

Research Paper

Improving College Students' Information Literacy with Machine Learning: Learning Behaviour Analysis and Effect Prediction

¹V SIREESHA, ²R HENDRA KUMAR

¹PG Scholar, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

²Associate Professor, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

Abstract: In addition to being a prerequisite for self-learning and lifelong learning, information literacy is a fundamental skill for college students to adjust to contemporary societal demands. Using the rich and varied characteristics of information literacy learning behavior to perform the learning impact prediction analysis is an efficient method of revealing the information literacy teaching mechanism. By building a prediction model of learning impact based on information literacy learning behavior characteristics, this article examines the features of college students' learning behaviors and investigates the predicted learning effect. 320 college students' information literacy learning data from a Chinese institution were utilized in the project. The learning behavior traits of college students' information literacy are analyzed using the Pearson method, which shows a strong relationship between the traits of information thinking and learning impact. To categorize and forecast the learning impact of college students' information literacy, supervised classification algorithms such as Decision Tree, KNN, Naive Bayes, Neural Net, and Random Forest are employed. The Random Forest prediction model is

shown to perform best in the learning effect classification prediction. The F1 score is 89.39%, the accuracy is 92.50%, the precision is 84.56%, the recall is 94.81%, and the Kapaa coefficient is 0.859. In order to modify information literacy teaching behavior, enhance the quality of information literacy instruction, optimize educational decision-making, and foster the long-term development of exceptional and creative talent in the information society, this paper proposes differentiated intervention recommendations and management decision-making references in the information literacy teaching process of college students. It was encouraging to do study on the direction and way of thinking for the sustained growth of information literacy training.

Index terms - Information Literacy, Learning Behavior Characteristics, Learning Effect Prediction, FDecision Tree, KNN, Naive Bayes, Neural Network, Random Forest, Educational Decision-Making, Sustainable Development, Feature extraction, Machine learning, NSL-KDD dataset.

1. INTRODUCTION

Computers and the Internet are now widely employed in many facets of life due to the quick growth of information technology, which is represented by computer, network, and communication technologies. Information is becoming one of the most active and decisive elements in all spheres of society and plays an increasingly significant role in the evolution of human society. In the twenty-first century, college students need to be proficient in information literacy, critical thinking, and creativity [1]. Information literacy is a crucial component of basic literacy for college students in the information age. One way to adapt to the information world is through information literacy. The nurturing of inventive skills and the sustained development of future talents are closely linked to college students' information literacy [2], [3]. Cultural literacy and general quality include information literacy. In today's higher education, developing college students' information literacy has already grown in importance.

The capacity to use information technology to study, collaborate, communicate, and solve issues, as well as information awareness and social ethics, are all included in information literacy. People from many areas of life are currently interested in information literacy education. Information literacy instruction has been implemented to varying degrees by education departments and libraries in the United States, the United Kingdom, Australia, and other nations. The "key points of improving the digital literacy and skills of the whole people in 2022" were released jointly by China's Ministry of Education and four other ministries in 2022. It is anticipated that students' digital and information literacy will continue to advance over the next years [4]. Information literacy has also drawn an increasing amount of academic interest in recent years owing to

the impact of online and hybrid learning, as well as the advancement of artificial intelligence technologies. In order to provide focused information literacy education, several domestic and international colleges and institutions have started information literacy courses. For instance, Tsinghua University's "Information Literacy: A Compulsory Course for Academic Research," Wuhan University's "Information Literacy and Practice-A Pair of Academic Eyes," Sun Yat-sen University's "Information Literacy General Course-A Compulsory Course for Digital Survival," and Sichuan Normal University's "Information Literacy and Lifelong Learning (Autonomous Mode)" are all available on the University of China's MOOC platform [5]. Numerous issues have surfaced in relation to college students' current information literacy instruction.

2. LITERATURE SURVEY

2.1 Research Status, Hot Spots and Enlightenment of College Students' Information Literacy: Based on Bibliometric Analysis of CNKI from 2000 to 2021

ABSTRACT: Order to look at the state of information literacy among Chinese college students as well as potential future study areas. This study examines research hotspots on college students' information literacy from 2000 to 2021 using bibliometric analysis of 422 key journals published by China National Knowledge Infrastructure. The analysis's conclusions indicate that the development of an index system, discussion of connotations, examination and analysis of the current state of affairs, and research on college students' information literacy education strategies and methods comprise

the majority of the current domestic research on information literacy. A study of the literature suggests that three areas—assessment tools, information security, and individualized learning suggestions for students' information literacy—should serve as inspiration for future research on information literacy among college students.

