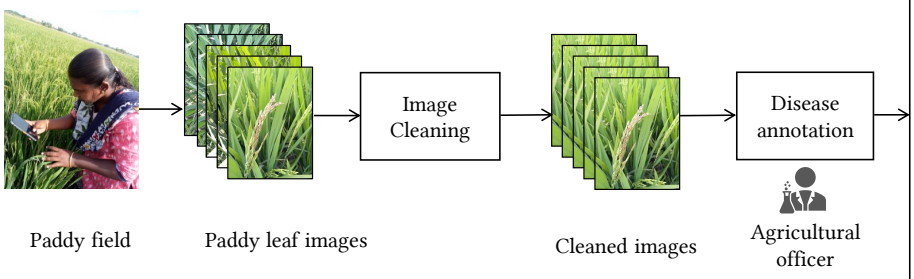
Paddy Doctor: A Visual Image Dataset for Automated Paddy Disease Classification and Benchmarking

Petchiammal A
Manonmaniam Sundaranar University
Tirunelveli, India
ampetchiammal@gmail.com

Briskline Kiruba S
Manonmaniam Sundaranar University
Tirunelveli, India
kiruba.briskline@gmail.com

D. Murugan
Manonmaniam Sundaranar University
Tirunelveli, India
dmurugan@msuniv.ac.in

Pandarasamy A
Singapore
mkusamy@gmail.com



SN	Paddy disease name	Count
1	Bacterial Leaf Blight (BLB)	648
2	Bacterial Leaf Streak (BLS)	505
3	Bacterial Panicle Blight (BPB)	450
4	Black Stem Borer (BSB)	506
5	Blast	2,351
6	Brown spot	1,257
7	Downy Mildew	868
8	Hispa	2,151
9	Leaf Roller	1,095
10	Tungro	1,951
11	White Stem Borer	1,273
12	Yellow Stem Borer	765
13	Normal paddy leaves	2,405
	Total	16,225

Figure 1: Data collection and annotation worflows of the *Paddy Doctor* dataset (https://paddydoc.github.io/).

ABSTRACT

One of the critical biotic stress factors paddy farmers face is diseases caused by bacteria, fungi, and other organisms. These diseases affect plants' health severely and lead to significant crop loss. Most of these diseases can be identified by regularly observing the leaves and stems under expert supervision. In a country with vast agricultural regions and limited crop protection experts, manual identification of paddy diseases is challenging. Thus, to add a solution to this problem, it is necessary to automate the disease identification process and provide easily accessible decision support tools to enable effective crop protection measures. However, the lack of availability of public datasets with detailed disease information limits the practical implementation of accurate disease detection systems. This paper presents *Paddy Doctor*, a visual image dataset for identifying paddy diseases. Our dataset contains 16,225 annotated paddy leaf images across 13 classes (12 diseases and normal leaf). We benchmarked the *Paddy Doctor* dataset using a Convolutional Neural Network (CNN) and four transfer learning based models (VGG16, MobileNet, Xception, and ResNet34). The experimental results showed that ResNet34 achieved the highest F1-score of 97.50%. We release our dataset and reproducible code in the open source for community use.

CODS-COMAD 2023, January 4–7, 2023, Mumbai, India

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Applied computing** → **Agriculture**.

KEYWORDS

Plant Disease Diagnosis, Paddy Diseases, Computer Vision, Deep learning, Transfer Learning.

ACM Reference Format:
Petchiammal A, Briskline Kiruba S, D. Murugan, and Pandarasamy A. 2023. Paddy Doctor: A Visual Image Dataset for Automated Paddy Disease Classification and Benchmarking. In *6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (CODS-COMAD 2023), January 4–7, 2023, Mumbai, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3570991.3570994

1 INTRODUCTION

Agriculture is one of the most important industries contributing to the majority of the national income in several countries. In India alone, 70% of the rural population relies on agriculture [5]. Paddy is a ubiquitous crop in most Asian countries, and India is the world's

Table 1: Comparison of open-source paddy leaf disease datasets.

