



Artificial Intelligence Prediction of Urban Spatial Risk Factors from an Epidemic Perspective

Yecheng Zhang^{1(✉)}, Qimin Zhang², Yuxuan Zhao¹, Yunjie Deng¹, Feiyang Liu³,
and Hao Zheng^{4(✉)}

¹ College of Architecture and Art, Hefei University of Technology, Hefei, China
575860760@qq.com

² School of Mechanical Engineering, Hefei University of Technology, Hefei, China

³ Wuhan Institute of Technology, Wuhan, China

⁴ Stuart Weitzman School of Design, University of Pennsylvania, Philadelphia, USA
zhhao@design.upenn.edu

Abstract. From the epidemiological perspective, previous research methods of COVID-19 are generally based on classical statistical analysis. As a result, spatial information is often not used effectively. This paper uses image-based neural networks to explore the relationship between urban spatial risk and the distribution of infected populations, and the design of urban facilities. We take the Spatio-temporal data of people infected with new coronary pneumonia before February 28 in Wuhan in 2020 as the research object. We use kriging spatial interpolation technology and core density estimation technology to establish the epidemic heat distribution on fine grid units. We further examine the distribution of nine main spatial risk factors, including agencies, hospitals, park squares, sports fields, banks, hotels, Etc., which are tested for the significant positive correlation with the heat distribution of the epidemic. The weights of the spatial risk factors are used for training Generative Adversarial Network models, which predict the heat distribution of the outbreak in a given area. According to the trained model, optimizing the relevant environment design in urban areas to control risk factors effectively prevents and manages the epidemic from dispersing. The input image of the machine learning model is a city plan converted by public infrastructures, and the output image is a map of urban spatial risk factors in the given area.

Keywords: Coronavirus disease 2019 · Spatial risk factor · Machine learning · Incidence prediction

1 Introduction

1.1 Research Background

In the post-epidemic era, scholars have conducted many relevant studies from the perspectives of pathology, molecular biology, and epidemiology. Still, most of them performed classical statistical analyses based on epidemiological surveys and classical statistical analyses. The exploration of the mechanism of association between disease transmission risk and spatial environmental factors has not been identified, since the types of data obtained from epidemiological surveys are primarily quantitative or qualitative indicators, such as numerical indicators used to measure the severity of the disease, the number of infected contacts, and the type of transmission of coronavirus. Furthermore, It is challenging to avoid covariance among multiple variables and quantify each factor's impact. Therefore, it is not easy to make accurate quantitative predictions based on epidemiological surveys.

1.2 Literature Review

With the growing popularity of machine learning in urban data [1, 3, 5, 10, 11], new technologies provide new solutions for epidemic prevention and control. Big-data analysis to accurately predict the risk of urban epidemic transmission provides the design improvement of urban-related facilities with new opportunities.

Machine learning in urban epidemic prevention research from an epidemic perspective focuses on the prediction of epidemic transmission risk and urban spatial optimization analysis. For the prediction of epidemic transmission risk, Cao Zhonghao et al. created an intelligent simulation model of the COVID-19 in Guangzhou city based on GIS technology, analyzed the transmission chain pathways by studying the Spatio-temporal trajectories of different individuals, and predicted the epidemic prevention and control trend, but lacked involving of urban environment and epidemic transmission mechanism [3], Li Zhao et al. proposed an epidemic transmission risk propagation model based on coupling LSTM algorithm and cloud model by the analysis of the June 2020 outbreak in Beijing, achieved the prediction of real-time and short-term epidemic transmission risk, but its data sample is exceptional and lack in universality [7]; Peng et al. studied the transmission trend of the COVID-19 by controlling and analyzing the real-time regeneration index of the epidemic in five countries, including China, Japan, and Italy, but the research granularity is large and fail to reveal profound mechanism behind transmission [9]. Analyzing optimizing urban space design is also an essential application in machine learning to prevent the epidemic. Xin Li et al. constructed an urban risk factor analysis model by analyzing the geographic location information of confirmed neighborhoods in Wuhan at the beginning of the epidemic outbreak, which provided an optimization plan for urban planning and architectural design [8, 12].

1.3 Problem Statement and Objectives

To sum up, most of the research granularity of machine learning models based on artificial intelligence technology and corresponding indicators at home and abroad are not

suitable for exploring urban environmental risks' characteristics and can not predict and verify the interaction between urban areas' environmental risks and epidemic spread. In this paper, an image-based neural network is proposed to study the relationship between urban spatial morbidity and urban facilities design, find out the correlation mechanism between urban epidemic spread risk and spatial environment factors, and simulate and verify the epidemic outbreak in Wuhan. By using open-source data and generative antagonism networks (GANs) to experiment with various input and model types, we can easily compare the influences of different factors in the urban environment and reveal the interaction between urban environmental risks and urban design. Generating an antagonistic neural network is innovative in data types and learning methods.

