International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I10P240
Volume 70 Issue 10, 408-414, October 2022

Original Article

Application of Hybrid Sampling and Stacked Deep Networks: Predicting the AI Convergence Education using Education Survey Data

Haewon Byeon[1,2]

[1]*Master Course, Department of Digital Anti-aging Healthcare (BK-21), Inje University, Republic of Korea.*
[2]*AI Convergence College, Inje University, South Korea, corresponding author*

[1,2]*Corresponding Author : bhwpuma@naver.com*

Received: 20 August 2022 Revised: 20 October 2022 Accepted: 21 October 2022 Published: 23 October 2022

Abstract *- This study proposed a method for predicting educational satisfaction based on hybrid sampling and stacked deep networks (SDN). We compared the performance of logistic regression (LR), support vector machine (SVM), decision tree (DT), and random forest (RF) machine learning algorithms with that of the SDN proposed in this study. The SDN method based on hybrid sampling proposed in this study had superior accuracy, recall, and precision performance to other ML algorithms. Especially when hybrid sampling was applied to the SDN algorithm, recall, and precision performance was greatly improved. Since SND has a feature extraction function and can be applied to unrefined data, additional studies should evaluate the performance by applying the proposed model to unrefined data with various attributes.*

Keywords - *Stacked deep networks, Logistic regression, Support vector machine, Hybrid sampling, Decision tree.*

1. Introduction

Across the globe, digital transformation is emerging in every industry. It is expected that the COVID-19 pandemic, which has had a catastrophic impact on social and economic sectors worldwide after 2020, will paradoxically accelerate this digital transformation [1]. Today, AI technology draws attention as digital transformation's core and main driver. With this background, AI experts have emerged as a key resource determining national competitiveness in the 21st century [2]. As a result, the demand for AI experts is also rising [2].

Digital transformation has received much attention owing to the advent of the digital age represented by the 4th industrial revolution and the sudden rise of AI technology [3]. The education sector has discussed distance learning for a long time without making much progress. However, the COVID-19 pandemic has spread AI education on a large scale in which a considerable proportion of the population from primary education to higher education participates [4,5]. Under this circumstance, the value and expectation of AI technology that will lead the digital transformation are increasing. McKinsey & Company [6] predicted that 70% of companies worldwide will utilize AI, adding more than $13 trillion to global GDP by 2030. Major global IT companies are rapidly positioning themselves as AI platform companies based on big data and the cloud [7]. They are also expanding their influence to all industrial fields [7]. It can be explained by the fact that the five largest market capitalization companies in the world (i.e., Apple, Microsoft, Google, Amazon, and Facebook) are ICT companies that have secured AI core algorithms and big data-based cloud platforms. The total market capitalization of these five companies exceeded $5.5 trillion in 2020 [8], four times higher than that of South Korean companies ($1.4 trillion) in the same period. In other words, AI technology has become a true core value beyond a supporter of corporate activities or a simple automation tool.

AI talent has emerged as a core resource that will determine national competitiveness in modern society due to the changing trend in the industry. Owing to the advancement of AI and intelligent automation, major developed countries such as the United States and Japan emphasized the need to provide AI convergence education for higher education personnel [9,10]. As a result, competition among countries, companies, and universities is intensifying worldwide to secure AI talent [11]. As society makes important public policy decisions related to AI technology, AI4K12[24], a guideline for artificial intelligence in the United States, raised the necessity of gaining experience in AI education from school days because people need to know the basics of AI as digital citizens. The demand for workers with AI literacy is

increasing in terms of job opportunities. China has developed high school textbooks related to AI education and applied them on a trial basis [13].

Moreover, China has developed an artificial intelligence site for students and distributed it for educational purposes. South Korea also made SW education mandatory for elementary/middle/high schools at the policy level [14]. Educational support measures (e.g., SW-centered university system and AI graduate school system) were prepared for university education. Especially nurturing high-quality AI experts and enhancing their qualities have been continuously discussed in university education [15].

