Detection of Microscopic Fungi and Yeast in Clinical Samples Using Fluorescence Microscopy and Deep Learning

Jakub Paplhám[1], Vojtěch Franc[1] and Daniela Lžičařová[2]
[1]*Department of Cybernetics, Czech Technical University in Prague, Prague, Czech Republic*
[2]*Second Faculty of Medicine, Charles University, Prague, Czech Republic*

Keywords: Filamentous Fungi, Yeast, Automated Detection, Fluorescence Staining, Poisson Image Editing.

Abstract: Early detection of yeast and filamentous fungi in clinical samples is critical in treating patients predisposed to severe infections caused by these organisms. The patients undergo regular screening, and the gathered samples are manually examined by trained personnel. This work uses deep neural networks to detect filamentous fungi and yeast in the clinical samples to simplify the work of the human operator by filtering out samples that are clearly negative and presenting the operator with only samples suspected of containing the contaminant. We propose data augmentation with Poisson inpainting and compare the model performance against expert and beginner-level humans. The method achieves human-level performance, theoretically reducing the amount of manual labor by 87%, given a true positive rate of 99% and incidence rate of 10%.

1 INTRODUCTION

Early detection of yeast and filamentous fungi in clinical samples is critical in treating patients predisposed to severe infections caused by these organisms. Fluorescence microscopy is a suitable method for this detection, where after application to a slide, the material is stained with a fluorescent dye (e.g., Calcofluor White), which binds to chitin contained in the fungal cell wall. This staining process is non-specific, as other structures that may occur accidentally in the sample (e.g., dust, pollen, arthropods) can also bind the dye. Clinical samples commonly consist of respiratory secretions or non-invasive tissue biopsy samples; the aforementioned foreign bodies are therefore routinely present.

Severe infections caused by filamentous fungi are sporadic but severe. Patients with a risk factor, therefore, undergo regular screening. It follows that a considerable number of slides must be carefully examined, most of which do not contain yeast or filamentous fungi.

This work was done in collaboration with Motol University Hospital in Prague, whose staff amassed a unique dataset of fluorescence microscopy images over a period of several years. The goal of the collaboration is to develop an automated system, which filters out samples that are easily distinguishable as negative and presents the remaining potentially positive samples to an expert for verification.

Currently, fungi and yeast cells are detected manually by trained personnel. Laboratory staff then spend a significant amount of time examining negative samples, which leads to job dissatisfaction, and the development of musculoskeletal disorders caused by repetitive stress injuries. Deployment of the system would therefore lead to reduction of the amount of manual labor and an increase in the quality of work.

This paper represents a feasibility study for automation and uses mostly standard deep learning methods. In actual deployment, the detector will be applied to a sequence of images obtained from an automated microscope rather than a single image.

Our main contributions are the following: (i) a detector based on convolutional neural networks (CNNs) is trained on a unique dataset,, (ii) a data augmentation technique specific to the detection task is proposed,, or (iii) performance of the model is evaluated and compared against expert and novice level humans.

2 RELATED WORKS

Recently, automated slide scanners were used for slide imaging of clinical samples and CNNs for evaluation of the data. This allows for fully automatic de-

Paplhám, J., Franc, V. and Lžičařová, D.
Detection of Microscopic Fungi and Yeast in Clinical Samples Using Fluorescence Microscopy and Deep Learning.
DOI: 10.5220/0011616100003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 777-784
ISBN: 978-989-758-634-7; ISSN: 2184-4321

tection and classification of microscopic organisms.

The approach has been successfully used for the detection of bacteria in Gram stains of blood culture, (Smith et al., 2018), and detection of intestinal protozoa in trichrome stained stool samples, (Mathison et al., 2020). Gram staining dyes were also applied to samples containing yeast and yeast-like fungi, allowing their successful classification, (Zieliński et al., 2020). Perhaps most similar to the goals of this work, (Gao et al., 2021) fully automate the process of scanning and classifying fluorescent dye-stained skin samples containing fungi.

All of the listed methods utilize the industry standard technique of fine-tuning a pre-trained CNN. They employ an automated scanner to obtain a large number of images from each sample, classify the images separately, then aggregate the result over the images to classify the sample. An identical approach can be observed in the entire field of automatic evaluation of digital microscopy.

The methods achieve human-level performance, motivating future large-scale deployment. Some novelty can be observed in the use of non-standard classifier heads, e.g., (Zieliński et al., 2020) replace the linear classification head with bag-of-words encoding followed by a support vector machine (SVM) classifier. To our knowledge, no notable domain-specific modifications of the standard techniques were used in these works.

