JOURNAL OF BASIC SCIENCE AND ENGINEERING

BREAST CANCER DETECTION USING DECISION TREE AND SVM CLASSIFIER

[1]Keerthi Asam, [2]B. Charishma
[1]PG Scholar, Department of CSE, Srinivasa Institute of technology and science, Kadapa.
[2]Associate Professor, HOD of CSE, Srinivasa Institute of technology and science, Kadapa.
[1]a.keerthireddy@gmail.com, [2]charishma.bavirisetty@gmail.com

ABSTRACT

Identification of breast cancer plays a major role in medical field nowadays. Women are facing different types of cancer and one among them is breast cancer which has severe impact. Breast cancer is of two types i.e. Malign or Benign type. Benign is given as a non-curable type of cancer and Malign is given as curable type of cancer. Breast cancer is symbolized by the modification of genes, persistent pain, changes in the measurement, change in shade (redness), and skin appearance of breasts. In the early days of identifying breast cancer is done by using different algorithms namely Support Vector Machine (SVM) algorithm, K Nearest Neighbor (KNN) algorithm, MLP algorithm, etc., By using these algorithms the accuracy of detecting the cancer is not met the extend. Our idea is to detect the breast cancer using Decision Tree algorithm [19]. The decision tree algorithm comes under the supervised learning technique. Our idea is to detect the breast cancer using Decision Tree algorithm. The tree algorithm comes under the supervised learning technique. The main advantage of this decision tree algorithm is identifying whether the predicted cancer is either malign or benign type by producing an 99% accuracy.

Keywords: Breast cancer, classification, decision tree algorithms, SVM, missing data imputation, Breast cancer, Data mining.

I. INTRODUCTION

In this paper we are using data science and machine learning concept. Nowadays technologies are developing a lot. In order to reduce human work, we are proposing a concept of machine learning and data science. Information Science is a mix of different apparatuses, calculations, and AI standards with the objective to find concealed examples from the crude information. Information Science is essentially used to settle on choices and forecasts making utilization of prescient causal examination, prescriptive investigation and AI [2]. Information Science is an increasingly forward-looking methodology, an exploratory path with the attention on breaking down the past or current information and foreseeing the future results with the point of settling on educated choices [1]. It responds to the open-finished inquiries about "what" and "how" occasions happen. As the data science consists of machine learning concept in it, we are trying to move forward with it to feed the machine. Machine learning is a process where the machine learns automatically with experience. Learning is nothing but recollection and assimilation of input data and the decision purely depends on equipped data. It is strenuous to take decision based on attainable inputs. To overcome this issue, certain algorithms were developed. In order to solve a specific task feed the algorithm with more specific data. In most cases a computer will use data as its source of information and compare its output to a desired output and then correct for it [1].

II. SYSTEM ANALYSIS

The Systems Development Life Cycle (SDLC), or Software Development Life Cycle in systems engineering, information systems and software engineering, is the process of creating or altering systems, and the models and methodologies that people use to develop these systems. In software engineering the SDLC concept underpins many kinds of software development methodologies [4].

i) Existing System

Tolga Ensar proposed a paper on Breast Cancer grouping utilizing Machine learning calculations. They have utilized Naïve Bayes calculation and K-Nearest Neighbor calculation to characterize the malignancy as either insult or favourable. They used 11 attributes in which they had id as one attribute which has been removed and thus have 9 criteria. The breast cancer classification they done had 9 classification which included Clump Thickness, Uniformity of cell size, Uniformity of cell shape, Marginal Adhesion, Single Epithelial cell size, Bare Nuclei, Bland Chromatin, Normal Nuclei, Mitosis. An Image processing concept and two machine learning algorithms. The algorithms are Logistic Regression (LR) and Back Propagation Neural Network (BPNN) to detect the breast cancer [3].

Disadvantages of Existing System

- They had 683 datasets. The classification is done using the respective formulae for the algorithms. By comparing both the algorithms they got an accuracy of 97.51% in K-Nearest Neighbor algorithm and 96.19% in Naïve Bayes algorithm [6].
- Thus, the accuracy rate for detecting the cancer using LR and BPNN algorithm exceeded 93% [6].

ii) Proposed System

Dataset: The dataset we are using is Wisconsin Breast Cancer dataset with 32 attributes and 569 data. The inputs are (1) id -(ID number), (2) diagnosis-(M = malignant, B = benign), (3) radius-(mean of distances from center to points on the perimeter), (4) texture-(standard deviation of gray-scale values), (5) perimeter-(mean size of the core tumor), (6) area, (7) smoothness mean-(mean of local variation in radius lengths), (8) concavity- (mean of severity of concave portions of the contour), (9) compactness-(mean of perimeter^2 / area - 1.0), (10) concave points mean-(mean for number of concave portions of the contour), (11) symmetry, (12) fractal dimension mean –(mean for "coastline approximation" – 1).Here, (1) is not considered and the rest are taken into account. For these attributes they calculate mean, standard error, worst.

