RESEARCH   **Open Access**

Dubhe: a deep-learning-based B5G coverage analysis method

Haoyan Xu, Xiaolong Xu[*] and Fu Xiao

* Correspondence: xuxl@njupt.edu.cn

Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts and Telecommunications, Nanjing, China

Abstract

In recent years, with the rapid development of various technologies such as the Internet of Things and the Internet, the demand for massive device connections and a variety of differentiated new business applications has continued to increase. In order to better cope with the rapid growth of mobile data in the future, 5G also came into being. Then, B5G was proposed and applied in industries such as traditional voice/video, smart city, automotive car or ship, unmanned aerial vehicle, marine monitoring, IoT, and intelligent industry. In these scenarios, B5G is required to achieve seamless global coverage. As these scenarios are complex and changeable, analysis of the coverage of 5G base stations has become a challenge. We decompose the environment around the base station into multiple grids, and analyze the signal strength of each grid. A signal propagation model needs to be constructed to predict whether each grid is covered. The commonly used wireless propagation model is an empirical model based on a mathematical formula for statistical analysis of a large amount of test data during the establishment of a 5G local area network. It has universal applicability, but has insufficient prediction accuracy for specific scenarios. Therefore, it is necessary to calibrate and modify the typical propagation model according to the specific environment to obtain an accurate propagation model that matches the current area. We improved the traditional wireless communication model, and proposed a deep-learning-based B5G coverage analysis method named *Dubhe* which is one of the planets of the Big Dipper. In a real cell scenario, the mean square error of the link budget of the typical UMa model is 17.9 dBm, while the mean square error of the proposed *Dubhe* model constructed in this article is only 6.78 dBm. The recognition rate of weak coverage can reach 42.86%.

Keywords: B5G, Link budget, Deep learning, Geographic information

1 Introduction

With the development of B5G technology [1], the application of B5G on a global scale continues to expand at the same time. In the process of deploying B5G networks, operators need to reasonably select base station sites within the coverage area to achieve strong coverage in all areas. In the entire wireless network planning process, efficient network expansion is of great significance for accurate B5G network deployment. Due to the complex propagation

environment of radio waves, it will be affected by various factors on the propagation path, such as plains, mountains, buildings, lakes, oceans, forests, atmosphere, and the curvature of the earth itself. This makes electromagnetic waves no longer propagate in a single way and path and produce complex transmission, diffraction, scattering, reflection, refraction, etc. Therefore, it is a very difficult task to establish an accurate model.

Existing wireless propagation models can be distinguished according to research methods, which are generally divided into empirical models, theoretical models, and improved empirical models. The empirical model is obtained by obtaining a fixed fitting formula from empirical data. Typical empirical models include Cost 231-Hata [2], Okumura-Hata [3], etc. Theoretical model is based on the electromagnetic wave propagation theory, considering the reflection, diffraction, and refraction of electromagnetic waves in space to calculate the loss. The more representative theoretical model is volcano model [4]. The improved empirical model can provide a calculation model for more detailed classification scenarios by introducing more parameters into the fitting formula. A typical improved empirical model is SPM [5].

In the propagation model modeling, in order to obtain a propagation model that meets the actual environment of the target area, a large amount of additional measured data, engineering parameters, and electronic maps need to be collected to correct the propagation model. In addition, wireless LTE networks have become popular all over the world, with billions of users around the world generating large amounts of data every moment. How to reasonably use these data to assist wireless network construction has become an important topic.

In recent years, big data-driven AI machine learning technology has made great progress, and it has been successfully used in the field of language and image processing. With the development of parallel computing architecture, machine learning technology also has the ability of online computing. Its high real-time and low complexity makes it possible to closely integrate with wireless communication. The original B5G wireless propagation model is an empirical model derived from mathematical formulas based on statistical analysis of a large number of test data under the establishment of a B5G local area network. It has universal applicability, but the prediction accuracy for specific scenarios is insufficient [6]. Therefore, it is necessary to calibrate and modify the typical propagation model according to the specific environment to obtain an accurate propagation model that matches the current area.

In response to the above problems, we propose a B5G coverage analysis method based on deep learning. Geographic information and signal characteristics are analyzed by the deep learning model named *Dubhe* to calculate the path loss and coverage of the signal in a complex environment. *Dubhe* can accurately predict the wireless signal coverage strength in the new environment, which can reduce network construction costs and improve network construction efficiency.

To summarize, the contributions of this paper are listed as follows:

1. Analyze and summarize typical empirical models, and analyze the shortcomings of these models in B5G scenarios.
2. We use deep learning methods to calibrate and modify the empirical models according to the specific environment to obtain a more accurate model that matches the current region.
3. We use deep learning technology to establish a signal coverage analysis model *Dubhe*, and accurately analyze the strength of wireless signal coverage in various scenarios. It can reduce the cost of base station construction, improve the efficiency of base station construction, and achieve full range of B5G signal coverage.

The rest of this paper is organized as follows: Section 2 introduces classical signal propagation model. Section 3 describes the dataset and the details of the proposed method. In Section 4, the experimental results are analyzed and compared. Section 5 closes with a summary and conclusion.

2 Typical empirical models

2.1 Cost 231-Hata

The Hata model is a classic model of signal link budget in cities. The COST-231 Hata model is an extended version of the Hata model, with an application frequency of 1500 ~ 2000 MHz [7]. The definition of Cost 231-Hata is as follows:

$$PL = 46.3 + 33.9 \log_{10}f - 13.82 \log_{10}h_b - \alpha + \left(44.9 - 6.55 \log_{10}d\right) + C_m \tag{1}$$

where PL is defined as the propagation path loss (dB), f is the carrier frequency (MHz), h_b is the effective height of the base station antenna (m), h_{ue} is the effective height of the user antenna (m), α is the user antenna height correction term (dB), d is the link distance (km), and C_m is the scene correction constant (dB).

The 5G signal frequency bands used in the People's Republic of China mainly include the following:

China Mobile: 2515 MHz–2675 MHz, a total of 160 MHz, the frequency band number is n41, 4800 MHz–4900 MHz, a total of 100MHz, the frequency band number is n79.

China Telecom: 3400 MHz–3500 MHz, a total of 100MHz, the frequency band number is n78.

China Unicom: 3500 MHz–3600 MHz total 100 MHz, frequency band number is n78.

