Expert Systems With Applications 189 (2022) 116052

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Picture-based and conversational decision support to diagnose post-harvest apple diseases

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

[a] Faculty of Computer Science, Free University of Bozen-Bolzano, Piazza Domenicani, 3, 39100 Bolzano, Italy
[b] Faculty of Science and Technology, Free University of Bozen-Bolzano, Piazza Università, 5, 39100 Bolzano, Italy
[c] Dipartimento di Informatica, Sistemistica e Comunicazione, University of Milano-Bicocca, Viale Sarca, 336, 20126 Milano, Italy
[d] Faculty of Technical Sciences, University Klagenfurt, Universitaetsstrasse 65, 9020 Klagenfurt, Austria

ARTICLE INFO

Keywords:
Expert system in agriculture
Picture-based user interface
Conversational interaction
Contextual multi-armed bandit

ABSTRACT

This article presents the development of an expert system to support the diagnosis of post-harvest diseases of stored apples. We propose a picture-based and conversational interaction with users, where sampled images depicting symptoms of apples with known diseases are presented to users to elicit their feedback on perceived similarities in order to determine the most likely diagnosis of a diseased target apple. This article makes, besides the description of the industrial application scenario, multiple contributions circled around three rounds of user studies: (i) an usability and effectiveness assessment of the approach, where three user interface configurations are put to a test and the effectiveness of different types of user feedback mechanisms is assessed; (ii) contextual multi-armed bandit approaches for dynamic selection of displayed images with symptoms of diseased apples, that clearly outperform random and greedy sampling baseline strategies; (iii) a comparison of two different strategies for determining the context representation of a contextual multi-armed bandit approach, namely based on PCA of image features and a gamified large-scale user study. We therefore provide design insights for the development of such diagnosis applications on diseases that manifest themselves through visual symptoms in general and, hence, the findings can be also valid for domains other than post-harvest fruit diseases.

1. Introduction

The domesticated apple (*Malus x domestica*) is the third most produced fruit in the world (behind bananas and watermelons) according to FAO, with an amount of more than 87 million metric tons in 2019 (Shahbandeh, 2021). In the same year, annual worldwide shipments of apples were valued at more than 7 billion USD (Workman, 2020). Apple trees are indeed the most common temperate fruit tree species and, due to their good storage properties, apples can be stored for prolonged periods of time under controlled atmosphere conditions. Nevertheless, physiological disorders and pathogenic microorganisms can deteriorate the quality and quantity of the produce and lead to considerable economic losses (Sutton et al., 2014). Our goal is therefore to develop an application for interactive decision support (Marakas, 1998) that helps users to correctly diagnose a disease in due time (i.e., independently from the development and the advancement stage of the disease), in order to decide on the prevention and the management of post-harvest diseases in apples. For instance, it depends on the exact pathogen species to decide on the right strategy for immediate damage containment or to recommend a plant protection scheme for the following year. This characteristic of the system is in line with the recent definition of agricultural decision support systems as provided by Zhai et al. (2020). Namely, "an agricultural decision support system can be defined as a human–computer system which utilizes data from various sources, aiming at providing farmers with a list of advice for supporting their decision-making under different circumstances". In fact, in order to reliably determine the nature of the disease, several macroscopic symptoms, such as appearance, texture, consistency and colour of the lesion, as well as the odour of the decaying fruit need to be considered. Thus, the crucial aspect of this system is the user interaction, where an interactive and highly useable interface needs to incrementally incorporate the users' feedback about observed symptoms in order to effectively guide the diagnosis workflow. The importance of the design and other factors influencing the uptake of such a type of tools has been extensively investigated by Rose et al. (2016). Thus, we propose *DSSApple*, a picture-based and conversational application system for the diagnosis of post-harvest

* Corresponding authors.
E-mail addresses: gsottocornola@unibz.it (G. Sottocornola), sanja.baric@unibz.it (S. Baric), Maximilian.Nocker@stud-inf.unibz.it (M. Nocker), fabio.stella@unimib.it (F. Stella), mzanker@unibz.it (M. Zanker).

https://doi.org/10.1016/j.eswa.2021.116052
Received 17 July 2020; Received in revised form 23 June 2021; Accepted 5 October 2021
Available online 20 October 2021

diseases in apple fruits. Pictures depicting symptoms of the diseased fruit or plant are the predominant source of information in systems for automated disease diagnosis in agriculture. Images represent a type of data that can be easily processed by machine, as well as understood by (non-expert) humans. Moreover, our system is designed in such a way that the user is actively involved in the diagnostic process. An extensive discussion on the motivations that influenced this decision is presented in Section 4. The approach of interacting with users in a way that mimics the human-to-human dialog has been referred in the recommender systems literature as *"conversational"* (Christakopoulou et al., 2016). By adopting this paradigm, users do not need to specify all their preferences and constraints upfront within a single "static" step, but the information is elicited during an interactive "conversation" where the system should properly react to user feedback (Jugovac & Jannach, 2017). The presented core framework can be easily extended such that it can also include other sources of expert information in the reasoning mechanism. In future work, we will allow users to interact with the system using additional structured knowledge and evidence elicitation mechanisms, next to images. Furthermore, this knowledge base will enable the system to illustrate the actual diagnosis with additional information to better guide users in the decision-making process concerning the treatment of suggested diagnosis. Nevertheless, the core methodological aspects for *DSSApple* user interface and the interactive diagnostic system are fully described in this paper.

In this work we explore a wide range of aspects that are worth to be investigated in the context of *DSSApple*. Namely, we address the following research questions:

- R.Q.1 How effective and useable is a picture-based expert system on the task of diagnosing post-harvest diseases of apple?
- R.Q.2 Which is the impact of different interface configurations in terms of effectiveness and usability of this system?
- R.Q.3 How much could a conversational interaction paradigm improve the effectiveness of the expert system?
- R.Q.4 How could the symptoms be represented in a machine-readable manner in order to improve the capability of the system adapting to the user feedback?

The listed research questions are explored and evaluated by means of an extensive set of user studies. Thus, the resulting contribution that we present in this article is many-fold, namely:

- We illustrate the design of a picture-based expert system for the novel and challenging task of identifying post-harvest diseases of apple fruit.
- We describe an original adaptation of contextual multi-armed bandit algorithms (Agrawal & Goyal, 2013; Li et al., 2010) to produce a dynamic interaction with the user, following the conversational paradigm (Jugovac & Jannach, 2017).
- We discuss and analyse two different definitions of the image context, based on automated image analysis and collected user interactions.
- We present three large scale user studies conducted to thoroughly evaluate different aspects of the system, such as usability and effectiveness.

The rest of the paper is organized as follows. In Section 2 the related work in the field of expert systems and machine learning classification are presented. In Section 3 we illustrate the detailed system architecture. In Section 4 the design choices are discussed in light of our specific diagnostic tasks. In Section 5 we present the adopted methodology for the conversational interface and image processing. In Section 6 an extensive evaluation of the different aspects of our system is conducted. The user studies are described and their results are discussed, before drawing conclusions and identifying avenues for future work in the final Section 7.

2. Related work

In this section, we review the related work in the area of diagnostic decision support systems with a particular focus on the agricultural domain. For better readability, we structure this section such that in the first part, the literature in the area of knowledge-based expert systems is presented, while in the second part, the focus is laid on machine learning and image processing techniques for automated disease classification.

2.1. Knowledge-based expert systems

Expert and knowledge-based systems attracted the attention of researchers and practitioners since the late '80s (Plant & Stone, 1991). These types of systems address complex diagnostic and decision support tasks based on encoded domain knowledge. In the agricultural domain, expert systems have been widely adopted to cope with the diagnosis and the prevention of diseases and disorders largely affecting the production result (Barbedo, 2016).

In the early years, Roach et al. (1987) proposed *POMME*, an expert system aiming at helping apple growers to manage their orchards. The knowledge-based diagnostic module for a common apple disease was just a part of the full system that advised growers with respect to control of disorders and the management of pesticide selection and application. Another earlier prototype of a rule-based expert system was presented by Boyd and Sun (1994). The system supported the identification of 17 different potato diseases, both pathogenic and non-pathogenic, based on 127 prolog-like rules. The core of the reasoning system was composed of 8 knowledge bases, developed together by knowledge engineers and domain experts and evaluated in fielded demos. Post-harvest potato diseases were also the focus of the work presented by Adams et al. (1990). The interactive system included a knowledge base encoded in the form of *if-then* rules, and an expert system interface asking users specific questions in order to provide the desired diagnosis. The user interacted with the system by answering questions with "yes"/"no"/"I don't know" or any number of multiple choice responses. According to authors, when fed with the correct answers, the system was effective in correctly identifying all the 35 different disorders. A similar methodology was taken by Yialouris and Sideridis (1996) for supporting decision making for tomato diseases. Authors exploited an *object-attribute-value* formalism in order to model the knowledge base and enhance it with fuzzy logic in order to deal with the uncertainty. From the same research group, Mahaman et al. (2003) presented an advisory expert system, called *DIARES-IPM*, with an integrated knowledge base, designed as a set of *if-then* inference rules. The objective of the system was to help non-experts to identify pests in crops and to suggest appropriate treatments. In the same year, the architecture and the implementation of a large web-based fish disease diagnostic system was also introduced (Li et al., 2002). Another important contribution was provided by *EXSYS*, an Expert system for diagnosing flowerbulb diseases, pests and non-parasitic disorders (Kramers et al., 1998). The system consisted of two components: the knowledge base and the inference engine. The knowledge was represented by frame structures with embedded rules and the inference engine used a backward chaining approach to find the most probable hypothesis given the set of input symptoms. After an evaluation with researchers and domain experts the system was fielded within flower industry.