These data are fed into the machine. These data are used for training the machine using Decision tree algorithm by doing so we can get the output as either Malign or Benign. Using the Decision tree algorithm the breast cancer is identified [8].

The decision tree algorithm is represented using a tree structure. By using this algorithm 99% accuracy is produced to identify whether the breast cancer is either Malign or Benign. The decision tree algorithm comes under the supervised learning technique. Structured project management techniques (such as an SDLC) enhance management's control over projects by dividing complex tasks into manageable sections. A software life cycle model is either a descriptive or prescriptive characterization of how software is or should be developed [6]. But none of the SDLC models discuss the key issues like Change management, Incident management and Release management processes within the SDLC process, but, it is addressed in the overall project management. In the proposed hypothetical model, the concept of user-developer interaction in the conventional SDLC model has been converted into a three-dimensional model which comprises of the user, owner and the developer [21]. In the proposed hypothetical model, the concept of user-developer interaction in the conventional SDLC model has been converted into a three-dimensional model which comprises of the user, owner and the developer. The one size fits all‖ approach to applying SDLC methodologies is no longer appropriate. We have made an attempt to address the above-mentioned defects by using a new hypothetical model for SDLC described elsewhere. The drawback of addressing these management processes under the overall project management is missing of key technical issues pertaining to software development process that is, these issues are talked in the project management at the surface level but not at the ground level [2].

III. SYSTEM DESIGN

a) Methodology

The diagram consists of the subsequent stages: Stage

Stage 1. Dataset Collection: In dataset collection we are collecting the data from the Wisconsin breast cancer dataset. After the collection is over the data is processed and move on to data cleaning. Stage

Stage 2. Data Cleaning: In which the attributes are checked for any null value process, if any null values are present then replace it with zero.

Stage 3: Data Analysis: In which the data collected has been processed and helps in making any decision.

Stage 4: Supervised Learning: Managed learning is the AI errand of learning a capacity that maps a contribution to a yield dependent on model info yield sets. It construes a capacity from marked preparing information comprising of a lot of preparing models. In regulated adapting, every model is a couple comprising of an info object (ordinarily a vector) and ideal yield esteem (additionally called the supervisory flag).

Stage 5: Classification Technique: In supervised learning there comes a regression and classification part. In which we are using a classification part to identify Malign or Benign type of breast cancer.

Stage 6: Decision Tree Algorithm: This algorithm is normally represented in a tree structure. By using this algorithm, we are providing an accuracy of 99% in our project.

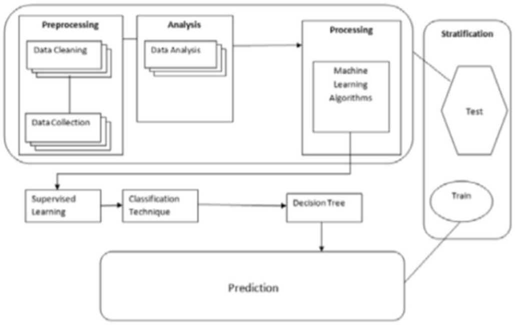


Figure: System architecture

b) Algorithm:

i) Decision Tree Algorithm

The decision tree algorithm comes under the supervised learning in machine learning algorithms. The decision tree is displayed in the tree structure. The input given to the decision tree is based on certain criteria's and the output is displayed as either true or false. This algorithm is said to be simpler and more successful. The values in the node are given by comparing each attribute. Based on the weight age of information the node gets spitted the final output is displayed in the leaf node. Entropy describes the importance of the node Algorithm.

Step 1: In order to do the process of learning training dataset is selected.

Step 2: Make a map of each individual attribute to respective classes.

Step 3: Catch all practicable values for each attribute that correlate with feasible classes.

Step 4: Compute values of every attribute which belongs to distinctive classes.

Step 5: Root node is generated to that attribute which has minimum number of values which reside in the unique classes.

Step 6: Comparably pick another attribute for next extent in decision tree from prevailing attributes based on least number of values which has distinctive classes.