However, the application frequency of Cost 231-Hata is 1500 ~ 2000 MHz, and the calculated PL can only be used as a reference feature [8].

2.2 UMa

The Uma [8, 9] model is a new propagation model defined in the 3GPP protocol that is suitable for the current 5G high frequency development trend. The channel measurement frequency range is 0.5G–100 GHz, and the signal propagation effective distance is 10–5000 m [10]. 3GPP protocol 38.901 defines UMa, and its empirical formula is as follows:

Under line of sight (LOS) conditions:

$$PL_{UMa-LOS} = \begin{cases} PL_1, 10\text{m} < d_{2D} \leq d'_{BP} \\ PL_2, d'_{BP} < d_{2D} \leq 5\text{km} \end{cases} \tag{2}$$

where

$$PL_1 = 28.0 + 22 \log_{10}\{d_{3D}\}_{\text{m}} + 20 \log_{10}\{f_c\}_{\text{GHz}} \tag{3}$$

$$PL_2 = 28.0 + 40 \log_{10}\{d_{3D}\}_{\text{m}} \\ + 20 \log_{10}\{f_c\}_{\text{GHz}} - 9 \log_{10}\left(\{d'_{BP}\}^2_{\text{m}} + \left(\{h_{\text{BS}}\}_{\text{m}} - \{h_{\text{UT}}\}_{\text{m}}\right)^2\right) \tag{4}$$

Under non-line-of-sight (NLOS) conditions:

$$PL_{\text{UMa-NLOS}} = \max(PL_{\text{UMa-LOS}}, PL'_{\text{UMa-NLOS}}), 10\text{m} \leq d_{2D} \leq 5\text{km} \tag{5}$$

$$PL'_{\text{UMa-NLOS}} = 13.54 + 39.08 \log10\{d_{3D}\}_{\text{m}} + 20 \log_{10}(f_c)_{\text{GHz}} - 0.6(\{h_{\text{UT}}\} - 1.5) \tag{6}$$

where $PL_{UMa-LOS}$ is the path loss under line-of-sight propagation. $PL_{\text{UMa-NLOS}}$ is the path loss under non-line-of-sight propagation. d_{2D} is the horizontal distance from the mobile terminal to the base station; d'_{BP} is the breakpoint distance set for the model. d_{3D} is the straight-line distance from the mobile terminal to the base station. f_c is the signal carrier frequency used by the base station. h_{BS} is the effective height of the base station antenna. h_{UT} is the effective height of the user terminal.

The path loss under non-line-of-sight propagation can also be simplified as

$$PL = 32.4 + 20 \log_{10}\{f_c\}_{\text{GHz}} + 30 \log_{10}\{d_{3D}\}_{\text{m}} \tag{7}$$

The initial parameter setting of this model is that the base station antenna height is 25 m, and the user terminal height range is 1.5–22.5 m. We need to optimize the design of the basic propagation model according to the actual base station layout, actual propagation environment and user terminal measurement conditions to adapt to the real regional signal propagation laws.

2.3 Limitations of typical empirical models

Typical empirical models are difficult to apply in complex signal propagation scenarios. Cost 231-Hata is only suitable for macro cellular systems with a cell radius greater than 1 km, while the UMA model is suitable for typical urban scenarios. These traditional models can only be used in specific environments. There are also restrictions on the application frequency, Cost 231-Hata is limited to 1500 ~ 2000 MHz, and Uma is limited to 0.5 G–100 GHz. In the case of line-of-sight and non-line-of-sight, the parameters of the propagation model also need to be changed. This makes these traditional models unable to analyze effectively when faced with complex environments. Observing the Cost 231-Hata and UMa models, it is not difficult to find that there are common parts in the calculation formula, and accurate analysis can be performed only by adjusting the parameters. In view of these problems, it is very important to propose a general model that can automatically adjust parameters.

3 Materials and methods

3.1 Materials

The data used in this article is a wireless signal propagation data set provided by Huawei. The data set contains measured data of 4000 cells, with a total of 12,011,833 pieces of data. It is divided into three parts: engineering parameter data, map data, and received power.

3.1.1 Engineering parameter data

The engineering parameter data records the engineering parameter information of the stations in a certain cell, and there are 9 fields in total. The corresponding meaning of each field is shown in Table 1.

In order to facilitate data processing, the map is rasterized, each grid represents an area of 5 m × 5 m. As shown in Fig. 1, (*Cell X*, *Cell Y*) records the coordinates of the upper left corner of the grid where the site is located. Other engineering parameters (height, azimuth, electrical downtilt, mechanical downtilt) are shown in Fig. 2. The *mechanical downtilt* is achieved by adjusting the bracket behind the antenna panel, which is a physical signal downtilt. The *electrical downtilt* is achieved by adjusting the coil inside the antenna, which is an electrical signal downtilt. The actual downward tilt angle of the signal line is the sum of the mechanical downward tilt angle and the electrical downward tilt angle.

3.1.2 Geographic information data

Geographic information data records information such as topography and landforms. There are 8 fields in total, and the corresponding meaning of each field is shown in Table 2. Considering the diversity and complexity of map types, actual features such as urban areas, rural areas, and lakes are abstracted into numbers, and these numbers are called feature type name numbers. In Table 3, you can see the actual feature type corresponding to the feature type name number.

Like the engineering parameter data, the map data are also rasterized. Each grid represents an area of 5 m × 5 m, where (*X*, *Y*) records the coordinates of the upper left corner of the grid where the map is located.