Later on, Kolhe et al. (2011) reported a web-based intelligent diagnostic system for oilseed-crops. The knowledge-based system built on fuzzy logic reasoning and supported an audio–visual–graphical user interface using also text-to-speech conversion tools. The system, which has been tested in three oilseed crops, resulted in drawing fast and acceptable diagnosis. Gonzalez-Andujar (2009) built an expert system for the identification of 9 weeds, 14 insects and 14 diseases in olive crops. Knowledge was gathered by a literature review and interviews

with experts, and the system adopted the conventional *if-then* knowledge representation. It was divided into three subsystems (diseases, pests and weeds), in such a way that only parts of the rules were active simultaneously. The system also employed 150 digital images to assist the user within the identification process. The same team of researchers proposed a similar system applied to the identification of weeds, insects, and diseases in pepper plants (Gonzalez-Diaz et al., 2009). A more recent example of an agricultural diagnostic system being supported by knowledge with images was the *Identificator* system (Pertot et al., 2012). The framework took advantage of macroscopic features (symptoms) of strawberries to diagnose potential diseases. Users thus selected a sequence of predefined images and descriptions of symptoms to finally get to the correct diagnosis. Our approach innovates these prior systems not only by building a diagnostic system for a novel application domain, namely post-harvest diseases of apple fruit, but by also realizing a conversational and picture-based interaction with users in order to effectively guide them.

2.2. Machine learning classification

The decision support problem in agriculture has been addressed in the scientific literature also through automated and machine learning methods, that do not require the mediation of the expert knowledge. Specifically, the major part of the presented techniques rely on image segmentation, image classification and feature extraction from images in order to support the diagnosis of the correct disease. Therefore, in the final part of this section the most influential deep learning techniques in this context are also covered.

One of the first examples is represented by the work by Pydipati et al. (2006) where a colour co-occurrence method was used to determine whether texture-based hue, saturation, and intensity colour features in conjunction with discriminant analysis classification allowed to distinguish between diseased and normal citrus leaves. Similarly, Camargo and Smith (2009) proposed a method, based on image transformations in order to identify visual symptoms of plant diseases. The transformed image is segmented by analysing the distribution in the intensities histogram. An alternative method to hue saturation intensity and rust colour index for segmenting infected areas in plants was investigated by Cui et al. (2010). Namely, soybean rust was detected by analysing the centroid of leaf colour distribution in the polar coordinate system. Greenness identification based on histogram analysis was also proposed (Romeo et al., 2013). The researchers investigated an additional approach that relies on fuzzy clustering, which achieved the best results. Another paper aimed at classifying different types of rice diseases by extracting features from the infected regions of the rice plant images (Phadikar et al., 2013). Fermi energy segmentation was used to isolate the infected region, then, diseases' symptoms were characterized using features like colour, shape and position of the infected portion. Finally, rules were mined exploiting these features. Other methods for leaf image segmentation relied on super-pixels in combination with Markov random fields (Ye et al., 2015) or K-means clustering (Zhang et al., 2018). Furthermore, Ma et al. (2017) presented a more complex image processing technique using colour information and region growing for segmenting greenhouse vegetable foliar disease spots images captured under real field conditions. Another method required the combination of colour histogram and textural features processed with principal component analysis for the detection of citrus diseases (Ali et al., 2017).

Other types of approaches were more focused on supervised classification. For example, artificial neural networks and decision trees were tested to classify the rottenness caused by *Penicillium* in citrus fruits (Gómez-Sanchis et al., 2012). Features were extracted from hyper-spectral images and selected through the Minimum Redundancy Maximal Relevance method. Finally classifiers were trained on manually labelled pixels to be assigned to the correct class. Automatic detection of citrus diseases was in the focus of another related work (Stegmayer et al., 2013). Traditional multi-class classification algorithms (i.e., decision tree classifier, multi layer perceptron, and naive bayes) were applied on engineered features to predict the correct disease. In the same direction went the work by Chaudhary et al. (2016), that proposed an improved version of random forest classifier for disease classification. Johannes et al. (2017) presented a novel image processing algorithm based on candidate hot-spot detection in combination with statistical inference methods to tackle wheat diseases identification in natural conditions, namely, with images taken in fields through mobile devices. A different methodology was illustrated in the paper by Zhang et al. (2017). After segmenting images with K-means clustering and extracting relevant features the leaf disease was classified using sparse representation.

More recently, the long wave of deep learning, pushed by the extraordinary successes in some computer vision tasks (He et al., 2016), swept also the field of agricultural diagnostic systems. One of the seminal works in the botanic field was *Deep-Plant*, a framework for plant species identification (Lee et al., 2015). A Convolutional Neural Network (CNN) was applied to learn features from leaf images in a unsupervised manner. An additional module of Deconvolutional Networks was employed as a visualization and explanation tool on the learned features. Following up on this, Sladojevic et al. (2016) presented a Deep Neural Network framework (again based on CNN) for the classification of 13 different types of plant diseases from leaf images. The same architecture was explored for the identification of potato (Oppenheim & Shani, 2017) and rice (Lu et al., 2017b) diseases. Fuentes et al. (2017) illustrated a robust deep learning framework for the real-time identification of tomato plant diseases and pests. Three different deep learning meta-architectures were tested and combined with two feature extractors. Image areas with disease symptoms were manually annotated and assigned to one of the 10 classes. Experiments on a large and heterogeneous image dataset proved the effectiveness of the proposed method. A similar procedure was performed by Lu et al. (2017a) for the in-field identification of wheat diseases on plant leafs. Finally, Ferentinos (2018) provided an extended work in which a comparison of different deep learning models was performed to predict 58 distinct classes of [plant, disease] combinations on a database of around 90 000 images.

3. System description

Our presented expert system, named *DSSApple*,[1] is designed to be an easy-to-use web application that allows also non-expert users to perform the diagnosis of apple diseases. The interaction with the system is conducted by simply clicking on pictures, representing the symptoms' variety of different diseases at the different stages of infection. The system interaction is conceptualized as a sequential process. At each round of interaction the user provides immediate feedback on a small set of images, depicting disease symptoms, based on the perceived similarity with the actually diseased target apple. In an earlier work (Nocker et al., 2018), we already demonstrated the usability of this design choice. Fig. 1 depicts a round of interaction, where users can provide feedback on any number of small-scale symptom images, before submitting their choices to the system. Given this setting, we could consider different types of feedback. An explicit *positive feedback* is provided by the user to suggest a strong similarity between the target apple and the depicted symptom. An explicit *negative feedback*, in contrast, suggests a (strong) dissimilarity between the target apple and the depicted symptom. Similarly, if a depicted symptom is *ignored* (i.e., neither positive nor negative feedback is given), the system also interprets it as a (weak) dissimilarity between the target apple and the depicted symptom.

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

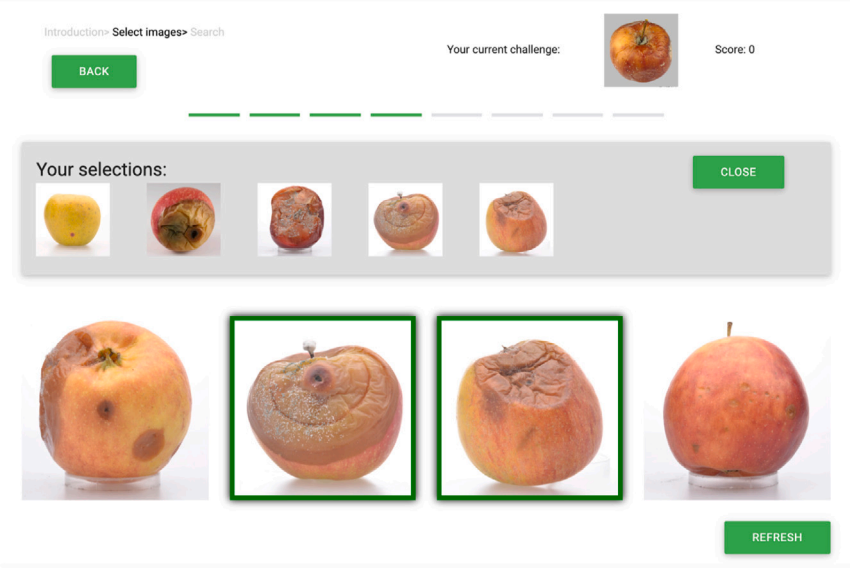

Fig. 1. A round of picture-based interactions with *DSSApple*.

The images of disease symptoms have been sampled from diseased apples at different storage houses. The ground truth (i.e., the actual disease) has been determined in a lab using microbiological and molecular diagnosis techniques. At the end of each round, the system automatically reloads alternative images based on the feedback provided by users. We induce some sort of "dialog" between the system (which suggests a set of images to be evaluated), and the user (who provides feedback with respect to the proposed pictures). Different reloading strategies can be applied on top of this interface, where a stratified random selection strategy, where images from candidate diseases are equally likely to be displayed, is the baseline. The goal is to develop an interactive "conversational" system that is able to adapt the selection of images to be displayed to users' previous feedback by exploiting similarities between the depicted candidate symptoms. The advantage of employing a conversational interaction over multiple rounds is two-fold: on the one side we reduce the cognitive load on users by letting them focus on a few depicted symptoms at a time; on the other side, we increase the capability of the system to adapt to user feedback. Indeed, after each round, we are able to elicit both explicit and implicit (i.e., images ignored) feedback to refine the belief of the system and propose more relevant images in the subsequent round. After a fixed number of rounds, the system stops feedback collection and recommends a set of candidate diseases that are ranked based on the received user feedback on the symptom images. An example is given in Fig. 2.

4. Discussion

For the task of diagnosing a post-harvest disease of apple, the employment of fully-automated machine learning techniques for image recognition appears to be insufficient up to now. The intra-disease variance is very high: 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 external symptoms, particularly at early stages of an infection.

In Fig. 3 we show three instances 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*) (Amaral Carneiro et al., 2021). 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.

Another desired property of our system emerges from the example shown in Fig. 3. The feedback that the user provides is not related to the actual disease of the selected images, but to the visual characteristics of the depicted symptoms. For instance, if the user selects the central image from the example in Fig. 3, she is requesting the system literally "show me more pictures of apples with a brownish circular rot of medium dimension at the top-side of the apple" instead of "show me more *Alternaria* images". In the next round, it is therefore more reasonable that *DSSApple* proposes the left-most image to be evaluated, rather than the right-most, even if this shares the same ground truth disease with the previously selected one. This mechanism allows the system to be more resilient to misleading feedback of users with respect to the actual ground truth disease and to guide them effectively towards a diagnosis.