Our work focuses on a different domain, namely fluorescent microscopy of human secretions. Further, we are presented with only a single image per sample and demonstrate that the method achieves sufficient performance for deployment even in this setting.

3 METHODS

3.1 Dataset

The dataset contains high-resolution images of samples collected by staff of Motol University Hospital in Prague from January 2018 to April 2021. The images are a priori assumed to be negative and are considered positive only when structures specific to microscopic yeast or fungi are detected, even if low-resolution and present only in a small portion of the image. An example of such a case, where a single yeast cell is present and covers only a very small portion of the image, is shown in Figure 3.

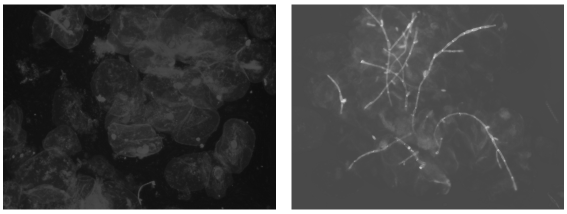

Figure 1: Showcase of randomly selected positive samples from the dataset.

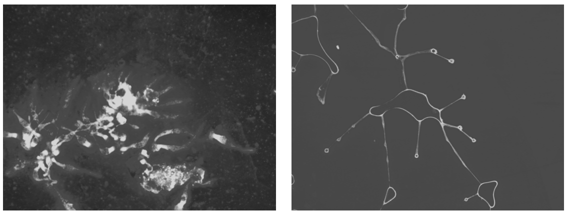

Figure 2: Showcase of randomly selected negative samples from the dataset.

The dataset contains a total of 1244 high-resolution images. The pixel dimensions of the images are not identical, see Table 1. However, the aspect ratio $\frac{\text{Width}}{\text{Height}} \approx 1.33$ is constant throughout the dataset, and each image captures the same field of view. The images can therefore be resized to a uniform shape, which allows for mini-batching of the data, improving the training efficiency. Annotations are given in the form of the binary label (positive/negative) for each image. In other words, the annotation provides no information on the location or size of the specimen within the image.

The main challenges associated with the dataset are twofold: (i) the amount of available data is relatively low, and (ii) positive and negative images have a high degree of similarity. Examples of positive and negative samples are shown in Figure 1 and Figure 2, respectively.

Table 1: Dimensions of images in the dataset.

Width × Height	Annotation	
	Positive	Negative
4140×3096	374	546
2040×1536	77	3
1360×1024	231	13
Total	682	562

Sample Preparation. Clinical Material[1] was smeared on a sterile slide and dried. The dried slides were dyed with Calcofluor White mixed 1 : 1 with 20% potassium hydroxide solution and immediately covered with cover slides and examined. Fluorescence microscopy was performed manually with the

[1] Specifically (i) sputum, (ii) endotracheal or bronchial aspirate, (iii) bronchoalveolar fluid or tissue, (iv) pleural fluid, (v) pericardial fluid, (vi) cerebrospinal fluid, or (vii) liquid or solid contents of pathological cavities.

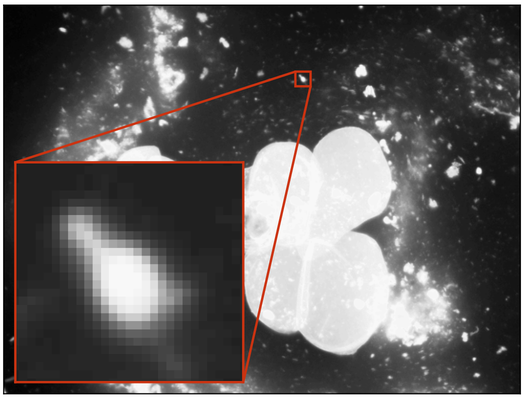

Figure 3: Except for a single budding yeast cell, there is no contaminant in the image. The yeast cell takes up only a small portion of the image, but it is entirely responsible for the final classification as a positive sample.

use of Olympus BX 53 fluorescence microscope, UplanFLN 20x objective lens, FN 26,5. The entire slide was examined, and a representative section of the slide was selected. The image of the selected section was captured using an Olympus DP72 microscope digital camera.

3.2 Model

Formal Definition. Let us denote the common pixel domain as $\mathcal{D} \subset \mathbb{Z}^2$, and the monochromatic domain as $\mathcal{M} \subset \{x \mid x \in \mathbb{R}\}$, where lower and upper bounds of the values and machine precision are ignored for simplicity. The set of all possible grayscale images is then $I = \mathcal{M}^{\mathcal{D}}$.