2 Data and Spatialization

2.1 Study Area and Data Sources

In this paper, we selected the central area within the third ring of Wuhan city as the study object, covering about 591 km² and the seven most infected districts back to February 30. In this study, we used a multi-source data fusion approach. We modified data according to the Baidu map app and obtained the epidemic distribution data from various public websites and the official website of the Wuhan Health and Wellness Commission. The epidemic's increase in Wuhan tended to be stable after February 30 (Fig. 1), thus it could be argued that the data used in this paper are well interpreted and representative. As can be seen from the figure, the number of new infections fluctuated wildly between February 17 and 25; the spatial distribution of infected patients in the four days during this interval shows that there is an apparent spatial aggregation of infected patients, primarily in densely populated areas such as hospitals and shopping malls. The spatial migration process of patients was roughly gathered in the central district and showed a decreasing trend to the peripheral regions. Later after February 23, only the central section and scattered peripheral areas still had many infections.

2.2 Epidemic Distribution

The spatial analysis method needs to spatialize the information of infected patients based on table records in advance. The way of transmission for the disease is close contact and the spread of drop. Besides, hospitals and neighborhoods are two known places in transmission viruses. In this paper, we selected the neighborhoods address information as the geographic location of patients. After obsoleting POI data outside the third ring, a total of 30,617 data were involved as samples for this study. By combining the community-level population data released by the population census, we can improve the heat map (incidence rate), which better reflects the region's overall epidemic situation.

This paper used the kernel function (one of the most widely used methods for analyzing spatial patterns) to calculate the quantity per unit area based on the point elements to fit each outbreak distribution point, setting a smooth cone-shaped surface. The heat map in Fig. 2 shows the rough result of kernel density estimation, which indicates that

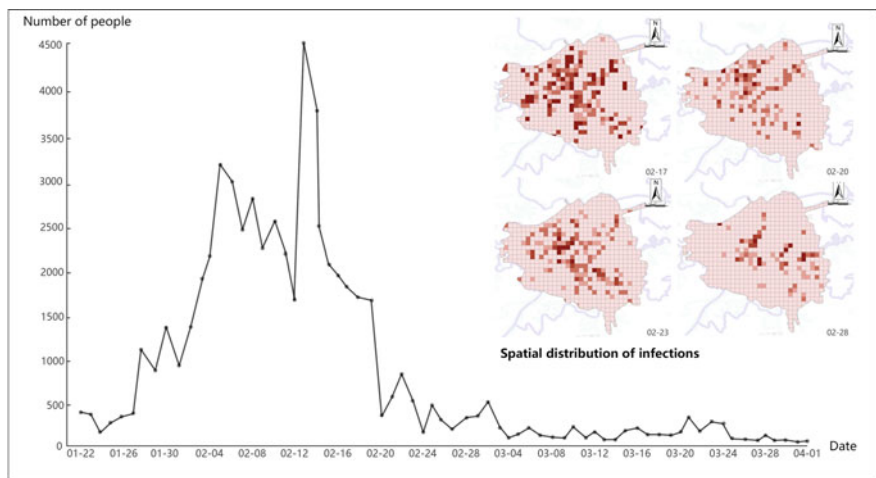


Fig. 1. Statistics of daily new infections in Wuhan.

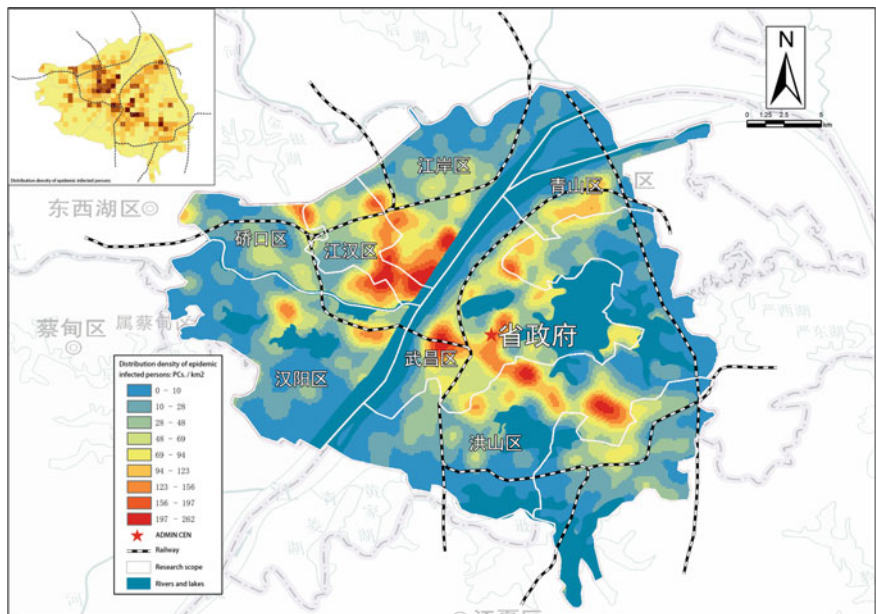


Fig. 2. Thermal and spatial risk factors of the epidemic in Wuhan.