Dataset	Image resolution	No. of images	No. of diseases	Names of paddy diseases
Gujarat, India [11]	2,848 x 4,288	120	3	Bacterial Leaf Blight (BLB), Brown spot, Leaf Smut
Indonesia [12]	1,440 x 1,920	240	3	Blast, Bacterial Leaf Blight (BLB), Tungro
Malaysia [2]	4,032 x 1,908	400	3	Blast, Brown spot, Hispa
Philippines [6]	300 x 300	552	3	Bacterial Leaf Blight (BLB), Blast, Brown spot
Gujarat, India (augmented) [1]	3,081 x 897	1,294	3	Bacterial Leaf Blight (BLB), Brown spot, Leaf Smut
Nigeria [9]	256 x 256	3,355	3	Blast, Brown spot, Hispa
Odisha, India [14]	300 x 300	5,932	4	Bacterial Leaf Blight (BLB), Blast, Brown spot, Tungro
Paddy Doctor	**1,080 x 1,440**	**16,225**	**12**	**Bacterial Leaf Blight (BLB), Bacterial Leaf Streak (BLS), Bacterial Panicle Blight (BPB), Black Stem Borer, Blast, Brown spot, Downy Mildew, Hispa, Leaf Roller, Tungro, White Stem Borer, Yellow Stem Borer**

second-largest producer of paddy. Paddy cultivation is a complex process affected by many diseases and pests. The early identification of these paddy diseases is a daunting task for agriculturists as well as for agriculture experts [15]. Traditionally, farmers employ manual techniques based on their experience and visual inspection to identify the paddy diseases, but this is highly inefficient, time-consuming, and error-prone [8]. At times, even experienced farmers and agriculture experts might fail to identify the crop diseases accurately due to the large variety of identical disease symptoms. Moreover, farmers apply a large quantity of fertilizer or pesticide without identifying the exact reason for disease manifestation, monitoring the depth of the disease, and measuring the micro-nutrient deficiency. Pesticides are well known for affecting both plants and the soil. It is increasingly important to automate the process of detection of the paddy disease at the earlier stage to reduce pesticide usage and subsequently minimize the loss in the yield [7].

With the advent of Information and Communication Technology (ICT), many researchers have proposed automated disease identification methods by leveraging computer vision techniques [4]. While the traditional methods usually involve manual feature engineering, the recent deep learning-based approaches automatically extract and analyze image features and improve performance. Convolutional neural networks are one of the widely used techniques. Moreover, the variations of convolutional neural network architecture such as DenseNet [18], AlexNet [20], and EfficientNet [17] have enabled the machines to understand critical patterns from the diseased part of the leaf, delivering even better performances than human analysis in many classification problems. Despite all these efforts, the lack of availability of labeled data from real paddy fields hinders the proliferation of these techniques into practical use.

This paper presents *Paddy Doctor*, a large-scale annotated dataset for automated paddy disease identification. The paddy leaf images were collected from real paddy fields using high-resolution smartphone cameras. The collected images were carefully cleaned and annotated with the help of an agricultural officer. The final dataset contains 16,225 leaf images across 13 classes (12 distinct diseases and healthy leaves). Furthermore, we benchmark our *Paddy Doctor* dataset using five advanced state-of-the-art deep-learning models and compare their performance. The models are Deep Convolutional Neural Network (DCNN) and four transfer learning-based models such as VGG16, MobileNet, Xception, and ResNet34[1]. Our experimental results revealed that ResNet34 achieved the highest F1-score of 97.50%. We release the *Paddy Doctor* dataset and reproducible code in the open source[2].

The rest of the paper is organized as follows. In Section 2, we review the related works. In Section 3, we describe our *Paddy Doctor* dataset in detail. In Section 4, we present our benchmarking study and results, followed by conclusions in Section 5.