The model is a binary image classifier, i.e., the mapping $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subset I$, $\mathcal{Y} = \{+1, -1\}$. Often the classifier h is further decomposed as $h = f \circ d$, where $f : \mathcal{X} \to \mathbb{R}$, and $d : \mathbb{R} \to \{+1, -1\}$, with the decoding mapping d defined as

$$d(x, \theta) = \begin{cases} +1 & \text{for } x \geq \theta, \\ -1 & \text{for } x < \theta, \end{cases}$$

for some fixed threshold $\theta \in \mathbb{R}$. We further define an ensemble of binary classifiers, as a binary classifier where the score function f is defined as

$$f(x) = \frac{\sum_{i=1}^{n} f_i(x)}{n}, \qquad f_i : \mathcal{X} \to \mathbb{R}. \tag{1}$$

Implementation. We choose ResNet-50 with a single linear output layer as our baseline binary classifier f and use the shallower variant, ResNet-18, to search for an optimal training setup, e.g., data augmentations and image preprocessing. We further train and compare performance of (i) ResNet-50x1-V2, (He et al., 2016), (ii) EfficientNet models, (Tan and Le, 2019), from B0 to B4, (iii) EfficientNet-V2-S model, (Tan and Le, 2021), and (iv) Vision Transformer ViT-B using 32 × 32 embeddings (Dosovitskiy et al., 2021). All models we use are pretrained on ImageNet.

3.3 Saved Time Metric

Assuming that manual examination of each sample takes constant time, the amount of human time saved by the model is directly proportional to the number of samples which need not be examined by human personnel. If a sample is to be classified as positive, manual confirmation is required. Therefore, the saved time is proportional to the number of samples classified as negative by the model.

The maximal value of such a metric can be achieved by classifying all samples as negative. This is clearly undesirable. Therefore, we define the saved time metric as the portion of samples classified as negative while guaranteeing that the true positive rate is higher than a specified level. The metric represents an alternative to the standard ROC curve. It directly measures the clinical utility of the model and can easily be explained to medical staff.

Formal Definition. We evaluate the prediction rule $h\colon \mathcal{X} \to \{+1, -1\}$ in terms of two metrics. First, the true positive rate (a.k.a. sensitivity) $\mathrm{TPR}(h) = \mathbb{E}_{x \sim p(x|y=+1)} \llbracket h(x) = +1 \rrbracket$, which is the probability that a positive sample is correctly classified as positive. Secondly, the saved time $\mathrm{ST}(h) = \mathbb{E}_{x \sim p(x)} \llbracket h(x) = -1 \rrbracket$, equal to the probability that any input sample is classified as negative. It is useful to rewrite the saved time as

$$\begin{aligned} \mathrm{ST}(h) = {} & \big[1 - p(y = +1)\big] \cdot \big[1 - \mathrm{FPR}(h)\big] \\ & + p(y = +1) \cdot \big[1 - \mathrm{TPR}(h)\big], \end{aligned} \tag{2}$$

where $p(y = +1)$ is the prior probability of the positive class, and $\mathrm{FPR}(h) = \mathbb{E}_{x \sim p(x|y=-1)} \llbracket h(x) = +1 \rrbracket$ is the false positive rate, i.e., the probability that a negative sample is incorrectly classified as positive. The equation (2) shows that the two metrics, $\mathrm{TPR}(h)$ and $\mathrm{ST}(h)$, are antagonistic, i.e., increasing one leads to a decrease of the other and vice versa.

Evaluating the Metric. As defined in Section 3.2, the model is a binary image classifier of the form

$$h(x; \theta) = \begin{cases} +1 & \text{for } f(x) \geq \theta, \\ -1 & \text{for } f(x) < \theta, \end{cases} \tag{3}$$

where $f\colon \mathcal{X} \to \mathbb{R}$ is a score function trained from examples and $\theta \in \mathbb{R}$ is a decision threshold used to tune the operating point of the model. With a slight abuse of notation, we use $\mathrm{TPR}(\theta)$ and $\mathrm{ST}(\theta)$ as a shortcut for $\mathrm{TPR}(h(\cdot;\theta))$ and $\mathrm{ST}(h(\cdot;\theta))$, respectively.