the outbreak in Wuhan before February 30 was distributed in north–south along the river and aggregated in Wuchang, Hankou, and Hengyang. Due to the influence of natural factors such as water bodies and mountains, epidemic distribution in Wuhan is patchy; as the data in infected communities is limited by the people flow and activities frequency, it cannot accurately reflect the situation of the epidemic in the region, thus we introduce the

definition of the incidence rate applied in the study of SARS transmission in Guangzhou [2]. Therefore, we can realize the spatial visualization from corresponding population census data.

2.3 Incidence Rate and Spatialization

Incidence Rate (IR) refers to the frequency of newly reported cases in a certain period, representing the regional distribution of cases. Furthermore, it can be visualized using the same criteria as the associated geographical factors. Since this paper focuses on the influence of various urban spatial risk factors at the microscopic scale, we pre-processed a grid cell-based incident rate map in the population spatialization with more presentable information.

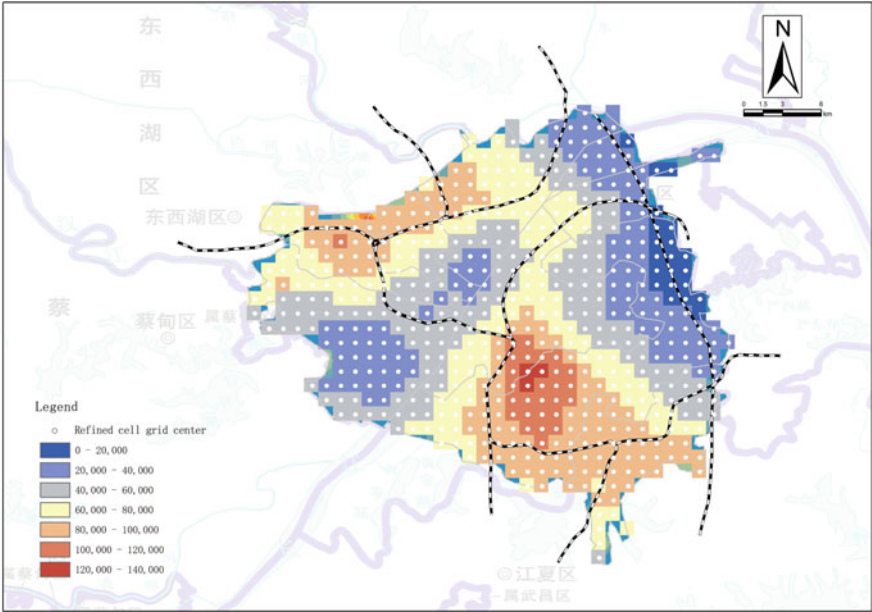
Based on the 7th national census data, taking the kriging surface interpolation model to assign the demographic data in 1 km*1 km fine grid cells (591 in total). Compared with the beginning of 2020, the spatial pattern of population distribution has not changed much due to policy control. Therefore, after correction, it could be said that the population distribution is the population distribution map on the grid by the end of 2019. The spatial distribution of population density on the grid is close to the actual distribution. Kriging interpolation involved the spatial autocorrelation characteristics of Wuhan population density in data gridding, making the calculation results of population distribution more in line with the actual situation.

In this paper, we chose the core density estimation method to estimate the spatial distribution density of infected patients in Wuhan on 1 km*1 km grid cells. The core density function is the normal Gaussian curve function. In this paper, the size of the GIS grid network is 1 km, and the kernel radius is determined to be 2.5 km after several trials, which can retain enough details and meanwhile reflect the overall trend of spatial distribution. Dividing the spatial density value of infected data within the grid element by corresponding population density, the result shows the spatial incidence rate map of COVID-19. Since the grid cells do not contain enough information to support the features needed for network training, we densified the midpoint grid points of the incidence rate again according to the above parameters. Finally, we obtained the incidence rate label sample.

It can be seen from Figs. 3 and 4 that the epidemic of Wuhan's epidemic is mainly aggregated in the central district of the urban area with residential quarters as the core, which is also the most active commercial area of Wuhan's economic activities and complex transportation system. Regarding the study, it was found that spatial risk factors such as schools, shopping malls, subway stations, hospitals, and hotels were also important factors influencing the high incidence of the New Coronary Pneumonia outbreak, so a modeling analysis of spatial risk factors and incidence rates was conducted.

