Copyright Protection of Computer Software Deep Learning based Patent Text Clustering

Xiaojuan Huang[1,2,*] and Bei Wang[3]

[1]Civil, Commercial and Economic Law School, China University of Political Science and Law, Beijing 100088, China, 1602020093@cupl.edu.cn
[2]School of Law and Public Administration, Qujing Normal University, Qujing, Yunnan 655011, China, 1602020093@cupl.edu.cn
[3]School of Art and Design, Wuhan University of Technology, Wuhan, Hubei 430000, China, etwang117@163.com

Corresponding Author: Xiaojuan Huang, 1602020093@cupl.edu.cn

Abstract. The calculation method of the program is the core part of the software function, while the copyright law cannot effectively protect the software operation method, but can only protect the software development. Therefore, the instrumental value of software, such as its operation method, creation process and concept expression, can be protected by patent, so as to improve the protection of software property rights. This paper focuses on the analysis of the legal protection countermeasures of computer software intellectual property, in order to further promote the sustainable development of the computer software development industry. Establish data intellectual property protection system through computer neural network. The protection of data property rights and intellectual property rights based on deep learning and computer aided design is established. Two methods are used to initialize the word embedding layer, one is to use the pre trained word vector, the other is to initialize randomly. Using the advantages of recursive CNN (convolutional neural network) in processing sequence data, it can automatically capture the order and other correlations between text sentences, and train patent text according to the results of topic clustering. The results show that the final convergence accuracy of the recursive CNN combined model is higher than that of the comparison model, and the accuracy of the recursive CNN combined model reaches 95.39%, which indicates that the performance of the combined model based on the computer-aided model is better than that of the single model. The establishment of data intellectual property protection system through computer neural network is more stable and effective. Therefore, this model improves the level of intellectual property protection of data property rights.

Keywords: Deep Learning; Computer Aided Design; Data Property Rights; Intellectual Property.
DOI: https://doi.org/10.14733/cadaps.2023.S7.120-130

1 INTRODUCTION

After the emergence of the era of big data, the data economy has advanced by leaps and bounds. Big data has evolved into a new type of resource and method to create great value, and data has been continuously developed into a new type of asset, and at the same time, it is increasingly endowed with great commercial value by the market. With the deepening development of digital economy, data disputes are also on the rise. Existing data disputes can be roughly divided into personal data processing legitimacy disputes and enterprise data property rights disputes [1]. At present, governments all over the world are vigorously promoting data legislation, aiming at strengthening the data protection and management level of Internet enterprises, and to some extent, strict personal data protection rules have been achieved. To establish the intellectual property protection system of data, so as to provide a predictable legal environment for the circulation, sharing and utilization of data [2].

Since the birth of the concept of big data, there have been various classifications of big data. To clarify the ownership of data, reasonable data classification standards should be adopted first. However, this classification standard [3] should not be used in this paper. From the perspective of the subject, most of the data held by the enterprise comes from the online behavior of individuals, and its content is identifiable, which can identify specific individuals [4]. Data right refers to the legal effect of the act that the data right subject requires the opposite party to acknowledge that it owns the data, or requests to return or acknowledge the data for some reasonable reason [5]. The legal attributes of data can be divided into adjacent right object theory, property object theory and data asset theory Letaief et al. [6] believes that although the current computer has special copyright law protection, the trademark law can provide certain protection methods from different commercial perspectives. Different business marks can protect the "works" of different computer software, but they do not protect the idea of software and its "functionality". Suzuki et al. [7] interpretation of computer software is the program and related documents in a computer system. Computer software refers to a set of program systems and documents that use computers and develop computer efficiency. The document is for the convenience of understanding the information required by the procedure. Moreover, within a certain geographical scope and time, the right holders of computer software enjoy exclusive exclusive use rights, that is, exclusivity. Yang et al. [8] believes that with the development of the network and the trend of knowledge globalization, the regional characteristics of software and other intellectual property rights are gradually weakening. The timeliness of computer software is legal, which means that the legal protection of software is not eternal. In particular, the rights related to property have certain time limits. However, the right of identity related to software developers and other personal rights generally have no time limit. Yeo et al. [9] believes that computer software is a special object of intellectual property. In the standards concerning the effectiveness, scope and utilization of intellectual property, computer programs are protected as objects of intellectual property. Computer program is a kind of object of intellectual property rights, which is classified as the object of copyright protection. Yuvalı et al. [10] believes that there is an important principle for the copyright protection of computer software, that is, only the expression or manifestation of software is protected. The expression form of software refers to the instruction sequence or statement sequence represented by numbers, words and symbols contained in the software. And this sequence can be fixed with a tangible carrier. For programs, whether embodied in source code or object code, they may be protected by copyright law. The LSTM based on tree structure is applied to semantic composition, which has better semantic understanding effect than previous methods and is helpful to better understand text.