The number of positive and negative samples in the available test set is approximately the same, which does not match the real distribution at the deployment time. According to Motol University Hospital's staff, approximately 9 out of 10 samples that arrive at the laboratory for examination are negative. Therefore, when evaluating the detector we assume that the incidence rate is $p(y=+1)=0.1$. However, the methodology as a whole is general and, if necessary, can be applied to any incidence rate. When evaluating the metric, we resolve the mentioned distribution mismatch as follows. Given a test set $\{(x_i, y_i) \in \mathcal{X} \times \{+1,-1\} \mid i = 1,\dots,n\}$, we compute the empirical estimates of $\mathrm{TPR}(\theta)$ and $\mathrm{FPR}(\theta)$,

$$\widehat{\mathrm{TPR}}(\theta) = \frac{1}{n_+}\sum_{i=1}^{n} \llbracket h(x_i;\theta)=+1 \wedge y_i=+1 \rrbracket, \qquad (4)$$

$$\widehat{\mathrm{FPR}}(\theta) = \frac{1}{n_-}\sum_{i=1}^{n} \llbracket h(x_i;\theta)=+1 \wedge y_i=-1 \rrbracket, \qquad (5)$$

where $n_+ = \sum_{i=1}^{n} \llbracket y_i = +1 \rrbracket$ and $n_- = \sum_{i=1}^{n} \llbracket y_i = -1 \rrbracket$. Then, we fix the positive class prior to the expert estimate of the incidence rate, $p(y=+1)=0.1$, and compute the empirical estimate of the saved time $\widehat{\mathrm{ST}}(\theta)$ by substituting $\widehat{\mathrm{TPR}}(\theta)$ and $\widehat{\mathrm{FPR}}(\theta)$ into equation (2). We evaluate the predictor h by a curve $\left\{\left(\widehat{\mathrm{TPR}}\,(\theta), \widehat{\mathrm{ST}}\,(\theta)\right) \mid \theta \in (-\infty,\infty)\right\}$ which summarizes the entire space of achievable true positive rates and saved times. As a reference, we also plot the best achievable saved time curve as a function of TPR, i.e., we plot the curve $\{(\mathrm{TPR}, \mathrm{ST}^*(\mathrm{TPR})) \mid \mathrm{TPR} \in (0,1)\}$ where $\mathrm{ST}^*(\mathrm{TPR}) = p(y=+1)\cdot[1-\mathrm{TPR}] + [1-p(y=+1)]$, which is obtained from equation (2) when assuming an ideal predictor with zero $\mathrm{FPR}(h)$.

In case we need to evaluate the predictor by a single scalar, e.g., when ranking different models, we report the saved time at desired true positive rate τ, which is defined as $\widehat{\mathrm{ST}}_\tau = \max_{\theta\in(-\infty,\infty)} \widehat{\mathrm{ST}}(\theta)$ subject to $\widehat{\mathrm{TPR}}(\theta) \geq \tau$.

3.4 Domain-Specific Data Augmentation

To enlarge the number of samples used for training, we utilize standard image data augmentations, namely (i) horizontal flip, (ii) vertical flip, (iii) rotation, and (iv) crop and resize. We also implement and evaluate the effects of a custom augmentation method described in the following text.

Motivation & Overview. While obtaining negative samples is simple, getting positive samples is comparatively complex and expensive. We propose an augmentation method specific to the detection task, where any negative image becomes positive if the contaminant (e.g., yeast or fungi) is introduced. The technique takes advantage of a large number of negative images and uses them to generate synthetic positive images by inpainting the positive contaminant into a negative background. The contaminant can further be rotated and shifted to create practically an unlimited number of positive samples.

To generate additional positive samples, we locate the fungi or yeast within the image either by (i) gradient-based localization, Grad-CAM (Selvaraju et al., 2017), which is a broadly applicable method with minimal prerequisites, or (ii) by exploiting the fluorescent staining process. We then augment the image by inpainting the located yeast and fungi into a negative background using Poisson image editing, (Pérez et al., 2003).

We also generate synthetic negative samples, to keep the augmentation symmetrical with respect to classes; preventing the model from associating potential inpainting artifacts with the positive class. In negative samples, we inpaint structures that are visually similar to the yeast and fungi, localized using Grad-CAM, (Selvaraju et al., 2017).

Overview of Poisson Inpainting. Consider the task of inpainting a portion of a source image s into a background b to form a resulting image r. The naive approach is to directly copy the pixel values from the source s to the background b. This, however, creates visible edges between the inpainted region and the background. Instead of copying values of the pixels, Poisson image editing, (Pérez et al., 2003), copies the gradient. A comparison with the naive procedure is shown in Figure 6.

