Privacy Preserving Method for Image Recognition based on Denoising Autoencoder

Yongshang Tian *

Department of Computer Science, Nanjing University, Nanjing, China

* Corresponding author email: 201840009@smail.nju.edu.cn

Abstract. Machine learning based on neural networks have obtained great achievements in diverse domains. Training models usually requires large, labeled datasets, which are often crowdsourced and consist of private information. There is no doubt that any private information about the datasets should not be let out by the models. To realize the goal, this article introduces a composed model based on denoising autoencoder for classification. The difference between the reconstructed images by the decoder and the actual images is then set as the optimization object while training the encoder and decoder parts, and multiple fully linked layers are added to the pretrained encoder to create a composite model for classification. The entire assembled model is then trained once more with the noised images. The constructed model produces a satisfactory result since the characteristics extracted by the encoder were what the model used to do classification. The experiment results demonstrate that the composed model can protect privacy at a low cost in model quality and accuracy compared to the baseline models which take raw images as input.

Keywords: Privacy Preserving; Denoising Autoencoder; Neural Network; Deep Learning.

1. Introduction

Recent progress in neural networks has contributed to lots of applications in diverse fields, including image recognition, natural language processing, etc. While these achievements are mainly based on access to large and labeled datasets for training models. There is a non-negligible probability that the datasets are crowd-sourcing and have a big chance to contain sensitive and private information. Their use requires techniques that can provide protection for users' data privacy. Besides, COVID-19 has become the most important issue during the last three years, the society has been affected a lot including people's daily routine, different industries and government policies. For instance, many shops and businesses have been shut down according to poor economic performance. Under this circumstance, thousands of Computed Tomography (CT) images of lung have been taken in order to confirm COVID-19 infection. These images include many personal details of the patients which to some extents are very sensitive and may do harm to private information safety. Therefore, how to diagnose based on patients' CT images of the lung while ensuring the data security is worthy of attention.

There is a great deal of work on CT images recognition and classification. In the early field of CT images recognition, a classification, people prefer to take a more medical way, by performing a retrospective single-center analysis on COVID-19 patients using a combination of lung CT imaging and acute inflammatory lesions affecting each lobe [1]. This type of study consumes quite a lot of time and energy though the final result is good and valuable. Recently, AI gradually comes into researchers' perspective, and applying Artificial Intelligence (AI) for medical use has gotten attention due to worldwide efforts on human-centric AI, model performance and network design. Prior research has concentrated on using convolutional neural networks to CT image classification tasks in an effort to increase the accuracy of CT image classification and speed up the training of classification models [2]. In [2], the researchers applied the CDBN model with Adam optimizer [3], and the training speed is a little bit faster than the normal gradient algorithm. Besides, some researchers are doing more specific work, they concentrate on early phase detection of COVID-19 on the basis of CT images [4]. In this work, many different types of algorithms have been applied for feature extracting, for instance Grey Level Co-occurrence Matrix (GLCM), Grey-Level Size Zone Matrix (GLSZM), etc. And using

Support Vector Machines (SVM) for classification according to the extracted features. Achieving the best accuracy as 99.68% with GLSZM feature extraction method and 10-fold cross-validation. However, all the works above have access to many patients' CT images, but none of them pay attention to privacy preservation. As is known to all, protecting privacy is a constant concern since the widespread use of the Internet has made it easier for attackers to conduct malicious database attacks, communication monitoring, and leaks of user personal information.

In this regard, this study designed a combined convolutional neural network to make classification with privacy preserved through Denoising Auto Encoder (DAE) [5]. A DAE network was firstly trained, after the loss between the original image and the restored image by decoder is minimized. The encoder network's settings are saved in this study, which later adds a number of untrained, fully linked layers to the encoder network to create a classifier. The CT scans of COVID-19 patients are used as an example in this article. This classifier takes images as input and decide whether the owner of the CT image is affected by COVID-19. To show the approach's effectiveness, this study chooses several baseline methods for comparison, for instance, resnet52, resnet152 (pretrained, only train fully connected layers here), simple CNN, etc. In the paper, the classifier takes noised images as input while all baseline methods take original image as input. Finally, the model in this study gets a 91% percent in testing, while simple CNN get 95% on the test dataset. It can be concluded that though the approach achieves a lower accuracy than simple CNN, but the approach protects the patients' accuracy for our approach take noised data as input and 91% is good enough for classification, another 4% increase in accuracy seems to be a slight advantage.