Table 1 Description of engineering parameter data

Name	Description	Unit
Cell index	Unique ID of the cell	–
Cell X	The grid position of the site to which the cell belongs, X coordinate	–
Cell Y	The grid position of the site to which the cell belongs, Y coordinate	–
Height	The height of the transmitter relative to the ground	m
Azimuth	Horizontal direction angle of cell transmitter	Deg
Electrical downtilt	Vertical electric downtilt of cell transmitter	Deg
Mechanical downtilt	Vertical mechanical downtilt of cell transmitter	Deg
Frequency band	Cell transmitter center frequency	MHz
RS power	Cell transmitter transmit power	dBm

Fig. 1 Coordinate description of rasterized map
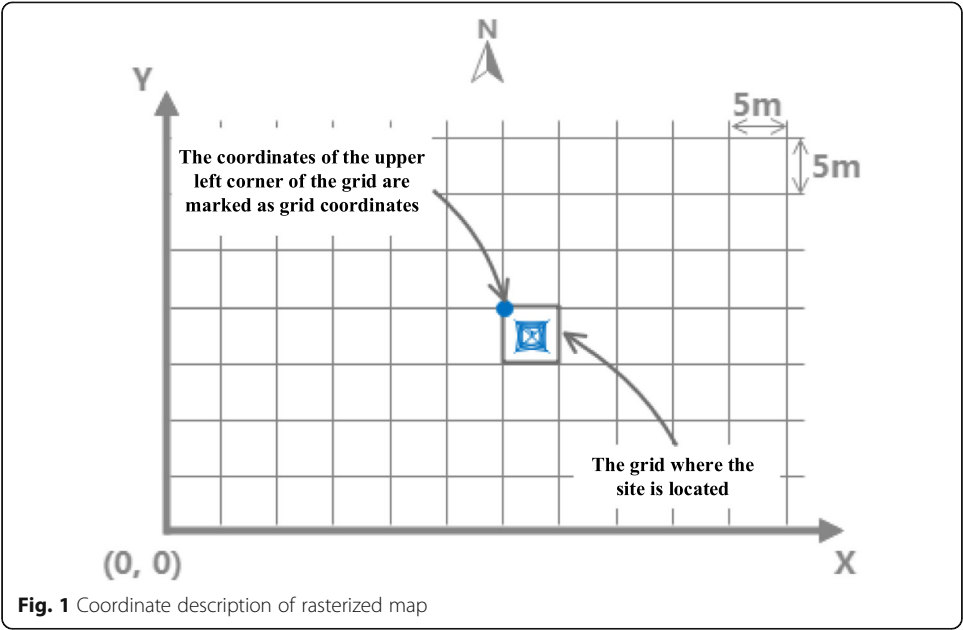
3.1.3 RSRP tag data

The RSRP tag data are used as the actual measurement result to be compared with the result predicted by the machine learning model in supervised learning. There is 1 field in total, and the corresponding meaning is shown in Table 4.

3.2 Data preprocessing

We transform the information obtained from the map into three feature maps required for training: height feature map, scene feature map, and signal feature map.
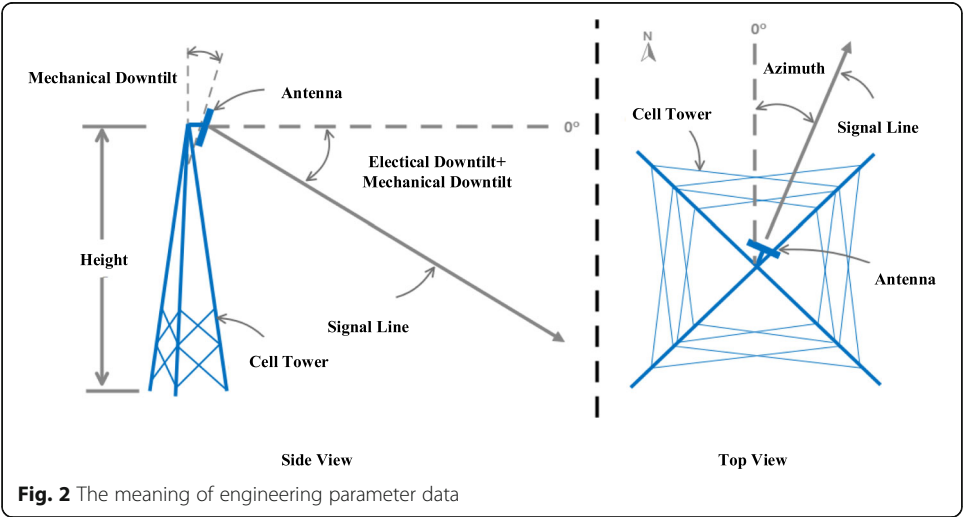

Fig. 2 The meaning of engineering parameter data

Table 2 Description of geographic information data

Name	Description	Unit
Cell building height	The building height of the grid (cell *X*, cell *Y*) where the cell site is located. If the grid has no buildings, it is 0	m
Cell altitude	The altitude of the grid (cell *X*, cell *Y*) where the cell site is located	m
Cell clutter index	The feature type index of the grid (cell *X*, cell *Y*) where the cell site is located	–
1.1.1.1.1.1.1.1. *X*	Grid position, *X* coordinate	–
1.1.1.1.1.1.1.2. *Y*	Grid position, *Y* coordinate	–
Building height	Grid position, the height of the building on the *Y* coordinate grid (*X*, *Y*), if the grid has no buildings, it is 0	m
Altitude	Altitude on the grid (*X*, *Y*)	m
Clutter index	The index of the feature type on the grid (*X*, *Y*)	–

3.2.1 Height feature map

The height feature map includes *height* (h_b), *cell building height, cell altitude, altitude,* and *building height.*, as shown in Fig. 3.

The relative height Δh_v of the grid and the signal line can be calculated through the *height* of the transmitter relative to the ground, the mechanical downtilt θ_{MD}, the vertical electrical downtilt θ_{MD}, the grid position of the transmitter and the target grid position, the *altitude* of the base station, and the altitude of the signal receiving place.

First, we need to calculate the distance *d* between the two grids.

$$d = 5 \times \sqrt{(cellX - X)^2 + (cellY - Y)^2} \tag{8}$$

The calculation method of Δh_v is as follows:

$$\Delta h = h_b + CellAltitude - (\tan(\theta_{MD} + \theta_{ED}) \cdot d + Altitude) \tag{9}$$

Table 3 Number meaning of feature type name

Clutter index	Description	Clutter index	Description
1	ocean	11	High-rise buildings in urban area (40m ~ 60 m)
2	lake	12	Medium- and high-rise buildings in urban area (20 m ~ 40 m)
3	Wetlands	13	High-density buildings < 20 m in urban area
4	Suburban open area	14	Multi-storey buildings < 20 m in urban area
5	Urban open area	15	Low-density industrial building area
6	Road open area	16	High-density industrial building area
7	Vegetation area	17	Suburb
8	Shrub vegetation	18	Developed suburban area
9	Forest vegetation	19	Countryside
10	Urban super high rise buildings (> 60 m)	20	CBD business circle