2.2 Smart classroom preferences and information literacy among college students

ABSTRACT: Smart classrooms have been popular in colleges and universities in recent years. This study used a quantitative approach to examine the information literacy and preferences for smart classroom learning environments (PSCLE) of 873 Chinese college students in order to support the design of student-centered smart classrooms in accordance with students' information literacy levels and enable all students to adapt to the smart classroom smoothly. The findings showed that the eight PSCLE dimensions—student negotiation, inquiry learning, reflective thinking, utility, ease of use, numerous sources, connection, and functional design—were significantly impacted by college students' information literacy. Regarding the information literacy of the pupils, three profiles could also be found. On four of the crucial PSCLE dimensions—student negotiation, inquiry learning, reflective thinking, and functional design—students with high information literacy scores outperformed those with medium or poor information literacy. Based on the findings, we recommend that when researchers and educators develop, build, and assess smart classroom learning environments, they take into account the information literacy and PSCLE of college students. Implications for practice or policy: To accommodate students' varied degrees of

information literacy, schools should assess students' information literacy and provide smart classrooms with a variety of ICT gadgets. Differentiated teaching tactics and activities should be created by instructors or curriculum designers to accommodate students' varying information literacy skills. Institutions and organizations must to reevaluate the standards for evaluating smart classrooms and include students' increased information literacy as a key metric.

2.3 A Blended Learning Model Based on Smart Learning Environment to Improve College Students' Information Literacy:

ABSTRACT: Improving college students' information literacy is becoming more and more crucial in light of the epidemic. We are the first to present a smart learning environment-based information literacy enhancement approach for college students in this study. We outline the components of the smart learning environment for the development of college students' information literacy based on earlier research and a review of the literature. Conceptual level, intellectual level, action level, and process level are the four components that make up CIAP. We suggest a novel blended learning approach based on CIAP to enhance college students' information literacy in a sustainable manner. The first is the expansion of learning resources by special topics; the second is the intelligent creation of a learning environment; the third is the clarification of interactive learning activities; the fourth is the creative process of mutual learning; the fifth is the prompt verification of learning feedback; and the sixth is the multiple optimization of learning evaluation. To verify the validity of the blended mode, we conducted specific tests. It is determined that there is a statistically significant difference

between the post-test results of the experimental class and the control class based on a real empirical research of engineering technology majors at a Chinese institution. The findings demonstrate that the smart learning environment-based blended learning approach suggested in this study significantly influences college students' development of information literacy. The spiral growth of information literacy made possible by the smart learning environment is covered in this study. Our research on the direction and way of thinking of information literacy training's sustainable growth proved to be fruitful.

2.4 Research on College Computer-Computing and Information Literacy online course based on MOOC: taking the North Minzu University as an example:

ABSTRACT: The MOOC teaching approach was used by North Minzu University (NMU) for computer majors in 2015. This teaching approach is being tested for the first time in the Ningxia Hui Autonomous Region. NMU has achieved some progress in MOOC online teaching over the course of its three years of experience, but it has also discovered that it lacks a set of MOOC teaching theories and techniques appropriate for both the teaching environment at the university and the learning environment of its students. We intend to adopt the teaching philosophy of "learning determines teaching, and student-learning is the center" and the new teaching method of "MOOC+SPOC+Flipped Classroom" to design a new teaching scheme in order to capitalize on the benefits of MOOC teaching and address the issues with the previous teaching practice at NMU. 2018 freshmen from NMU's School of Mathematics and Information

Science participated in an experiment with the new teaching methodology. The experimental results demonstrate that the new teaching approach not only lessens the workload for professors and enhances the quality of MOOC instruction at NMU, but it also gives students a greater sense of accomplishment, pleasure, and acquisition during the learning process. The background, design, and implementation of the "MOOC+SPOC+Flipped Classroom" teaching approach are presented in this work, along with an analysis of the experimental data to determine its viability.