Many South Korean universities currently offer AI courses such as software applications for all students (especially freshmen) to nurture AI talents. AI education aims to help students, regardless of their majors, deal with AI and foster their ability to communicate with it [16]. Regarding the AI curriculum, basic liberal arts education requirements related to AI are offered from the first semester of the freshman year, when they are exposed to AI for the first time [25]. In addition to the curriculum, they are also systematically operating AI education-related competency-strengthening training by planning and running various education support programs such as training for core lecturers to improve AI class design ability, advanced competency reinforcement, and specialized training [18]. Despite South Korean universities' administrative and financial support for AI education, there are not enough analyses on consumer demand for AI convergence education for all students (all majors) and basic studies on educational satisfaction.

This study examined the perception, demand, and educational satisfaction of education consumers (university students) for university-led AI convergence education for all students, which was conducted to enhance competitiveness, such as AI literacy, following digital transformation. This study proposed a method for predicting educational satisfaction based on hybrid sampling and stacked deep networks (SDN). The structure of this study is as follows. Section 2 describes the educational satisfaction prediction model based on the proposed hybrid sampling and SDN. Section 3 examines the performance of the proposed model using actual educational satisfaction survey data. Section 4 discusses conclusions and future research directions.

2. Materials and Methods

2.1. Configuration of the educational satisfaction prediction system

Figure 1 shows the overall configuration of the proposed educational satisfaction prediction system. A deep learning model is designed by (1) conducting data balancing to balance the class of the education-satisfied group and that of the dissatisfied education group, (2) dividing them into learning data and validation data for creating a model, and then (3) generate a model with the proposed SDN.

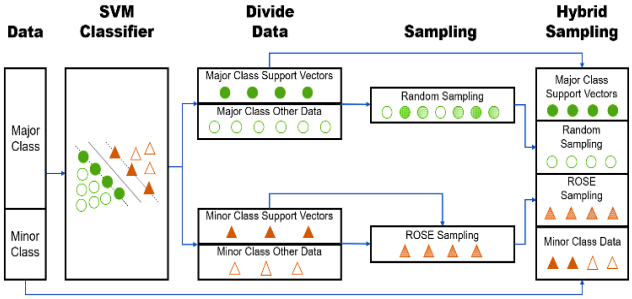

Fig 1. Configuration of the educational satisfaction prediction system proposed by this study

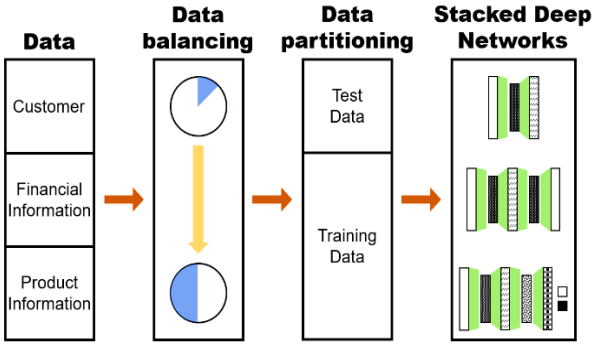

Fig 2. Flowchart of hybrid sampling

2.2. Data balancing

When the balance of the subjects to be classified is biased to one side, the classification will be conducted to increase the accuracy of the major class because the classification accuracy is affected by the class with larger classification targets. Therefore, it is necessary to increase the classification accuracy for a small class, and the method applied to achieve this is called data balancing. Sampling is mainly used as a method for data balancing. Representative methods are under-sampling to reduce the number of data in a large class, over-sampling to increase the number of data in a small class, and hybrid sampling that mixes these two methods.

This study used hybrid sampling as a way to overcome class imbalance (Lee [19]). This method maintains the strengths of under-sampling and over-sampling while minimizing their shortfalls by extracting data from the large class with considering the overall distribution and generating some new data from a small class (please see Lee[19] for further details). Figure 2 presents the flowchart of hybrid sampling used in this study.