To inpaint a region of the source into the background using Poisson inpainting, (i) pixels on the border of the source region are set to match the neighboring pixels in the background, and (ii) the remaining pixel values of the inpainted source region are found by solving the Poisson equation with the condition of preserving the gradient of the source image. I.e., the color of the inpainted region is modified to match the color of the background, but the relative color difference between pixels is preserved.

Formal Definition of Poisson Inpainting. Here, we briefly review our usage of the method famously introduced by (Pérez et al., 2003). Let us denote a background image, a source image, and a resulting image as $b, s, r \in I$, respectively. By $\Omega \subset \mathcal{D}$, let us denote a region within the source image s which is to be inpainted into the background b to form the resulting image r. Let us further denote by $p \in \mathcal{D}$ a pixel position and by s_p, b_p, r_p values of the pixel within the source, background, and result images, respectively.

To seamlessly inpaint the region Ω of the source image into the background image, we solve the optimization problem

$$\min_{r} \sum_{\langle p,q \rangle \cap \Omega \neq \emptyset} (r_p - r_q - s_p + s_q)^2, \tag{6}$$

subject to

$$r_p = b_p, \forall p \in \delta\Omega, \tag{7}$$

where $\langle p,q \rangle$ is a pixel neighbor pair.

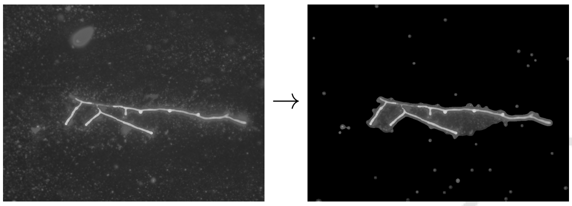

Figure 4: Thresholding a fluorescent stained positive sample to obtain the Ω regions for Poisson inpainting.

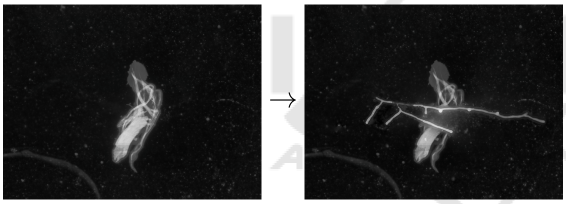

Figure 5: Synthetic sample created by inpainting region containing the positive class into a negative background.

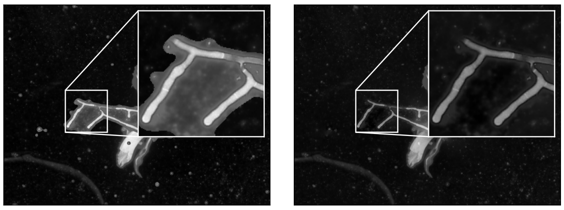

Figure 6: Comparison of the inpainting methods. The inpainted region has a distinct background color when the value of pixels is copied directly. The Poisson image editing ensures that the cutout seamlessly blends into the negative image, creating a more realistic result.

Generating Synthetic Positive Samples. The augmentation procedure requires (i) a positive source image s, (ii) a localization mask defining the region Ω, and (iii) a negative background image b. To localize the fungi and yeast, we exploit the fact that due to the fluorescent staining process, the pixel values of the yeast and fungi are always greater than those of the background. Therefore, we can obtain a rough localization mask by thresholding the pixel values, e.g., by employing Yen's non-parametric thresholding algorithm, (Yen et al., 1995). It must be noted that the fluorescent dye is non-specific, however, and other structures, such as dust, pollen, or arthropods, may also bind the dye. Yeast or fungi is therefore always inpainted, but some miscellaneous structures are inadvertently inpainted as well. This, however, does not harm the creation of new positive examples.

If the localization mask contains multiple connected components, we interpret each connected component as a region Ω and inpaint it separately with random rotation and random position. This modification is especially suited for yeast.

The following steps summarize the augmentation: (i) Localize regions Ω of a positive source image s which contain yeast or fungi., (ii) Select a negative background image b at random., or (iii) Inpaint each region of s discovered in step (i) into the negative image b to form the resulting image. The position and rotation of the inpainted region within the result are selected at random.