Table 4 Description of RSRP label data table

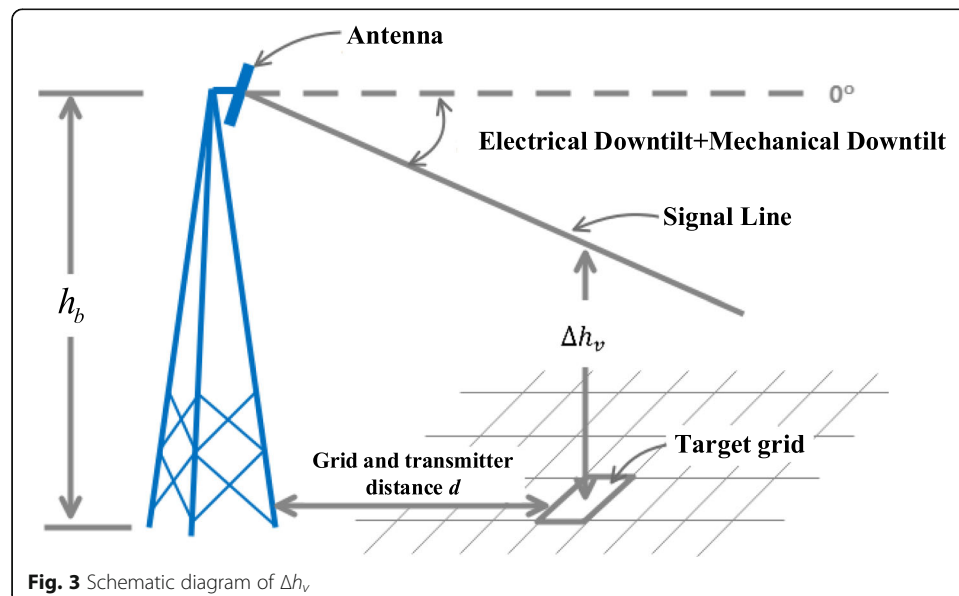| Name	Description	Unit
RSRP	Reference signal receiving power of grid (X, Y), label column	dBm

3.2.2 Scene feature map

B5G signal is sent by the transmitter in a straight line. According to the position of the transmitter and the signal receiver, the grid number of 20 scenes that the signal passes through is counted to form the scene feature map.
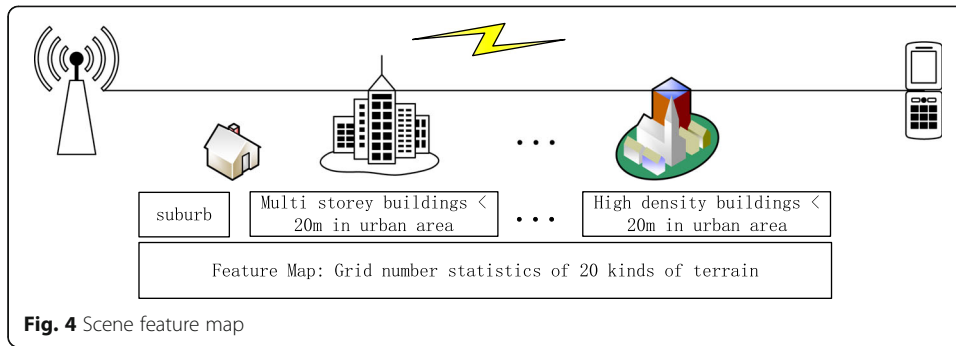
The scenes includes: ocean, lake, wetlands, open suburban areas, urban open areas, road open area, vegetation area, shrub vegetation, forest vegetation, urban super high-rise buildings (> 60 m), high-rise buildings in urban area (40 m ∼ 60 m), medium- and high-rise buildings in urban area (20 m ∼ 40 m), high-density buildings < 20 m in urban area, multi-storey buildings < 20 m in urban area, low-density industrial building area, high-density industrial building area, suburb, developed suburban area, countryside, and CBD business circle. The specific scene feature map is shown in Fig. 4.

Through the inequality $\frac{\left|\frac{CellX \cdot x}{CellY} - y - \frac{CellY \cdot X}{CellX} + Y\right|}{\sqrt{\frac{y}{x} + 1}} < 5$, the terrain with a radius of five grids for each point on the straight line from the transmitter to the signal receiving end can be calculated. Among them, (Cell X, Cell Y), (X, Y) and (x, y) are the transmitter coordinates, the receiver coordinates and the terrain coordinates of the pathway, respectively. Counting all (x, y) terrain numbers is the feature map of the terrain.

3.2.3 Signal feature map

The signal characteristic graph is composed of distance from transmitter to signal receiving end, signal transmission power of transmitter, and transmitter frequency.

Fig. 3 Schematic diagram of Δh_v

Xu *et al. EURASIP Journal on Advances in Signal Processing* (2021) 2021:65

Page 9 of 18

Fig. 4 Scene feature map

3.3 Dubhe: 5G link budget model

Comprehensive formulas (1)–7), the path loss of different areas and different terrains can be simplified to the following formula:

$$PL = K_1 + K_2 \log_{10}\{f_c\}_{\text{GHz}} + K_3 \log_{10}\{d_{3D}\}_{\text{m}} \tag{10}$$

where K_1, K_2, and K_3 are the unknown weight. This paper fully considers each terrain through which the le1 signal propagates, and obtains different weights for each terrain through the deep learning method, which further improves the applicability of the model. *Dubhe* uses a deep learning model to assign values to K_1, K_2, and K_3.

3.3.1 K_1

From the UMa model, it can be seen that there are three values for K_1:

$$K_1 = \begin{cases} 28.0 \\ 28\text{–}9 \log_{10}\left(\{d'_{BP}\}_{\text{m}}^2 + \left(\{h_{\text{BS}}\}_{\text{m}}\text{–}\{h_{\text{UT}}\}_{\text{m}}\right)^2\right) \\ 13.54\text{–}0.6(\{h_{\text{UT}}\}\text{–}1.5) \end{cases} \tag{11}$$

K_1 can be regarded as the sum of a scene correction constant K_{1_1} and a function K_{1_3} related to height, distance, and scene.

K_{1_1} is mainly determined by the scene through which the signal propagates. We calculate a scene correction constant K_{1_1} according to the scene feature map through the convolutional neural network.

