	0.278	0.278	**0.357**
grey_mold	18	0.417	**0.463**	**0.463**	0.217
side_rot	10	0.083	**0.417**	**0.417**	0.0

Tables 6 and 7 show how the four diagnostic methods perform in terms of F2-score with respect to each ground truth target disease, in the simulated and real-world challenge respectively. In the simulated challenge evaluation, reported in Table 6, it is easy to notice how the hybrid method emerges as the best performing approach for half of the diseases (5 out of 10), whereas image-based method outscores all other methods in two situations (for *alternaria_rot* and *side_rot*), and expert in a single case (for *fusarium_rot*). For 3 diseases the hybrid method is tied with one of the two simpler components (once with image, once with expert, and once with both). In two situations, for *alternaria_spot* and *bulls_eye* user-made diagnosis is better than the one performed by our DSS for a +10% F2-score. Important to mention that in the simulated challenge, the number of diagnosis provided for each target disease (*#diag*) is controlled by the system and thus it is balanced across the 10 different candidate diseases. We noticed that, when users are already quite confident about the diagnosed disease (like in the cases of the two *Alternaria* diseases or *bitter_rot*), the image-based version of the system achieves better results. A peculiar situation can be observed in the *alternaria_spot* case, where users made an effective diagnosis (93% of F2-score), with similarly high results for image-based diagnosis (83% of F2-score), while the expert method is scoring one of the lowest result, with just 28% F2-score. Vice versa, among the diseases which are the hardest to be identified by the user, the expert-based method shines. This happens, for instance, in the case of *fusarium_rot*, where the user scores 20% F2-score and the expert 38%, or in the case of *black_rot*, where the user scores 36% F2-score and the expert 56%.

Different insights can be derived from the F2-score results for the real-world challenge, presented in Table 7. As a limitation, just seven out of the ten candidate diseases were considered for the real-world challenge. In the time period of the challenge no apples infected by *Alternaria* spp. or *Mucor* spp. could be obtained from storage houses in the Bolzano province. It is important to note that for this study, the number of instances (*#diag*) for each disease is not balanced, but it is again subjected to natural constraints (i.e., the number of apples with each disease, available in our lab). In fact, two of the cases in which *DSSApple* performed worse than the user selection, the number of diagnoses taken by users is low (three for *fusarium_rot* and five for *bulls_eye*).

Hence, the results may be conditioned by the bias of such a limited test set. Indeed, *blue_mold* shows a very high performance of the image-based module (73% F2-score), while the expert suffers of poor results, below 20% (similarly as in the simulated challenge). *grey_mold* and *black_rot* report performances which are close to their behaviour in the simulated challenge, whereas *bitter_rot* registers a significant drop in the performances of all the DSS components, getting just a 9% F2-score. Finally, for *side_rot* the image-based component, as well as the user se-

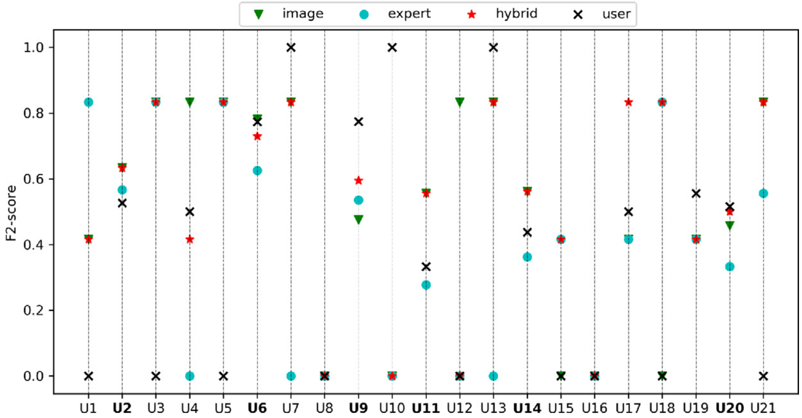


Fig. 8. F2-score achieved by the four diagnostic methods, i.e., *image*, *expert*, *hybrid*, and *user* diagnosis, conditioned on every user (anonymized on the x-axis) participating in the simulated challenge. In bold, we highlight users who made more than two diagnoses (median value).