Generating Synthetic Negative Samples. The augmentation procedure requires (i) a negative image, serving as both a source s and background b, and (ii) a localization mask defining the region Ω. We use the gradient-based Grad-CAM localization, (Selvaraju et al., 2017), to discover the Ω regions. Grad-CAM provides a course heatmap from which we generate a binary mask by thresholding. It must be mentioned that the heatmap specifies the relative magnitude of activations of the learned convolutional filters. If there are no structures in the image that are similar to fungi or yeast, the activations across the entire image are of similar magnitude and the resulting mask covers the entire image. We, therefore, do not augment the sample in such a case. The following steps summarize the augmentation: (i) Localize regions Ω of a negative image that are visually similar to yeast or fungi., or (ii) Inpaint each region discovered in step (i) into the image to form the resulting image. The position and rotation of the inpainted region within the result are selected at random. This step can be repeated multiple times.

4 EXPERIMENTS & RESULTS

Setup. Unless explicitly stated otherwise, we train and evaluate the models using 30-fold cross-validation with training, validation, and test sets containing 80%, 10%, and 10% of the total data set, respectively. We use the SGD optimizer with 0.9 Nesterov momentum. The initial learning rate is set to 0.001 and reduced by a factor of 3 upon reaching 33% and 66% of the 150 total training epochs. We train using a batch size of 10 images due to hardware limitations. We use images of a uniform size of 952 × 716 pixels. We show the effects of image size on performance of a ResNet-18 model in Figure 7. A larger image size results in better performance; however, an increase over 680 × 512 yields only marginal improvements.

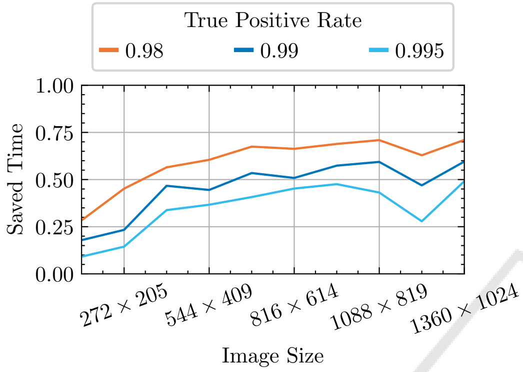

Figure 7: Dependence of saved time metric on image size.

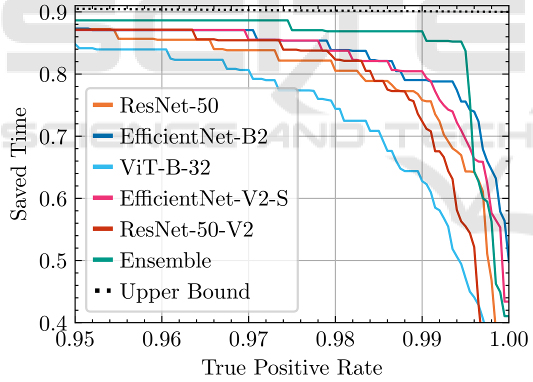

Figure 8: Comparison of different model architectures.

4.1 Model Architecture

We train multiple state-of-the-art models of similar complexity as ResNet-50. It is well known that ensembling techniques result in improved performance at the cost of increased computational complexity. We, therefore, also produce an ensemble of the architectures by averaging their predictions.

The achieved saved time metric values are displayed in Table 2 and the curve of achievable values is shown in Figure 8. The performance of all the models is within a 5% margin, except for an outlier, the Vision Transformer, which performs significantly worse. The best performance is achieved by EfficientNet-B2 models, reaching both the highest value of the saved time metric and the lowest standard deviation between folds, i.e., the architecture performs the best consistently. From the EfficientNet family of models, we only report the best performer, EfficientNet-B2. By ensembling, the performance can further be improved, resulting in a theoretical reduction of manual labor (saved time) of 87%, given a true positive rate of 99%. We show the ROC curve in Figure 9. The ensemble comprises (i) ResNet-50, (ii) EfficientNet-B2, (iii) ViT-B-32, (iv) EfficientNet-V2-S, (v) ResNet-50-V2. Score function of the ensemble is computed as the mean of score functions of the component models.

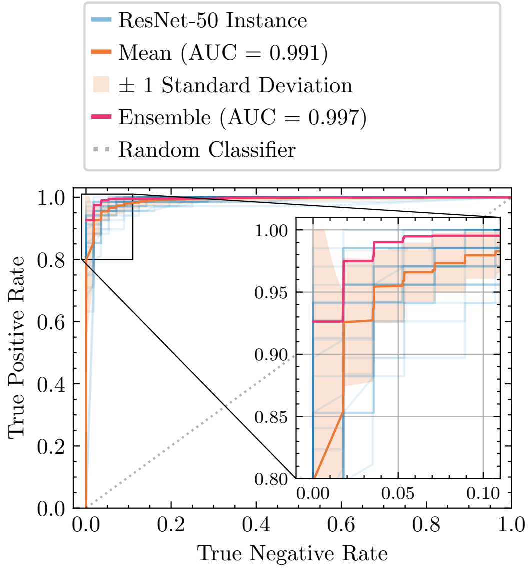

Figure 9: ROC curve (receiver operating characteristic) of ResNet-50 models. The **orange** curve shows the mean ROC over all folds, with the light orange area marking the standard deviation between folds. The **pink** curve shows the ROC for an ensemble of different model architectures.