2.4 Spatial Risk Factors and Model Constructs

According to the first law of geography, the object of geographic space affects regional correlation, which is an essential basis for the spatial spread of epidemics. Scholars who studied the direction of Spatio-temporal modeling of infectious diseases have pointed out that the spatial spread of SARS and other infectious diseases is closely related to



factors such as population, the environment and scope of human life, and the distribution of various other spatial influencing factors [7]. Here in this paper, with the help of Li Xin's team's definition of spatial risk factors for the spread of COVID-19 [7], we make an association test under normal distribution for its most critical urban facilities (Fig. 5).

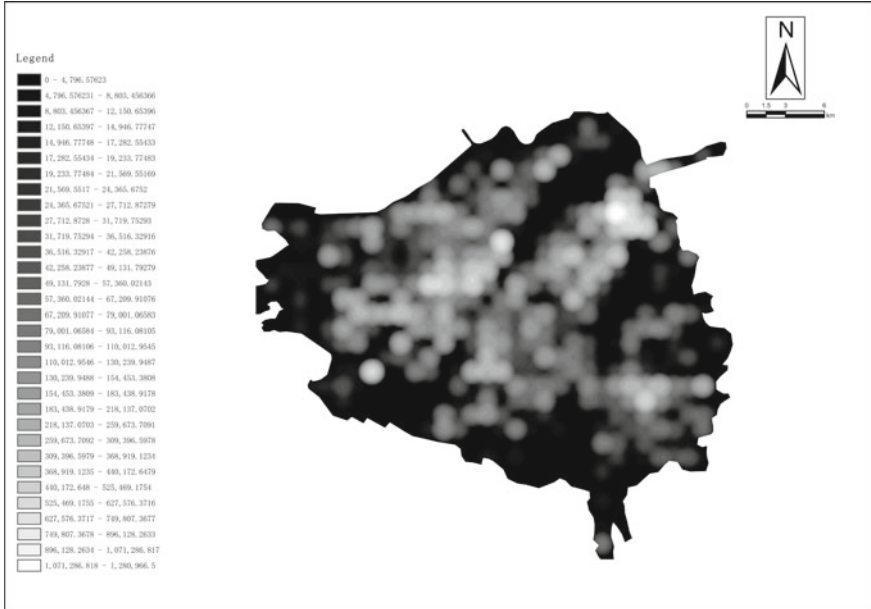


Fig. 5. Sample labels of incidence rate.

Comparing the incidence rate of COVID-19 and the density of various spatial risk factors (Fig. 6), it is evident to see a positive correspondence between them. We used 591 grid cells as samples and the Pearson index to the incidence rate and the density of various spatial risk factors using Statistical Product and Service Solutions. The results showed a significant positive correlation between each risk factor and the incidence rate of coronavirus in the two-tailed test with a significance level $P = 0.000 (<0.05)$. The strongest associations are present in schools, supermarkets, subway stations, hospitals, parks and squares, and hospitals. These areas contained the following two characteristics: high population density and high population circulation; hospitals, as the main shelter for patients earlier, had a higher risk exposure for health workers on their way to treat patients, and thus became the areas with the highest incidence rates. In contrast, spots such as subway stations, government, and park squares have a high degree of overlap in the spatial distribution of outbreaks, particularly for the locations of the highest outbreak points in these areas that almost overlapped.

Therefore, this paper used the above correlation weights with poi data distribution to build an urban spatial risk factor map with a resolution of $11,871 \times 12,630$ as a feature sample for GAN training (Fig. 7). By image processing of the incidence rate label sample (Fig. 5), we used the PIL library in Python, cut the whole incidence rate map

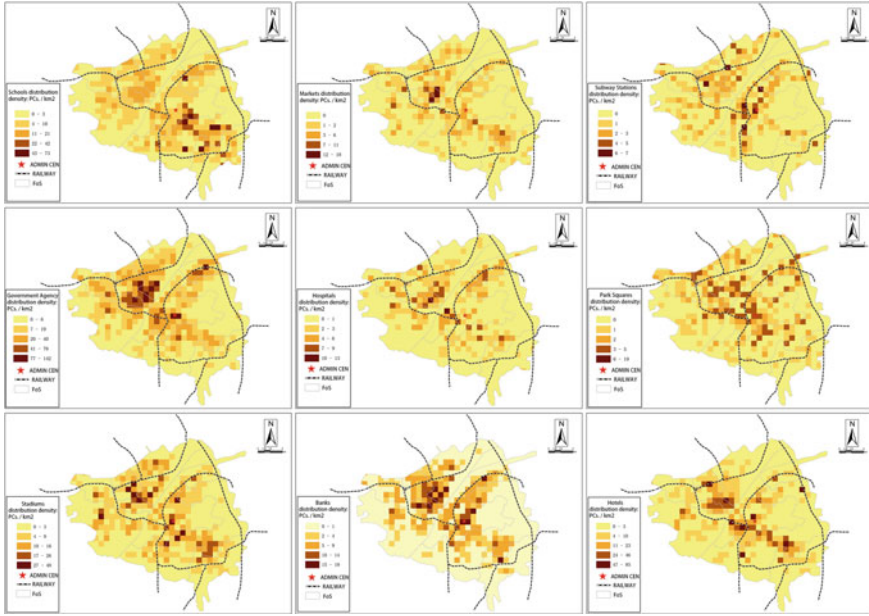


Fig. 6. Density distribution of COVID-19 spatial risk factors in Wuhan.