The total data is D, and Equation (1) presents the formula. D_{ma} and D_{mi} stand for data belonging to the major class and those belonging to the minor class, respectively.

$$D = D_{ma} + D_{mi} \quad (1)$$

A support vector indicating the boundary of classes (Equation (2)) shall be calculated by applying the SVM classifier to the entire data. Among the support vectors (SV), the support vector belonging to the major class is $SV_{ma,}$ and that belonging to the minor class is SV_{mi} (Equation (3)).

$$SV = SV_{ma} + SV_{mi} \quad (2)$$

$$SV_{ma} \in D_{ma},\ SV_{mi} \in D_{mi} \quad (3)$$

Among data belonging to a large class (Equation (4)), D_{mar} is obtained by applying random sampling to data D_{mas}, not support vectors (Equation (5)). The number does not exceed the number of the support vectors of a large class (Equation (6)).

$$D_{mas} = D_{ma} - SV_{ma} \quad (4)$$

$$D_{mar} = rand(D_{mas}) \quad (5)$$

$$n(D_{mar}) < n(SV_{ma}) \quad (6)$$

Random over-sampling example methods (Equation (7)) shall be applied to the data belonging to a small class to balance the number of the new large class data and that of the new small class data [20]. The formula is presented in Equation (8).

$$D_{mir} = ROSE(D_{mi}) \quad (7)$$

$$n(D_{mir}) + n(D_{mi}) = n(D_{mar}) + n(SV_{ma}) \quad (8)$$

New data are constructed by combining the support vectors of the large class, random sampling data, data belonging to the small class, and data generated by additional random sampling. The formula is presented in Equation (9).

$$D_n = SV_{ma} + D_{mar} + D_{mi} + D_{mir} \quad (9)$$

Hybrid sampling reduces the time required for modeling because the generated data is balanced between classes, and the number of data is smaller than the number of original data. Moreover, it decreases the possibility of generating unrealistic data because the number of newly generated data is similar to that of the original small class.

2.3. SDN

Convolutional neural networks (CNN) and recursive neural networks (RNN), deep learning algorithms, have difficulties in applying them directly to corporate data because the forms of data are limited [19]. Therefore, deep networks are composed by increasing the layers of the multi-layer perceptron. Still, it has a problem that the performance does not increase much even if the number of a network's layers is increased. It is due to the vanishing gradient, a neural network learning method. When adjusting the weight of a network, if a factor affecting learning has a deeper network, the influence becomes 0.

The vanishing gradient conducts network learning only in one or two layers close to the output layer. Therefore, the learning time increases drastically when the number of network layers increases. However, it does not change its performance much. SDN is a method to resolve vanishing gradient problems [19]. SDN imitates the learning process of CNN. It learns one layer each rather than learning deep networks all at once.

SDN consists of two stages: hidden layer learning and fine-tuning. First, the hidden layer learning step learns and accumulates one layer per time. Autoencoder [21] is used as an algorithm for this step. Autoencoder is an unsupervised learning algorithm in which the input and output layers are the same. The hidden layer of the model generated by the learning results of feature extraction regulates the number of features while maintaining the attributes of the input data. It is possible to increase or decrease the number of features by increasing or decreasing the number of nodes in the hidden layer.

Moreover, the output of the hidden layer means the hidden features of the input data. Since one node of the hidden layer has a complete connection form connected to all input data nodes, new hidden features are extracted because they are non-linearly transformed for the input data. The formula of the autoencoder is as Equation (10) [19].

$$X = D_1 \cdot E_1(X) \quad (10)$$

The hidden feature is expressed as Equation (11). It means the output of the encoder. The second hidden layer learns E2(H_1) an autoencoder (Equation (13)) as input data using H1, the hidden output of Equation (12), which is the output of the first hidden layer.

$$E_1(X) = H_1 \quad (11)$$

$$H_1 = D_2 \cdot E_2(H_1) \quad (12)$$

$$E_2(H_1) = H_2 \quad (13)$$

Afterward, it was combined with the network to perform fine-tuning (Equation (14)).

$$X = D_1 \cdot D_2 \cdot E_2 \cdot E_1(X) \quad (14)$$

After repeating this process as many as the desired number of hidden layers, the output data are learned by applying the perceptron as shown in Equation (15).

$$O = F(H_n) \quad (15)$$

Finally, the final deep learning model learns by performing fine-tuning for the entire network (Equation (16)).

$$O = F \cdot E_n \cdot E_{n-1} \cdot \ \ldots \cdot E_2 \cdot E_1(X) \quad (16)$$

2.4. Experimental Data

The experimental data of this study were 470 college students enrolled in five universities in Busan and Gyeongsangnam-do, the Republic of Korea. This study collected samples using nonprobability convenience sampling considering subjects' majors (SW-related majors or non-SW-related majors) and grades (freshmen, sophomores, juniors, or seniors). Data were collected through an online questionnaire, and this study finally analyzed 438 subjects after excluding 32 non-respondents.