2 RELATED WORK

A few public datasets are available to experiment with and develop automated paddy disease classification systems. Table 1 compares the image resolution, number of images, number of diseases, and list of paddy diseases present in the existing public datasets. In [11], authors have prepared a database of 120 paddy leaf images (40 samples each for three diseases) in Gandhinagar, Gujarat, India. An augmented version of the same dataset containing 1,294 images is available in [1]. Similarly, authors from Indonesia [12], Philippines [6], and Nigeria [9] have also created a public dataset of 240, 552, and 3,355 images, respectively, across three disease classes. It is to be noted that each image in these datasets contains a close-up view of a single paddy leaf, showing disease symptoms on white background, captured in a controlled environment using high-resolution professional cameras. Unlike this, the researchers from Malaysia [2] have created a dataset of 400 images, across three disease classes, by capturing the paddy leaf images from real paddy fields. Similarly, in [14], a large public dataset with 5,932 images from Odisha, India, is presented. They used a high-resolution professional camera for data collection and then extracted the patches (300x300) of the diseased portion from the original large images to create the final annotated dataset.

Inline with the ongoing efforts towards dataset creation, quite a few research groups have also developed machine learning and deep learning based techniques to detect and classify rice diseases [10, 13, 16, 19]. In [13], a CNN model was used to classify

[1] https://keras.io/api/applications/
[2] https://paddydoc.github.io/

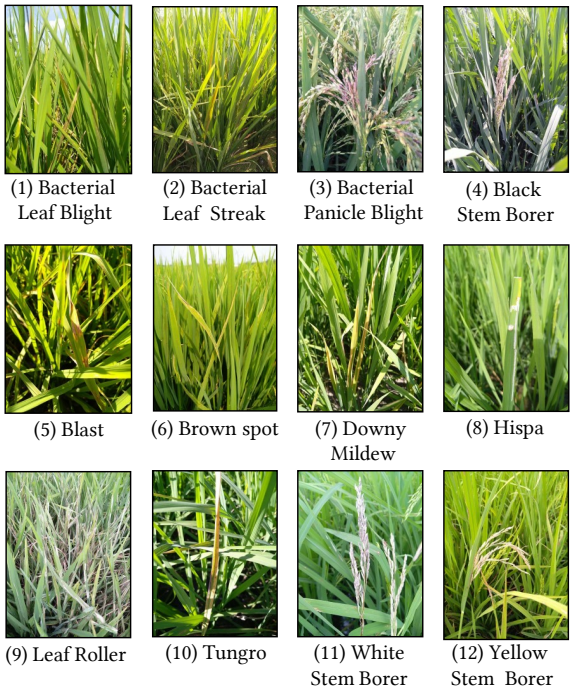


Figure 2: Sample disease images from *Paddy Doctor* dataset.

three types of rice diseases. The proposed CNN model has achieved an accuracy of 94.12%. In [19], the authors proposed an Attention-based Depthwise Separable Neural Network - Bayesian Optimization (ADSNN-BO) model that has achieved an accuracy of 94.65%. In [3], AlexNet model was used for rice disease identification and achieved an accuracy of 96.5%. In [10], CNN and Multilayer Perceptron (MLP) models were proposed to achieve 81.03% and 91.25% accuracy respectively.

Due to lack of availability, most of the prior work used relatively smaller datasets with fewer paddy diseases (See Table 1). In contrast, we present a large paddy leaf disease dataset containing 16,225 annotated images with 13 classes (12 diseases and normal leaf) and also benchmark the performance of several off-the-shelf deep-learning models.

3 PADDY DOCTOR DATASET

The data collection and annotation process of the *Paddy Doctor* dataset is shown in Figure 1. We collected RGB images of paddy leaves from real paddy fields in a village near the Tirunelveli district of Tamilnadu, India. The data collection happened from February to April 2021, when the age of the paddy crop was between 40 to 80 days. We used the CAT S62 Pro smartphone with a built-in camera to capture high-resolution RGB images. Our initial dataset contained approximately 30,000 images in JPEG format with a pixel resolution of 1,080 (width) by 1,440 (height). Next, we carefully examined each sample and removed the inferior and duplicate images. After image cleaning, we are left with 16,225 images in our dataset.

Next, we manually annotated each image, with the help of an agricultural officer, based on the presence of disease symptoms and assigned a diseased class label. After annotation, the final dataset had 13 classes, corresponding to 12 diseases and healthy leaves. The annotated paddy diseases are as follows - Bacterial Leaf Blight (BLB), Bacterial Leaf Streak (BLS), Bacterial Panicle Blight (BPB), Black Stem Borer (BSB), Blast, Brown spot, Downy Mildew, Hispa, Leaf Roller, Tungro, White Stem Borer, Yellow Stem Borer, and Normal leaf (See Figure 1).