Therefore, we believe that this problem needs to be approached by combining automated image processing with user-in-the-loop feedback elicitation. The decision support system thus guides the users towards correct diagnoses by proposing more refined choices, based on previous feedback and automated image-similarity computation.

5. Methodology

In this section we deepen the methodology behind the computational aspects of *DSSApple*. Namely, we present the *Contextual Multi-Armed Bandit* (CMAB) algorithm devoted to the sampling of images after each conversational round. Furthermore, we present and discuss the proposed method to produce a fast-to-compute contextual vector for each diseased apple image within the system.

5.1. Contextual multi-armed bandit

Foremost, we highlight that the bandit algorithm is not used as a predictive engine to directly map the user feedback to a diagnosis. This methodology is, instead, exploited to take advantage of user feedback and propose more relevant images (i.e., more significant for the current diagnosis task) in every future round of interaction. The multi-armed bandit also ensures a certain degree of exploration (i.e., tolerance to incorrect feedback). Thus, we formalize the sampling of the k images

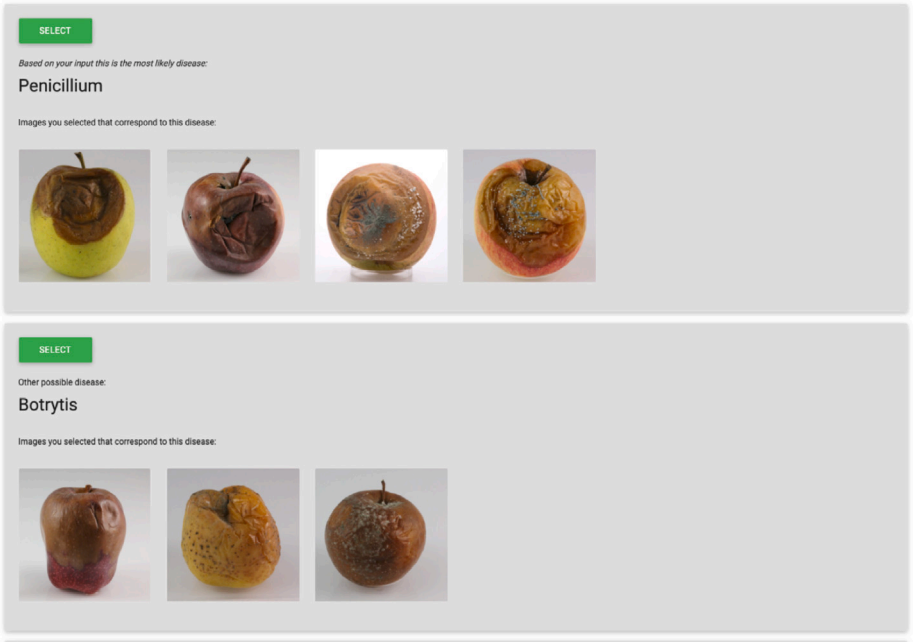

Fig. 2. List of recommended diagnosis.

Fig. 3. How difficult is it to classify the correct disease? — The left-most apple is infected by *Neofabrea*, while the others are infected by *Alternaria*.

per round as a CMAB problem. Each of the N candidate symptom images is associated with an arm. In analogy to human perception, the arms are not independent, but every arm, indexed by i, is associated with a d-dimensional vector $\boldsymbol{b}_i$ (i.e., the context) defined by the visual representation of symptoms termed features such as fructification of spores, shape and dimension of rotten spots, etc. We assume that a user, given an instance of an infected target apple to be diagnosed, embeds a d-dimensional parameter $\boldsymbol{\mu}^*$ which drives her selection strategy with respect to the d components of the context. Furthermore, when the ith image is shown to a user, she provides a stochastic reward r_i to the associated context $\boldsymbol{b}_i$, that could be positive (i.e., the image is selected) or negative (i.e., the image is ignored or discarded). Thus, we use a Gaussian likelihood function to update the belief of the system towards the optimal parameter $\boldsymbol{\mu}^*$. More formally, at each round t, the system updates a Gaussian multivariate distribution $\mathcal{N}(\hat{\boldsymbol{\mu}}(t), \boldsymbol{B}(t)^{-1})$, where the parameter $\hat{\boldsymbol{\mu}}(t)$ represents an online-learned approximation of $\boldsymbol{\mu}^*$ and $\boldsymbol{B}(t)$ is the variance–covariance matrix of the distribution. The posterior update of the parameters $\hat{\boldsymbol{\mu}}(t)$ and $\boldsymbol{B}(t)$ is achieved similarly to the Gaussian update described by Agrawal and Goyal (2013). Intuitively, the update of $\hat{\boldsymbol{\mu}}(t)$ depends on the user reward r_i given to every shown context vector $\boldsymbol{b}_i$ up to round t. Namely, a positive reward will move $\hat{\boldsymbol{\mu}}(t)$ into the direction of $\boldsymbol{b}_i$, while a negative one will move it into the opposite direction. We decided to consider also a negative reward to let the update function converge faster (i.e., with less interactions) to the optimal choice parameter $\boldsymbol{\mu}^*$. The equation for the iterative parameters update at time $t+1$ and are formally expressed as follows.

$$\boldsymbol{B}(t+1) = \boldsymbol{B}(t) + \boldsymbol{b}_i \boldsymbol{b}_i^{\mathrm{T}},$$

$$\boldsymbol{f}(t+1) = \boldsymbol{f}(t) + \boldsymbol{b}_i r_i,$$

$$\hat{\boldsymbol{\mu}}(t+1) = \boldsymbol{B}(t+1)^{-1} \boldsymbol{f}(t+1),$$

where $\boldsymbol{b}_i$ is the context associated to the ith image for which a new feedback reward r_i is provided at time $t+1$.

Thus, the problem of sampling a set of k arms $A(t+1) = \{a(t+1)_1, a(t+1)_2, \ldots, a(t+1)_k\}$ for round $t+1$ is translated into an optimization problem. A simple linear payoff is considered in our model: the expected reward for sampling arm $i = a(t+1)_j$, with $j = 1, \ldots, k$, associated with context $\boldsymbol{b}_i$ given the parameter $\hat{\boldsymbol{\mu}}(t)$ is computed as:

$$E[r(t+1)_i] = \boldsymbol{b}_i^{\mathrm{T}} \hat{\boldsymbol{\mu}}(t).$$

Finally, we select the k arms such that:

$$a(t+1)_j \leftarrow \underset{i}{\operatorname{argmax}}\, E[r(t+1)_i].$$

In the literature two main policies of CMAB are known that differ in how they achieve the exploration–exploitation trade-off. The two methods are *Thompson Sampling* (TS) (Agrawal & Goyal, 2013) and *Upper Confidence Bound* (UCB) (Li et al., 2010). In the former, $\tilde{\boldsymbol{\mu}}(t) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}(t), \boldsymbol{B}(t)^{-1})$, sampled from the Gaussian distribution at each round t, substitutes $\hat{\boldsymbol{\mu}}$ as an heuristic for the expected reward computation. The TS formulation of the expected reward becomes:

$$E[r(t+1)_i]_{TS} = \boldsymbol{b}_i^{\mathrm{T}} \tilde{\boldsymbol{\mu}}(t).$$

In the latter, the predictive standard deviation of the expected reward $\boldsymbol{b}_i^\intercal \hat{\boldsymbol{\mu}}(t)$, expressed by $\sqrt{\boldsymbol{b}_i^\intercal \boldsymbol{B}(t)^{-1} \boldsymbol{b}_i}$, is considered. The UCB formulation for the expected reward becomes

$$E[r(t+1)_i]_{UCB} = \boldsymbol{b}_i^\intercal \hat{\boldsymbol{\mu}}(t) + \alpha \sqrt{\boldsymbol{b}_i^\intercal \boldsymbol{B}(t)^{-1} \boldsymbol{b}_i},$$

where α is the only free parameter of the model which controls the exploration–exploitation trade-off.

5.2. Image-based context computation

In a CMAB setting the context is an important component for the model. In the proposed application, the desired property of the context is the one of representing images in a multidimensional space where closer vectors are mapped to symptoms that are considered visually similar by users. This is a non-trivial task, since similar symptoms may appear with varying shapes in different areas of the apple. Different cultivar types add additional variety to the appearance of symptoms. In principle the task of computing a reliable context should require the mediation of expert knowledge. On the other hand, we need to process hundreds of images and adapt our model to new incoming images from the lab, which makes the creation of expert-defined features hardly feasible. Thus, we adopt a fast to implement and easy to generalize method that simulates in an effective way the perception of users. The whole process is summarized in Fig. 4. We start with a set of RGB 1000×1000 images converted to grey-scale. We apply histogram normalization in order to maximize image contrast and equalize lighting. We resize the images to 32×32 pixels and flatten them into a set of vectors of dimension 1024, rescaled to $[0, 1]$. Finally, we apply a Principal Component Analysis (PCA) with L2-norm (Tipping & Bishop, 1999) in order to further reduce the dimensionality of the vectors while maintaining most of the information.

A similar approach for the automated computation of the context would employ deep learning techniques for image processing, namely Convolutional Neural Networks and Autoencoders (Goodfellow et al., 2016). We favoured the simpler approach described so far over a deep learning one, mainly due to the dimensionality of the dataset. With just a few hundred instances of apple disease symptoms it has been unfeasible to build a reliable highly parameterized model, able to learn a significant representation of the data structure. Finally, an alternative process for the context computation would be to exploit collected data about users' perceived similarities. This method aims to infer the similarities among different symptom images from past user interactions and derive the context from these observational data. This method has been implemented and will be described in Section 6.3.