2. Methods

2.1 Dataset Description and Preprocessing

In this paper, the author applied a CT image data set about COVIID-19 provided by some researchers of UCSD [6]. 349 COVID-19 CT pictures from 216 patients and 463 non-COVID-19 CTs from normal are included in the original dataset. These CT pictures are also based on various sizes. The average minimum and greatest heights are 153, 491 and 1853, respectively. 124, 383, and 1485 are the typical minimum and maximum widths.

The data preprocessing used in this study consists of three main parts. First, the author manually divides the origin data set into a data set for training and one for testing. After dividing, the number of training set is 678 and that of testing set is 103, some of abnormal images are abandoned. Second, the images are all resized to 224×224 to ensure that the image size is consistent and can be uniformly processed by the model. Last, the author normalizes the images to avoid the influence of some extreme values.

2.2 Background and Proposed Approach

2.2.1 Convolutional Neural Network (CNN)

A grid-based machine learning model called CNN is used for data processing and research [7], since it has been applied in many domains successfully [8-10]. CNN is made to automatically and adaptively learn the space structure of features, from basic to complex patterns. Convolution, pooling, and fully linked layers are the three primary types of layers found in CNN. Convolution layers and pooling layers, the first two types, work as feature extractors; the fully connected layer then maps the extracted features on the output.

2.2.2 Denoising Auto Encoder (DAE)

Vincent first proposed the Denoising Autoencoder (DAE) concept in [5]. The original autoencoder was altered to create a denoising autoencoder shown in Figure 1,2, which was intended to create a clean output from a noisy or contaminated input. First, a corrupted version of the original input $\tilde{x}$ is obtained by dint of a random mapping $\tilde{x} \sim M_S(\tilde{x}|x)$. The researchers considered the following corrupting process during their experiments, destruction proportion v is firstly chosen: A fixed

number of components are randomly selected for each input x and their values are set to 0, while all other values remain the same. The original input image for the chosen elements is completely erased, and the autoencoder will be trained to fill in these fictitiously created blank spaces. Considering alternative corrupting noises, the corrupted input $\tilde{x}$ is then mapped to a hidden representation $y = f_\theta(\tilde{x}) = s(W\tilde{x} + b)$ from which the reconstructed result $z = g_{\theta'}(\tilde{x}) = s(W'y + b)$ is obtained. Then the parameters in the network are trained to minimize the average reconstruction error $L_E(x, z) = E(D_x||D_z)$ (i.e., E means entropy stands for Bernoulli distribution, L represents reconstruction cross-entropy) on a training dataset, i.e., to make z as similar to the raw input as possible. Many types of loss function can be used to show the difference between original input x and the reconstructed output z, here in this article, Mean Squared Error (MSE) is applied.