into small images with a resolution of 512×512 fragment. After the process, the size of each image became more suitable for machine learning. Meanwhile, the overlap the spatial risk factor feature samples with the incidence rate label samples to get the input of feature samples, acquiring 552 sets of slices of 23×24 in total. Since the validation scope used the area within the three-ring urban area, the regional boundary data were not highly reliable. After removing these data, we got 275 sets of feature and label slices. In addition, the uneven distribution of urban spatial elements results in few features and labels in rivers, greenery, and other areas. This paper manually eliminates unnecessary identification elements and obtains 225 sets of slices in total.

To describe the training model for forecasting incidence rate more conveniently, we used the value domain of incidence rate on fine spatial cells to establish its weights and implement data vectorization in rhino with the tools such as grasshopper (Fig. 7 right). The algorithm starts with the fine grid points (591 in total) as the center and the epidemic incidence value at this point as the search weight, traversing the remaining points within the connected range. The urban incidence rate grayscale map can be extracted from the grid values to achieve a better visual representation in grasshopper.

3 Neural Network Training and Data Analysis

3.1 Model Selection

The machine learning model is responsible for learning the relationship between the input urban facility layout with spatial risk information and the output of COVID-19

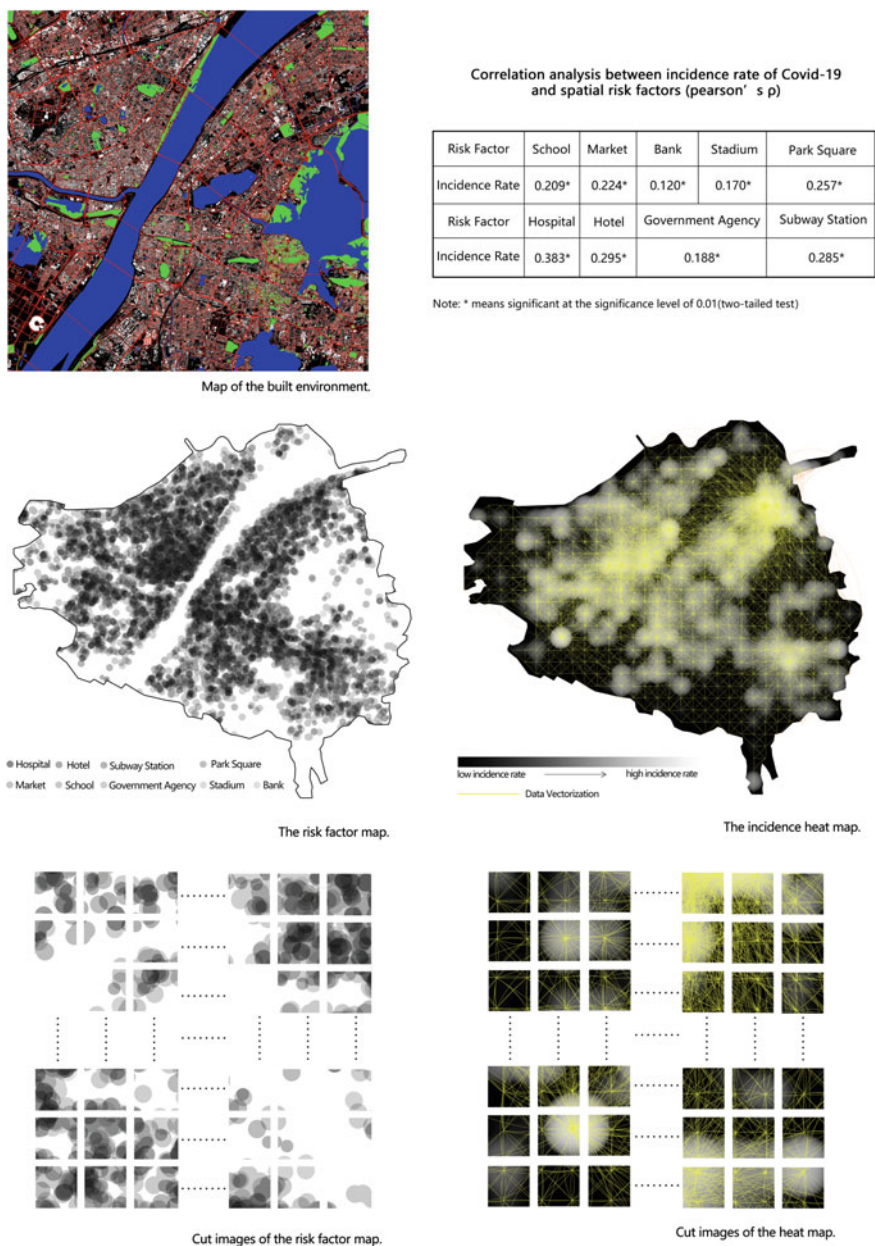


Fig. 7. Processing of features and labeled samples.

epidemic incidence distribution, trained based on an image-based GAN framework with convolution and deconvolution kernels. Conditional GAN from Goodfellow et al. [4] and pix2pixHD (an open-source project) from Isola et al. [6] were used to develop the

algorithms for this study. pix2pixHD implies a pixel-to-pixel transformation where the size of the input and output images remain constant.