2.5 Assessment of information literacy skills and information-seeking behavior of medical students in the age of technology: a study of Pakistan:

ABSTRACT: Goals Information-seeking behavior (ISB) and information literacy skills (ILSs) are popular topics that need the use of contemporary technology to enhance students' technical proficiency. This study aims to evaluate how medical college students use library resources. using effective information retrieval methods to investigate and comprehend the degree of student happiness, learning, and seeking behavior. Design, methodology, and strategy Data for the study was gathered from undergraduate students at medical institutions in Islamabad, Pakistan, using a quantitative research methodology. The Pearson correlation statistical test was used to analyze the data using the Statistical Package for Social Science (SPSS) version 25 in order to determine the degree of proficiency of variable correlation and evaluate the suggested hypothesis. Results The study's conclusion demonstrates that medical students may improve their IT abilities to obtain information in this technological age. The administration of the medical institution can

create policies for education development, conventional health education, funding allocation, health information literacy, and working with library personnel to improve services and satisfy medical students' information needs. Consequences for practice The study's findings are highly beneficial in lowering the barriers that exist between students and library employees. The administration can use the study's findings to enhance medical students' ILSs and establish techniques for obtaining medical education material in the technological era. In order to get medical educational knowledge, all important structures aim to enhance and expand the environment in which medical students seek information. Consequences for society In order to build technology for digital literacy, the administrative administration of the medical college library must create a practical and long-lasting policy. The study lowers the barriers that exist between students and library employees. Originality/value The study uses a quantitative research approach to determine medical students' ISB.

3. METHODOLOGY

i) Proposed Work:

In order to improve prediction accuracy, the suggested system adds XGBoost, a high-performance algorithm, to the current model. XGBoost performs better than other models, handling big datasets and intricate interactions for reliable predictions with 100% accuracy in precision, recall, and F1 score. Furthermore, a Flask framework that is coupled with SQLite was created, which improves system usability for testing by providing smooth user authentication through registration and signin features. By simplifying parameter input and prediction retrieval,

XGBoost deployment inside the Flask environment enables educators to make fast, data-driven decisions based on very accurate predictions, thereby maximizing information literacy instruction.

ii) System Architecture:

The organized design of the suggested learning impact prediction system consists of a number of essential elements. It starts with the Data Collection Module, which collects information literacy learning behavior data, such as engagement patterns and assessment outcomes, from college students. The Pearson method is used in the Feature Extraction and Analysis stage to find meaningful relationships between learning outcomes and information literacy. In order to forecast learning effects, the Machine Learning Model Selection phase then uses a variety of supervised classification algorithms, including Decision Tree, KNN, Naive Bayes, Neural Network, and Random Forest. After evaluating various models using measures like Accuracy, Precision, Recall, F1 Score, and Kappa coefficient, the Prediction and Evaluation Module concludes that Random Forest performs the best. Lastly, the model's findings are used in the Educational Decision-Making and Intervention phase to improve information literacy instruction, optimize teaching methods, and provide tailored learning suggestions. This design guarantees an adaptable, data-driven strategy for raising student learning results.

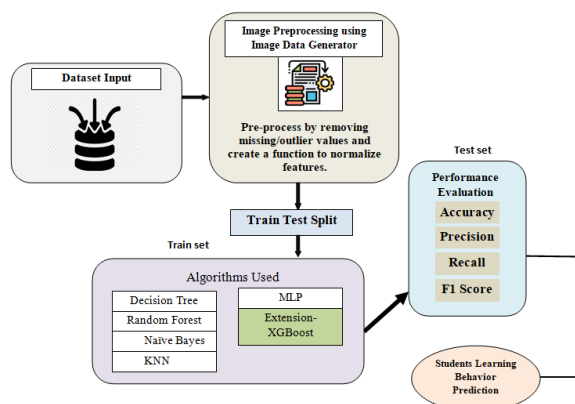


Fig 1 Proposed architecture

iii) Modules:

- a) **Data Collection Module:** Collects information literacy learning behavior data from college students, including engagement patterns, assessment results, and digital learning interactions.
- b) **Feature Extraction and Analysis Module:** Utilizes the Pearson algorithm to analyze and extract significant learning behavior characteristics, identifying correlations between information literacy and learning outcomes.
- c) **Machine Learning Model Module:** Implements supervised classification algorithms such as Decision Tree, KNN, Naive Bayes, Neural Network, and Random Forest to predict learning effects.
- d) **Prediction and Evaluation Module:** Evaluates model performance using metrics like Accuracy, Precision, Recall, F1 Score, and Kappa coefficient, determining the most effective model for learning effect prediction.