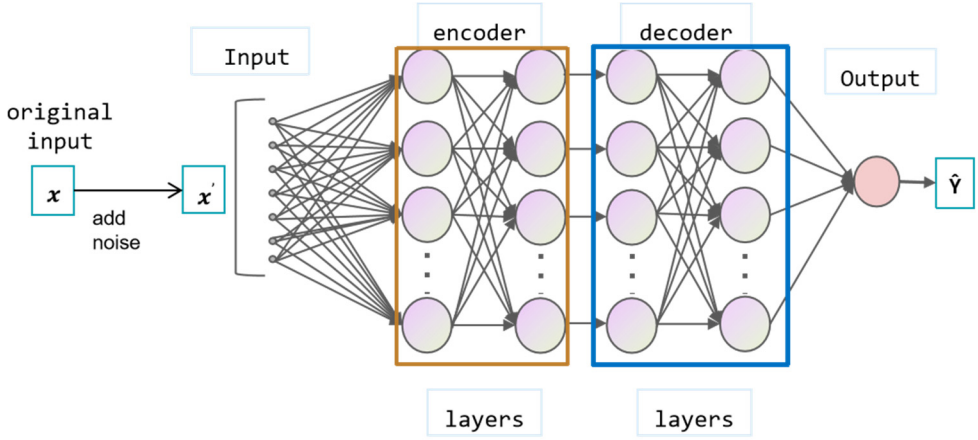


Figure 1. The general structure of DAE

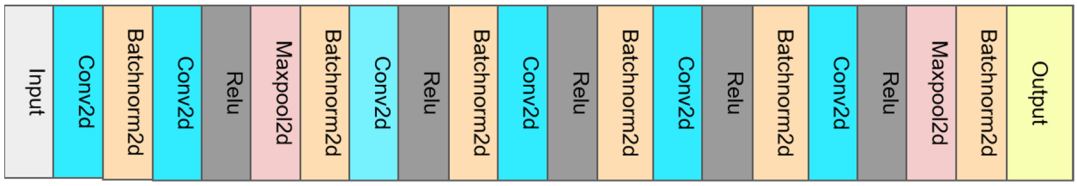


Figure 2. Encoder configuration

2.2.3 From Encoder to Classifier

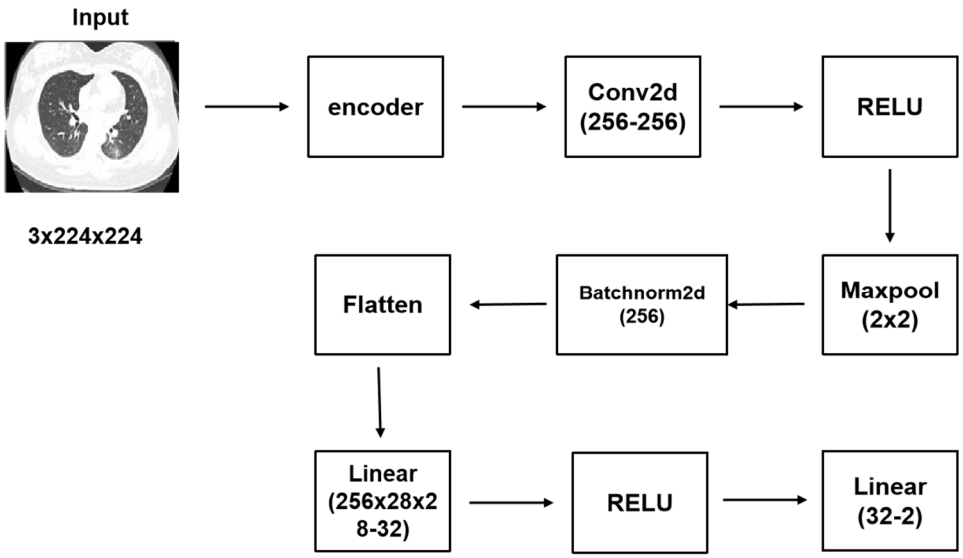


Figure 3. CNN configuration

In this article, the author intends to utilize the advantage of DAE that it can take noised images as input and can reconstruct a similar picture according to the extracted features. So, several untrained fully connected layers are added to the final layer of the encoder part in DAE to modify it into a classifier. This approach can be applied in two steps. First, DAE network is designed and trained to minimize the difference between original input and the reconstructed output. Second, freezing the pretrained encoder and adding some untrained fully connected layers, then training the composed neural network with noised training data and real class labels.

2.3 Implementation Details

In this paper, Gaussian noise has been chosen to add to original images due to its simplicity and practicability The core parameter in Gaussian noise equals to 0.2 in this experiment. A demo picture group is listed in Figure 3.