6. Experiments

To validate the different aspects of *DSSApple*, we performed a set of experiments conducted in the form of user studies. They were designed in the form of a gamified challenge, called *Bad Apple Challenge* (Sottocornola et al., 2020), in order to encourage true engagement of participants. During the challenge users interacted with the application to accomplish the diagnostic task, and afterwards, the logged interaction data of users was analysed. The effects of gamification in boosting user interaction has been extensively studied by Hamari et al. (2014). The goal of the challenge is to simulate the in-field usage of the application without involving costly simulation with domain experts. These challenges allowed us to collect large-scale data and to evaluate the performances of the application system. We simulate the scenario in which a person needs to diagnose the disease of an apple in her hands. For this purpose, we randomly select an image of a target infected apple as a proxy for a real target apple to be diagnosed. The challenge participants have to use the *DSSApple* application to diagnose the potential disease of the target apple. A positive score is provided to users who manage to identify the correct disease. Given the flexibility of the system, we can allow participants to take multiple sessions of the challenge, where different target apples are requested to be diagnosed. We conducted the bad apple challenge within different small groups of people (i.e., with around 50 people). The participants were invited to enter the challenge at the same time and to complete it within a given time-frame. At its end we announced the leaderboard with the best scoring participants within the respective group.

In the remainder of this section three user studies, with different goals, are illustrated and the collected results are discussed. In the first set of trials, the impact of different interface invariants is analysed. In the second, the multi-armed bandit reloading strategies are compared with baseline strategies to test their effectiveness. Finally, in the third, the user perception of symptom similarity is measured in order to generate a better context for symptom images.

6.1. Experiments on user interface configurations

We first implemented a user study to measure the impact of different interface variants to assess effectiveness and usability of *DSSApple*. We utilized the *Bad Apple Challenge* with the abovely described task to identify the disease of a target image, randomly sampled from 12 different images taken from a web portal, called *Frudistor* (Zanella et al., 2021).[2] For this challenge, a set D of four different fungal diseases, namely the pathogen genera *Alternaria Botrytis*, *Mucor*, and *Penicillium* was employed. The main objective to be investigated was if the number of images per round, that are displayed to the user, influences the accuracy in diagnosing the correct disease. We employed in this experimental study design three different interface configurations that varied the number of images shown per each round. At the beginning of each challenge, one of these configurations is randomly assigned to the participant, ensuring a balance among the three experimental conditions of 4, 8 and 12 displayed images per round. However, for three conditions, the total number of displayed images was fixed to 24 in order to control for the potential amount of user feedback on displayed images for the diagnosis task and leading to a different numbers of rounds for each condition, namely, 4×8, 6×4, and 12×2 rounds. The images were randomly sampled from the pool of candidate images exploiting a stratified sampling over all different diseases. This approach ensured that during each round an equal number of images belonging to each of the four diseases is shown. For this first experiment we invited the user to explicitly provide both positive and negative feedback. The user should use the green '+' button to mark the images with similar symptoms to the target one and the red 'x' button to mark the images with totally different symptoms to the target one. The user is allowed to provide any number of feedback per round and submit its selection to forward the challenge to the next round (or to the end of it if the maximum number of rounds is reached). Fig. 5 shows a screenshot of the user interface with 12 images per round.

At the end of the selection rounds, a ranked list of diagnoses is presented to users. The ranking is based on the score for each disease, which reflects the feedback provided by the user. The score for disease $d_i \in D$ is computed as: $S(d_i) = \sum_j f_j(d_i)$. Where $f_j(d_i)$ represents a user feedback on the jth displayed image for disease d_i. If the feedback was positive $f_j(d_i) = +1$, if the feedback was negative $f_j(d_i) = -0.9$, and if no feedback was provided $f_j(d_i) = -0.1$.

6.1.1. Results

The experiment was performed within three classes of Computer Science students at the Free University of Bozen-Bolzano and the University of Milano-Bicocca, at the end of November and beginning of December 2018. 133 people participated in this user study, 18 were female and 115 were male. The average age of the participants was 23.5, the median age was 23. In total 305 challenges were completed

[2] http://www.frudistor.de.

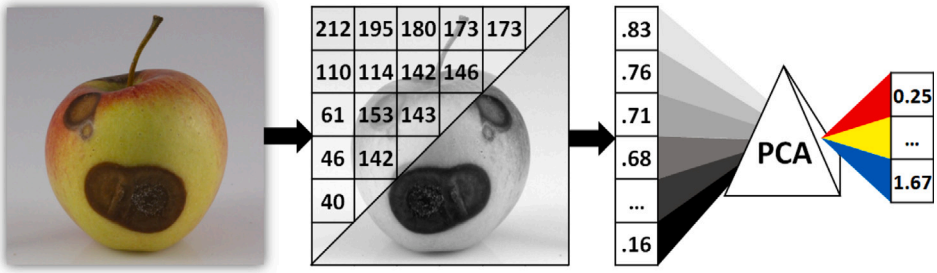

Fig. 4. Automatic context creation from images.

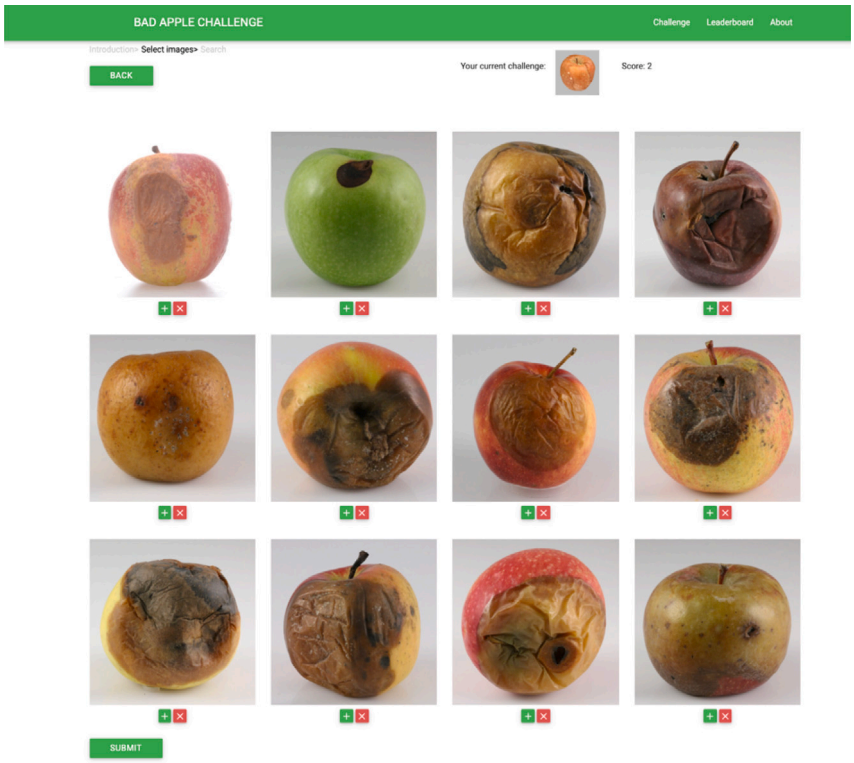

Fig. 5. User interface with 12 images per round.

by all participants, with around 2.3 challenges completed by each individual.

The collected results of the challenge present an overall success rate of the diagnosis of 44%. The results show that the different settings with 4, 8, and 12 images per round have no significant effect on the diagnosis outcomes. The user interface with 8 images has the highest success rate with 49%, followed by the 4 images with 43%. The lowest performance is registered by the setting with 12 images with 41%. In Table 1 we report the detailed results of the diagnosis task divided for the different configurations. We performed a χ^2-test on the contingency table to verify that there is no significant difference in the diagnosis performance among different interface configurations. As expected, the test does not reject the null hypothesis of the diagnosis outcome (*# success, # failures*) being independent of the setting (*4-images, 8-images, 12-images*) with a significance level of $p = 0.05$.

A *System Usability Score* (SUS) (Brooke, 1996) analysis was also performed by a subset of participants, in order to assess the usability of our application. A standard SUS questionnaire was submitted to 32 users after completing the challenge. The average score achieved by the system is above 73%, demonstrating the applicability and high usability of a picture-based approach in this novel application domain (Nocker et al., 2018).

Table 1
Challenge results over the different interface configurations.

	4-images	8-images	12-images
# success	49	42	43
# failures	65	44	62
success rate	0.43	0.49	0.41

In Fig. 6 we plot the fine-grained results for the precision and recall of the diagnosis for the four diseases. It becomes immediately clear that, in this setting, there is a strong bias with respect to the selected/target disease. *Alternaria* achieves by far the highest precision (around 85%) and recall (around 56%). This is an indication of the higher capability of a user to identify an *Alternaria* infection by its symptoms and clearly distinguish it from the 3 other diseases involved in this first study. Also *Botrytis* shows a similar recall of around 52% and a good precision, close to 50%. This means that a *Botrytis* infected apple was correctly diagnosed around half of the times, and users correctly indicated *Botrytis* as the right disease around half of the times. *Mucor* and *Penicillium*, instead, result in much worse performances. Namely, both diseases get a similar recall of around 31%; *Mucor* gets an acceptable precision of 38%, while *Penicillium* registers the worst

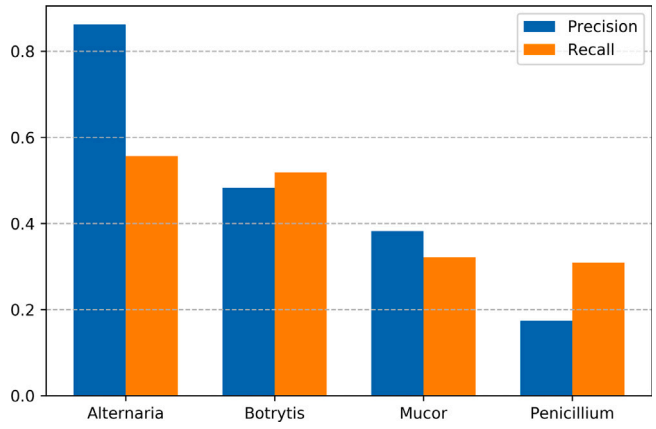

Fig. 6. Precision and recall achieved for each disease.

Table 2
Probability of a positive feedback on each disease conditioned on the target disease of the challenge.

TARGET	*FEEDBACK*			
	Alternaria	Botrytis	Mucor	Penicillium
Alternaria	**0.33**	0.16	0.22	0.29
Botrytis	0.13	**0.34**	0.23	0.30
Mucor	0.13	**0.31**	0.27	0.29
Penicillium	0.17	0.27	0.27	**0.28**

precision of 17%. The reason for this is two-fold: on one side *Penicillium* was least sampled as target disease during the challenge (namely, 30% less than the most frequently sampled disease); on the other side, it was most commonly selected by users, because its symptoms seem to be very similar to the external appearance of other diseases.