First, we use a multi-layer BP neural network to calculate a value K_{1_2} based on the height feature map. Then use the BP neural network to calculate the frequency band, distance, RS power, and K_{1_1} to obtain a scene weight W about K_{1_2}. Multiply W and K_{1_2} to get the value of K_{1_3}. The value of K_1 is K_{1_1} plus K_{1_3}.

3.3.2 K_2 and K_3

The scene feature map, the height feature map and the signal feature map are spliced into a one-dimensional vector and the values of K_2 and K_3 are calculated using a convolutional neural network.

The pseudo code of the model is as follows:

ALGORITHM: *Dubhe*

Input:	$Data_1 = [Scene\ feature\ map]$
	$Data_2 = [Signal\ feature\ map]$
	$Data_3 = [Height\ feature\ map,\ Data_2\ [1]]$
	$Data_4 = [Scene\ feature\ map,\ Height\ feature\ map,\ Signal\ feature\ map]$
Output:	Result = RSRP // The subscript of the category with the highest probability is output

```
1:     Def K1( Data₁ , Data₂ , Data₃ ):
2:        C = Conv( Data₁ *W)
3:        B = Batch normalization(C)
4:        A = Activation(B)
5:        D = Dropout(A)
6:        P = Pooling(D)
7:        K₁_₁ = FullyConnected(P)
8:        W = BP([ K₁_₁ , Data₂ ])
9:        K₁_₂ = BP( Data₃ )
10:       K₁_₃ = W* K₁_₂
11:       K₁ = K₁_₂ + K₁_₃
12:       return  K₁
13:    Def K2_K3( Data₄ ):
14:       C = Conv( Data₄ )
15:       B = batch normalization(C)
16:       A = Activation(B)
17:       D = Dropout(A)
18:       P = Pooling(D)
19:       [ K₂ , K₃ ]= FullyConnected(P)
20:       return [ K₂ , K₃ ]
21:    K₁ = K1( Data₁ , Data₂ , Data₃ )
22:    K₂ , K₃ =K2_K3( Data₄ )
23:    f_c = Data₂ [0]
24:    d_3D = Data₂ [1]
25:    P_t = Data₂ [2]
26:    PL = K₁+K₂ log₁₀ {f_c}_GHz +K₃ log₁₀{d_3D}_m
27:    RSRP= P_t -PL
```

The overall calculation process is shown in Fig. 5. The overall model consists of multiple modules.

3.3.3 Input layer

The input data are four vectors. Vector 1 is a topographic feature map. Vector 2 is the terrain feature value obtained by convolution of the signal feature map and the terrain feature map. Vector 3 is a vector composed of height feature map and distance. Vector 4 is a vector composed of terrain feature map, height feature map, and signal feature map.

3.3.4 Convolution

The convolutional layer contains 16 convolution kernels of length 4. In the conv layer, the input sequence of each convolution kernel and the boundary of the input layer is 0. Make the length of the input sequence and output sequence the same. When the scene feature map is convolved, the parameters of all convolution kernels are made non-negative, and it is ensured that the parameter W of the first layer is positively correlated with the output of the last layer. Furthermore, through the analysis of W, the influence of each scene on the signal can be analyzed.

3.3.5 Batch normalization layer

Batch normalization [11] is to change the distribution of the input value of any neuron in each layer of the neural network to a standard normal distribution with a mean value of 0 and a variance of 1. This makes the activation input value fall in a region with a
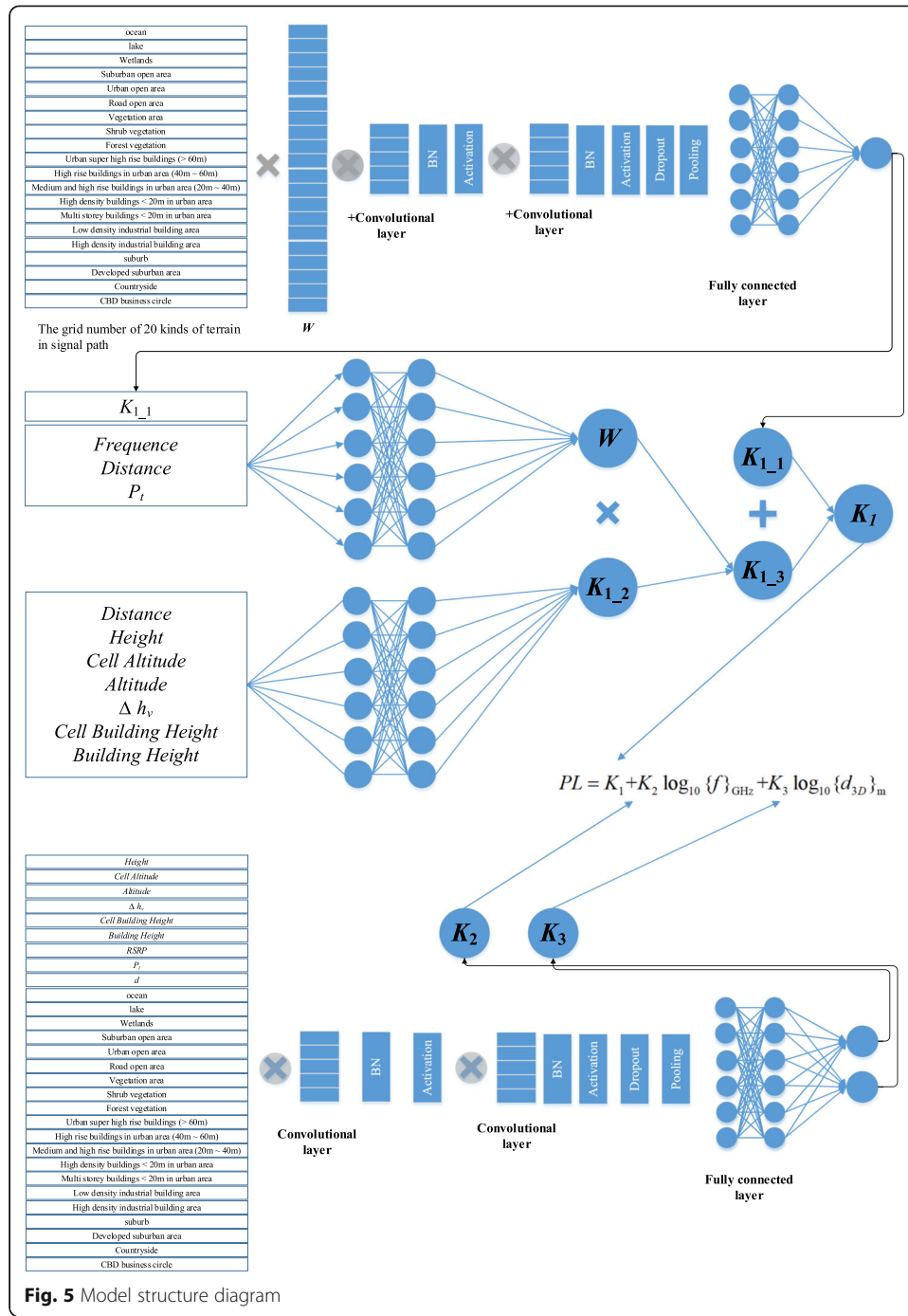
Fig. 5 Model structure diagram