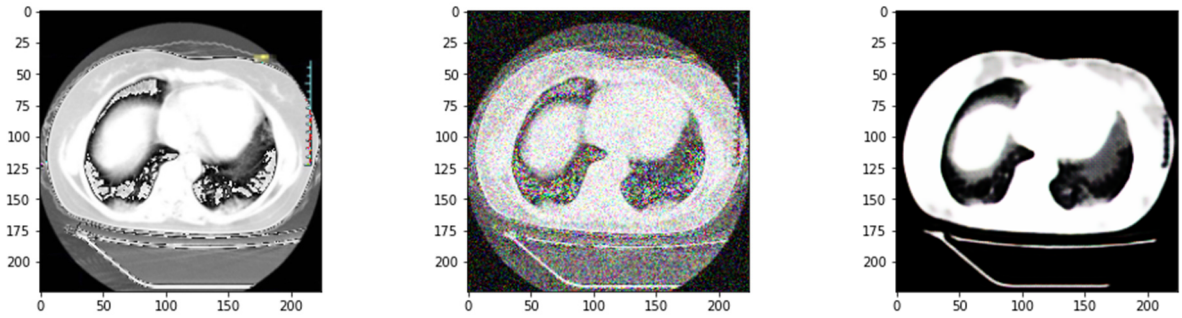


Figure 4. Demo images

During the training process, two main loss functions have to be minimized in different stages. First is the MSE loss between the original image and the reconstructed one. In order to achieve the best reconstructed result, the MSE loss should be minimized while training DAE. The MSE loss L_1 is defined as follows:

$$L_1 = \mathrm{MSE}(y_1 - y_0) = \frac{\sum_{i=1}^{n}(y_{1i} - y_{0i})^2}{n} \quad (1)$$

In the equation above, y_1 represents the data of the reconstructed image, y_0 represents the data of the original image, n refers to the amount of data involved in computing.

Cross-entropy loss serves as a measure of the performance of classification models. Cross-entropy loss decreases when the result the classification model provides becomes closer to the labels. The Cross-entropy loss L_2 is defined as follows:

$$L_2 = \sum -(y \log(y^{\sim}) + (1 - y)\log(1 - y^{\sim})) \quad (2)$$

In the equation, y represents the true label of the current image (COVID-19 or NonCOVID-19), $y^{\sim}$ represents the possibility that the model predicts that the sample is positive (COVID-19).

Before training, several hyper-parameters should be set (optimizer=Adam optimizer, batch size=20, learning rate=5×10^{-4} for DAE, learning rate=3×10^{-4} for classifier).

3. Results and Discussion

3.1 Experiments Results

By adjusting noise scale, the raw images are corrupted in different degree. In Figure 5, it can be concluded that the noised images blurred all the information in the raw images. However, the key characteristics of the raw images can still be seen in the reconstructed images based on the suggested model, demonstrating the efficacy of the technique employed in this study. But some details in the

original images are discarded and ignored while reconstructing the images. In the experiments in this article, $\sigma = 0.2$ is applied.