Table 2: Saved time metric comparison for different model architectures. The first value indicates the mean saved time metric; the second is the standard deviation between folds.

	True positive rate		
Model	98%	99%	99.5%
RN-50	0.81 (0.12)	0.76 (0.18)	0.64 (0.20)
EN-B2	**0.84** (**0.05**)	0.79 (**0.10**)	**0.76** (**0.10**)
ViT-B-32	0.74 (0.12)	0.63 (0.18)	0.47 (0.22)
EN-V2-S	0.84 (0.05)	**0.81** (0.12)	0.70 (0.12)
RN-50-V2	0.82 (0.08)	0.72 (0.19)	0.56 (0.28)
Ensemble	**0.87** (**0.03**)	**0.87** (0.16)	**0.84** (0.16)

4.2 Poisson Augmentation

We construct a learning curve, shown in Figure 10, to verify that the model can benefit from a larger dataset. The performance steeply improves with additional data, motivating further data augmentations beyond the standard techniques.

We train ResNet-50 with additional synthetic positive samples, which were created either (i) by Poisson inpainting, or (ii) by directly copying the pixel values. The results are shown in Figure 11 and demonstrate that the Poisson inpainting is crucial, as the naive technique does not result in any performance improvements.

We, therefore, train ResNet-50 with additional synthetic samples, both positive and negative, created with Poisson inpainting. Models trained using the augmentation of both classes consistently outperform the baseline. The results are shown in Figure 12.

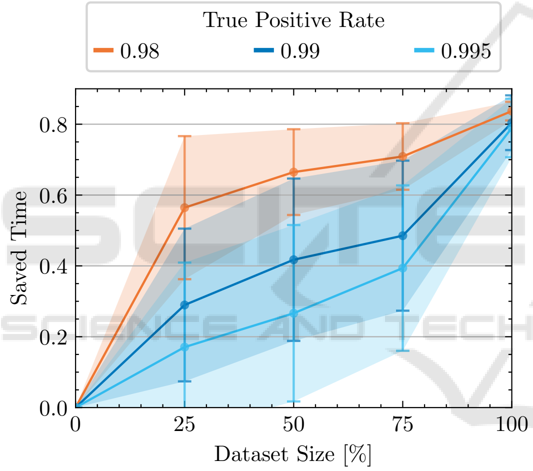

Figure 10: Learning curve for ResNet-50. The transparent area shows ±1 standard deviation. Result on 11 folds.

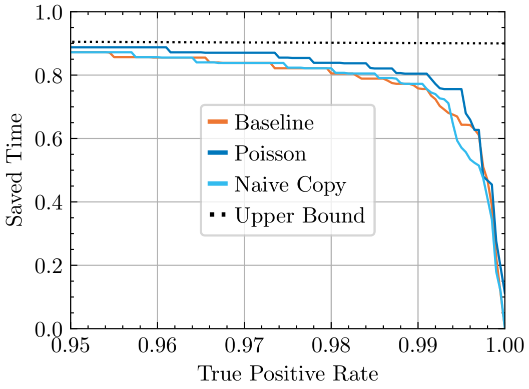

Figure 11: Training with synthetic positive samples created with **Poisson** inpainting results in better performance than **naively copied** pixel values. Result on 20 folds.

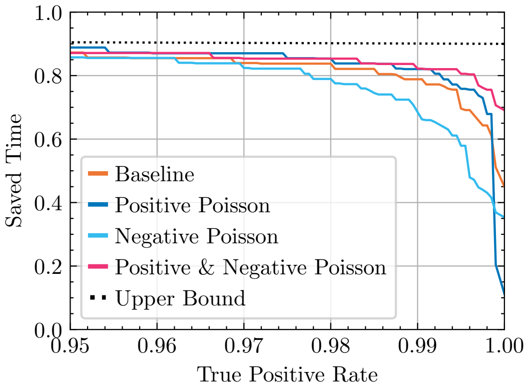

Figure 12: Training with both, **positive** and **negative**, synthetic samples results in significant increase in performance. Result on 15 folds.

