WIP : Fundamentals on Cyber Fraud Detection and Investigation: Empowering High School Students for a Secure Digital Future

Gahangir Hossain
Information Science
University of North Texas
Denton, TX, USA
Gahangir.Hossain@unt.edu

Tylor Hurt
Computer Inforation Systems
West Texas A&M University
Canyon, TX, USA
taylor.jackson.hurt@gmail.com

Mikyung Shin
Special Education
Illinois State University
Normal, IL, USA
mshin2@ilstu.edu

***Abstract*— The main goal of this work-in-progress paper is to develop a high school cybersecurity curriculum with introduction to the fundamental concepts of cyber forensics and investigation. This includes understanding various forms of multimedia data, digital evidence, ethical considerations, and the investigative process. The aim of this research paper is to empower students to critically analyze digital incidents, identify potential threats, and adopt responsible online behavior. The paper outlines a comprehensive curriculum structure, discusses the incorporation of practical exercises, case studies, and real-world scenarios to enhance students' hands-on experience and problem-solving skills. The proposed teaching methodology focuses on interactive and engaging approaches including learning technology platform (e.g. canvas or blackboard). Online modules and prerecorded video lectures will be developed so that the students can explore them as their interest. After each moule, there will be some assessments on the topics in the form of quizzes, the student will have to pass the quizzes with 80% score to go through the next modules. Recognizing the potential challenges of introducing cyber forensics at the high school level, the paper addresses concerns such as resource limitations, teacher training, and ethical considerations. It offers viable solutions and strategies to overcome these challenges, emphasizing the importance of collaboration between educators, industry professionals, and policymakers. The proposed curriculum not only equips students with essential technical skills but also fosters a culture of responsible digital citizenship.**

Keywords— Cybersecurity Education; High School Curriculum; Cyber Fraud Detection; Digital Citizenship; Proactive Cybersecurity

I. INTRODUCTION

In the age of increasing cyber threats and the ubiquitous nature of technology, preparing high school students with the necessary cybersecurity skills to understand, investigate, and mitigate cyber incidents becomes imperative. The goal of this work-in-progress paper is to early educate about computer forensics methods and fraud investigation process, and the ability to safely interact with digital world. The course aims to prepare them effectively with essential cybersecurity literacy that will be helpful for their personal and professional lives encouraging a proactive response to various cyber risks. Computer forensics is continually evolving, with new technologies expanding the array of tools available to investigators. A survey on various methods can also be adopted from [2].

In this research, we have performed a case study on credit card fraud dataset. In the credit card industry, debtors are not always honest about their current financial situations, intent, and their identity. Even for large credit card companies, the cost of a debtor not paying back what they owe is high, especially considering this cost can be mitigated. Using machine learning, credit card companies could become better at making data-driven decisions regarding their application approvals/disapprovals.

Fraud hurts not only the credit card company, but also sellers and people whose identity is stolen. It's important that creditors perform their due diligence as best as they can, and for current times, that involves the use of machine learning tools. The dataset for this project is from Kaggle.com. It's a large dataset with less than 10% of the occurrences being fraudulent. This large imbalance of fraudulent transactions versus non-fraudulent transactions creates an issue for creating a machine learning model. To resolve this issue, SMOTE (synthetic minority oversampling technique) can be applied to the dataset, which will create more fraudulent transactions for training the machine learning model. Based on the case study, it is observed that both large and small entities who are involved in lending money to debtors should utilize and implement machine learning models into their decision-making processes. SMOTE has proven to be a viable solution to dealing with class imbalance in the case of fraud detection, as such, lenders should include this technique in their machine learning models. Furthermore, SMOTE could be applied to occurrences which are heavily imbalanced in any industry. Such examples could be electric line fault prevention, customer churn, employee churn, injury prevention.

Section II provides a fundamental definition of cyber forensics, including examples, investigative techniques, and types of investigations. Section III outlines the detailed steps involved in the cyber fraud investigation process. In Section IV, a case study on credit card fraud is presented to illustrate a practical application of cyber fraud investigation. Finally, Section V concludes the paper, summarizing the key points discussed.

II. CYBER FORENSICS

A. What is Cyber Forensics?

Cyber forensics, also known as computer or digital forensics, involves the systematic collection and analysis of

digital data for use as evidence in legal cases related to civil rights, criminal activities, and security threats. This process requires specialized computing equipment, such as workstations in secure laboratories or robust and secure servers.

B. Cyber Investigations: Public vs. Private

The purpose of a cyber investigation is to conduct various forensic analyses and investigative tasks on systems that may contain evidence. The investigative triad (figure 1) comprises three components: (1) risk management and vulnerability/threat assessment, (2) testing and integrity verification of stand-alone workstations and network servers, and (3) network intrusion detection and incident response.