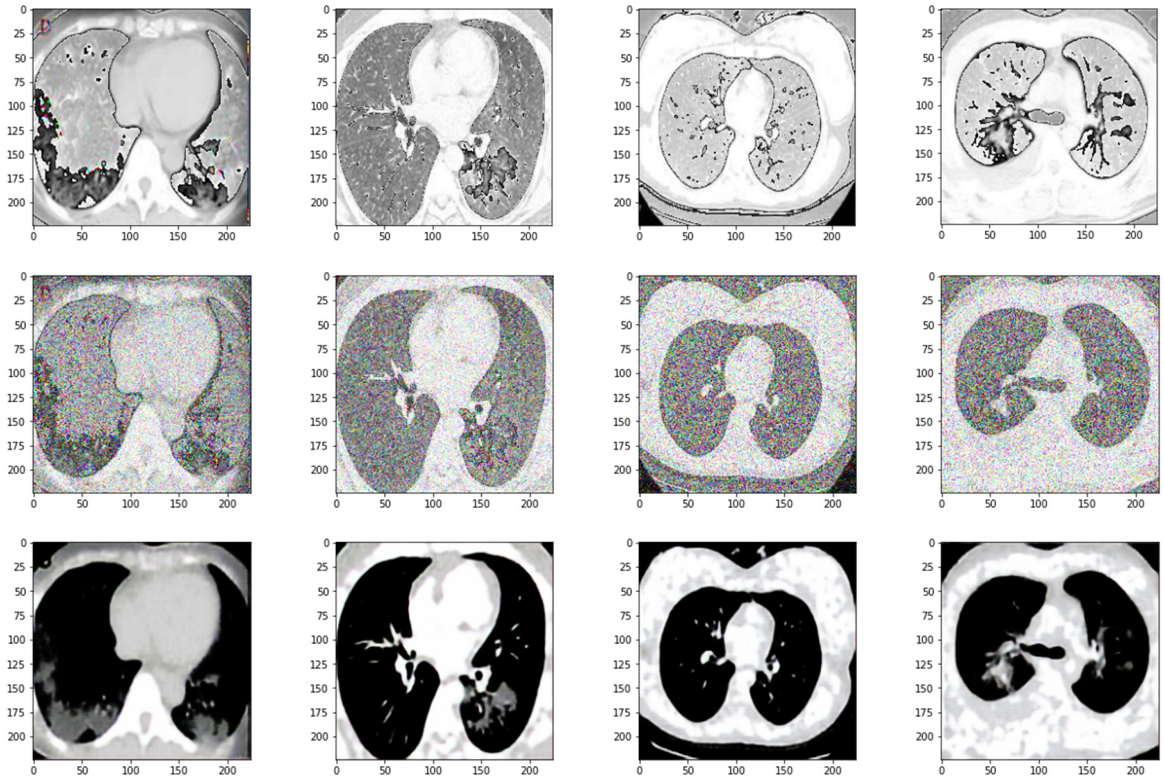


Figure 5. Demo images (the first row are raw images, the second are noised images, the third are reconstructed images. Noise scale of the left two pictures is $\sigma = 0.2$,that of the right two is $\sigma = 0.3$)

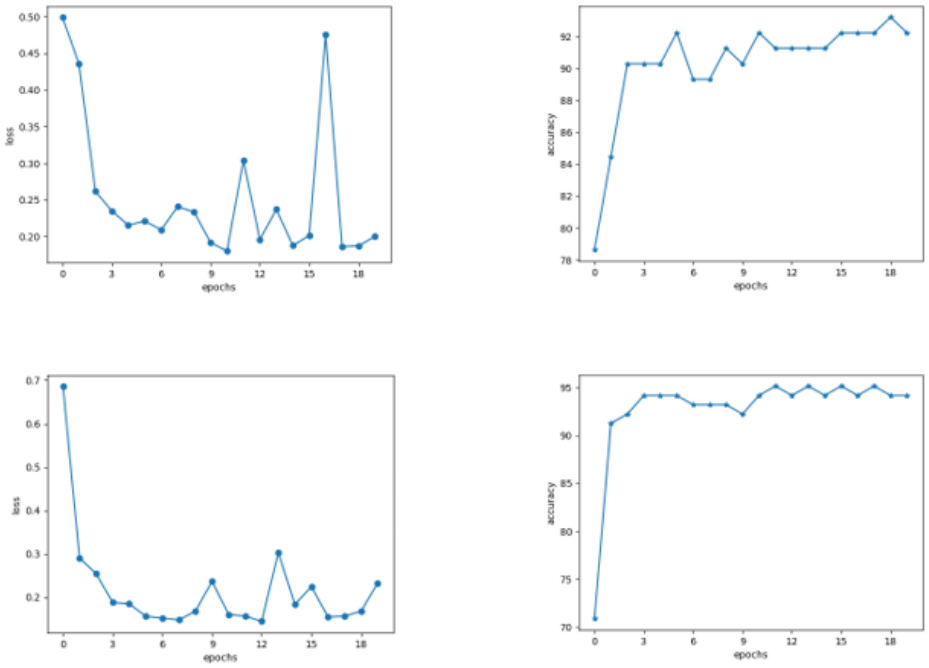


Figure 6. Loss and accuracy (The accuracy of the model on the test set is shown in the top right, while the loss of the model introduced in this article is shown in the upper left. The accuracy of the SimpleCNN model on the test set is shown in the lower right, while the loss of the model is shown in the lower left.)

Due to the diverse noised scale, the classification accuracy is affected to some extent. The performance of the model based on encoder with extra fully connected layers (upper left and upper right) is shown in Figure 6. Here the SimpleCNN model is taken as the baseline model for comparison. From the figure, though there are fluctuations in loss, the loss is decreasing in general. Moreover, the accuracy first increases rapidly and gradually become stable at a small interval.