4.3 Human-Machine Comparison

To assess the difficulty of the task and further evaluate the model performance, we select 100 positive and 100 negative images from the dataset at random and compare the performance of the model with the performance of humans on the classification.

The images were shown to 4 expert microbiologists. They were prompted to classify the images as either positive or negative. A group of 3 beginners was also shown the images after a brief training session that included a showcase of 20 representative positive and 20 negative samples. The task was verbally explained, and special attention was given to the specific structures of the contaminants.

For each of the images, the automated classification was produced by an ensemble model, which contained the given image in its test set.

All expert microbiologists perform similarly, achieving a true positive rate of 89%, 89%, 90% and 94% with a saved time of 91.1%, 91.1%, 89.2% and 90.6% respectively. The experts achieve values of the saved time at the theoretical upper bound, i.e., the experts achieve a false positive rate of FPR ≈ 0. It should be noted that the presented task is significantly different from the standard operating procedure of the experts. In the usual setting, the expert is presented with an entire slide and can freely move between portions of the slide and search for the contaminant. In this experiment, the view is locked, and the expert is presented with only a single image.

The beginner-level humans perform significantly worse than the automated model. They either (i) do not achieve sufficient true positive rate, or (ii) achieve sufficient true positive rate, but a low value of the saved time metric. The ensemble of models performs at the same level or a better level than the expert humans. The result is shown in Figure 13.

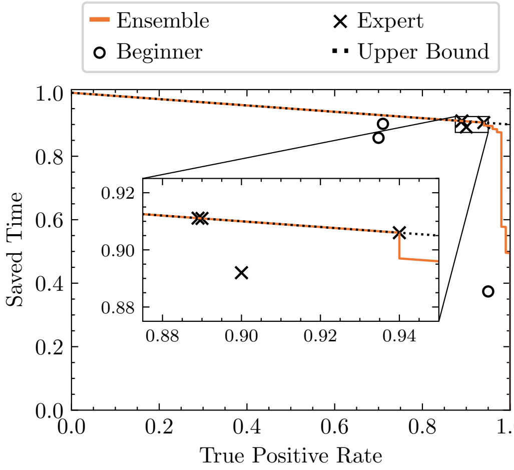

Figure 13: Human-machine performance comparison. Results on a randomly selected set of 100 positive and 100 negative samples. Medical experts perform significantly better than beginner humans, who are outperformed by the automated model by a significant margin.

5 CONCLUSION

The results indicate that the detection of microscopic yeast and fungi in clinical samples can be tackled by standard deep-learning methods, employing an ensemble of convolutional neural networks. The developed model consistently performs on par or better than a human expert and, if deployed, should reduce the amount of manual labor by approximately 87% when operating at a true positive rate of 99%. The results are achieved with annotations only on the image level, i.e., the network was not instructed what part of the image is responsible for the classification.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics. The paper also acknowledges project No. VJ02010041 supported by the Ministry of the Interior of the Czech Republic. Special thanks belongs to MUDr. Daniela Lžičařová for collecting the dataset and to MUDr. Kamila Dundrová, MUDr. Vanda Chrenková, Bc. Karla Kaňková, Vladimír Kryll, RNDr. Pavlína Lysková, Ph. D. for filling out the human-machine comparison survey.

REFERENCES

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Gao, W., Li, M., Wu, R., Du, W., Zhang, S., Yin, S., Chen, Z., and Huang, H. (2021). The design and application of an automated microscope developed based on deep learning for fungal detection in dermatology. *Mycoses*, 64(3):245–251.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks.

Mathison, B. A., Kohan, J. L., Walker, J. F., Smith, R. B., Ardon, O., Couturier, M. R., and Pritt, B. S. (2020). Detection of intestinal protozoa in trichrome-stained stool specimens by use of a deep convolutional neural network. *Journal of Clinical Microbiology*, 58(6):e02053–19.

Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Smith, K. P., Kang, A. D., and Kirby, J. E. (2018). Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. *Journal of Clinical Microbiology*, 56(3):e01521–17.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. *ArXiv*, abs/2104.00298.

Yen, J.-C., Chang, F.-J., and Chang, S. (1995). A new criterion for automatic multilevel thresholding. *IEEE Transactions on Image Processing*, 4(3):370–378.

Zieliński, B., Sroka-Oleksiak, A., Rymarczyk, D., Piekarczyk, A., and Brzychczy-Włoch, M. (2020). Deep learning approach to describe and classify fungi microscopic images. *PLOS ONE*, 15(6):1–16.