Step 7: Stop

ii) Input Design

Input design is a part of overall system design. The main objective during the input design is as given below:

- To produce a cost-effective method of input.
- To achieve the highest possible level of accuracy.
- To ensure that the input is acceptable and understood by the user.

iii) Output Design

Outputs from computer systems are required primarily to communicate the results of processing to users. They are also used to provide a permanent copy of the results for later consultation. The various types of outputs in general are:

- External Outputs, whose destination is outside the organization
- Internal Outputs whose destination is within organization and they are the
- User’s main interface with the computer.
- Operational outputs whose use is purely within the computer department.
- Interface outputs, which involve the user in communicating directly.

IV. IMPLEMENTATION

i) Design

The software system design is produced from the results of the requirements phase. Architects have the ball in their court during this phase and this is the phase in which their focus lies. This is where the details on how the system will work is produced. Architecture, including hardware and software, communication, software design (UML is produced here) are all part of the deliverables of a design phase.

ii) Implementation

Code is produced from the deliverables of the design phase during implementation, and this is the longest phase of the software development life cycle. For a developer, this is the main focus of the life cycle because this is where the code is produced. Implementation my overlap with both the design and testing phases. Many tools exist (CASE tools) to actually automate the production of code using information gathered and produced during the design phase [15].

V. TESTING

Testing is the process where the test data is prepared and is used for testing the modules individually and later the validation given for the fields. Then the system testing takes place which makes sure that all components of the system property function as a unit. The test data should be chosen such that it passed through all possible condition. The following is the description of the testing strategies, which were carried out during the testing period [10].

During testing, the implementation is tested against the requirements to make sure that the product is actually solving the needs addressed and gathered during the requirements phase. Unit tests and system/acceptance tests are done during this phase. Unit tests act on a specific component of the system, while system tests act on the system as a whole. So, in a nutshell, that is a very basic overview of the general software development life cycle model. Now let’s delve into some of the traditional and widely used variations [13].

i) System Testing

Testing has become an integral part of any system or project especially in the field of information technology. The importance of testing is a method of justifying, if one is ready to move further, be it to be check if one is capable to with stand the rigors of a particular situation cannot be underplayed and that is why testing before development is so critical. When the software is developed before it is given to user to user the software must be tested whether it is solving the purpose for which it is developed. This testing involves various types through which one can ensure the software is reliable. The program was tested logically and pattern of execution of the program for a set of data are repeated. Thus, the code was exhaustively checked for all possible correct data and the outcomes were also checked [14].

ii) Module Testing

To locate errors, each module is tested individually. This enables us to detect error and correct it without affecting any other modules. Whenever the program is not satisfying the required function, it must be corrected to get the required result. Thus, all the modules are individually tested from bottom up starting with the smallest and lowest modules and proceeding to the next level. Each module in the system is tested separately. For example, the job classification module is tested separately. This module is tested with different job and its approximate execution time and the result of the test is compared with the results that are prepared manually [16]. Each module in the system is tested separately. In this system the resource classification and job scheduling modules are tested separately and their corresponding results are obtained which reduces the process waiting time.

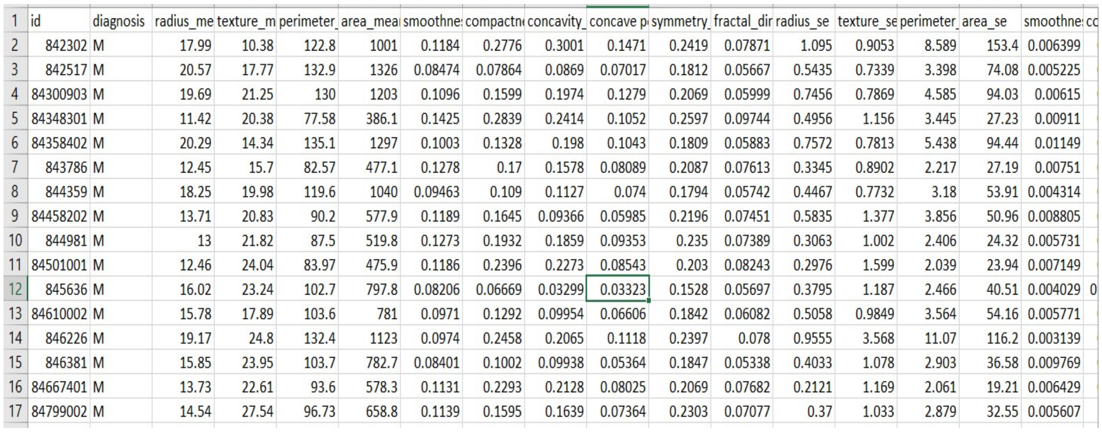
iii) Integration Testing

After the module testing, the integration testing is applied. When linking the modules there may be chance for errors to occur, these errors are corrected by using this testing. In this system all modules are connected and tested. The testing results are very correct. Thus, the mapping of jobs with resources is done correctly by the system [14].

iv) Acceptance Testing

When that user fined no major problems with its accuracy, the system passers through a final acceptance test. This test confirms that the system needs the original goals, objectives and requirements established during analysis without actual execution which elimination wastage of time and money acceptance tests on the shoulders of users and management, it is finally acceptable and ready for the operation [18].