In this training, the input feature sample is a spatially weighted distribution map of various facilities in the city with spatial risk information, and the output label sample is a COVID-19 onset distribution map generated by kriging interpolation. Since the data used is open source data of a single city in Wuhan, among the 275 sets of slices obtained, 80% (179 groups) were put into the training sample set and 20% (46 groups) into the test set to verify the accuracy of the model. After training, the model can predict the epidemic distribution with an unknown spatial factor map (Fig. 8).

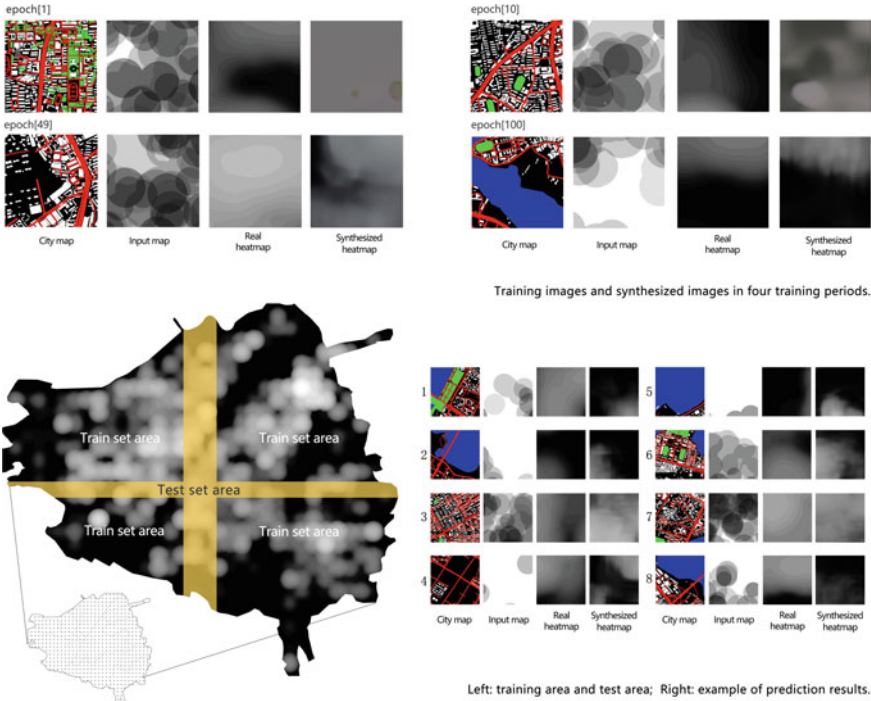


Fig. 8. Model training process.

3.2 Accuracy Measures

The training set should include all areas of the Third ring urban area of Wuhan. When the images of the test set are identified to represent different areas of the city, to reduce the influence of uneven spatial elements, we select two vertical columns and one horizontal column from the middle of the whole data area, totaling 46 groups of slices. This prediction result should be accurate because the training set includes 179 groups of screening slices in various areas of the Third Ring Road. To verify the accuracy of the model training, four ways are set to verify the fitting degree of the model. First, we

Traversed the pixel value $m(m = r*0.299 + g*0.587 + b*0.114)$ of each pixel of 46 groups of samples (512*512 resolution) in the whole test set by getpixel method in the PIL library and obtained the average difference between the predicted value and the true value of all pixel points in each group of slices. The average Generation Accuracy of this training model test set is 0.7895. This judgment method can better explain that the absolute difference between the predicted value and the true value is small. Second, to test whether the model has formed a good fit to the spatial distribution, we set the randomly generated pixel value to compare with the true value and got the average Random Accuracy of 0.6959, which shows that the predicted value of the model is productively improved compared with the random situation. Third, by inverting the true pixel value, we calculated the difference between the inverted value and the true value and got the average Inversion Accuracy of 0.6401. It indicates the situation that the predicted values are entirely false compared to the real value. Last, we took the maximum difference between 0 and 255 and the true value to maximize the relative error rate, and get the average Lowest Accuracy of 0.3201. The comparison of the Generation Accuracy, Random Accuracy, Inversion Accuracy, and Lowest Accuracy is shown in Fig. 9, in which the accuracy of our generated values is much higher than the comparative groups. When the mathematical expectation of the Random Accuracy is regarded as 0.5, and the Inverse accuracy is regarded as 0, the predicted value of our model is increased by 167.74% compared to the random guess.