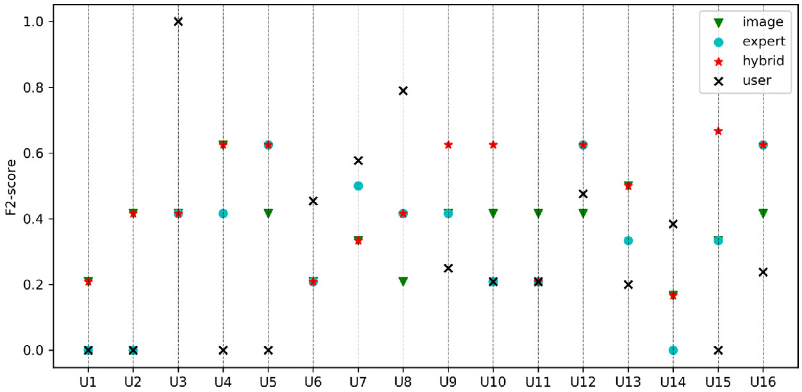


Fig. 9. F2-score achieved by the four diagnostic methods, i.e., *image*, *expert*, *hybrid*, and *user* diagnosis, conditioned on every user (anonymized on the x-axis) participating in the real-world challenge.

lection, got a very low F2-score, while the expert-based method kept the same performance of the simulated challenge.

To conclude, from this analysis conditioned on the target ground truth disease we could identify different kind of behaviour of *DSSApple* and user selection. For instance, *alternaria_spot* and *bulls_eye* seem to be easier to be diagnosed by the user than by our system. Vice versa, *blue_mold* and *alternaria_rot* can be effectively identified by the image-based method, while *black_rot*, *fusarium_rot*, and *grey_mold* required the mediation of an expert model. For some other diseases, like *bitter_rot* and *side_rot* we had conflicting evidence from the two evaluation scenarios. Nevertheless, we can generally conclude that hybridizing the two streams of information elicitation (i.e., image and expert) is an effective way to achieve a good trade-off for every target disease.

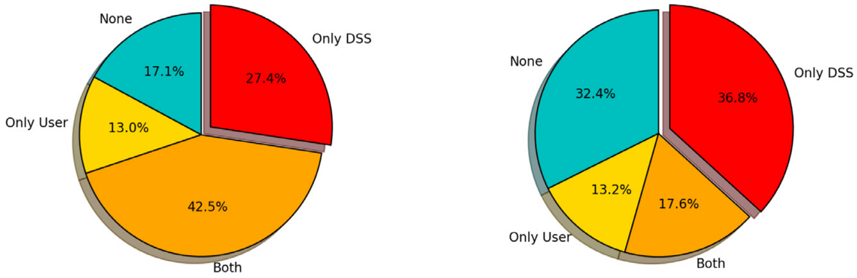
3.4. User-related results

In this section, we report the results related to the performance of each user in the diagnostic task. In particular, we aim at comparing the hybrid version of *DSSApple* with respect to the performances of its components (i.e., the image-based and the expert-based model) and the user capability of diagnosing apple diseases.

In Fig. 8, we depict the performances for each user involved in the simulated challenge. Specifically, the graph reports each user F2-score for the *image*-based, the *expert*-based, and the *hybrid* models, as well as the score for the *user*-made diagnosis. Important to mention that, among the 21 users just seven performed more than 2 diagnoses, namely U2, U6, U9, U11, U14, U20, and U21 (in bold in the graph). Hence, we should consider that the other 14 users might exhibit a larger variance in their results due to the limited number of interactions. Generally speaking, in 12 out of 21 cases at least one of the DSS methods significantly outperformed the user selection. Of these, in a single case (for U1) the best performing method was the expert one, in two cases (for U4 and U12) the best method was the image-based, while for the remaining nine participants the hybrid model achieved the best score, in most cases tied with one of the two simpler methods. For the users which outperformed *DSSApple*, just in five cases the score is significantly higher (greater than 0.1) with just one user with more than two interactions (U9). For the remaining users, the hybrid model was tied or close to the user-made diagnosis performance, including two degenerative cases in which the user was not able to get to any correct diagnosis (U8 and U16).

Similar considerations could be derived for the real-world challenge, as shown in Fig. 9. In this second challenge, we ensured that each user had at least four interactions (i.e., target apple diagnosis) with the system. Also notice that the user ids (e.g., U1, U2, etc.) do not correspond to the ones in the simulated challenge. From this real-world test, we can see how the hybrid model emerges to be the best performing diagnostic method for 10 out of 16 users. In three cases hybrid was tied with expert method, in four cases with image method and in three cases without any tie with simpler methods. For the remaining six users, five of them outperformed our DSS with their decision, while in a single scenario (U11) hybrid was tied with user decision, but image-based diagnosis alone performed better. Finally, we can conclude that for more than half of the tested users (more than two third in the real-world scenario) the *DSSApple* application is able to boost their capability in identifying the correct post-harvest apple disease.