larger gradient, avoids the problem of gradient disappearance, and accelerates the convergence of the loss function.

3.3.6 Activation function

The activation function of Relu [12] can be used for more effective gradient descent and back propagation to avoid gradient explosion and gradient disappearance. The calculation process is simplified and is not affected by other complex activation functions

(such as exponential functions). The discrete nature of activities reduces the overall computational cost of neural networks.

$$relu = \ \max(0, x) \tag{12}$$

3.3.7 Dropout layer
Dropout refers to temporarily discarding part of the neural network unit from the network according to a certain probability during the model training process to speed up the model training speed and prevent overfitting [13].

3.3.8 Pooling layer
Max pooling is to take the point with the largest value in the local acceptance domain. Pooling can keep the main features while reducing the parameters and calculations to prevent overfitting.

$$m_i = \ \max(\{Y_i, Y_{i+1}, Y_{i+2}, Y_{i+3}\}) \tag{13}$$

3.3.9 Fully connected layer and band gate
The fully connected layer is used to map the feature space to the label space. First, all the extracted features are transformed into a one-dimensional vector. Then connect all the eigenvalues to the neurons in the fully connected layer. After each calculation, the non-linearity is improved by the relu activation function, and finally connected to one or more neuron output.

3.3.10 Adam optimizer
Adam is considered to be a better optimizer by default in many cases. Comprehensively consider the first moment estimation and the second moment estimation to calculate the update step. The implementation is simple, the calculation is efficient, and the memory requirement is small. The update of the parameters is not affected by the scaling transformation of the gradient. Hyperparameters are very explanatory and usually do not need to be adjusted or require very little fine-tuning. The learning rate can be adjusted automatically. It is suitable for large-scale data and parameter scenarios.

4 Results and discussion
4.1 Performance evaluation index
4.1.1 Poor coverage recognition rate (PCRR)
In the process of forecasting, if weak coverage areas can be effectively identified, it can better help operators to accurately plan and optimize networks to improve customer experience. Therefore, in addition to RMSE as an effective test target, the accuracy of weak coverage recognition is also a very valuable evaluation index.

We divide the data into four types, as shown in Table 5. The value of the weak coverage decision threshold P_{th} is set to $- 103$ dBm. If the predicted or measured RSRP value is less than P_{th}, it is weak coverage and marked as 1, and if it is greater than or equal to P_{th}, it is non-weak coverage and marked as 0. According to the difference between the weak coverage and the non-weak coverage obtained by comparing the predicted value with the measured value, the following parameters can be counted:

Table 5 Definition of TP, FP, FN, and TN

		Real result	
		True (weak coverage)	False (non-weak coverage)
Predicted result	Positive (weak coverage)	TP	FP
	Negative (non-weak coverage)	FN	TN

- True positive (TP): the true value is weak coverage, and the predicted value is also weak coverage.
- False positive (FP): the true value is non-weak coverage, and the predicted value is weak coverage.
- False negative (FN): the true value is weak coverage, and the predicted value is non-weak coverage.
- True negative (TN): the true value is non-weak coverage, and the predicted value is also non-weak coverage.

PCRR comprehensively considers the goals of Precision and Recall, and its calculation formula is as follows:

$$\text{PCRR} = 2*\frac{\text{Precision}*\text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Precision can be understood as the probability that the grid that is predicted to be weakly covered is actually also weakly covered, which is defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

Recall can be understood as the probability that the grid whose real result is weak coverage is predicted to be weak coverage. It is defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

4.1.2 Root mean squared error (RMSE)

RMSE is an index to evaluate the overall deviation between the predicted value and the actual measured value, and its magnitude intuitively expresses the simulation accuracy. Calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P^{(i)} - \hat{P}^{(i)}\right)^2} \tag{17}$$

where $P^{(i)}$ is the RSRP predicted value of the ith evaluation data set by the machine learning model, and $\hat{P}^{(i)}$ is the actual RSRP measured value of the i-th evaluation data set.

4.2 Experimental results

4.2.1 Model validity evaluation

The parameters of model training are shown in Table 6.

Epoch represents the number of rounds of training in the training set. Learning rate represents the step size of each update parameter. beta_1 and beta_2 represent the

Table 6 Experimental parameters

Parameter	Value
Epoch	20
Learning rate	0.0001
beta_1, beta_2	0.9, 0.999
Dropout	0.05
Batch size	1280

exponential decay rate of the first and second estimates. Dropout makes the neurons with a certain probability not to participate in the calculation. Batch size represents the size of each training batch.

According to the standards established by Huawei, signal energy lower than − 103 dbm is divided into weak coverage areas, and signal energy higher than 110 dbm is classified into strong coverage areas. Select 10% of the data set as the test set, and the sample size of positive is 150,631. The sample size of negative is 868,976.

According to Table 7, it can be calculated that the precession = 30.55%, recall = 71.74%

The validity of the model can be judged by PCRR. In the standards developed by Huawei, the model's weak coverage recognition rate (PCRR) must be greater than or equal to 20% in order to be considered a valid model. The PCRR of our model can reach 42.86%.

4.2.2 Model performance evaluation

We fully considered all the geographic information through which the signal propagated and analyzed the map features constructed by geographic information. Finally, we used the artificial intelligence model to modify the Uma model, verify it on the real collected data set, and predict the received signal power. The final average RMSE of each sample is 6.78 dBm. Figure 6 shows *Dubhe*'s training process for each step on the test set.