Data and information often come together and depend on each other. In addition, different countries and regions use different legal systems and legal terms for data protection, which makes data and information often used vaguely. Unlike the personal data related to the data content and the object as the identification standard, enterprise data is a concept determined by the judgment of the subject collecting and forming data. The data generated by the enterprise itself, the data

legally collected and the data products processed are all enterprise data. The diversity of content determines the complexity of its rights and interest rules. In the era of big data, the utilization and protection of enterprise data has become a practical problem, which is related to the resource utilization and innovative development of enterprises. Therefore, based on DL and computer aided design, this paper analyzes the legitimacy foundation of enterprise data property right protection from the perspective of enterprise data protection path, and defines the object definition, right content, right limitation and other specifications of enterprise data property right construction, so as to realize the balanced development of data information governance.

2 RESEARCH METHOD

2.1 Basic Construction of Data Property Rights

How to design or handle this interest relationship between users and operators legally becomes the basic premise of whether the current data economy and data assetization can be effectively and reasonably carried out. Unfortunately, China's legislation has not provided a clear and reasonable solution to this problem so far. By setting the user agreement, the user is guided to establish an authorization relationship about personal information. This method was quickly widely recognized in practice. Data activity stakeholders are allowed to establish the debt relationship of data collection and utilization through user agreement (personal information authorization contract), sometimes supplemented by some management norms. However, these methods, whether alone or combined, can't reasonably meet the complex needs of the current adjustment of data interests.

As a new form of property rights, data property rights reflect the trend of the times of the property system. Data property right is the right of the right subject to control the data property and exclude the interference of others. For anonymous data sets, enterprises have data rights, while for non-anonymous data, because the personal information contained in the data has not been lost, and the right subject has not changed, it cannot enter the circulation field. The data enterprise itself provides the network platform and specific network services to specific information providers, while the information providers pay the corresponding consideration by using the network platform and receiving services. The data enterprise itself needs to pay labor and cost for providing the platform. If the right is given to individuals, it will undoubtedly increase the transaction cost in the use process, which will greatly hinder the development of data economy.

The content structure of different property rights is not static, and the content of data property rights also has its uniqueness, which is shown as the type of power closely related to data acquisition rights, portability rights, use rights and income rights, as shown in Figure 1:

Figure 1: Type of power.

Ignore the protection of the right subject's personal rights and interests, and the protection of the right subject is precisely the primary pursuit goal of building the right property right system. We know that data can be divided into national data, public data and personal data in terms of its master, while the commercial data controlled by enterprises by secret means is only one part of a

huge database. The trade secret theory only pays attention to this part of data, and this understanding of data rights is one-sided.

2.2 The Application of DL and Computer Aided Design in Text Processing

In the traditional concept, information should not be owned by private individuals, but belongs to the common wealth of society. To some extent, the theory of labor rights affirms the intellectual labor paid by enterprises to collect, store and utilize data. Starting from the labor right, the value-added labor paid by data enterprises can get the corresponding right protection, but this right protection has a prerequisite: being out of the natural state. Data enterprises can make use of market strategies to obtain reasonable economic returns. For some leading data enterprises in the industry, the first-Mover advantage is enough to ensure that they can get enough benefits in data production and processing. It is doubtful whether the property right system can bring additional incentives. The key to solve this problem lies in adjusting the content of enterprise data rights, and balancing enterprise data monopoly and social data flow.

MH (Minimum hash) algorithm solves the problem of high time and space complexity when dealing with large-scale data, and has the effect of dimension reduction. The similarity measure of A, B sets is converted into the probability that the MH value of set A, B is equal after hash conversion, and the formula is shown in (1):

$$P_r\{h_{\min}(A) = h_{\min}(B)\} = J(A, B) \tag{1}$$

h represents the hash algorithm, and the hash conflict is ignored; $h_{\min}(A)$ represents the smallest element after hash conversion of all elements in the set A, and $h_{\min}(B)$ is the same. The model diagram of Skip-Gram model of Word2Vec is shown in figure 2.