Figure 2 shows the sample images of the leaves having 12 distinct diseases. In addition to the RGB images, we manually collected additional metadata for each leaf image, such as the variety and age of the paddy crop when these images were logged. The entire dataset development process spanned approximately 500 man-hours.

4 BENCHMARKING

We benchmark our Paddy Doctor dataset using five contemporary deep-learning models and compare their performance in classifying paddy disease images. The experimented models include a deep CNN model and transfer learning with four pre-trained models: VGG16, MobileNet, Xception, and ResNet34. The details of these models are presented below.

4.1 Deep Convolutional Neural Network

The architecture of the CNN model is shown in Figure 3. It consists of five 2D convolutional layers and a final dense layer. The first convolutional layer is filtered with 32 kernels of size 3 × 3. Then, a 3×3 max-pooling layer is added after the first convolutional layer. The next convolutional layer contains 64 convolution kernels of size 3×3. We have used a batch normalization layer to automatically standardize the inputs in a model and improve the accuracy and stability of neural networks. Similarly, the subsequent three layers use filters of sizes 64, 128, and 128, respectively. The last layer is composed of a max-pooling layer. Two dense connectivity strategies improve the usage efficiency of feature maps, enhancing the diagnostic performance for paddy leaf diseases and a 13-way Softmax layer.

4.2 Transfer Learning

In addition to the CNN model, we also apply four existing deep learning models to our dataset and evaluate their performance. Though the pre-training models can be used as feature extractors and predictors, we fine-tuned them to perform better. As shown in Figure 4, the fine-tuning approach involves keeping most of the existing pre-trained convolutional layers but training only the last few layers and a custom fully connected layer. The selected pre-trained models are VGG16, MobileNet, Xception, and ResNet34. We initialized the weights of these models using ImageNet[3]. Therefore, the training phase of these models involved assigning new weights to the last few layers and the final fully connected layer. The code repository of the *Paddy Doctor* contains more details about their configurations.

[3] https://image-net.org/

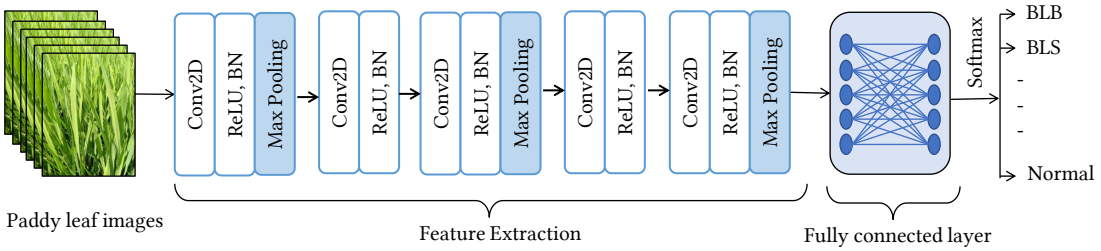


Figure 3: A six-layer deep CNN model for paddy disease classification.

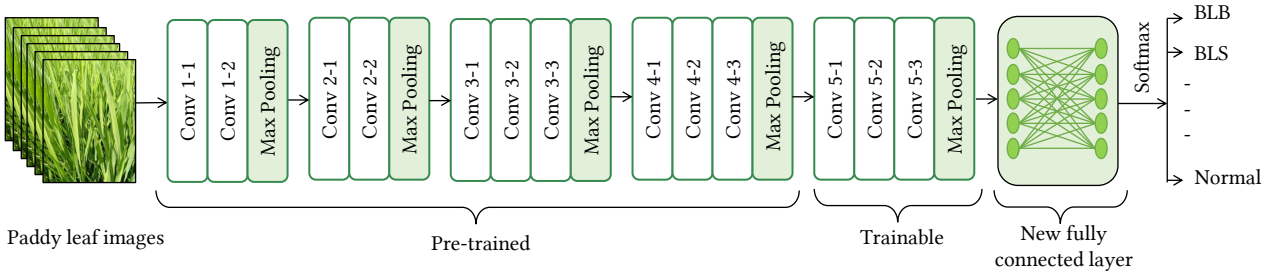


Figure 4: Fine-tuning of VGG16 model for paddy disease classification.