With Cost 231-Hata, the RMSE of the sample is 18.6 dBm. In Method 1, using the deep learning method to predict the RMSE of RSRP is 7.384 dBm [14]. Method 2 uses multiple linear regression models to predict RSRP [14]. Method 3 uses the XGBoost model to predict RSRP [14]. In addition, we calculated the RMSE of the traditional cost231-Hata and the uncorrected Uma model. The mean square error of cost231-Hata is 18.6 dBm, while the mean square error of the Uma model is 17.9 dBm. The comparison of these six methods is shown in Fig. 7.

It can be seen from Table 8 that the performance of the proposed method has been greatly improved, compared with the traditional empirical models Uma and Costco31-Hata. Compared with other existing work, it has also improved a lot.

Table 7 Number of TP, FP, FN, and TN in test data set

	TP	FN	TN	FP
Number of samples	108,063	42,568	623,403	245,573

Fig. 6 Model training results
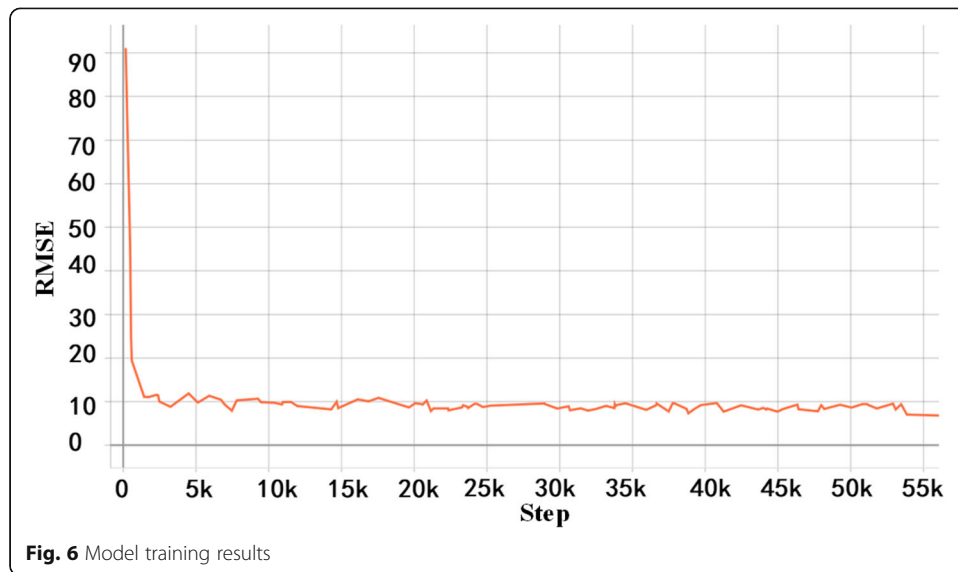
4.2.3 Versatility of the model

The model we designed is more versatile and can be used in more environments. The 20 scenarios mentioned in this article can all use Dubhe to analyze the propagation loss and coverage of 5G signals. In addition, Dubhe analyzed the terrain of all grids of 5 m × 5 m through which the signal passed. Therefore, Dubhe can also be used well in complex scenes. Dubhe can analyze signal propagation loss and coverage well when the scene changes or in complex composite scenes.

5 Discussion

In the Dubhe, the weight matrix W multiplied by the number of each scene helps us to further analyze the influence of each scene on signal propagation. The weight of each terrain is shown in Table 9.
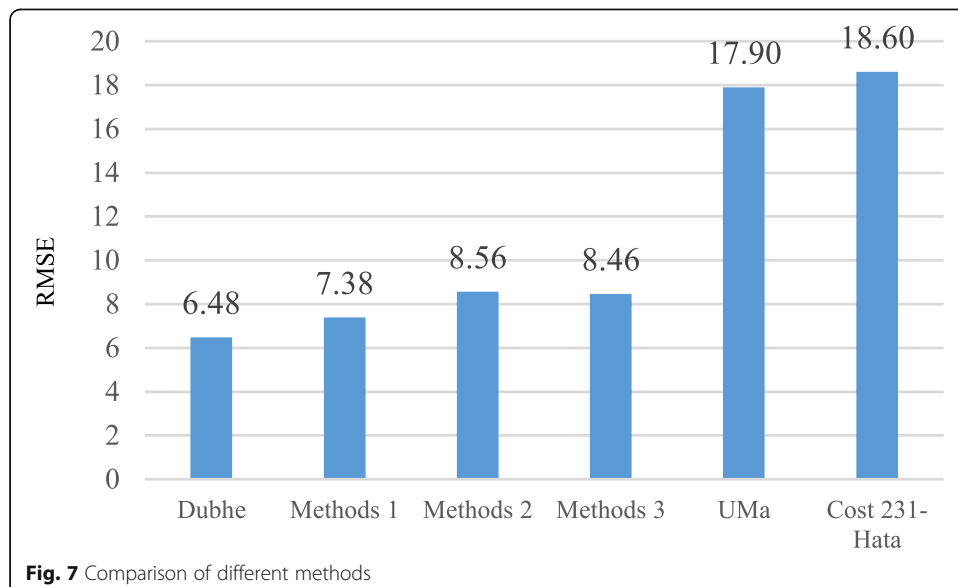

Fig. 7 Comparison of different methods

Table 8 Comparison of various methods

Method	RMSE
Deep learning	7.38 dBm
linear regression	8.56 dBm
XGBoost	8.46 dBm
cost231-Hata	18.6 dBm
Uma	17.9 dBm
Dubhe	6.48 dBm

Since the objective function is the signal strength of the area and the shared positive parameters are used later, the larger the weight, the stronger the positive correlation with the signal strength, and the smaller the number, the greater the signal loss. From Fig. 8, we can see that topography 1, 4, 7, 8, and 10 cause the least loss of signal propagation, while topography 3, 13, 14, and 18 are not conducive to signal propagation and will increase the loss. In areas with large signal loss, the construction of base stations should be strengthened to better cover the area.

After the model training is completed, the signal receiving strength of the target grid can be effectively analyzed according to the location of the base station, the location of the target grid, and the various terrains through which the signal propagates. It is judged whether the target grid is a weak coverage area, and if it is a weak coverage area, it is necessary to strengthen the construction of base stations in this area.