Table 1. Accuracy table (only train the last several fc layers of pretrained ResNet models

accuracy / model	Training Accuracy (%)	Testing Accuracy (%)
Encoder +fc layers	100	92.21
SimpleCNN	100	95.14
Resnet52 (pretrained)	71.38	63.10
Resnet152 (pretrained)	75	67.3

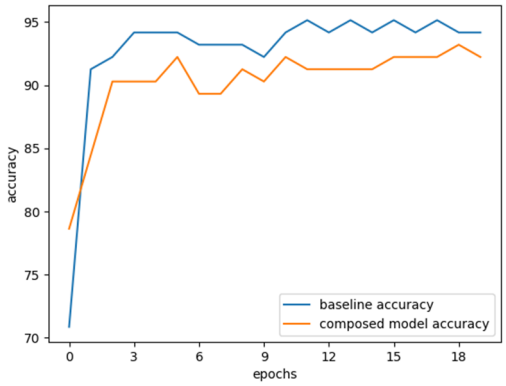


Figure 7. Accuracy comparison (the blue line represents the baseline accuracy; the yellow line shows the composed model accuracy)

In Figure 7 and Table 1, it can be illustrated that the designed model in this article achieve an accuracy of approximately 92%, which is 3% fewer than that of the baseline model. Though the performance of the composed model is not as well as the baseline model, a 92% accuracy is quite enough for classification tasks, besides the model takes noised images as input while the baseline model takes original images as input. In other words, the approach introduced in this article can achieve a high accuracy in classification task while having privacy preserved. Pretrained ResNet model are based on imagenet dataset, which mainly serves models that aim to simulate human' s recognition system. As a result, ResNet cannot perfectly extract features from CT images, which leads to low accuracy.

3.2 Discussion

From all the experiments results above, this seems that applying several fully connected layers to the pretrained encoder will make an effective classifier. What lies behind this phenomenon is that the characters extracted from the noised images by the encoder can be applied for classification. After adding noise to an image, some details are obscured while the main characters are remained for the decoder to reconstruct the image. From the experiments results showed in fig, the reconstructed images contain most information of the raw images with few details abandoned. So, the characters extracted by the encoder contains most information about the original image. As a result, applying the encoder as a feature extractor will help the composed model make classification on noised images, which means that the patients' privacy is preserved while training.

4. Conclusion

In this work, a method of combining denoising autoencoder and CNN for classification has been applied to address privacy preserving problem on images. In the experiment, several fully connected layers are added to the pretrained encoder to make a model for COVID-19 infection status recognition. The model achieved 92% accuracy on the testing set with noised. For comparison, extensive experiment based on the SimpleCNN model for classification has been conducted and achieves 95% accuracy, which is a little bit better than the combined model. The combined model can provide privacy preservation at the cost of little decrease in accuracy comparing to the baseline method. Many avenues for further work can be attractive. In particular, the research orientation may pay more attention to adapting the combined model to address diverse types of tasks which require privacy preservation. The experience in applying the combined model in image recognition should be helpful, many opportunities for new research are seen, for instance, Med3D model designed for 3D medical images can be improved by applying the techniques introduced in this article to protect privacy for the images' owners. In addition, obtaining additional improvements in accuracy will be an interesting question. Many training datasets are larger than COVID-19 dataset; larger size should benefit model accuracy. Last but not least, applying other types of noise will be a possible research orientation, for instance salt and pepper noise.

References

[1] Li, Kunwei, et al. CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). European radiology 30.8, 2020, 4407-4416.

[2] Zhang, Yuezhong, et al. CT image classification based on convolutional neural network. Neural Computing and Applications 33.14, 2021, 8191-8200.

[3] Kingma, Diederik P et al. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[4] Barstugan, Mucahid, Umut Ozkaya, and Saban Ozturk. Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:2003.09424, 2020.

[5] Vincent, Pascal, et al. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning. 2008.

[6] Yang, Xingyi, et al. COVID-CT-dataset: a CT scan dataset about COVID-19. arXiv preprint arXiv: 2003. 13865, 2020.

[7] Yamashita, Rikiya, et al. Convolutional neural networks: an overview and application in radiology. Insights into imaging 9.4, 2018, 611-629.

[8] Sarvamangala, D. R et al. Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence 15.1, 2022, 1-22.

[9] Yu, Q et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 2022, 103323.

[10] Chen, Samuel Yen-Chi, et al. Quantum convolutional neural networks for high energy physics data analysis. Physical Review Research 4.1, 2022, 013231.