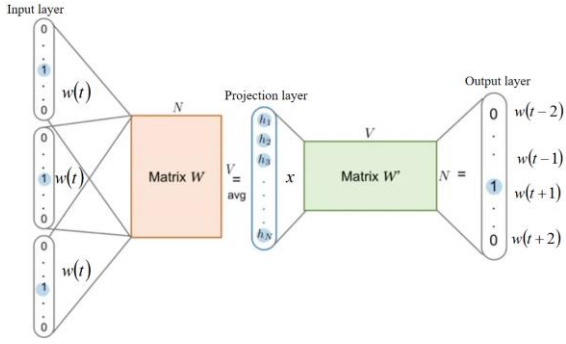

Figure 2: Skip-Gram model diagram.

Skip-Gram model predicts the context vocabulary vector according to the current vocabulary. Around a certain head word, the probability calculation of the context vocabulary vector is shown in formula (2).

$$P(w(t-c), \cdots, w(t-1), w(t+1), \cdots, w(t+c) | w(t)) \tag{2}$$

The convolution layer mainly performs convolution operation. After completing the calculation of each convolution kernel, a new feature graph is obtained. Let the original image be $P(x, y)$ and the convolution kernel be $Q(x, y)$, and the final result H is shown in formula (3):

$$H = P(x,y) * Q(x,y) = \sum i \sum j P(i,j) Q(x-i, x-j) \tag{3}$$

Only those functions that meet the nonlinear, differentiable and monotonic properties can be used as activation functions. Non-linearity is mainly used to ensure that the network meets the non-linear changes; The mathematical formula of Sigmoid function is shown in (4).

$$y = \frac{1}{1+e^{-x}} \tag{4}$$

LR (Logistic regression) model is a classical classification algorithm in statistics. Although it is called regression, it is actually a model for classification. Its dependent variables are classified into two categories and multiple categories.

$$F(x) = P(X \le x) = \frac{1}{1+e^{-\frac{x-\mu}{\gamma}}} \tag{5}$$

$$f(x) = F'(x) = \frac{e^{-\frac{x-\mu}{\gamma}}}{\gamma\left(1+e^{-\frac{x-\mu}{\gamma}}\right)^2} \tag{6}$$

γ is an open polymorphic parameter, and μ is a positional parameter.

$$x_{1:m} = x_1 \oplus x_2 \oplus \cdots \oplus x_m \tag{7}$$

$\oplus$ represents the splicing operation, and $x_{i:j}$ represents the splicing of the i to j word vectors in the patent text.

Extract the local features of $x_{1:m}$ in different dimensions, and the formula is as follows:

$$c_i = f\left(w^* x_{i:i+h-1} + b\right) \tag{8}$$

w is the parameter of convolution kernel, h is the height of convolution kernel, $w \in R^{h*d}$, b is offset, $b \in R$, $f(\cdot)$ is Relu activation function.

2.3 Design of Patent Text Classification Model

The special value-added mechanism of enterprise data determines that it should be promoted to open and share. The realization of enterprise data value depends on the full mining of the information content it carries, which is different from traditional property. After the enterprise data is "aggregated" through flow and sharing, its economic value and social value can be significantly improved. One of the important sources of enterprise data is the legally collected personal data. The inherent attribute of personal data "identification" determines that it is closely related to the personal dignity and personal freedom of data subjects. Data security is an organic part of national security and public security, and enterprise data is bound to be closely related to national security and public security. At the same time, based on the important role of enterprise data in the economy and management of the information society, it also indirectly affects the economic development, public management and social welfare of the information society.

In text processing, neural network model shows great advantages in text classification, semantic analysis and other tasks. In this paper, DL and computer aided design are used to optimize CNN model. The model structure and key parameters are shown in Figure 3. We use two ways to initialize the word embedding layer, one is to use pre-trained word vectors, and the other is to use random initialization.