The recognition of AI convergence education and the education requirement for AI convergence education, which were the features used in this study, were investigated using a questionnaire (a Likert 5-point scale) based on previous studies [22,23] (Table 1). Two professors who majored in AI carried out a content validity analysis for the learning elements to examine the content validity of the learning element items of AI convergence education (e.g., data collection and analysis, understanding the principles of artificial intelligence using unplugged activities, and learning using a text-type programming language). The imbalance of the data was 0.2.

2.5. Performance Evaluation of the Developed SDN

This study compared the performance of logistic regression (LR), support vector machine (SVM), decision tree (DT), and random forest (RF) machine learning algorithms with that of the SDN proposed in this study. This study fixed the random seed to "467134" for models containing randomness, such as RF. Performance was evaluated by accuracy, recall, and precision. This study compared the predictive performance of each model, and the model with the highest accuracy was defined as the model with the best predictive performance.

3. Results

Figures 3 and 4 compare the ML accuracy of the test data to that of the raw data without correcting data imbalance and that of the hybrid sampling data. It was found that, among the test data, DT among the ML algorithms had the lowest accuracy (87.5%) against the raw data, while SDN showed the highest performance (91.1%) (Figure 3). When data balance was improved by hybrid sampling (Figure 4), all ML algorithms using the test data showed 0.3 to 0.7% improved accuracy compared to the raw data. When SDN applied hybrid sampling to the test data, it improved accuracy by 0.7% compared to the raw data.

Table 1. Attributes of data

Items	Scale
Interest in AI	5-point Likert scale
Knowledge on AI	
Awareness of the impact of AI on society	
Opportunities owing to AI	
Degree of understanding of AI convergence education	
Methods of applying AI convergence education to the curriculum	
Degree of interest in AI convergence education	
Awareness of AI convergence education learning elements: principles of artificial intelligence (e.g., machine learning and deep learning)	
History of AI	
AI and social impact	
Data collection and analysis	
AI and ethics	
Understanding the principles of AI using unplugged activities	
Experiencing AI using the platform	
Evaluation activities using AI	
Experiencing and implementing AI using block coding	
Experiencing AI using text-type programming languages such as Python	
AI convergence education related to curriculum	

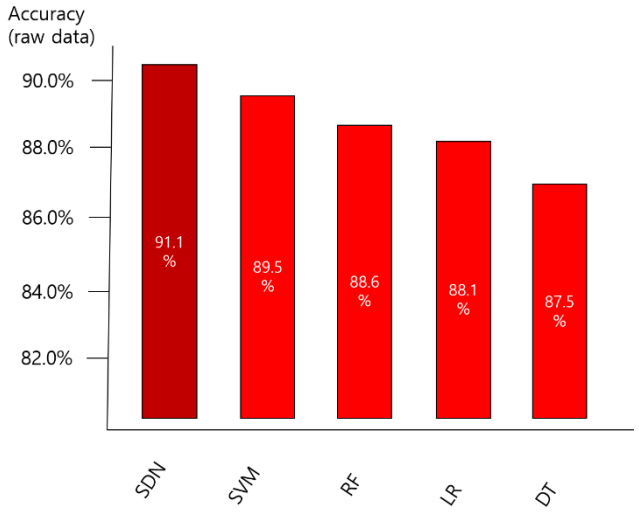

Fig. 3.Comparison of ML algorithms' accuracy using the raw data (%)