Table 2: Comparison of classification performance of five deep learning models on our *Paddy Doctor* dataset. The ResNet34 model achieved the best performance with an F1-score of 97.50%.

S.No.	Model	Metrics (%)			
		Accuracy	Precision	Recall	F1-score
1	DCNN	88.84	89.22	88.84	88.81
2	MobileNet	92.42	92.63	92.42	92.39
3	VGG16	93.19	93.49	93.19	93.20
4	Xception	96.58	96.61	96.58	96.57
5	ResNet34	97.50	97.52	97.50	97.50

4.3 Experimental Results

We implemented the five deep learning models in Python framework using Keras and TensorFlow libraries and conducted all the experiments on the Google Collab environment with GPU. We split the entire *Paddy Doctor* dataset into two sets: training and testing. The training set had 12,980 (80%), and the test set had 3,245 (20%) images out of the total 16,225 images. In addition, we extracted a validation set consisting of 2956 (20%) images from the training set itself. Therefore, the final training set had 10,384 images. We also used image data augmentation during model training by applying different image transformation techniques. The operations include rotation (5°), shear intensity (0.2°), zoom (0.2), width and height shift (5%), and horizontal flip. Moreover, all images were resized into 256x256 pixels and normalized. All models used a learning rate of 0.001, 100 epochs, and a batch size of 32 during training.

Table 2 compares the classification performance of five deep learning models using four evaluation metrics: accuracy, precision, recall, and F1-score. We observed that the ResNet34 model achieved the highest F1-score of 97.50%. This is followed by Xception (96.57%), VGG16 (93.20%), and MobileNet (92.39%). Comparatively, the DCNN model achieved the lowest F1-score of 88.81%. These results demonstrate the usability of our *Paddy Doctor* dataset for automated paddy disease classification tasks. Additionally, we plan to evaluate other pre-trained models leveraging different transfer learning strategies in the future.

5 CONCLUSION

Manual identification of paddy diseases is a challenging task for farmers. Hence, there is an increasing need to develop automated solutions that can scale to many diseases and plants. The lack of availability of public datasets with annotated disease names was a major bottleneck to benchmarking the recent deep learning-based models and wider adoption of the solutions. In this paper, we presented the *Paddy Doctor* dataset for automated paddy disease detection. It contains 16,225 annotated paddy leaf images across 13 classes (12 diseases and normal leaf). The presented dataset was benchmarked using five deep learning-based models and we compared their performance across each other. The results demonstrate that ResNet34 achieved a superior accuracy of 97.5% followed by 96.58% with Xception based model. Finally, plans are underway to expand our *Paddy Doctor* dataset by collecting fine-grained data, such as infrared and hyper-spectral images, about paddy diseases and pests and benchmark them using additional deep learning models.

REFERENCES

[1] Md. Sabbir Ahmed. 2020. UCI - Rice Leaf Diseases Data Set (augmented). https://www.kaggle.com/datasets/badhon7432/paddyleafdiseaseuci. [Online; accessed 2022-11-24].

[2] Bifta Sama Bari, Md Nahidul Islam, Mamunur Rashid, Md Jahid Hasan, Mohd Azraai Mohd Razman, Rabiu Muazu Musa, Ahmad Fakhri Ab Nasir, and Anwar P.P. Abdul Majeed. 2021. A real-time approach of diagnosing rice leaf disease using deep learning-based faster R-CNN framework. *PeerJ Computer Science* 7 (April 2021), e432. https://doi.org/10.7717/peerj-cs.432

[3] R Jeya Bharathi. 2020. Paddy Plant Disease Identification and Classification of Image Using AlexNet Model. *Int. J. Anal. Exp. modal Anal.* 12, 0886 (2020), 1094–1098.

[4] Konstantinos P. Ferentinos. 2018. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture* 145 (Feb. 2018), 311–318. https://doi.org/10.1016/j.compag.2018.01.009

[5] Food and Agriculture Organization (FOA). 2022. India at a glance. https://www.fao.org/india/fao-in-india/india-at-a-glance/en/. [Online; accessed 2022-11-24].