In order to better understand the decision biases of the user we had a deeper look into the feedback (i.e., positive and negative explicit feedback) provided by the user conditioned on the target disease $d_i \in D$ selected for the challenge. We compute the probability $P(d_i^+|T = d_i)$ of a user providing positive feedback on a candidate apple infected by disease d_i given that the target disease is $T = d_i$. The average probability of a correct positive feedback is 31%. This is significantly higher than the probability of a wrong positive feedback (i.e., a positive feedback on any disease other than the target one), which occurs on average 23% of the times. This is computed as $P(d_j^+|T = d_i)$, with $d_j \in D$ and $d_i \neq d_j$. In Table 2 we report the probability of a user providing positive feedback on each candidate disease given the target disease.

Table 2 depicts that there are clear differences in user perceptions depending on the target disease to be diagnosed. *Alternaria* and *Botrytis* seem to be the most distinguishable diseases, while they have a 33% and 34% chance to receive a true positive feedback by the user, while the probability of a false positive feedback is below 23% for all other diseases except for *Penicillium* (around 29%). Vice versa, *Mucor* has a higher chance to be confused for *Botrytis* (31%) and *Penicillium* (29%), than to be positively clicked itself (27%). Finally, *Penicillium* is hard to diagnose and also easy to be mistaken for other diseases. Namely, an image showing *Penicillium* symptoms has almost the same chance to be positively clicked (between 28% and 30%) independently of the target disease. Similar observations can be drawn from the computed probability of a user providing negative feedback on an image of disease $d_i \in D$, conditioned on the target disease $d_j \in D$. This is computed as the probability $P(d_i^-|T = d_j)$ of a user assigning negative feedback on an apple infected by disease d_i given that the target disease belongs to another class $T = d_j$, and $d_i \neq d_j$. The average probability of a correct negative feedback is 26%. Nevertheless, the average probability of a wrong negative feedback (i.e., providing negative feedback on images sharing the same disease with the target) $P(d_i^-|T = d_i)$ is above 22%. Both probabilities are very close to the random guess of 25%. Results for each disease are reported in Table 3.

Table 3
Probability of a negative feedback on each disease conditioned on the target disease of the challenge.

TARGET	*FEEDBACK*			
	Alternaria	Botrytis	Mucor	Penicillium
Alternaria	0.22	**0.29**	0.26	0.23
Botrytis	**0.32**	0.19	0.26	0.23
Mucor	**0.30**	0.22	0.24	0.24
Penicillium	**0.31**	0.23	0.22	0.24

Again, *Alternaria* appears to be the most distinguishable disease, since it receives a correct negative feedback in more than 30% of cases with another target disease. Also *Botrytis* is well identifiable, since it is the disease which has an even lower chance to get a wrong negative feedback (around 19%). Finally, *Mucor* and *Penicillium* register the worst performances (see third and fourth columns), given that they both have a 24% chance to be wrongly indicated as negative, when they actually are the target disease.

Furthermore, we measure the impact of positive and negative feedback on the final diagnosis outcome. The positive feedback was used 2944 times, while the negative feedback was used even 3872 times. However, in 82 challenges the negative feedback was not used, which represented only 27% of all challenges. In challenges where the negative feedback was used, the average success rate was 43%, while in challenges where no negative feedback was given, the average success rate was slightly higher with 48%. Considering these reported observations, we decided to discard the negative feedback from the future development of the system due to its marginal or even misleading effect. This decision will have the positive side effect on users to further alleviate their cognitive load.

6.2. Experiments on sampling algorithms

In order to understand the effectiveness of a multi-armed bandit sampling procedure for symptom images we designed another experiment. Based on the *Bad Apple Challenge* the participants' goal is to diagnose the correct disease of target apples. Here we controlled for five different fungal diseases, each one represented by three target image instances, depicting different disease symptoms taken again from the web portal *Frudistor* (Zanella et al., 2021) to avoid overfitting to our set of ground truth images. For each challenge one target image was randomly selected to simulate the real-world diagnostic task. A variety of 30 symptom images for each disease was selected by a pool of domain experts. These are displayed to participants over multiple rounds in order to elicit user feedback as mentioned. We generate a context of 64 dimensions, such that PCA is able to retain 95% of the variance of the original vectors. Due to the outcome of the experiments with different interface configurations, we decided on 8 rounds of interaction, where in each round a set of 4 images is needed to be evaluated by the user. Due to its ineffectiveness the negative feedback option was removed, the users could therefore either provide positive feedback (i.e., clicking on similar images) or ignore dissimilar images. The reloading strategy after each round is the main focus of investigation. We include two alternative CMAB policies, namely, UCB (with $\alpha = 1$) and TS, and two baselines, namely a random selection stratified over diseases, and a greedy policy, that fully exploits user feedback, without any exploration. The selection strategy is manipulated within participants and randomly selected for each challenge. The strategy determines the reloading of images for each round. At the end of the 8 rounds of feedback elicitation, the users are presented with a recommendation list of possible diagnoses, which are ranked based on the number of coherent feedback provided for each disease. From

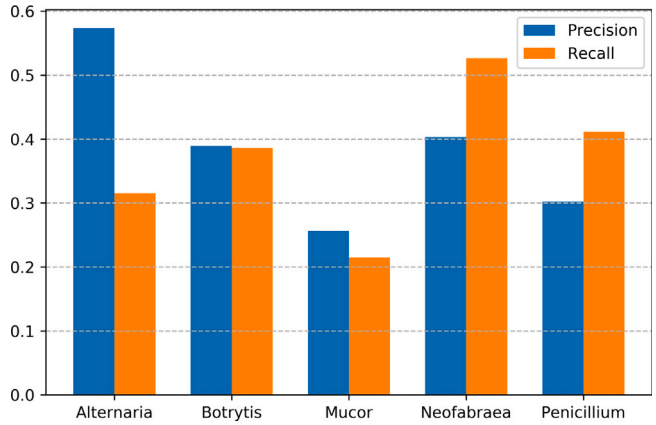

Fig. 7. Precision and recall achieved for each disease.

Table 4
Confusion matrix over the diseases.

SELECTED	*ACTUAL*				
	Alternaria	Botrytis	Mucor	Neofabraea	Penicillium
Alternaria	35	5	5	14	2
Botrytis	3	44	41	6	19
Mucor	8	7	20	12	31
Neofabraea	45	14	7	50	8
Penicillium	20	44	20	13	42

this list, users have to pick one disease as their final diagnosis. Note that participants can subsequently run multiple challenges and will therefore experience multiple selection strategies over different target images.

6.2.1. Results

We ran the controlled user study within two Computer Science bachelor classes and collected 163 participants in total consisting of 147 males and 16 females. The average age of the participants is 22.25, the median is 20. The total number of performed challenges is 591 (i.e., 3.6 challenges were on average performed by each individual). We cleaned the data by removing challenge entries where users did not provide any explicit feedback (i.e., they did not select any image). Thus, we ended up with 515 challenges from which we derived the following reported results.

Out of 515 challenges a total of 191 identified the correct diagnosis, with a success rate of 37%. In Table 4 we show the confusion matrix over the five diseases and in Fig. 7 we aggregate average precision and recall results by each disease.

It becomes clear from Table 4 that, depending on the disease to be diagnosed, a different degree of difficulty applies. For instance, *Alternaria* achieves the highest diagnostic precision of 57%, while the disease is easily confused with *Neofabraea*. *Botrytis* scores an average success rate of around 39% for both precision and recall and is easily confused with *Penicillium*. *Mucor* has been the hardest to diagnose, it is frequently mistaken for *Botrytis* and *Penicillium* and therefore scores the lowest average recall (21%) and precision (26%) values. Finally, *Neofabrea* is the disease with the highest recall (52%) and *Penicillium* the one that is frequently confused with other diseases (139 selections and a precision of 30%).

In Table 5 we present the results of the study for each algorithm. What clearly emerges from the online experiment is that the two CMAB sampling policies significantly improve the capability of the user to get to the correct disease; χ^2-test rejected the null hypothesis of the diagnosis outcome (*# success, # failures*) being stochastically independent of the algorithm (*TS, UCB, greedy, random*) with significance level $p = 0.05$. Specifically, UCB gets the best result with a 46% success rate,

Table 5
Results of the challenge per selection algorithm.

	TS	UCB	greedy	random
# success	55	53	42	41
# failures	80	63	78	103
success rate	0.41	**0.46**	0.35	0.28

Table 6
Precision for each disease conditioned on the sampling algorithm.

	TS	UCB	greedy	random
Alternaria	**0.85**	0.55	0.30	0.55
Botrytis	0.36	0.29	**0.58**	0.25
Mucor	0.21	**0.45**	0.07	0.24
Neofabraea	0.50	**0.60**	0.45	0.24
Penicillium	0.35	**0.48**	0.24	0.18

Table 7
Recall for each disease conditioned on the sampling algorithm.

	TS	UCB	greedy	random
Alternaria	**0.50**	0.33	0.43	0.23
Botrytis	0.41	**0.56**	0.32	0.40
Mucor	0.21	**0.27**	0.11	0.17
Neofabraea	0.54	**0.68**	0.47	0.43
Penicillium	0.41	**0.56**	0.32	0.40

followed by TS, which achieves a diagnostic success rate of 41%. These two results are significantly better than the 28% success rate that the system achieves with a simple random stratified reloading technique. Interesting to note, that a purely greedy strategy, that just exploits user feedback, does not outperform the bandit strategies (35% success). This is due to the fact that, in such a sensitive context where symptoms from different diseases can be easily confused, users are often misled to provide incorrect feedback. Thus, a purely exploitative strategy does not pay off and a specific degree of exploration is needed to, at least partially, alleviate the misleading effect.

A more fine-grained analysis of the results is provided with the computation of the diagnosis performances for each disease class conditioned on the algorithm used during the challenge.