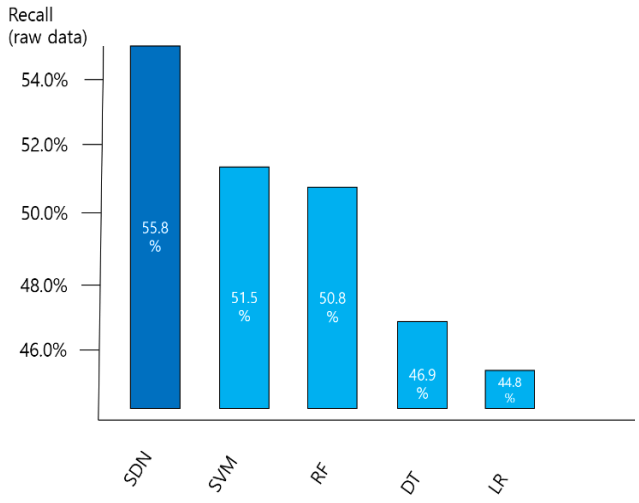

Fig. 6 Comparison of ML algorithms' recall using hybrid sampling (%)

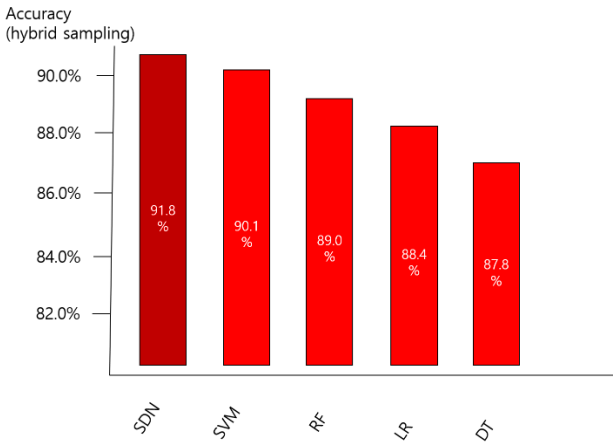

Fig. 4 Comparison of ML algorithms' accuracy using hybrid sampling (%)

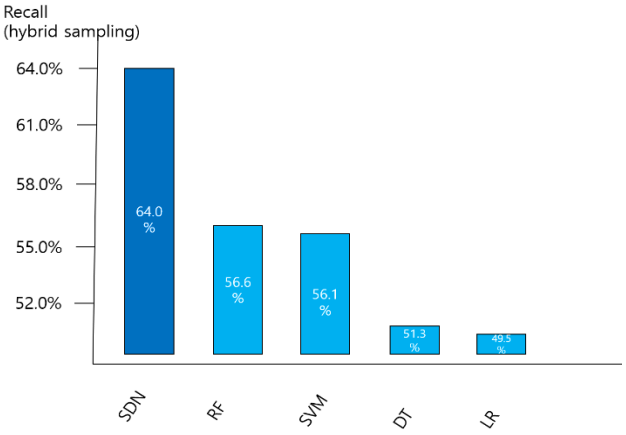

Fig. 7 Comparison of ML algorithms' precision using the raw data (%)

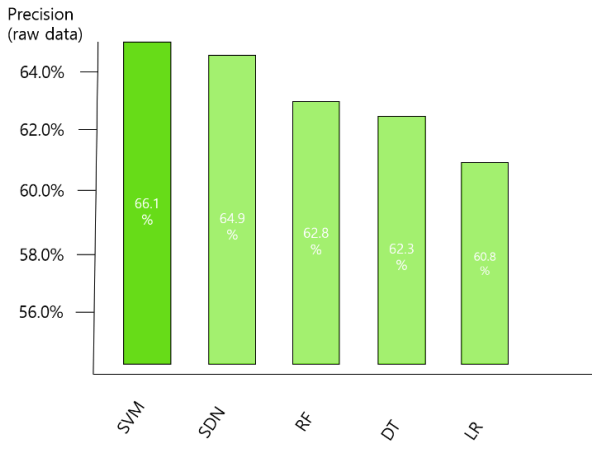

Fig. 5 Comparison of ML algorithms' recall using the raw data (%)