This thesis is further supported by the pie charts represented in Fig. 10. These show the percentage of interactions where the user or the DSS were able to correctly identify the target post-harvest disease, in the simulated (Fig. 10a) and the real-world (Fig. 10b) challenge. For these graphs, we restricted the DSS diagnosis to the top-2 diseases selected by the hybrid model only. In the simulated challenge, the user alone was able to correctly identify the ground truth disease in more than 55% of the cases, while, with the help of *DSSApple* this percentage increased to around the 70%, with more than 27% of the tested



(a) Simulated challenge (b) Real-world challenge

Fig. 10. Percentage of correct diagnoses made by the DSS or the user, for the simulated (a) and the real-world challenge (b).

apples recovered from a wrong diagnosis. A proportion of 17% of the apples failed to be identified anyhow, while the user is losing 13% of correct diagnoses by trusting the DSS instead of her intuition. The same percentage of correct diagnoses made by the user only is shown in the real-world scenario. Nevertheless, in this case the user alone was able to correctly identify just a bit more than 30% of the diseases, while the system was able to boost this percentage up to 54%, with more than 36% of refined diagnoses. The increased difficulty of a real-world in-field diagnosis is testified by the fact that around one third of the apples were wrongly diagnosed with both decision models.

4. Conclusions

DSSApple is an interactive decision support system diagnosing post-harvest diseases of apple based on the observed macroscopic symptoms. The application is designed with a practical web-based interface to elicit information about the unknown disease on a target apple from both expert and non-expert users. Specifically, our application allows for a two-stream hybrid interaction, based on both expert-defined questions and visual stimuli referring to observable symptoms. The system has been thoroughly tested by means of two real-world experiments involving semi-expert users. They were challenged to use the *DSSApple* application in order to correctly diagnose two sets of infected apples, one set of apples simulated by photos, and the second set of real infected apples provided by storage houses. This evaluation proved the diagnostic effectiveness of the hybrid system, which generally performed better than its two single components (i.e., *expert-based* and *image-based*), as well as the user-made diagnosis as another baseline. In particular, we registered an increment of +14% in terms of recall and +8% in terms of F2-score for the simulation with apple images, and +23% recall and +15% F2-score for the physical apples, compared to the self-reported diagnosis of the users. In the real-world environment, hybrid *DSSApple* was able to correct the wrongly assigned diagnoses by the user in more than 36% of the test cases. Thus, we demonstrated that *DSSApple* is a valid tool to support both expert and non-expert users in the diagnosis of an apple infected by an unknown post-harvest disease. Furthermore, the presented methodology progressed the current state-of-the-art by (i) introducing an adaptive hybrid interface which leverages both images and expert-based questions to support both expert and non-expert users and to boost diagnostic accuracy, and (ii) presenting a BN-based reasoning system which is able to deal with uncertainty in knowledge elicitation and diagnosis computation. Finally, (iii) a practical algorithm able to explain the suggested diagnosis in the light of the feedback provided, was also introduced for the first time in such a context.

4.1. Future directions

Future directions of the presented work are manifold. For instance, a major limitation of the current approach is represented by the fact that the model has been built with the support of a single domain expert. A natural extension would be to involve more experts in a further refinement of the knowledge base. Hence, BN parameters will be derived from a panel of experts and will thus be more robust due to consensus among multiple experts [11]. More live experiments are also needed to further validate the developed model in more realistic environments (e.g., in packing-houses or in quality control). Another relevant issue of such a knowledge-based system is the one of transferability [22], namely how to effectively transfer the developed expert model to different environments (e.g., users with different expertise levels in the field of phytopathology). We are currently investigating an approach founded on the concept of *likelihood evidence* [26], to bridge the gap between the expert model and the users' perception in order to further improve the diagnostic performance. Finally, another development under consideration is to increase the reliability of the diagnosis by including additional sources of information such as automated classification of microscopic images of fungal growth.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the Free University of Bozen-Bolzano for funding the project “Development of a decision support system for the determination of postharvest disorders and diseases of apple – DSSApple” (Project Code IN2067 - ID Call 2017) and Greice Amaral Carneiro for providing part of the images and diagnoses necessary to establish the system.

References

[1] R.A. Baeza-Yates, B.A. Ribeiro-Neto, Modern Information Retrieval, ACM Press / Addison-Wesley, 1999. http://www.ischool.berkeley.edu/~hearst/irbook/glossary.html

[2] J. Barbedo, Expert systems applied to plant disease diagnosis: survey and critical view, IEEE Lat. Am. Trans. 14 (2016) 1910–1922, doi:10.1109/TLA.2016.7483534.