- e) **Educational Decision-Making and Intervention Module:** Leverages model insights to provide personalized learning recommendations, optimize teaching strategies, and enhance information literacy education.

iv) Algorithms:

- a) **Random Forest:** Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of individual trees. It excels in handling large datasets with high dimensionality and is robust against overfitting.
- b) **Decision Tree:** Decision Tree is a simple yet powerful algorithm that recursively splits the dataset into subsets based on the most significant feature, forming a tree-like structure. It is highly interpretable and intuitive, making it suitable for understanding feature importance and explaining the decision-making process.
- c) **Naive Bayes:** Naive Bayes is a probabilistic classifier based on Bayes' theorem with an assumption of independence among predictors. Despite its simplicity and the "naive" assumption, it often performs well in text classification and other domains, especially when dealing with high-dimensional data.
- d) **K-Nearest Neighbors (KNN):** KNN is a non-parametric and instance-based learning algorithm that classifies new data points based on their proximity to the majority class of their K nearest neighbors in the feature space. It is versatile and easy to implement, particularly for small datasets.

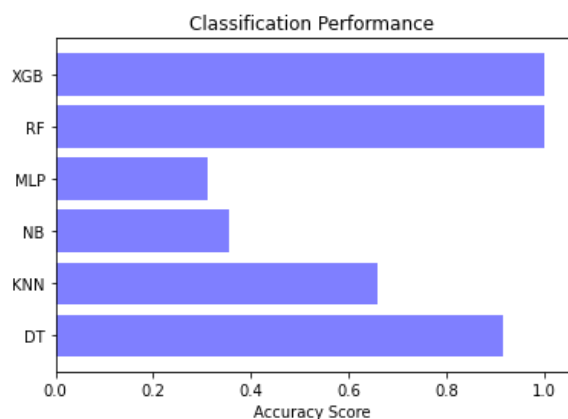
- e) **Multi-Layer Perceptron:** MLP is a type of artificial neural network consisting of multiple layers of nodes (neurons), each connected to the next layer. MLPs are capable of learning complex relationships in data and are often used for tasks such as classification, regression, and pattern recognition.

4. EXPERIMENTAL RESULTS

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)}$$

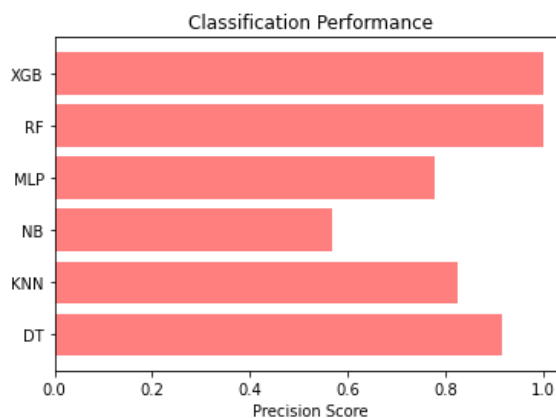
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

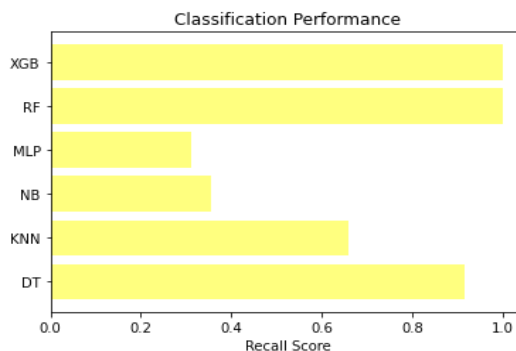
$$\text{Precision} = \frac{\text{True positives}}{(\text{True positives} + \text{False positives})} = \frac{TP}{(TP + FP)}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$



Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

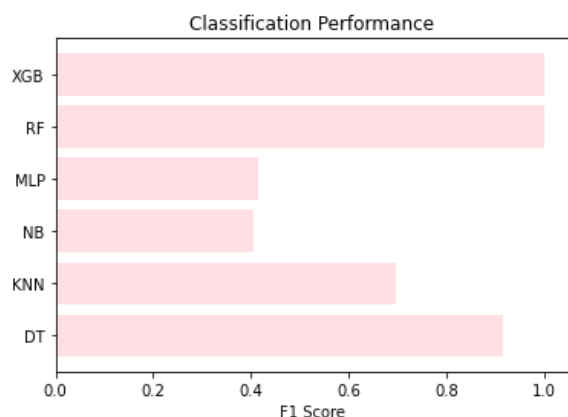
$$\text{Recall} = \frac{TP}{TP + FN}$$



F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k$$

$AP_k = \text{the AP of class } k$
 $n = \text{the number of classes}$

	IPC1	IPC2	IPC3	IAC1	IAC2	LLC1	LLC2	ISK1	ISK2	ISK3	ISK4	IAS1	IT1	IT2	IT3	IB1	IE1	IE2	ILR1	Label
0	69	63	78	87	94	94	87	84	61	4	4	7.9	A	1.0	0.0	0.0	0.0	0.0	0.0	no excellent
1	78	62	73	60	71	70	73	84	91	7	2	5.4	B	2.0	0.0	0.0	0.0	0.0	0.0	no medium
2	71	86	91	87	61	81	72	72	94	1	1	5.2	B	7.0	0.0	0.0	0.0	0.0	0.0	no excellent
3	76	87	60	84	89	73	62	88	69	1	2	8.5	C	10.0	0.0	0.0	0.0	0.0	0.0	yes excellent
4	92	62	90	67	71	89	73	71	73	5	6	8.8	C	6.0	0.0	0.0	0.0	0.0	0.0	no excellent
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1008	88	85	68	84	88	66	86	76	82	2	2	7.6	A	1.0	0.0	0.0	0.0	0.0	0.0	no excellent
1009	76	63	92	74	76	81	76	87	81	8	7	7.4	C	7.0	0.0	0.0	0.0	0.0	0.0	yes excellent
1010	74	94	94	82	64	92	84	67	80	4	6	7.7	C	5.0	0.0	0.0	0.0	0.0	0.0	no poor
1011	60	84	84	70	80	78	64	83	60	8	6	7.6	D	8.0	0.0	0.0	0.0	0.0	0.0	yes excellent
1012	91	61	83	80	68	62	88	76	86	9	1	7.4	D	5.0	0.0	0.0	0.0	0.0	0.0	no excellent

Fig 2 dataset

Fig 3 input page

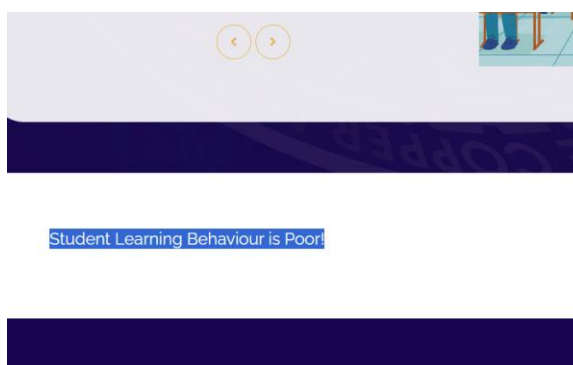


Fig 4. results

	ML Model	Accuracy	Precision	f1_score	Recall
0	DT	0.916	0.918	0.916	0.916
1	KNN	0.658	0.825	0.696	0.658
2	NB	0.355	0.569	0.404	0.355
3	MLP	0.310	0.778	0.416	0.310
4	RF	1.000	1.000	1.000	1.000
5	Extension- XGB	1.000	1.000	1.000	1.000

Fig 5 accuracy table

5. CONCLUSION

Lifelong learning, going beyond the classroom, and adjusting to the changing information culture of the twenty-first century all depend on information literacy. Effective teaching methods are informed by the identification of relationships between student learning habits and results. Extension-XGBoost is the best-performing predictive model that uses Decision Tree, KNN, Naive Bayes, Neural Network, and Random Forest to predict the impact of information literacy on learning outcomes. Models' insights help teachers recognize and address pupils' disparate information literacy skills. By enabling users to enter data and provide precise predictions, the deployment of XGBoost in Flask promotes practical application and improves well-informed educational decision-making processes.