In Table 6 we summarize the precision values for each disease conditioned on the sampling algorithm. UCB is the best performing algorithm for three diseases (i.e., *Mucor*, *Neofabraea*, and *Penicillium*), Thompson Sampling is the most precise one for a single disease (i.e., *Alternaria*) likewise the greedy algorithm for *Botrytis* disease. The worst algorithms for each disease are three times Stratified Sampling and twice the Greedy algorithm.

In Table 7 we report the recall for each disease conditioned on the sampling algorithm. UCB is the best algorithm in recalling the correct disease for all the diseases except for *Alternaria*, where Thompson Sampling achieves best recall values. The worst algorithm is three times Stratified Sampling and once the Greedy algorithm. These results confirm that the proposed contextual multi-armed bandit algorithms clearly outperform the baselines in the identification of the correct diagnosis independently from the disease taken under consideration.

6.3. User study for context computation

While in the previous study we got confirmation that the contextual multi-armed bandit strategies are superior to the baseline algorithms, we designed an additional user study to further the develop the context representation. This study seeks to make the contextual representation even more accurate in mimicking the users' perceptions of similar symptoms. We adapt the *Bad Apple Challenge* application to let participants focus more on the similarity of symptoms, rather than identifying the correct disease. The challenge is organized in a single-round trial, in which the task is to identify the images that depict symptoms similar

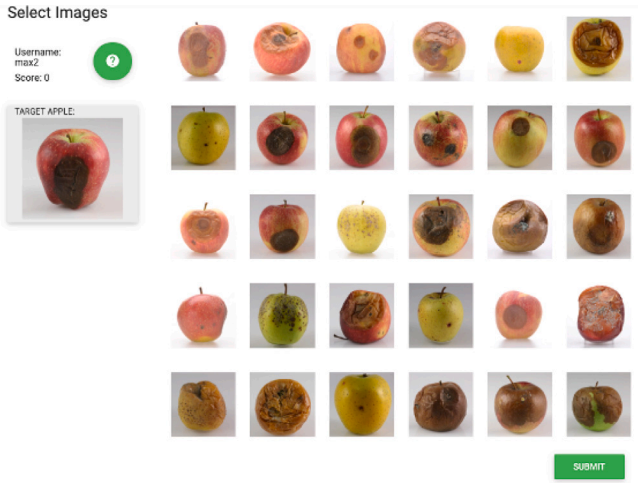

Fig. 8. System interface for context computation challenge.

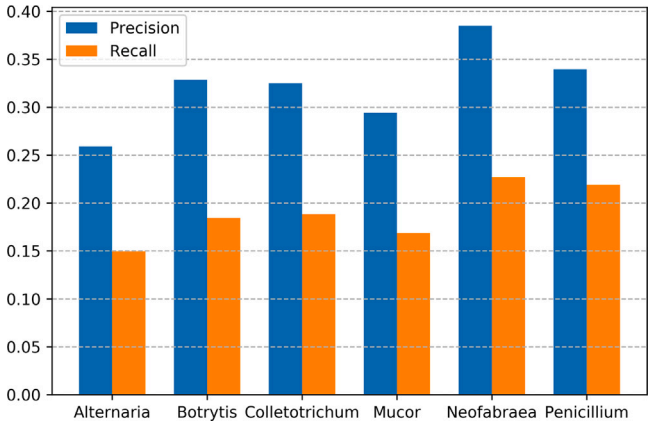

Fig. 9. Precision and recall of user feedback, achieved for each disease.

to the randomly selected target one. The user can select any number of symptoms, she considers related to the target one, and then submit her choices. After the submission, a score is computed based on the number of coherent symptoms guessed (i.e., symptoms belonging to the same disease of the target image). We assign +1for each correct image clicked and −0.2 for each wrong one, thus, discouraging random clicking to boost the score. The application allows the participants to do the challenge multiple times with new randomly drawn target images for each trial. The cumulative score for each participant is reported in a leaderboard. At the end of a time period of several days the top3 participants were awarded with a symbolic prize.

The number of candidate images that are shown in each challenge and can be selected is 30, organized in a 6×5 matrix of small-scale images, as depicted in Fig. 8. For this challenge, we added a sixth disease to the system (i.e., *Colletotrichum*), in addition to the five diseases of the previous study (i.e., *Alternaria*, *Botrytis*, *Mucor*, *Neofabraea*, and *Penicillium*). Each disease is represented by a set of 30 images, thus, the total number of symptom images is 180. We sampled from this set for both the target image and the candidate images, since our goal is to estimate the similarity of the symptoms as they are perceived by the users. We apply a random stratified sampling strategy over the diseases in order to display five random images of every disease during each challenge. Furthermore, we remove the a priori bias by ensuring that each pair of target-candidate image is displayed in a balanced way across all challenges. Namely, the pairs that have been selected fewer times up to now have higher chances to be sampled next. By doing simple math, we identify that the total number of possible pairs is 32 220. Given that in a single challenge 30 candidate images are shown (i.e., 30 pairs with the target image), the number of challenges to be completed in order to have each pair to be shown at least once is therefore 1074.

6.3.1. Results

The contest was announced in the Free University of Bozen-Bolzano newsletter, where an email was sent to all the students (i.e., approximately 4200 people) inviting them to take part in the challenge for the period of two weeks, from the 2019-06-03 to the 2019-06-17. In total, 175 people participated in the user study and 4357 challenges were completed. On average around 25 challenges were completed by each user. Nearly one-third of all challenges were made by two participants. The median age of the participants is 23 years and their mean age is 24.3 years.

We report the results of this user study in terms of precision and recall of the user feedback. We define precision as the percentage of correct feedback (i.e., click on a candidate image that belongs to the same disease of the target image) on the total number of user feedback. We define recall as the percentage of correct feedback on the total number of relevant images (i.e., the number of candidate images that belong to the same disease of the target image). At the end of the data collection of this user study, we measured an overall recall of 19% and an overall precision of 32%. In Fig. 9 we summarize the results of precision and recall for the six diseases. The images clicked with least precision are the one related to *Alternaria* (26%). The other diseases achieve an average precision above 29%, with a maximum result of 38% for *Neofabraea*. Similar results are registered for the recall. Again, *Alternaria* is the hardest disease to correctly identify for the user, with a recall of 15%. A set of three diseases, namely *Botrytis*, *Colletotrichum*, and *Mucor*, get a similar recall between 16% and 19%. *Neofabraea* and *Penicillium* achieve the highest recall, around 22%.

6.3.2. Context improvement

Consequently, the goal of this third round of user study was to create a reliable context for the symptom images, i.e., a context representation that properly mimics the users' perceptions of similar symptoms. Given the number of challenges completed (i.e., 4357) and the balanced policy adopted to sample the pairs, we ensured that each pair of images target-candidate was shown at least four times. Given the log data of the user study, the similarity matrix of each pair of target-candidate symptom image is computed as $S(i)_j = click(i)_j / show(i)_j$. Where $S(i)_j$ represents the similarity of candidate image i to the target image j, $click(i)_j$ represents how many time image i was clicked given image j as a target, and $show(i)_j$ represents how many times image i was shown given image j as a target. Please notice that this similarity matrix is not symmetric, since $S(i)_j \neq S(j)_i$. To remove sparsity from the context matrix we applied PCA and retain 64 principal components, similarly to the procedure applied for the context derived from image processing. We ended up with an unbiased and more reliable context, able to represent the symptoms' similarity as perceived by the user.

In order to prove this assumption we empirically compare the effectiveness of the new context (i.e., *user-based context*) with respect to the one described in Section 5.2, fully based on image processing (i.e., *image-based context*). We build a similarity matrix for the two context representations, by applying cosine similarity on the two context matrices. Thus, we are able to rank the images according to their similarities with respect to a target one. We randomly selected five target images and retrieved the five most similar images for each, according to the two context representations.

The qualitative improvement of the new context emerges clearly in this analysis when comparing Figs. 10 with 11. In some cases, namely for T1 and T4, the image-based context in Fig. 10 struggles to capture the symptom appearance. A dark-red apple could be mistaken by a rotten one, and the similarity computation relies more on the external

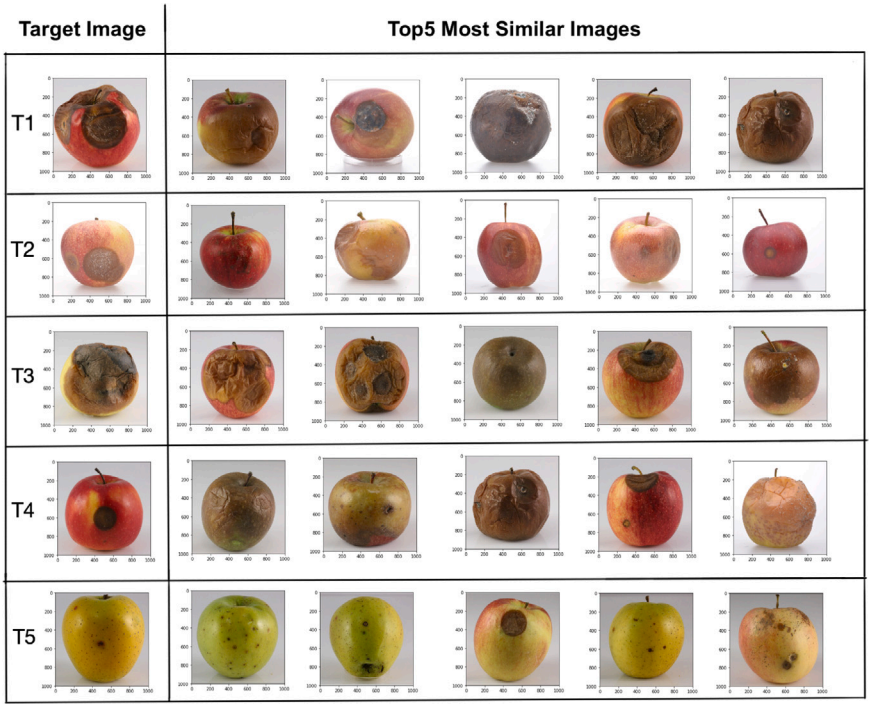

Fig. 10. Most similar images by image-based context for five target images.