[6] Aldrin Kein G. Francisco. 2019. Rice-Disease-DataSet. https://github.com/aldrin233/RiceDiseases-DataSet. [Online; accessed 2022-11-24].

[7] Amritha Haridasan, Jeena Thomas, and Ebin Deni Raj. 2023. Deep learning system for paddy plant disease detection and classification. *Environmental Monitoring and Assessment* 195, 1 (2023), 1–28.

[8] B. Leelavathy and Ram Mohan Rao Kovvur. 2020. Prediction of Biotic Stress in Paddy Crop Using Deep Convolutional Neural Networks. In *Proceedings of International Conference on Computational Intelligence and Data Engineering*. Springer Singapore, 337–346. https://doi.org/10.1007/978-981-15-8767-2_29

[9] Nizor Ogbezuode. 2022. Rice Leaf Images. https://www.kaggle.com/datasets/nizorogbezuode/rice-leaf-images. [Online; accessed 2022-11-24].

[10] Rutuja R. Patil and Sumit Kumar. 2022. Rice-Fusion: A Multimodality Data Fusion Framework for Rice Disease Diagnosis. *IEEE Access* 10 (2022), 5207–5222. https://doi.org/10.1109/access.2022.3140815

[11] Harshadkumar B. Prajapati, Jitesh P. Shah, and Vipul K. Dabhi. 2017. Detection and classification of rice plant diseases. *Intelligent Decision Technologies* 11, 3 (Aug. 2017), 357–373. https://doi.org/10.3233/idt-170301

[12] Rukhsar and Santosh Kumar Upadhyay. 2022. Rice Leaves Disease Detection and Classification Using Transfer Learning Technique. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2151–2156. https://doi.org/10.1109/icacite53722.2022.9823596

[13] G K Sagarika, SJ Krishna Prasad, and S Mohana Kumar. 2020. Paddy Plant Disease Classification and Prediction Using Convolutional Neural Network. In *2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. IEEE, 208–214. https://doi.org/10.1109/rteict49044.2020.9315634

[14] Prabira Kumar Sethy, Nalini Kanta Barpanda, Amiya Kumar Rath, and Santi Kumari Behera. 2020. Deep feature based rice leaf disease identification using support vector machine. *Computers and Electronics in Agriculture* 175 (Aug. 2020), 105527. https://doi.org/10.1016/j.compag.2020.105527

[15] Vimal K Shrivastava, Monoj K Pradhan, Sonajharia Minz, and Mahesh P Thakur. 2019. Rice plant disease classification using transfer learning of deep convolution neural network. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* XLII-3/W6 (July 2019), 631–635. https://doi.org/10.5194/isprs-archives-xlii-3-w6-631-2019

[16] R Swathika, S Srinidhi., N Radha, and K Sowmya. 2021. Disease Identification in paddy leaves using CNN based Deep Learning. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. IEEE, 1004–1008. https://doi.org/10.1109/icicv50876.2021.9388557

[17] Ümit Atila, Murat Uçar, Kemal Akyol, and Emine Uçar. 2021. Plant leaf disease classification using EfficientNet deep learning model. *Ecological Informatics* 61 (March 2021), 101182. https://doi.org/10.1016/j.ecoinf.2020.101182

[18] Ruchi Verma and Varun Singh. 2022. Leaf Disease Identification Using DenseNet. In *Artificial Intelligence and Speech Technology*. Springer International Publishing, 500–511. https://doi.org/10.1007/978-3-030-95711-7_42

[19] Yibin Wang, Haifeng Wang, and Zhaohua Peng. 2021. Rice diseases detection and classification using attention based neural network and bayesian optimization. *Expert Systems with Applications* 178 (Sept. 2021), 114770. https://doi.org/10.1016/j.eswa.2021.114770

[20] Salma Zakzouk, Mohamed Ehab, Silvana Atef, Retaj Yousri, Rania M Tawfik, and M. Saeed Darweesh. 2021. Rice Leaf Diseases Detector Based on AlexNet. In *9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*. IEEE, 170–174. https://doi.org/10.1109/jac-ecc54461.2021.9691435