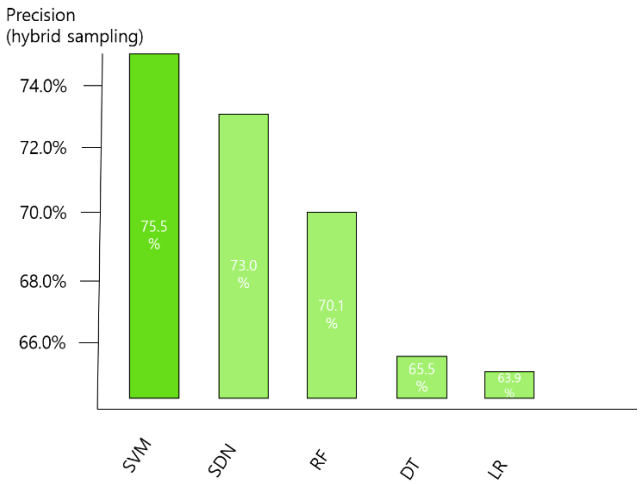

Fig. 8 Comparison of ML algorithms' precision using hybrid sampling (%)

Figures 5, 6, 7, and 8 compared the recall and precision of ML algorithms for the test data regarding the raw data and the hybrid sampling data. Regarding the recall of the raw data, SDN among the ML algorithms had the highest recall, and SVM had the highest precision. When the data balance was improved by hybrid sampling, all ML algorithms improved the recall by 5-8% and the precision by 3-10% compared to the raw data. When hybrid sampling was applied, the SDN proposed in this study improved the precision and recall by 8% or higher than the raw data.

5. Conclusion

As digital transformation accelerates, the demand for AI education also continues to increase in the education sector. Therefore, to enhance AI education, it is necessary to understand the needs of education consumers, improve education quality, and increase satisfaction. This study proposed an ML model which accurately predicted educational satisfaction by combining hybrid sampling and SDN.

This study improved the performance by balancing the number of data between classes using hybrid sampling to overcome the problem of decreased recall and precision compared to the accuracy that unbalanced classes of data could induce. Moreover, this study proposed a method to increase AI education satisfaction classification performance using SDN. This study applied the proposed method to the questionnaire data and showed excellent performance in accuracy, recall, and precision in predicting educational satisfaction.

The SDN method based on hybrid sampling proposed in this study had superior accuracy, recall, and precision performance to other ML algorithms. Especially when hybrid sampling was applied to the SDN algorithm, recall, and precision performance was greatly improved. It is believed that the overall recall and precision were improved because the recall and precision of the small classes, which were lower, were improved owing to the corrected data imbalanced by hybrid sampling proposed by us. Since SND has a feature extraction function and can be applied to unrefined data, additional studies should evaluate the performance by applying the proposed model to unrefined data with various attributes.

Acknowledgment

The 2021 Inje University research grant supported this work.

References

[1] G. Vial, “Understanding Digital Transformation: A Review and a Research Agenda, Managing Digital Transformation,” *The Journal of Strategic Information Systems*, vol. 28, no. 2, pp. 114-118, 2019.

[2] A. C. Davies, A. Davies, A. Wilson, H. Saeed, C. Pringle, I. Eleftheriou, and P. A. Bromiley, “The Health Information Workforce,” ser. Working as a Health AI Specialist, Cham, Switzerland: *Springer*, pp. 247-268, 2021.

[3] M. A. Boden, “*AI: its Nature and Future*,” Great Britain, United Kingdom: Oxford University Press, pp. 37-78, 2016.

[4] X. Zhai, X. Chu, C. S. Chai, M. S. Y. Jong, A. Istenic, M. Spector, J. B. Liu, J. Yuan, and Y. Li, “A Review of Artificial Intelligence (AI) in Education From 2010 to 2020,” *Complexity*, vol. 2021, pp. 18, 2021.

[5] S. Zain, “*Future Directions in Digital Information*,” D. Baker, and L. Ellis, Ed. Cambridge, United Kingdom: Chandos, pp. 223-234, 2021.