However, due to the limitations of the data set, the model designed in this paper also has limitations. The data set used in this paper uses static discrete locations for data collection, so real-time, continuous, and dynamically changing RSRP analysis of the location cannot be realized. It can only predict the signal strength received from a designated base station in a static target area.

6 Conclusions

The main contribution of this paper is to establish an intelligent wireless propagation model based on deep learning, which can quickly predict the average signal reception rate in a specific environment. The application of traditional wireless propagation models and wireless intelligent propagation models based on

Table 9 The weight of each scene

Clutter index	Weight	Clutter index	Weight	Clutter index	Weight	Clutter index	Weight
1.	2.4067097	6.	0.9826381	11.	1.0766366	16.	1.3850743
2.	1.6685416	7.	2.2179184	12.	1.853667	17.	− 0.6082736
3.	− 2.0274625	8.	3.0624619	13.	− 1.8139887	18.	− 1.2811766
4.	3.192918	9.	− 0.04044187	14.	− 1.971452	19.	1.3171432
5.	1.2578012	10.	2.3548207	15.	0.2523321	20.	1.4282409

Fig. 8 Weight of each scene
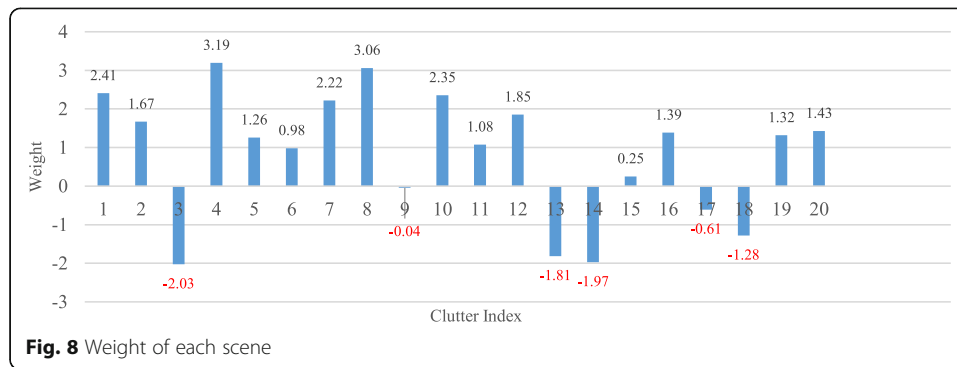
deep learning methods are analyzed. The traditional wireless propagation model is an empirical model that is statistically summarized on the basis of measured data, and the intelligence based on deep learning constructed in this article. The wireless propagation model will automatically adjust the model according to the geographic information through which the signal propagates, making the propagation model more applicable and accurate.

Abbreviations
5G: 5th generation wireless systems; A: Artificial intelligence; B5G: Beyond 5G; FN: False negative; FP: False positive; LTE: Long-term evolution; PCRR: Poor coverage recognition rate; RMSE: Root mean squared error; RSRP: Reference signal receiving power; SPM: Standard propagation model; TN: True negative; TP: True positive

Authors' contributions
All authors made contributions in the discussions, analyses and implementation of the proposed method. Haoyan Xu also contributed in writing the manuscript. All authors read and approved the final manuscript.

Availability of data and materials
Get data access URL:https://download.csdn.net/download/m0_37864755/18363841.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The manuscript does not contain any individual person's data in any form (including individual details, images, or videos) and therefore the consent to publish is not applicable to this article.

Competing interests
The authors declare that they have no competing interests.

References
1. E. Serpedin, A.R. Ekti, A. Boyac, M.A. Imran, S. Yarkan, The advances of fronthaul and backhaul communication for 5g and beyond. *Physical Communication* **36**, 100–819 (2019)
2. V. S. Abhayawardhana, I. J. Wassell, D. Crosby, M. P. Sellars, and M. G. Brown, Comparison of empirical propagation path loss models for fixed wireless access systems, in *Vehicular Technology Conference,* 2005. *VTC 2005-Spring. 2005 IEEE 61st,* 2005.
3. A. Medeisis and A. Kajackas, On the use of the universal okumurahata propagation prediction model in rural areas, in IEEE Vehicular Technology Conference, 2000, pp. 1815–1818.
4. G. Goel, S. H. Melvin, Y. Lostanlen, and D. Hatzinakos, Connectivity analysis of indoor wireless sensor networks using realistic propagation models, in *Proceedings of the 17th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems,* 2014, pp. 13–20.

5.   S. I. Popoola, A. A. Atayero, N. Faruk, C. T. Calafate, L. A. Olawoyin, and V. O. Matthews, Standard propagation model tuning for path loss predictions in built-up environments, in International Conference on Computational Science and Its Applications. Springer, 2017, pp. 363– 375.
6.   P.K. Gkonis, P.T. Trakadas, D.I. Kaklamani, A comprehensive study on simulation techniques for 5g networks: State of the art results, analysis, and future challenges. *Electronics* **9**(3), 468 (2020)
7.   E. Omoze, F. Edeko, Statistical tuning of cost 231 hata model in deployed 1800mhz gsm networks for a rural environment. *Niger. J. Technol.* **39**(4), 1216–1222 (2020)
8.   T. Zugno, M. Polese, N. Patriciello, B. Bojovic, S. Lagen, and M. Zorzi, ´Implementation of a spatial channel model for ns-3, *arXiv preprint arXiv:2002.09341*, 2020.
9.   M. Almahadeen and A. M. Matarneh, Performance assessment of throughput in a 5g system, *Jordanian Journal of Computers and Information Technology* (*JJCIT*), vol. 6, no. 3, pp. 303–316, 2020.
10. A. M. El-Hajj and T. Naous, Radiation analysis in a gradual 5g network deployment strategy, in *2020 IEEE 3rd 5G World Forum* (*5GWF*), 2020, pp. 448–453.
11. K. Oh, Y.-C. Chung, K.W. Kim, W.-S. Kim, I.-S. Oh, Author correction: classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Sci. Rep.* **10**(1), 1–1 (2020)
12. X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks. *J. Mach. Learn. Res.* **15**, 315–323 (2011)
13. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)
14. H. Shangxian, *Application of deep learning method in intelligent wireless propagation model* (Hebei University of economics and trade. M.Eng. Thesis, Hebei, 2020) https://d.wanfangdata.com.cn/thesis/D01964251

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.