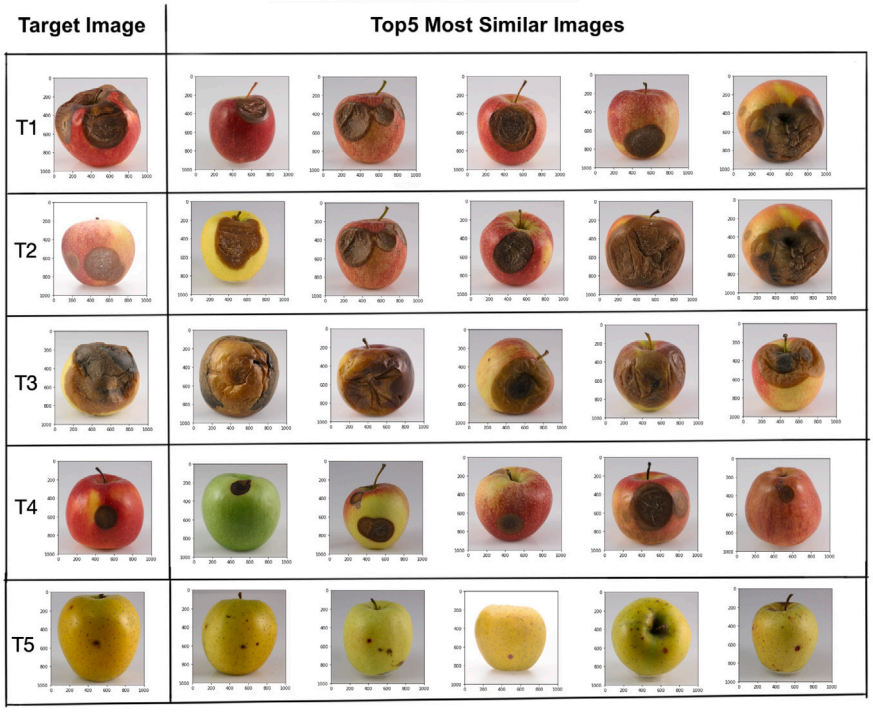

Fig. 11. Most similar images by user-based context for five target images.

shape of the apple other than the symptom appearance. Nevertheless, in some cases even the image-based context achieves good results, in terms of symptom similarity and cultivar diversity (see for instance T3 and T5). Another interesting aspect is caught by the example of T2: the image-based context computation is somehow dependent on the enlightenment conditions. In this case all the retrieved results, except for the first one, show a very strong enlightenment. With the user-based context the qualitative improvement in the capability of represent symptom characteristics is evident (see Fig. 11). T4 presents the most interesting results: retrieved images share the same symptom appearance (i.e., a black round slightly-sunken rot), independently from the cultivar or the specific position of the rot. This capability is hard to achieve with a fully automated image-processing technique. Also the enlightenment problem is not present in this case as demonstrated by

example T2. Nevertheless, also in this case some shortcomings of the context computation emerge. For instance, the last retrieved images in the top5 list for T1 and T2 do not represent fully coherent symptoms. This may be due to the struggle, also for the users, to identify properly related symptoms during the challenge.

7. Conclusions

We presented *DSSApple*, a conversational picture-based expert system aiming at supporting users in the identification of post-harvest apple diseases. The system is designed as an easy-to-use web application that allows also non-expert users to perform the diagnosis task. The interaction with the system is conducted by the user clicking on pictures, representing the symptoms variety of different diseases at different stages of infection. We conceptualized an interactive diagnostic session with the system as a sequence of conversational rounds to avoid cognitive overload of users. At each round of interaction, the user provides immediate feedback on a small set of images, depicting disease symptoms, based on the perceived similarity with the actually diseased target apple. The sequence of rounds allows the system to incrementally refine its knowledge towards the characteristics of the infected apple. Thus, the feedback is exploited by the system to "dialog" with the user by proposing more relevant images in subsequent rounds, while keeping a certain degree of exploration. This goal is achieved through an ad-hoc contextual multi-armed bandit algorithm for sampling of images. We described alternative methods to construct a reliable image context, based on both image processing and user interactions data. The effectiveness of the diagnosis support is evaluated throughout three user studies, designed as gamified challenges, in which users compete in identifying the correct diseases of target apples by the help of our system. In the first experiment, different interface configurations are tested in order to define the best setup (i.e., the one who leads to better precision, while minimizing the cognitive overload for the user). In the second experiment, the multi-armed bandit reloading strategy is compared against greedy and random baselines. The two contextual multi-armed bandit algorithms were proved to significantly outperform the baselines. The third and last user study is conducted in order to construct more reliable image contexts, based on the user perceived similarity among symptom images.

The proposed approach could be easily adapted and extended to similar domains where a dynamic interaction of the user with an image-based knowledge base could be exploited for supporting disease diagnosis. For instance, in agriculture domain, for the diagnosis of rice (Lu et al., 2017b), wheat (Lu et al., 2017a), potato (Oppenheim & Shani, 2017), citrus (Ali et al., 2017), tomato (Fuentes et al., 2017), olive (Gonzalez-Andujar, 2009), strawberry (Pertot et al., 2012) diseases, or in human healthcare, for the diagnosis of dermatological diseases (Prabhu et al., 2019) or cancer detection (Lee & Chen, 2015). Furthermore, the methodological findings and insights are valid and could be exploited in totally different contexts, where the interaction with the user is conducted by means of visual or graphical stimuli, to dynamically elicit her preferences. For example, in the area of tourism recommendation systems (Neidhardt et al., 2015; Ricci et al., 2005), or in image-mediated retrieval systems (Goodrum, 2000).

The main open challenge for the future development of the system concerns the injection of structured expert knowledge to the reasoning mechanism of *DSSApple*. The direction of investigation so far concern the translation of a crafted domain ontology (Niederkofler et al., 2019), into a Bayesian Network (Koller & Friedman, 2009). The Bayesian Network will allow the system to include incremental expert-based evidence provided by the user in order to refine the online belief over each disease. The reasoning mechanism is based on Bayesian inference to deal with knowledge uncertainty. An engine founded on mutual information will be capable of identifying the most discriminative information to be asked to the user at a given point in time (i.e., based on the acquired knowledge so far). Furthermore, the likelihood computation of the evidence provided by the user in the light of the suggested diagnosis, will naturally converge in a expert-based explanation, increasing the transparency of the system and, hence, better supporting the decision-making process of the user. The Bayesian model is under refinement by means of interviews with domain experts (Sottocornola et al., 2021). This step is required to quantify a set of conditional probability tables representing the likelihood of observing certain symptoms under specific diseases and degrees of infection.

CRediT authorship contribution statement

Gabriele Sottocornola: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing. **Sanja Baric:** Supervision, Resources, Data curation, Project administration, Funding acquisition, Writing – review & editing. **Maximilian Nocker:** Software, Investigation, Data curation. **Fabio Stella:** Methodology, Supervision, Writing – review & editing. **Markus Zanker:** Supervision, Validation, Project administration, Funding acquisition, Writing – review & editing.

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.

Acknowledgement

The authors acknowledge the Free University of Bozen-Bolzano, Italy for funding the project "Development of a decision support system for the determination of postharvest disorders and diseases of apples – DSSApple" (Project Code IN2067 - ID Call 2017) and Greice Amaral Carneiro for providing part of the images and diagnoses necessary to establish *DSSApple*. This work was supported by the Open Access Publishing Fund of the Free University of Bozen-Bolzano, Italy.

References

Adams, S. S., Stevenson, W. R., Delhotal, P., & Fayet, J. (1990). An expert system for diagnosis of post-harvest potato diseases. *EPPO Bulletin*, *20*(2), 341–347. http://dx.doi.org/10.1111/j.1365-2338.1990.tb01217.x.

Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th international conference on international conference on machine learning* (pp. 127–135). JMLR.org.

Ali, H., Lali, M. I., Nawaz, M. Z., Sharif, M., & Saleem, B. A. (2017). Symptom based automated detection of citrus diseases using color histogram and textural descriptors. *Computers and Electronics in Agriculture*, *138*, 92–104. http://dx.doi.org/10.1016/j.compag.2017.04.008.

Amaral Carneiro, G., Walcher, M., Storti, A., & Baric, S. (2021). Phylogenetic diversity and phenotypic characterization of Phlyctema vagabunda (syn. Neofabraea alba) and Neofabraea kienholzii causing postharvest bull's eye rot of apple in northern Italy. *Plant disease*, http://dx.doi.org/10.1094/pdis-04-21-0687-re.

Barbedo, J. (2016). Expert systems applied to plant disease diagnosis: Survey and critical view. *IEEE Latin America Transactions*, *14*, 1910–1922. http://dx.doi.org/10.1109/TLA.2016.7483534.

Boyd, D. W., & Sun, M. K. (1994). Prototyping an expert system for diagnosis of potato diseases. *Computers and Electronics in Agriculture*, *10*(3), 259–267.

Brooke, J. (1996). Usability evaluation in industry. CRC Press.

Camargo, A., & Smith, J. S. (2009). An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosystems Engineering*, *102*(1), 9–21. http://dx.doi.org/10.1016/j.biosystemseng.2008.09.030.

Chaudhary, A., Kolhe, S., & Kamal, R. (2016). An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*, *3*(4), 215–222. http://dx.doi.org/10.1016/j.inpa.2016.08.002.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 815–824). New York, NY, USA: ACM, http://dx.doi.org/10.1145/2939672.2939746.

Cui, D., Zhang, Q., Li, M., Hartman, G. L., & Zhao, Y. (2010). Image processing methods for quantitatively detecting soybean rust from multispectral images. *Biosystems Engineering*, *107*(3), 186–193. http://dx.doi.org/10.1016/j.biosystemseng.2010.06.004.

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*(September 2017), 311–318. http://dx.doi.org/10.1016/j.compag.2018.01.009.

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

Gómez-Sanchis, J., Martín-Guerrero, J. D., Soria-Olivas, E., Martínez-Sober, M., Magdalena-Benedito, R., & Blasco, J. (2012). Detecting rottenness caused by Penicillium genus fungi in citrus fruits using machine learning techniques. *Expert Systems with Applications*, *39*(1), 780–785. http://dx.doi.org/10.1016/j.eswa.2011.07.073.