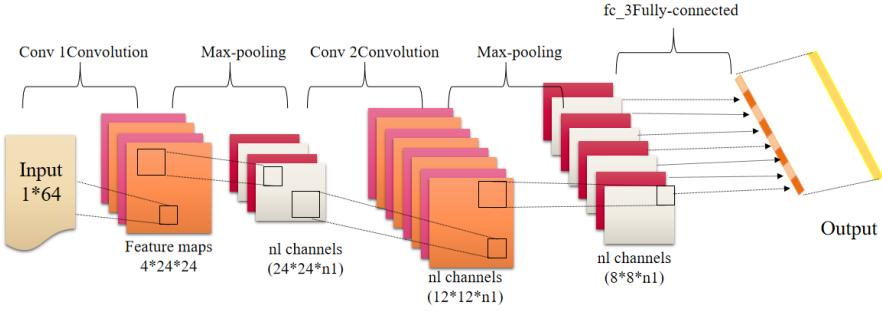

Figure 3: CNN model optimization structure.

CNN model has achieved good results in extracting and representing patent text features, but CNN model also has some shortcomings, such as the inability to capture the long-term dependence of sequence data and the difficulty in determining.

$$c_l(w_i) = f\left(W^l c_l(w_{i-1}) + W^{sl} e(w_{i-1})\right) \tag{9}$$

$$c_r(w_i) = f\left(W^r c_r(w_{i+1}) + W^{sr} e(w_{i+1})\right) \tag{10}$$

The attention weight α_j at each moment is weighted and averaged to h_j to obtain the characteristic matrix $h_j^{'} \in R^{m\times 2f}$, the calculation formula of which is shown in (11).

$$h_j^{'} = \sum_{j=1}^{T} \alpha_j^{'} h_j \tag{11}$$

The relationship weight between the subject v_i and the patent text feature vector p is calculated by the attention mechanism, and its calculation formula is shown in (12).

$$x_i = soft\max\left(s_1^T \tanh\left(Wei_1 * concat[v_i; p] + a_1\right)\right) \tag{12}$$

$$score\left(\bar{h}, h_i\right) = w^T \tanh\left(A\bar{h} + Bh_i + b_i\right) \tag{13}$$

$$a_i = \frac{\exp\left(score\left(\bar{h}, h_i\right)\right)}{\sum_j \exp\left(score\left(\bar{h}, h_j\right)\right)} \tag{14}$$

$$T = \sum_{i=0}^{n} a_i h_i \tag{15}$$

Where h_i is the output value of the ith hidden layer, $\bar{h}$ is the patent text vector, and b_i is the offset term.

Using the negative log likelihood of the correct label as the training loss, the calculation is shown in Formula (16):

$$L = -\sum_{d} \log p_{dj} \tag{16}$$

Where j is the label of the text d. p_{dj} indicates the probability that the text d belongs to the effect j.

3 ANALYSIS AND DISCUSSION OF RESULTS

This paper mainly builds the model through the third-party platform Anaconda. The model is implemented by TensorFlow, a deep neural network framework, which is efficient, convenient and extensible. Tensor stands for multidimensional array. When new modules need to be added, just write new classes or functions according to the existing modules, so it is more suitable for researchers to quickly turn their ideas into reality. In order to enable the algorithm to be applied to practical problems, this paper uses the powerful Django framework to connect the foreground interface with the background algorithm, and uses sqllite3 database to store data. When product designers need to analyze the effects of this patent and find examples of patents with the same effects to provide ideas for their own innovation, they need to classify this patent based on effects. Obtain the candidate themes of this patent, and the results of selecting some themes of patents to be classified are shown in Figure 4.

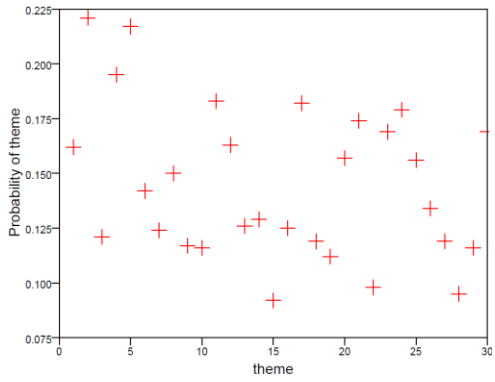

Figure 4: Partial selection result of candidate node of patent text to be classified.

Calculate the correlation between candidate topics and patents to be classified, and the results are shown in Table 1.

Candidate theme	*Weight*	*Is it greater than the threshold*
4	0.612	Yes
5	0.294	No
9	0.472	Yes
12	0.439	Yes
17	0.658	Yes

23	0.374	Yes
27	0.339	Yes
30	0.586	Yes

Table 1: Correlation results of patent texts of candidate topics.