VI. OUTPUT SCREENS

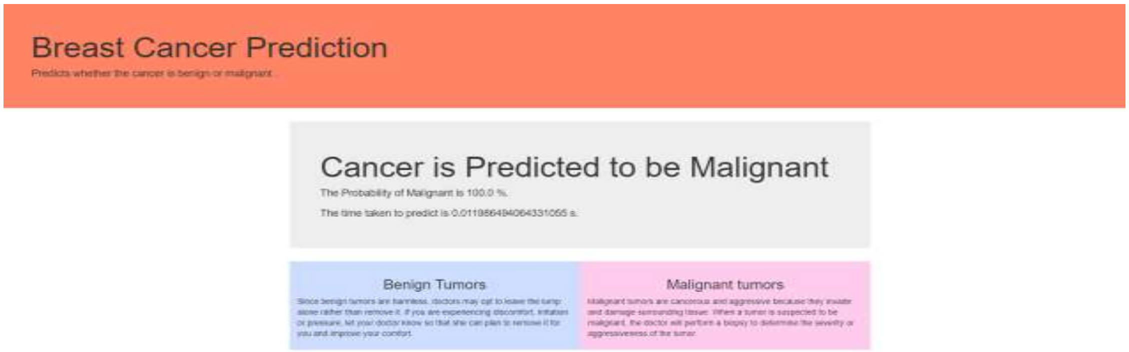


1	id	diagnosis	radius_me	texture_m	perimeter_	area_mea	smoothne:	compactn	concavity_	concave p	symmetry_	fractal_dir	radius_se	texture_se	perimeter_	area_se	smoothne:	cc
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	
17	84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	

Figure: Dataset

Breast Cancer Prediction Using Machine Learning

Predicts whether the cancer is benign or malignant .



Breast Cancer Prediction

Predicts whether the cancer is benign or malignant

Cancer is Predicted to be Malignant

The Probability of Malignant is 100.0 %.

The time taken to predict is 0.011986494064331055 s.

Benign Tumors

Since benign tumors are harmless, doctors may opt to leave the lump alone rather than remove it. If you are experiencing discomfort, irritation or pressure, let your doctor know so that she can plan to remove it for you and improve your comfort.

Malignant tumors

Malignant tumors are cancerous and aggressive because they invade and damage surrounding tissue. When a tumor is suspected to be malignant, the doctor will perform a biopsy to determine the severity or aggressiveness of the tumor.

VI. CONCLUSION

The identification of breast cancer is done here. The identification is based on whether the patient is affected by either Malign or benign type of cancer. The type of cancer is predicted by using decision tree algorithm which comes under the supervised learning technique, by doing so the accuracy of 99% is obtained. This system may be increased by adding extra data of the affected patient like adding some more attributes. Further, we will be enhancing our project by identifying at which stage it is present and also providing the preventive measures for the patient.

REFERENCES

[1] V. Fotedar, S. Fotedar, P. Takur, S. Vats, A. Negi, and L. Chanderkant, “Knowledge of breast cancer risk factors and methodsfor its early detection among the primary health-care workers in shimla, himachal pradesh,” Journal of Education and Health Promotion, vol. 8, p. 265, 2019.

[2] M. S. Simon, T. A. Hastert, A. Barac et al., “Cardiometabolic risk factors and survival after cancer in the women’s health initiative,” Cancer, vol. 127, no. 4, pp. 598–608, 2021.

[3] H. F. Pasha, R. H. Mohamed, M. M. Toam, and A. M. Yehia, “Genetic and epigenetic modifcations of adiponectin gene: p,” Te Journal of Gene Medicine, vol. 21, no. 10, p. e3120, 2019.

[4] D. Hong, A. J. Fritz, S. K. Zaidi et al., “Epithelial to mesen chymal transition and cancer stem cells contribute to breast cancer heterogeneity,” Journal of Cellular Physiology, vol. 233, no. 12, pp. 9136–9144, 2018.