Gonzalez-Andujar, J. (2009). Expert system for pests, diseases and weeds identification in olive crops. *Expert Systems with Applications*, *36*(2), 3278–3283. http://dx.doi.org/10.1016/j.eswa.2008.01.007.

Gonzalez-Diaz, L., Martinez-Jimenez, P., Bastida, F., & Gonzalez-Andujar, J. (2009). Expert system for integrated plant protection in pepper (Capsicum annuun L.). *Expert Systems with Applications*, *36*(5), 8975–8979. http://dx.doi.org/10.1016/j.eswa.2008.11.038.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodrum, A. (2000). Image information retrieval: An overview of current research. *Informing Science*, *3*(2), 63–66.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - a literature review of empirical studies on gamification. In *47th Hawaii International Conference on System Sciences (HICSS)* (pp. 3025–3034). IEEE Computer Society.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

Johannes, A., Picon, A., Alvarez-Gila, A., Echazarra, J., Rodriguez-Vaamonde, S., Navajas, A. D., & Ortiz-Barredo, A. (2017). Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. *Computers and Electronics in Agriculture*, *138*, 200–209. http://dx.doi.org/10.1016/j.compag.2017.04.013.

Jugovac, M., & Jannach, D. (2017). Interacting with recommenders - overview and research directions. *ACM Trans. Interact. Intell. Syst.*, *7*(3), 10:1–10:46.

Kolhe, S., Kamal, R., S. Saini, H., & Gupta, G. (2011). A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops. *Computers and Electronics in Agriculture*, *76*(1), 16–27. http://dx.doi.org/10.1016/j.compag.2011.01.002.

Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. *Adaptive computation and machine learning*, MIT Press.

Kramers, M., Conijn, C., & Bastiaansen, C. (1998). EXSYS, an expert system for diagnosing flowerbulb diseases, pests and non-parasitic disorders. *Agricultural Systems*, *58*(1), 57–85. http://dx.doi.org/10.1016/S0308-521X(98)00046-8.

Lee, S. H., Chan, C. S., Wilkin, P., & Remagnino, P. (2015). Deep-plant: Plant identification with convolutional neural networks. *Proceedings - International Conference on Image Processing, ICIP*, *December*, 452–456. http://dx.doi.org/10.1109/ICIP.2015.7350839.

Lee, H., & Chen, Y.-P. P. (2015). Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications*, *42*(12), 5356–5365. http://dx.doi.org/10.1016/j.eswa.2015.02.005.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on the world wide web (WWW)* (pp. 661–670). New York, NY, USA: ACM, http://dx.doi.org/10.1145/1772690.1772758.

Li, D., Fu, Z., & Duan, Y. (2002). Fish-expert: a web-based expert system for fish disease diagnosis. *Expert Systems with Applications*, *23*(3), 311–320. http://dx.doi.org/10.1016/S0957-4174(02)00050-7.

Lu, J., Hu, J., Zhao, G., Mei, F., & Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. *Computers and Electronics in Agriculture*, *142*, 369–379. http://dx.doi.org/10.1016/j.compag.2017.09.012.

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

Ma, J., Du, K., Zhang, L., Zheng, F., Chu, J., & Sun, Z. (2017). A segmentation method for greenhouse vegetable foliar disease spots images using color information and region growing. *Computers and Electronics in Agriculture*, *142*, 110–117. http://dx.doi.org/10.1016/j.compag.2017.08.023.

Mahaman, B., Passam, H., Sideridis, A., & Yialouris, C. (2003). DIARES-IPM: a diagnostic advisory rule-based expert system for integrated pest management in Solanaceous crop systems. *Agricultural Systems*, *76*(3), 1119–1135.

Marakas, G. M. (1998). *Decision support systems in the twenty-first century*. USA: Prentice-Hall, Inc..

Neidhardt, J., Seyfang, L., Schuster, R., & Werthner, H. (2015). A picture-based approach to recommender systems. *Journal of IT & Tourism*, *15*(1), 49–69.

Niederkofler, A., Baric, S., Guizzardi, G., Sottocornola, G., & Zanker, M. (2019). Knowledge models for diagnosing postharvest diseases of apples. In *Proceedings of the joint ontology workshops (JOWO)*. http://ceur-ws.org/Vol-2518/paper-ODLS6.pdf.

Nocker, M., Sottocornola, G., Zanker, M., Baric, S., Carneiro, G. A., & Stella, F. (2018). Picture-based navigation for diagnosing post-harvest diseases of apple. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 506–507). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3240323.3241616.

Oppenheim, D., & Shani, G. (2017). Potato disease classification using convolution neural networks. *Advances in Animal Biosciences*, *8*(2), 244–249. http://dx.doi.org/10.1017/S2040470017001376.

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

Phadikar, S., Sil, J., & Das, A. K. (2013). Rice diseases classification using feature selection and rule generation techniques. *Computers and Electronics in Agriculture*, *90*, 76–85. http://dx.doi.org/10.1016/j.compag.2012.11.001.

Plant, R. E., & Stone, N. D. (1991). *Knowledge-based systems in agriculture*. USA: McGraw-Hill, Inc..

Prabhu, V., Kannan, A., Ravuri, M., Chaplain, M., Sontag, D., & Amatriain, X. (2019). Few-shot learning for dermatological disease diagnosis. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of the 4th machine learning for healthcare conference, vol. 106* (pp. 532–552). Ann Arbor, Michigan: PMLR.

Pydipati, R., Burks, T. F., & Lee, W. S. (2006). Identification of citrus disease using color texture features and discriminant analysis. *Computers and Electronics in Agriculture*, *52*(1–2), 49–59. http://dx.doi.org/10.1016/j.compag.2006.01.004.

Ricci, F., Wöber, K., & Zins, A. (2005). Recommendations by collaborative browsing. In A. J. Frew (Ed.), *Information and communication technologies in tourism 2005* (pp. 172–182). Vienna: Springer Vienna.

Roach, J., Virkar, R., Drake, C., & Weaver, M. (1987). An expert system for helping apple growers. *Computers and Electronics in Agriculture*, *2*(2), 97–108. http://dx.doi.org/10.1016/0168-1699(87)90020-2.

Romeo, J., Pajares, G., Montalvo, M., Guerrero, J. M., Guijarro, M., & De La Cruz, J. M. (2013). A new expert system for greenness identification in agricultural images. *Expert Systems with Applications*, *40*(6), 2275–2286. http://dx.doi.org/10.1016/j.eswa.2012.10.033.

Rose, D., Sutherland, W., Parker, C., Winter, M., Lobley, M., Morris, C., Twining, S., Ffoulkes, C., Amano, T., & Dicks, L. (2016). Decision support tools for agriculture: Towards effective design and delivery. *Agricultural Systems*, *149*, 165–174. http://dx.doi.org/10.1016/j.agsy.2016.09.009.

Shahbandeh, M. (2021). Global fruit production in 2019. https://www.statista.com/statistics/264001/worldwide-production-of-fruit-by-variety/, Accessed: 20 June 2021.

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience*, *2016*, http://dx.doi.org/10.1155/2016/3289801.

Sottocornola, G., Baric, S., Stella, F., & Zanker, M. (2021). Case study on the development of a recommender for apple disease diagnosis with a knowledge-based bayesian network. In V. W. A. et al. (Ed.), *Workshop proceedings of the 3rd edition of knowledge-aware and conversational recommender systems (KaRS) and the 5th edition of recommendation in complex environments (ComplexRec)*. CEUR-WS.org, http://ceur-ws.org/Vol-2960/paper13.pdf.

Sottocornola, G., Nocker, M., Stella, F., & Zanker, M. (2020). Contextual multi-armed bandit strategies for diagnosing post-harvest diseases of apple. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 83–87). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3377325.3377531.

Stegmayer, G., Milone, D. H., Garran, S., & Burdyn, L. (2013). Automatic recognition of quarantine citrus diseases. *Expert Systems with Applications*, *40*(9), 3512–3517. http://dx.doi.org/10.1016/j.eswa.2012.12.059.

Sutton, T. B., Aldwinckle, H. S., Agnello, A., & Walgenbach, J. F. (Eds.), (2014). *Compendium of apple and pear diseases and pests* (2nd ed.). APS Press.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *61*(3), 611–622.

Workman, D. (2020). Apples exports by country. http://www.worldstopexports.com/apples-exports-by-country/, Accessed: 20 June 2021.

Ye, M., Cao, Z., Yu, Z., & Bai, X. (2015). Crop feature extraction from images with probabilistic superpixel markov random field. *Computers and Electronics in Agriculture*, *114*, 247–260. http://dx.doi.org/10.1016/j.compag.2015.04.010.

Yialouris, C., & Sideridis, A. (1996). An expert system for tomato diseases. *Computers and Electronics in Agriculture*, *14*(1), 61–76. http://dx.doi.org/10.1016/0168-1699(95)00037-2.

Zanella, A., Neuwald, D. A., Bühlmann, A., Folie, I., Kittemann, D., Klein, N., Köpcke, D., Prunier, C., Rossi, O., & Stürz, B. (2021). Frudistor: eine app zur vorbeugung von lagerungsverlusten. *Laimburg Journal*, *3*, http://dx.doi.org/10.23796/LJ/2021.001.

Zhai, Z., Martinez, J. F., Beltran, V., & Martinez, N. L. (2020). Decision support systems for agriculture 4.0: survey and challenges. *Computers and Electronics in Agriculture*, *170*, http://dx.doi.org/10.1016/j.compag.2020.105256.

Zhang, S., Wang, H., Huang, W., & You, Z. (2018). Plant diseased leaf segmentation and recognition by fusion of superpixel, k-means and phog. *Optik*, *157*, 866–872. http://dx.doi.org/10.1016/j.ijleo.2017.11.190.

Zhang, S., Wu, X., You, Z., & Zhang, L. (2017). Leaf image based cucumber disease recognition using sparse representation classification. *Computers and Electronics in Agriculture*, *134*, 135–141. http://dx.doi.org/10.1016/j.compag.2017.01.014.