It can be seen that, because this patent uses FET effect to improve the signal control device, some themes in the thematic expression of this patent are not highly related to FET effect, such as theme 5 and theme 27. However, because of the high correlation between these themes and patents, they are also extracted as thematic expressions of the patents.

A series of experiments are carried out on the recurrent CNN model by using three kinds of word embedding layer initialization methods. Figure 5 summarizes the broken line of the accuracy loss of the verification set of the three word embedding initialization methods. Table 2 collates the converged data and counts the accuracy.

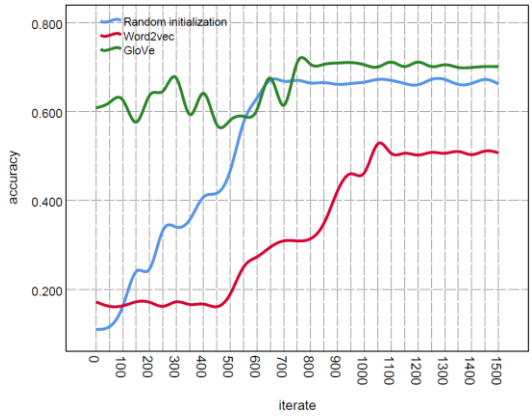

Figure 5: Accuracy loss curve comparison.

Word embedding mode	*After 800 iterations, the data are averaged*	*Number of iterations 1500*
Random initialization	0.668	0.668
Word2vec	0.472	0.502
GloVe	0.707	0.704

Table 2: Statistics of model training results.

The method of introducing recurrent CNN model is compared with the classical machine learning algorithms SVM (support vector machine), MH, LSTM, LR and TextCNN respectively, and the results are shown in Table 3 and Figure 6:

Method	*Accuracy*	*Recall*	*F1 value*
SVM	0.59	0.604	0.492

MH	0.6	0.452	0.568
LSTM	0.563	0.452	0.489
LR	0.674	0.543	0.606
TextCNN	0.637	0.567	0.684
our	0.793	0.684	0.791

Table 3: Comparison of text classification results.

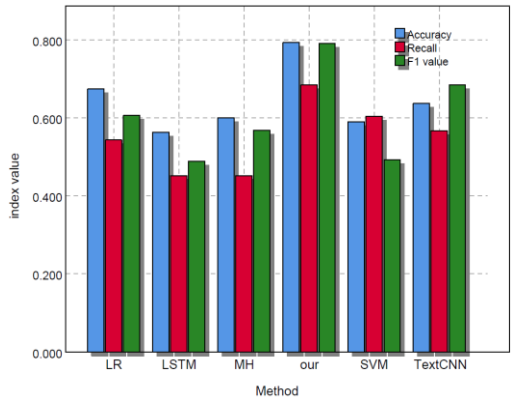

Figure 6: Statistical chart of text classification results.

The results show that DL-based text classification is better than machine learning-based text classification. From the results, it can be seen that the recurrent CNN result proposed in this paper is about 3.06% higher than SVM and about 2.18% higher than LSTM. The main reason is that when the former combines the output vectors, the attention mechanism gives different weights to each output vector, which makes the model focus on the vectors that are more important to the patent text classification task, and reduces the role of irrelevant vectors. To sum up, it shows that the classification method based on recurrent CNN model is effective to determine the category of patents.

After the experiment, the broken-line change diagram of loss rate of six Chinese patent text classification models is obtained as shown in Figure 7:

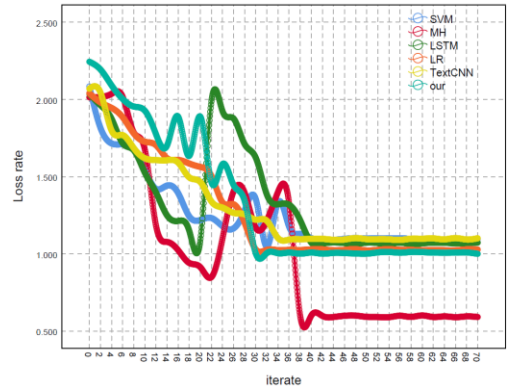

Figure 7: Line chart of loss rate of model.