[5] J. Zhong, D. Sun, W. Wei et al., “Contrast-enhanced ultra sound-guided fne-needle aspiration for sentinel lymph node biopsy in early-stage breast cancer,” Ultrasound in Medicine and Biology, vol. 44, no. 7, pp. 1371–1378, 2018.

[6] K. Babic, C. Siguan-Bell, M. Hee, and S. C. Lin, “Ocular g: ultrasound B-scan assessment of

retained surgical sponge after ahmed valve surgery: a case r," Journal of Glaucoma, vol. 26, no. 10, pp. e239–e241, 2017.

[7] A. Yala, P. G. Mikhael, F. Strand et al., "Toward robust mammography-based models for breast cancer risk," Science Translational Medicine, vol. 13, no. 578, Article ID eaba4373, 2021.

[8] G. Zheng, X. Liu, and G. Han, "Medical image computer aided detection and diagnosis system review," Journal of Software, vol. 29, no. 5, pp. 1471–1514, 2018.

[9] E. Y. Huang, S. Knight, C. R. Guetter et al., "Telemedicine and telementoring in the surgical specialties: a narrative review," TeAmerican Journal of Surgery, vol. 218, no. 4, pp. 760–766, 2019.

[10] Q. Wu,B. T.Wang,H.R.Rao,Y.Jiang, and C. Liu, "Breast cancer and benign disease cells form metrology research," Journal of Anhui Medical University, vol. 31, no. 02, pp. 91–93, 1996.

[11] Q. Shen, F. Shao, and R. Sun, "Prediction model of breast cancer based on xgboost," Journal of Qingdao University (Natural Science Edition), vol. 32, no. 1, 2019.

[12] Z. Deng, B. Su, and K. Zhan.g, "Breast cancer classifcation based on ensemble learning," China Medical Devices, vol. 35, no. 12, 2020.

[13] M. Monirujjaman Khan, S. Islam, S. Sarkar et al., "Machine learning based comparative analysis for breast cancer pre diction," Journal of Healthcare Engineering, vol. 2022, Article ID 4365855, 15 pages, 2022.

[14] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W.Ibrahim,"Tree-based and machinelearning algorithm analysis for breast cancer classifcation," Computational In telligence and Neuroscience, vol. 2022, Article ID 6715406, 6 pages, 2022.

[15] H. Dong and L. Ma, "Prediction model of triple negative breast cancer based on machine learning," Journal of Yunnan University, vol. 39, no. 1, pp. 111–115, 2017.

[16] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," European Journal of Operational Research, vol. 267, no. 2, pp. 687–699, 2018.

[17] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms," Expert Systems with Applications, vol. 41, no. 4, pp. 1476–1482, 2014.

[18] X. S. Jia, X. Sun, and X. Zhang, "Breast cancer identifcation using machine learning," Mathematical Problems in Engi neering, vol. 2022, Article ID 8122895, 8 pages, 2022.

[19] S. Singh, S. K. Jangir, M. Kumar et al., "Feature importance score-based functional link artifcial neural networks for breast cancer classifcation," BioMed Research International, vol. 2022, Article ID 2696916, 8 pages, 2022.

[20] T. R. Mahesh, V. Vinoth Kumar, V. Muthukumaran, H. K. Shashikala, B. Swapna, and S. Guluwadi, "Performance analysis of xgboost ensemble methods for survivability with the classifcation of breast cancer," Journal of Sensors, vol. 2022, Article ID 4649510, 8 pages, 2022.

[21] O. L. Mangasarian and W. H. Wolberg, Cancer Diagnosis Via Linear Programming, University of Wisconsin-Madison De partment of Computer Sciences, Madison, WI, USA, 1990.

[22] M. K. Das, A. Chaudhary, A. Bryan, M. H. Wener, S. L. Fink, and C. Morishima, "Rapid screening evaluation of sars-cov-2 igg assays using z-scores to standardize results," Emerging

Infectious Diseases, vol. 26, no. 10, pp. 2501–2503, 2020.

[23] X. Su and Y. Wang, “Feature selection algorithm for high dimensional unbalanced medical data,” Small Microcomputer Systems, https://kns.cnki.net/kcms/detail/21.1106.TP. 20221123.1608.034.html, 2022.

[24] P. Chen, F. Li, and C. Wu, “Research on intrusion detection method based on Pearson correlation coefcient feature se lection algorithm,” Journal of Physics: Conference Series, vol. 1757, no. 1, Article ID 012054, 2021.

[25] C. Lu and H. Shen, “Css: handling imbalanced data by im proved clustering with stratifed sampling,” Concurrency and Computation: Practice and Experience, vol. 34, no. 2, 2020.