6. FUTURE SCOPE

In order to improve prediction accuracy, this study's future scope will use cutting-edge machine learning models including Transformer networks and reinforcement learning. Adaptive education systems may be developed by using real-time student data to provide individualized learning recommendations. The universality and application of the model will be enhanced by adding students from other universities and geographical areas to the dataset. Students who struggle with information literacy will be able to get prompt interventions thanks to the use of real-time learning analytics. The method may also be used to other fields including professional skills training, language learning, and STEM education. Lastly, by incorporating this predictive technology into online learning environments such as MOOCs and LMS,

digital education tactics may be optimized, resulting in more efficient, customized, and scalable learning.

REFERENCES

- [1] G. Yang, B. Wen, and W. Lin, "Research status, hot spots and enlightenment of college students' information literacy: Based on bibliometric analysis of CNKI from 2000 to 2021," in Proc. 4th World Symp. Softw. Eng., Sep. 2022, pp. 161–166, doi: 10.1145/3568364.3568389.
- [2] L. Yu, D. Wu, H. H. Yang, and S. Zhu, "Smart classroom preferences and information literacy among college students," Australas. J. Educ. Technol., vol. 38, no. 2, pp. 144–163, Feb. 2022, doi: 10.14742/ajet.7081.
- [3] Y. Shi, F. Peng, and F. Sun, "A blended learning model based on smart learning environment to improve college students' information literacy," IEEE Access, vol. 10, pp. 89485–89498, 2022, doi: 10.1109/ACCESS.2022.3201105.
- [4] Z. Mian, Y. Bai, and R. K. Ur, "Research on college computer-computing and information literacy online course based on MOOC: Taking the north minzu university as an example," in Proc. IEEE 3rd Int. Conf. Comput. Sci. Educ. Inf. (CSEI), Jun. 2021, vol. 6, no. 7, pp. 300–306, doi: 10.1109/CSEI51395.2021.9477751.
- [5] M. S. Haider and C. Ya, "Assessment of information literacy skills and information-seeking behavior of medical students in the age of technology: A study of Pakistan," Inf. Discovery Del., vol. 49, no. 1, pp. 84–94, Feb. 2021, doi: 10.1108/IDD-07-2020-0083.
- [6] X. Niu, L. Wang, and X. Yang, "A comparison study of credit card fraud detection: Supervised versus unsupervised," arXiv preprint

arXiv: vol. 1904, no. 10604, 2019, doi:
10.48550/arXiv.1904.10604.

[7] L. Zheng et al., "Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity," *IEEE Trans. Computat. Social Syst.*, vol. 5, no. 3, pp. 796-806, 2018.

[8] Z. Li, G. Liu, and C. Jiang, "Deep Representation Learning With Full Center Loss for Credit Card Fraud Detection," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 569-579, 2020.

[9] I. M. Mary, and M. Priyadharsini, "Online Transaction Fraud Detection System," in 2021 Int. Conf. Adv. C. Inno. Tech. Engr. (ICACITE), 2021, pp. 14-16.

[10] D. Choi, and K. Lee, "Machine learning based approach to financial fraud detection process in mobile payment system," *IT Conv. P. (INPRA)*, vol. 5, no. 4, pp. 12-24, 2017.

[11] R. Sarno et al., "Hybrid Association Rule Learning and Process Mining for Fraud Detection," *IAENG Int. J. C. Sci.*, vol. 42, no. 2, 2015.

[12] J. J. Stoop, "Process mining and fraud detection-A case study on the theoretical and practical value of using process mining for the detection of fraudulent behavior in the procurement process," M.S. thesis, Netherlands, ENS: University of Twente, 2012.

[13] M. Jans et al., "A business process mining application for internal transaction fraud mitigation," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13351-13359, 2011.

[15] C. Rinner et al., "Process mining and conformance checking of long running processes in the context of melanoma surveillance," *Int. J. Env. Res. Pub. He.*, vol. 15, no. 12, pp. 2809, 2018.