After the introduction of Attention mechanism, although the calculation cost has increased, the accuracy of patent text classification has increased, which proves that the introduction of Attention mechanism in the hidden layer can improve the accuracy of patent text classification to a certain extent.

Data is consistent with the intrinsic nature of intellectual property objects. The legality of data sources should be regarded as an important prerequisite for its protection. Data collection involves many subjects, including data source, data collector, data processor, secondary processor and so on. The author believes that the principle of hierarchical boundary rights should be adopted, and different subjects should enjoy different rights according to their investment in data property. At the same time, because of the immaterial characteristics of data, it can be used by multiple agents at the same time. Therefore, different data collectors can collect the same data item under the premise of authorization and enjoy relevant rights. The data processor has invested in the data set because of its processing behavior, and enjoys the data rights for the new data set or processing results.

4 CONCLUSION

The main way of information dissemination. Data and information often appear together and depend on each other. In addition, different countries and regions adopt different legal systems and legal terms for data protection, which makes data and information often used fuzzily. This paper studies the protection of intellectual property rights of data property based on DL and computer aided design. A patent text classification model is established for the interests by intellectual property law. The results show that the final convergence accuracy of the recurrent CNN combined model is higher than that of the comparison model, and the accuracy of the recurrent CNN combined model reaches 95.39%, which shows that the performance of the combined model is better than that of the single model. After the introduction of Attention mechanism, although the calculation cost has increased, the accuracy of patent text classification has increased, which proves that the hidden layer can improve the accuracy of patent text classification to a certain extent.

Xiaojuan Huang, https://orcid.org/0000-0002-3730-4177
Bei Wang, https://orcid.org/0000-0002-4672-6750

REFERENCES

[1] Duvigneau, R.: CAD - consistent adaptive refinement using a NURBS - based discontinuous Galerkin method, International Journal for Numerical Methods in Fluids, 92(9), 2020, 1096-1117. https://doi.org/10.1002/fld.4819
[2] Fujita, H.: AI-based computer-aided diagnosis (AI-CAD): the latest review to read first, Radiological physics and technology, 13(1), 2020, 6-19. https://doi.org/10.1007/s12194-019-00552-4
[3] Hamaguchi, T.; Onodera, M.; Yokohari, T.: Development of technique for checking insulation distance for large-scale CAD data using voxel meshes, Mechanical Engineering Journal, 5(3), 2018, 17-00322-17-00322. https://doi.org/10.1299/mej.17-00322
[4] Han, Y.-S.; Lee, J.; Lee, J.; Lee, W.; Lee, K.: 3D CAD data extraction and conversion for application of augmented/virtual reality to the construction of ships and offshore structures, International Journal of Computer Integrated Manufacturing, 32(7), 2019, 658-668. https://doi.org/10.1080/0951192X.2019.1599440
[5] Imen, B.; Moncef, H.; Moez, T.; Nizar, A.: Generation of disassembly plans and quality assessment based on CAD data, International Journal of Computer Integrated Manufacturing, 33(12), 2020, 1300-1320. https://doi.org/10.1080/0951192X.2020.1815852

[6] Letaief, M.-B.; Tlija, M.; Louhichi, B.: An approach of CAD/CAM data reuse for manufacturing cost estimation, International Journal of Computer Integrated Manufacturing, 33(12), 2020, 1208-1226. https://doi.org/10.1080/0951192X.2020.1815842
[7] Suzuki, S.; Ohtake, Y.; Suzuki, H.: Fitting CAD data to scanned data with large deformation, Journal of Computational Design and Engineering, 7(2), 2020, 145-154. https://doi.org/10.1093/jcde/qwaa013
[8] Yang, C.; Weng, Y.; Huang, B.; Ikbal, M.: Development and optimization of CAD system based on big data technology, Computer-Aided Design and Applications, 19(S2), 2021, 112-123. https://doi.org/10.14733/cadaps.2022.s2.112-123
[9] Yeo, C.; Kim, B.-C.; Cheon, S.; Lee, J.; Mun, D.: Machining feature recognition based on deep neural networks to support tight integration with 3D CAD systems, Scientific reports, 11(1), 2021, 1-20. https://doi.org/10.1038/s41598-021-01313-3
[10] Yuvalı, M.; Yaman, B.; Tosun, Ö.: Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets, Mathematics, 10(3), 2022, 311. https://doi.org/10.3390/math10030311