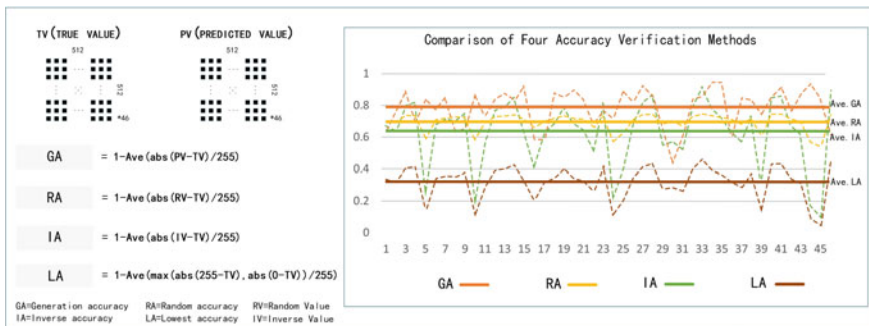


Fig. 9. Verification of model accuracy.

3.3 Layout Optimization and Prediction Application

With the trained model, we can continuously adjust design elements in urban planning to achieve the lowest public health risk by optimizing urban design plans for specific areas and predicting the spatial risk through neural networks (Fig. 10 left and middle). Changing the spatial attributes of the POI source data and testing implies that by increasing schools and hospitals and decreasing hotels and sports grounds, the predicted incidence rate has been changed evidently. In addition, assigning virtual negative correlation POI also corresponds to decreasing incidence rates, which indicates that this model learns

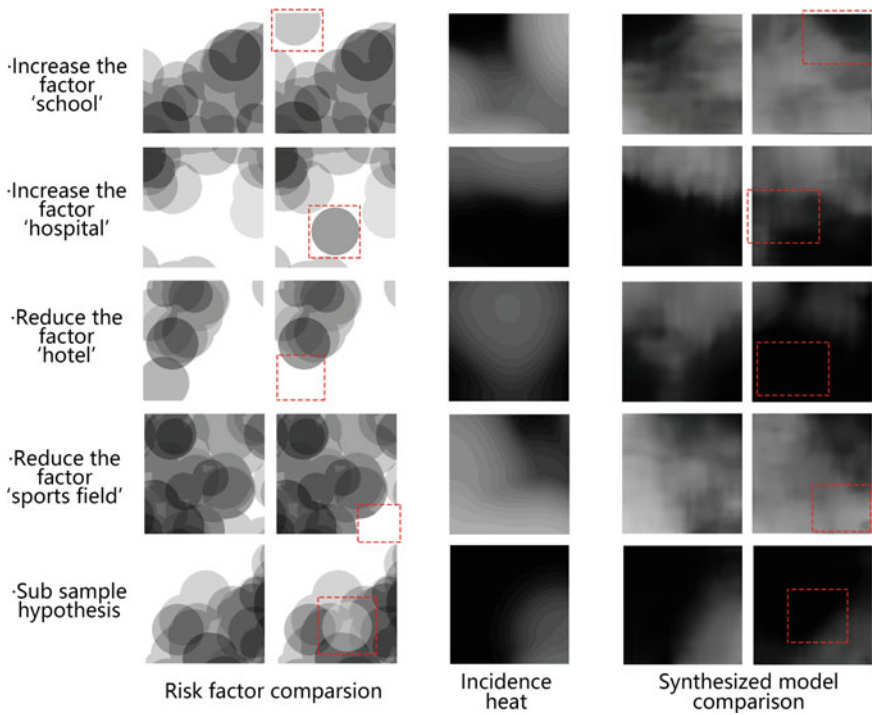


Fig. 10. The interplay between urban spatial risk and urban design.

better a particular mapping relationship between poi distribution and incidence rates and learns various risk factors weighting relationships between them.

Therefore, we can verify the interplay between urban environmental risk and urban design, revealing the potential optimization through combining multiple learning models in a logically closed loop, which quantitatively predict the epidemic incidence by modifying the figure-ground relationship and consequently iterating the urban design. It suggests that machine learning models can innovate advisors and supervisors to refine urban designers in planning and design and quantitatively achieve public health and safety.