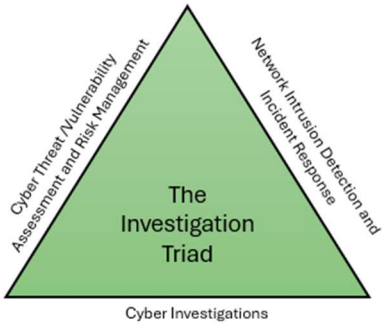

Fig 2. Investigation triad (figure modified and redrawn from [5])

There are two types of investigations: public sector investigation and private sector investigation. In public sector investigations, the Fourth Amendment of the United States Constitution, which restricts government search and seizure, takes precedence. These investigations are conducted by government agencies to determine whether an individual has violated public policies. Public sector investigators must be well-versed in laws related to standard court procedures, search and seizure regulations, and investigation filing and reporting norms [5]. Private sector investigations, on the other hand, are conducted by businesses and attorneys dealing with legal disputes and policy violations. Crimes in the private sector can include email harassment, data fabrication, age and gender discrimination, embezzlement, sabotage, and industrial espionage [6].

Both public and private sector investigations may involve digital forensics. In public sector cases, a search warrant is typically required before examining digital evidence [5]. It is recommended in [5], to adopt a methodical approach when conducting investigations and to carefully consider the case's requirements, nature, and methods for acquiring evidence when planning a case.

III. Fraud Investigation Process

The cyber fraud investigation process incorporates methods like traditional investigation processes, but with the addition of a technology-rich environment. Therefore, having at least one expert knowledgeable in the latest computing and cyber technologies is essential for the investigation team. Below, (figure 2) we outline a sample set of five steps that can be followed during the investigation process [6].

Step #1: Identification Phase - In the identification phase, the investigative team must locate and identify evidence in the form of data, such as images, on the devices. This includes gathering detailed information about the location and format of this data.

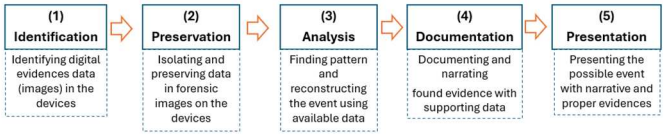

Fig 2. Cyber Fraud Data Classification and Investigation Process

Step #2: Preservation Phase - During the preservation phase, the data is isolated, protected, and preserved to ensure it can be copied for analysis when needed. This process, known as "imaging" a device, maintains the original evidence intact, making it admissible in court.

Step #3: Analysis Phase - In the analysis phase, the team reconstructs all fragments of data to develop a comprehensive narrative of the event under investigation. This step is crucial for understanding the context and details of the incident.

In this phase an investigator can use machine learning to find patterns. For instance, Weka machine learning tools [7][8] can be used in fraud data processing. The initial logo of Weka is shown in figure 3 (top).

After loading the data in Weka explorer, we can select the classification algorithms as shown in figure 3 (down). In particularly, in the figure 3 (down), a decision table-based classifier is used that shown corresponding Classification Result with, Correctly Classified Instances, Incorrectly Classified Instances, True Positive (TP) rate, False Positive (FP) rate, Precision, Recall, and F-measures. More details of machine learning at high school level can be explored from [7].

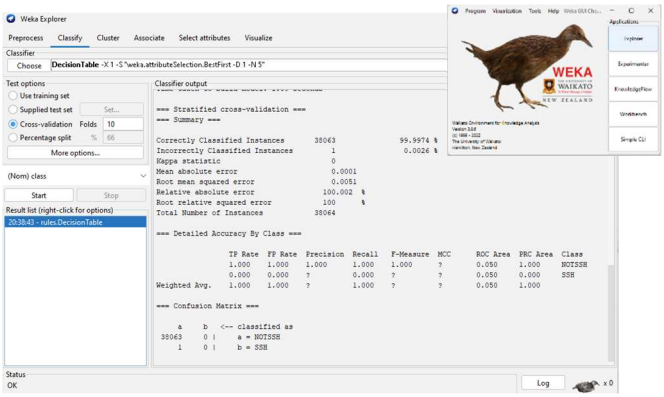

Fig 3. Cyber Data Classification with Weka

Step #4: Documentation Phase - The documentation phase involves organizing and preparing the evidence in a clear and proper format. This ensures that the legal team can easily comprehend and act upon the evidence during judicial proceedings.