[6] S. Li, “How Does COVID-19 Speed the Digital Transformation of Business Processes and Customer Experiences?,” *Review of Business*, vol. 41, no. 1, pp. 1-14, 2021.

[7] M. D. Jones, S. Hutcheson, and J. D. Camba, “Past, Present, and Future Barriers to Digital Transformation in Manufacturing: A Review,” *Journal of Manufacturing Systems*, vol. 60, pp. 936-948, 2021.

[8] Y. Li, “Apple Inc. Analysis and Forecast Evaluation,” *Proceedings of Business and Economic Studies*, vol. 4, no. 4, pp. 71-78, 2021.

[9] N. T. H. Giang, P. T. T. Hai, N. T. T. Tu, and P. X. Tan, Exploring the Readiness for Digital Transformation in a Higher Education Institution Towards Industrial Revolution 4.0,” *International Journal of Engineering Pedagogy*, vol. 11, no. 2, pp. 4-24, 2021.

[10] X. Yang, “Accelerated Move for AI Education in China,” *ECNU Review of Education*, vol. 2, no. 3, pp. 347-352, 2019.

[11] G. Rodríguez-Abitia, and G. Bribiesca-Correa, “Assessing Digital Transformation in Universities,” *Future Internet*, vol. 13, no. 2, pp. 52, 2021.

[12] Mrinalini Smita, "Logistic Regression Model For Predicting Performance of S&P BSE30 Company Using IBM SPSS," *International Journal of Mathematics Trends and Technology*, vol. 67, no. 7, pp. 118-134, 2021.

[13] China AI Textbook, 2018. [Online]. Available: https://item.jd.com/12347925.html

[14] M. Ryu, and S. Han, “AI Education Programs for Deep-Learning Concepts,” *Journal of the Korean Association of Information Education*, vol. 23, no. 6, pp. 583-590, 2019.

[15] H. J. Kim, “Digital Transformation of Education Brought by COVID-19 Pandemic,” *Journal of the Korea Society of Computer and Information*, vol. 26, no. 6, pp. 183-193, 2021.

[16] D. Lee, J. Y. Hwang, Y. Lee, and S. W. Kim, “Informatics and Artificial Intelligence (AI) Education in Korea: Situation Analysis using the Darmstadt Model,” *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 2, pp. 427-444, 2022.

[17] Gurdip Kaur Saminder Singh, "Understanding Students' Insights of Using Audio Technology in a Hybrid-Mode Higher Education Institution," *SSRG International Journal of Humanities and Social Science*, vol. 6, no. 2, pp. 23-27, 2019. Crossref, https://doi.org/10.14445/23942703/IJHSS-V6I2P104.

[18] K. S. Choi, "Opportunities for Higher Education of Artificial Intelligence in Korea," *International Journal of Engineering Research and Technology*, vol. 13, no. 11, pp. 3428-3430, 2020.

[19] H. Lee, "A Method of Bank Telemarketing Customer Prediction Based on Hybrid Sampling and Stacked Deep Networks," *Journal of Korea Society of Digital Industry and Information Management*, vol. 15, no. 3, pp. 197-206, 2019.

[20] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A Review on Imbalanced Data Handling using Undersampling and Oversampling Technique," *International Journal of Recent Trends in Engineering & Research*, vol. 3, no. 4, pp. 444-449, 2017.

[21] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," arXiv preprint, 2020.

[22] S. W. Lee, "Convergence Education of Elementary School Teachers and Pre-Service Teachers," *Journal of Korean Practical Arts Education*, vol. 34, no. 1, pp. 1-17, 2021.

[23] M. Y. Ryu, and S. K. Han, "The Educational Perception on Artificial Intelligence by Elementary School Teachers," *Journal of the Korean Association of Information Education*, vol. 22, no. 3, pp. 317-324, 2018.

[24] touretzkyds/ai4k12, 2019. [Online]. Available: https://github.com/touretzkyds/ai4k12/wiki/Resource-Directory

[25] M. Y. Ryu, and S. K. Han, "The Direction of AI Classes using AI Education Platform," *Journal of the Korea Society of Computer and Information*, vol. 27, no. 5, pp. 69-76, 2022.