[3] R. Barkai-Golan, Postharvest Diseases of Fruits and Vegetables: Development and Control, Elsevier Science B.V., Amsterdam, Netherlands, 2001.

[4] O. Biran, C.V. Cotton, Explanation and justification in machine learning : asurvey, 2017.

[5] D.W. Boyd, M.K. Sun, Prototyping an expert system for diagnosis of potato diseases, Comput. Electron. Agric. 10 (3) (1994) 259–267.

[6] J. Brooke, “SUS-A Quick and Dirty Usability Scale.” Usability Evaluation in Industry, CRC Press, 1996. ISBN: 9780748404605

[7] G. Chen, H. Yu, Bayesian network and its application in maize diseases diagnosis, in: D. Li (Ed.), Computer And Computing Technologies In Agriculture, Volume II, Springer US, Boston, MA, 2008, pp. 917–924.

[8] M.J. Druzdzel, L.C. Van Der Gaag, Elicitation of probabilities for belief networks: combining qualitative and quantitative information, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 141–148.

[9] K.P. Ferentinos, Deep learning models for plant disease detection and diagnosis, Comput. Electron. Agric. 145 (September 2017) (2018) 311–318, doi:10.1016/j.compag.2018.01.009.

[10] A. Fuentes, S. Yoon, S.C. Kim, D.S. Park, A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition, Sensors (Switzerland) 17 (9) (2017), doi:10.3390/s17092022.

[11] L.C. Gaag, S. Renooij, C.L.M. Witteman, B.M.P. Aleman, B.G. Taal, How to elicit many probabilities, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 647–654.

[12] J. Gonzalez-Andujar, Expert system for pests, diseases and weeds identification in olive crops, Expert Syst. Appl. 36 (2, Part 2) (2009) 3278–3283, doi:10.1016/j.eswa.2008.01.007.

[13] L. Gonzalez-Diaz, P. Martínez-Jimenez, F. Bastida, J. Gonzalez-Andujar, Expert system for integrated plant protection in pepper (*Capsicum annuun* L.), Expert Syst. Appl. 36 (5) (2009) 8975–8979, doi:10.1016/j.eswa.2008.11.038.

[14] A.T. Ihler, J.W. Fischer III, A.S. Willsky, Loopy belief propagation: convergence and effects of message errors, J. Mach. Learn. Res. 6 (2005) 905–936.

[15] C.E. Kahn, L.M. Roberts, K.A. Shaffer, P. Haddawy, Construction of a Bayesian network for mammographic diagnosis of breast cancer, Comput. Biol. Med. 27 (1) (1997) 19–29, doi:10.1016/S0010-4825(96)00039-X.

[16] U.B. Kjaerulff, A.L. Madsen, Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, first ed., Springer Publishing Company, Incorporated, 2010.

[17] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324, doi:10.1016/S0004-3702(97)00043-X.

[18] S. Kolhe, R. Kamal, H. S. Saini, G. Gupta, A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops, Comput. Electron. Agric. 76 (1) (2011) 16–27, doi:10.1016/j.compag.2011.01.002.

[19] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, Adaptive Computation and Machine Learning, MIT Press, 2009. https://books.google.co.in/books?id=7dzpHCHzNQ4C

[20] P.M. Kuhnert, T.G. Martin, S.P. Griffiths, A guide to eliciting and using expert knowledge in Bayesian ecological models, Ecol. Lett. 13 (7) (2010) 900–914, doi:10.1111/j.1461-0248.2010.01477.x.

[21] C. Lacave, F.J. Díez, A review of explanation methods for Bayesian networks, Knowl. Eng. Rev. 17 (2) (2002) 107–127, doi:10.1017/S026988890200019X.

[22] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, G. Zhang, Transfer learning using computational intelligence: asurvey, Knowl. Based Syst. 80 (2015) 14–23.

[23] Y. Lu, S. Yi, N. Zeng, Y. Liu, Y. Zhang, Identification of rice diseases using deep convolutional neural networks, Neurocomputing 267 (2017) 378–384, doi:10.1016/j.neucom.2017.06.023.

[24] B. Mahaman, H. Passam, A. Sideridis, C. Yialouris, DIARES-IPM: a diagnostic advisory rule-based expert system for integrated pest management in solanaceous crop systems, Agric Syst. 76 (3) (2003) 1119–1135.

[25] P. Maxin, M. Williams, R.W. Weber, Control of fungal storage rots of apples by hot-water treatments: a northern European perspective, Erwerbs-Obstbau 56 (2014) 25–34.