Step #5: Presentation Phase - In the presentation phase, the investigator summarizes the documentation, presenting a sequential pattern of evidence to explain the conclusions drawn about the event. This summary helps clarify the findings and supports the case during decision-making.

This work has been partially supported by the Office of Naval Research (ONR), USA, Award Number N00014-23-1-2454.

IV. CURRICULUM DESIGN : CYBER FRAUD INVESTIGATION

Designing and developing an innovative cybersecurity curriculum for high school students that effectively engages them is a demanding and challenging task. Douglas et al. [1] present a digital forensics course for high school students, highlighting best practices in curriculum design and evaluation while identifying challenges in introducing digital forensics education at this level. Table I, adapted from Nelson's book on cyber forensics [5], suggests a college-level syllabus that can be tailored for a high school cybersecurity curriculum with 14 weeks. This table provides a framework for developing comprehensive and age-appropriate coursework in cybersecurity for high school students.

TABLE I. SAMPLE CYBER FORENSICS AND FRAUD INVESTIGATION CURRICULUM (ADOPTED FROM [5])

Week	Topics
1	Understanding the Digital Forensics Profession and Investigations
2	The Investigator's Office and Laboratory
3	Data Acquisition
4	Processing Crime and Incident Scenes
5	Working with Windows and CLI Systems
6	Current Digital Forensics Tools
7	Linux and Macintosh File Systems
8	Recovering Graphics Files
9	Digital Forensics Analysis and Validation
10	Virtual Machine Forensics, Live Acquisitions, and Network Forensics
11	Email and Social Media Mobil Device Forensics
12	Cloud Forensics
13	Report Writing for High Tech Investigations Expert Testimony in High Tech Investigations
14	Ethics for the Investigator and Expert Witness

In addition, the learning activities related to cyber forensics and fraud investigation can be articulated and assessed through the model outlined in [13][15][16]. This approach ensures that the curriculum is both comprehensive and effectively evaluates students' understanding and skills. In evaluating the curriculum's effectiveness, educators and experts can employ assessment tools such as the Likert-scale-based System Usability Scale (SUS) to measure the effectiveness, efficiency, and satisfaction of the cybersecurity curriculum in supporting the topics of cyber fraud investigation. Furthermore, a brief survey measuring content validity ratio (CVR) to check the relevance of the content (e.g., 5 points = essential, 1 point = not necessary) aligned with the regional and national high school STEM and computer science curriculum can be implemented. In this case, the following CVR formula can be used to measure the curriculum relevancy: $CVR = (N_e - N/2)/(N/2)$, where the N_e is the number of experts stating "essential" and N is the total number of experts.

Table I illustrates the integration of a cyber forensics data analysis and fraud investigation process within project-based instruction in a high school computing curriculum. Principle I, as shown in Table I, focuses on incorporating problem-based learning into data visualization instruction. Principles II through V address the teaching and learning of forensic data identification and processing to uncover patterns related to cybersecurity incidents. Principles VI through IX define the roles of students and teachers within the high school context. Furthermore, Principle X, as suggested by [9], acts as a meta-principle advocating for action rather than mere theory. These ten principles, abbreviated as I through X, along with brief one-sentence descriptions, are outlined in Table I.

TABLE II. CYBER FRAUD INVESTIGATION INSTRUCTIN DESIGN

#	Category	Topic	Single sentence description
I	Course	Instruction-based	Course structure: problem-based learning module and exercises
II	Teaching	Cognition	Teaching cyber forensics data analysis and fraud investigation with "What is cyber forensics?", "What is forensics investigation?" and "How to find, orient, report and present digital evidence connected to fraud events?"
III		Adjustment	Adjusting the cyber forensics and fraud investigation course in school data science and computing curricula
IV		Projection	Designing a project in taking input data from fraud event and analyze, mine to identify important and evident patterns
V		Connectivity	Connecting cyber forensics, fraud data processing, investigation, and reporting, and other connected cybersecurity topics with real-world problems
VI	People	Evaluation	Understanding the effective fraud investigation process and data analysis to discover patterns of evidence will be evaluated as part of students secure computing or cyber data science and analytics learning
VII		Listening	Listening to students experience and feedback on learning the cyber forensics data identification, processing and fraud investigation topics
VIII		Reflection	Students' reflection and progress on fraud investigation and cyber forensic evidence learning
IX		Coaching	Coach or teachers' conception on the cyber forensics and fraud investigation instruction in the school level
X	Meta	Inspiration	Inspiring the fraud data processing, examination, reporting and presentation processes connected to advanced data science and machine learning project development and implementation methods

