# FungiTastic: A multi-modal dataset and benchmark for image categorization

## Motivation

**For what purpose was the dataset created?**
The dataset was created to facilitate research on image classification, particularly in the context of multimodal classification, fine-grained recognition, open-set classification, few-shot learning, and domain shift.

**Who created this dataset and on behalf of which entity?**
The dataset was created by: Lukas Picek🍄, Klára Janoušková🍄, Milan Šulc🍄, and Jiří Matas🍄 affiliated with 🍄University of West Bohemia & INRIA, 🍄CTU in Prague, and 🍄Second Foundation.

**Who funded the creation of the dataset?**
This project was partially founded by the Technology Agency of the Czech Republic, project No. SS73020004 (FunDive).

**Any other comments?**
None.

## Composition

**What do the instances that comprise the dataset represent?**
The instance data consists of photographs, microscopic images, and meta-information about the surrounding environment. The textual metadata represents the location of the specimen observation, the substrate/habitat, timestamp, camera settings, and other attributes such as the taxonomy categorization or whether the image is microscopic. The DNA sequence of some of the observations is also included. Satellite images represent the Earth's surface captured by satellites at the observation location. Meteorological data represent climatic variables such as precipitation or temperature.

**How many instances are there in total?**
There are 344,422 observations in total. Each observation consists of a variable number of images and other data.

**Does the dataset contain all possible instances or is it a sample?**
The data are *a sample* of existing European fungi, but as it originates primarily from Denmark, it has a strong bias towards Danish species.

**What data does each instance consist of?**
Camera photographs, textual data, microscopic images, DNA sequencing, satellite images, meteorological data, and segmentation masks.

**Is there a label or target associated with each instance?**
Each instance is associated with the expert-quality taxon label in the form of a species name.

**Is any information missing from individual instances?**
Segmentation, meteorological data, and textual metadata are partially annotated.

**Are relationships between individual instances made explicit?**
No.

**Are there recommended data splits?**
Yes. We provide recommended training/validation splits that are split in a time-aware fashion.

**Are there any errors, sources of noise, or redundancies in the dataset?**
The data's quality is among the highest anyone can achieve, as it undergoes multiple verifications and checks. In terms of labels, all the taxonomy labels are provided by taxon experts (specialists in that domain with years of expertise). We also provide a subset of the data where the label was determined based on the DNA sequence, the 'ultimate' ground truth. Besides, data were searched for duplicities, which, if found, were removed.

**Is the dataset self-contained, or does it link to external resources?**
The dataset is self-contained.

**Does the dataset contain data that might be considered confidential?**
No.

**Does the dataset contain data that might be offensive or cause anxiety?**
Not really. It is unsuitable just for people suffering from fungophobia.

**Does the dataset relate to people?**
The images may contain parts of the human body (in particular, hands) in the background.

**Does the dataset identify any subpopulations?**
No.

**Is it possible to identify individuals from the dataset?**
Not to the best of our knowledge.

**Does the dataset contain data that might be considered sensitive?**
Not to the best of our knowledge.

**Any other comments?**
None.

## Collection Process

**How was the data associated with each instance acquired?**
Based on its location and time.

**What mechanisms or procedures were used to collect the data?**
The dataset is based on citizen science data, where users seemingly randomly collect the data. Therefore, there was no standardized process in the data collection. However, when citizen scientists submit bad-quality data, they are usually lectured by experts on how to collect and report the data better.

**If the dataset is a sample from a larger set, what was the sampling strategy?**
Not applicable.

**Who was involved in the data collection process and how were they compensated?**
Labels were provided by volunteers and citizen scientists. The authors extracted and processed the data during their paid working hours.

**Over what timeframe was the data collected?**
The data was collected over 1970-2023, with the majority between 2011 and 2023.

**Were any ethical review processes conducted?**
No.

**Does the dataset relate to people?**
No. The images may contain parts of the human body (in particular, hands) in the background.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?**
No. Irrelevant.

**Were the individuals in question notified about the data collection?**
No. Irrelevant.

**Did the individuals in question consent to the collection and use of their data?**
No. Irrelevant.

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** Irrelevant.

**Has an analysis of the potential impact of the dataset and its use on data subjects**
No.

**Any other comments?**
None.

## Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done?**
Already clean dataset was check for duplicated datapoints, which were removed.

**Was the ”raw” data saved in addition to the preprocessed/cleaned/labeled data?**
Yes. Anyway, due to the large size of the data, we do not provide it to the “end user”.

**Is the software used to preprocess/clean/label the instances available?**
Yes. The code is available through GitHub.

**Any other comments?**
None.

## Uses

**Has the dataset been used for any tasks already?**
Not this dataset, but a dataset sampled from the same source has already been provided by FungiCLEF competition.

**Is there a repository that links to any or all papers or systems that use the dataset?**
No.

**What (other) tasks could the dataset be used for?**
The dataset is designed to support (i) standard close-set classification, (ii) open-set classification, (iii) multi-modal classification, (iv) few-shot learning, (v) domain shift, and many more.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

Not to the best of our knowledge.

**Are there tasks for which the dataset should not be used?**
No.

**Any other comments?**
None.

## Distribution

**Will the dataset be distributed to third parties outside of the entity?**
The primary intention behind the publication of this dataset is to make it publicly available.

**How will the dataset be distributed?**
The dataset is distributed through multiple channels. Kaggle, Project GitHub, and Documentation.

**When will the dataset be distributed?**
The dataset is already available on Kaggle.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
Yes, the dataset is published under the GPL license.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.
**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
No.

**Any other comments?**
None.

## Maintenance

**Who is supporting/hosting/maintaining the dataset?**
The authors of the paper will maintain the dataset and provide additional support. The dataset is hosted on Kaggle and is also publicly available on a server hosted in Prague at the CMP.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
Owners and dataset maintainers can be contacted via email and the Kaggle forum. All email addresses are provided on Kaggle and GitHub.

**Is there an erratum?**
No.

**Will the dataset be updated?**
Yes, the dataset is planned to be updated on a monthly basis.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?**
Irrelevant.

**Will older versions of the dataset continue to be supported/hosted/maintained?**
New dataset versions will most likely include bug fixes, etc. The older versions of the dataset will be hosted and available but should not be used.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
Anyone can propose a *Merge Request*, *Feature Request*, or *Bug Report* through the Kaggle forum or GitHub.

**Any other comments?**
None.