3.4 Discussion

To more accurately visualize and predict morbidity in different urban design areas, research needs to further discuss the credibility of urban spatial risk factors. The more credible entropy correlation and ridge regression can be used to determine the correlation coefficient between urban space facilities and the epidemic distribution. At the same time, it is necessary to consider the precise quantification of the impact of various urban facilities on urban life by existing scholars. Under different circumstances The influence scope of various poi facilities is different. This model can be optimized by combining the existing quantitative analysis in the post-epidemic era and more accurate

measurement research on the epidemic, and further inputting multi-dimensional natural environment elements such as ventilation environment and street built environment such as street openness, as a complement to space risk. In addition, the experience of epidemic prevention and control in countries around the world has proved that different social policy environments and human factors have a greater impact on the prevention and control of epidemic transmission. If we can start from spatial individuals, we can understand the process of epidemic transmission at the micro-geographical scale from the perspective of evolution and emergence, will be able to better explain the dynamic, sudden, self-organized and other complex characteristics of the epidemic. Therefore, it is necessary to supplement the explanatory variables for epidemic risk factors from local grid dynamic models such as urban constrained CA.

After the model is optimized above, it can be considered that a model for predicting the current generation of new coronary pneumonia has been obtained through GAN training. The adjustment process of streets and buildings in Fig. 10 is equivalent to simulating an urban planner using the training model in this paper to obtain feedback on the prediction of epidemic incidence and adjust the design solution to ultimately achieve a reduction in the distribution of the new crown in the city. In terms of accuracy, the optimization model of this paper, which further integrates multi-source data based on the characteristics of epidemic transmission, can quantitatively predict the incidence of new crowns and provide a usable model for urban infectious disease prevention and control planning.

4 Conclusion

This paper developed a suitable neural network model using Wuhan epidemic distribution data and census plot-level data to explore and validate the following. (a) The spread of an epidemic is due to the interaction of the infectious disease's driving forces and the multiple spatial factors and population movements in the environment. (b) The top three urban facility types in terms of increasing infection are hospitals, hotels, and subway stations, which address the shortcomings of subjective evaluation and the difficulties of data acquisition compared to previous methods. (c) The generated machine learning models could be extensively used for big data, as well as provide new capabilities for studying environmental behavior. The mapping relationship between urban facilities and epidemic onset can be learned by neural networks and used to predict the spread of epidemics in other cities. (d) By building multiple learning models, not only can the principles of interactions and determinants of behavior be derived. Nevertheless, real-time feedback on the assessment results can also support urban designers in improving designs to achieve some level of safe optimality in the urban environment. This study proposes new methods for risk assessment based on refined scales, which can provide fresh ideas for future disease risk assessment, decision support for epidemic prevention and control, and security for the people. Future research will base on a multi-scale epidemic prevention system. By incorporating other dimensional elements such as environmental elements, ventilation and sunlight, and social elements such as population movement, the research will be systematically developed for epidemic transmission.

References

1. Cao S, Zheng H (2021) A POI-based machine learning method for predicting residents' health status. Proceedings of the 3rd international conference on computational design and robotic fabrication (CDRF), Shanghai, China, 139–147
2. Cao Z, Wang J, Gao Y, Han W, Feng X, Zeng G (2008) Risk factors and autocorrelation characteristics on severe acute respiratory syndrome in guangzhou. *Acta Ecol Sinica* (9):981–993
3. Cao Z, Zhang J, Yang M, Jia L, Deng S (2021) The city agent model of COVID-19 based on GIS and application: a case study of Guangzhou. *J Geo-Inf Sci* 23(02):297–306
4. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv Neural Inf Process Syst*. MIT Press, Cambridge, MA, 2672–2680
5. He J, Zheng H (2021) Prediction of crime rate in urban neighborhoods based on machine learning. *Eng Appl Artif Intell* 106:104460
6. Isola P, Zhu J, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. arXiv preprint. <https://arxiv.org/abs/1611.07004>
7. Li X, Zhou L, Jia T, Liu F, Zou Y (2021) Decoding the impact of potential urban risk factors on the COVID-19 situation: a case study of Wuhan. *City Planning Rev* 45(08):78–86
8. Li, Gao H, Dai X, SunHai H (2021) Epidemic spread risk prediction model coupled with LSTM algorithm and cloud model. *J Geo-Inf Sci* 23; No.171(11):1924–1935
9. Peng ZH, Song WY, Ding ZX et al, Linking key intervention timings to rapid declining effective reproduction number to quantify lessons against COVID-19. *Front Med*.<https://doi.org/10.1007/s11684-020-0788-3>
10. Shou X, Chen P, Zheng H (2021) Predicting the heat map of street vendors from pedestrian flow through machine learning. Proceedings of the 26th international conference on computer-aided architectural design research in Asia (CAADRIA), Hong Kong, China, pp 2.569–578
11. Sun Y, Jiang L, Zheng H (2020) A machine learning method of predicting behavior vitality using open source data. Proceedings of the 40th annual conference of the association for computer aided design in architecture (ACADIA), Philadelphia, USA, pp 160–168
12. Yao Y, Yin H, Li X, Guo Z, Ren S, Wang R, Guan Q (2021) Fine-scale risk assessment of COVID-19 in Wuhan based on multisource geographical data. *Acta Ecologica Sinica* 41(19):7493–7508

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