[26] A.B. Mrad, V. Delcroix, S. Piechowiak, P. Leicester, M. Abid, An explication of uncertain evidence in Bayesian networks: likelihood evidence and probabilistic evidence - uncertain evidence in Bayesian networks, Appl. Intell. 43 (4) (2015) 802–824, doi:10.1007/s10489-015-0678-6.

[27] A. Niederkofler, S. Baric, G. Guizzardi, G. Sottocornola, M. Zanker, Knowledge models for diagnosing postharvest diseases of apples, in: Proceedings of the Joint Ontology Workshops 2019 Episode V: The Styrian Autumn of Ontology, Graz, Austria, September 23–25, 2019, in: CEUR Workshop Proceedings, vol. 2518, CEUR-WS.org, 2019. http://ceur-ws.org/Vol-2518/paper-ODLS6.pdf

[28] M. Nocker, G. Sottocornola, M. Zanker, S. Baric, G.A. Carneiro, F. Stella, Picture-based navigation for diagnosing post-harvest diseases of apple, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 506–507, doi:10.1145/3240323.3241616.

[29] D. O'Rourke, Economic Importance of the World Apple Industry, Springer International Publishing, Cham, 2021, pp. 1–18.

[30] L. Palou, J. Smilanick, Postharvest Pathology of Fresh Horticultural Produce, 2019. 10.1201/9781315209180.

[31] J. Pearl, G. Shafer, Probabilistic reasoning in intelligent systems: networks of plausible inference, Synthese-Dordrecht 104 (1) (1995) 161.

[32] I. Pertot, T. Kuflik, I. Gordon, S. Freeman, Y. Elad, Identificator: a web-based tool for visual plant disease identification, a proof of concept with a case study on strawberry, Comput. Electron. Agric. 84 (2012) 144–154.

[33] C. Pérez-Ariza, A. Nicholson, M. Flores, Prediction of coffee rust disease using Bayesian networks, in: Proceedings of the 6th European Workshop on Probabilistic Graphical Models, PGM 2012, 2012.

[34] S. Renooij, C. Witteman, Talking probabilities: communicating probabilistic information with words and numbers, Int. J. Approx. Reason. 22 (3) (1999) 169–194, doi:10.1016/S0888-613X(99)00027-4.

[35] J. Roach, R. Virkar, C. Drake, M. Weaver, An expert system for helping apple growers, Comput. Electron. Agric. 2 (2) (1987) 97–108, doi:10.1016/0168-1699(87)90020-2.

[36] D. Rose, W. Sutherland, C. Parker, M. Winter, M. Lobley, C. Morris, S. Twining, C. Ffoulkes, T. Amano, L. Dicks, Decision support tools for agriculture: towards effective design and delivery, Agric. Syst. 149 (2016) 165–174, doi:10.1016/j.agsy.2016.09.009.

[37] G. Sottocornola, S. Baric, M. Nocker, F. Stella, M. Zanker, Picture-based and conversational decision support to diagnose post-harvest apple diseases, Expert Syst. Appl. 189 (2022) 116052, doi:10.1016/j.eswa.2021.116052.

[38] G. Sottocornola, M. Nocker, F. Stella, M. Zanker, Contextual multi-armed bandit strategies for diagnosing post-harvest diseases of apple, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 83–87, doi:10.1145/3377325.3377531.

[39] D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen, R.G. Cowell, Bayesian analysis in expert systems, Stat. Sci. 8 (3) (1993) 219–247. http://www.jstor.org/stable/2245959

[40] Compendium of Apple and Pear Diseases and Pests, T.B. Sutton, H.S. Aldwinckle, A. Agnello, J.F. Walgenbach (Eds.), APS Press, 2014.

[41] N. Xie, G. Ras, M. van Gerven, D. Doran, Explainable deep learning: a field guide for the uninitiated, J. Artif. Intell. Res. 73 (2022) 329–396.

[42] C. Yialouris, A. Sideridis, An expert system for tomato diseases, Comput. Electron. Agric. 14 (1) (1996) 61–76, doi:10.1016/0168-1699(95)00037-2.

[43] F.M. Zanzotto, Human-in-the-loop artificial intelligence, J. Artif. Intell. Res. 64 (2019) 243–252.

[44] Z. Zhai, J.F. Martínez, V. Beltran, N.L. Martínez, Decision support systems for agriculture 4.0: survey and challenges, Comput. Electron. Agric. 170 (2020) 105256, doi:10.1016/j.compag.2020.105256.