In [16], Shuchi Grover makes a compelling case for formative assessment and teacher formative assessment literacy in K-12 computer science, particularly beneficial for cybersecurity education. Integrating formative assessment into classroom computing activities can play a critical role in teaching cyber forensics and fraud investigation. Figure 4 illustrates how teachers' pedagogical content knowledge (PCK [17]) is interwoven with classroom assessment and practical habits, highlighting its significance in both teacher training and student preparation in K-12 education. This PCK can be developed through multi-dimensional efforts that is based on experiences empowered over time. Towards cybersecurity learning assessment, we propose to adopt PCK model as shown in fig 6. As shown in the Ven diagram, a person (teacher)-centered approach, relevant content knowledge, practices, and formative assessment provide continuous information to the teacher to

grow professionalism and enhance the specific teaching subject area.

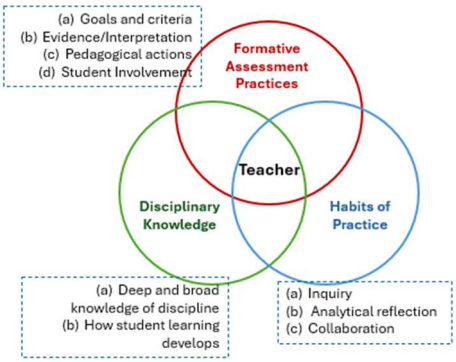

Fig 4. Teacher PCK is intertwined with classroom assessment and habits of practice (Image concept from [16][17]

We can also apply some summative approaches in learning assessment with learning impact and learning growth outcomes.

A. Learning Impacts Estimation for Binary Data

To examine the learning impact, we can use a link mixed model (LMM) in binary outcome. Researchers can consider the following logit model (Equation 1):

$$\text{logit}(y_{ij}) = \alpha_j + \beta X_{ij} + \theta T_i \quad (1)$$

Where, y_{ij} represents the outcome of interest for student i in course j; αj is a course-specific intercept; X_{ij} is a vector of characteristics of student i in course j; T_j is an indicator for random assignment to either the treatment or control group; and β and θ are parameters to estimate, with robust standard errors clustered at the course level. In this framework, the θ term represents the impact of attending the cyber forensics course. To account for the different treatment probabilities for other computing and cybersecurity-related courses, a weighted impact estimation can be adopted based on the inverse of each course's difficulty probabilities.

B. Measuring Learning Growth on Ordinal Responses Data

By participating in cyber forensics courses, through face-to-face, online, and hybrid modes, it is assumed that students can improve their self-efficacy toward cybersecurity and related technology. With the emphasis on data-based decision making in schools, a valid tool to observe and measure students' behaviors is required.

In many clinical and educational settings, teachers often evaluate students' perceptions of their learning behaviors through 5-point Likert scales (1 = strongly disagree to 5 = strongly agree). In this case, researchers can employ a cumulative link mixed model (CLMM) to examine the effect of cybersecurity forensics programs and see how students evaluate their own perspectives toward their learning over time.

For these ordinal response data, we can also use a logit link function to predict the cumulative log odds of being above a certain level of the ordinal outcome. Time-related variables such as measurement occasions (days, weeks, or months) at level 1 are nested within students at level 2 (Equation 2).

$$\text{Level 1: logit}\left[\pi_{kij}(Y > k)\right] = \ln\left(\frac{\pi(Y_{ij}>k)}{\pi(Y_{ij}\le k)}\right) = \left(\beta_{0j} + \beta_{1j} Time_{ij}\right) \quad (2)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Group}_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\left(\text{Group}_j\right)$$

in which logit (π_{kij}) is the logit link for the cumulative probability, $\pi_{kij}(x) = P(Y > k \mid x_1, x_2, \ldots, x_p)$, of being above a particular cut point k (k = 1, ··· 4) on the ith measurement occasion (i = 1, 2, 3 ··· I) for the jth student (j = 1, ··· J). β_{0j} and β_{1j} are logit coefficients, where each represents estimates for the intercept and the *Time* predictor, respectively. The intercept represents reversed (−) threshold coefficients. In the level 2 equation, γ_{00} is the overall logit or log odds of being above a particular outcome score on self-efficacy toward cybersecurity and related technology for the control or comparison group at the pretest which assumed to vary across students. γ_{01} indexes the fixed effect of level 2 predictor related to students (dummy coded as "0" for the control and "1" for the group receiving forensics programs), and γ_{10} and γ_{11} represent cross-level interactions between level 1 and 2 predictors. u_{0j} is the random deviation.

V. Case Study : Credit Card Fraud Detection

The global cost of credit card fraud is estimated to reach $35 billion in 2022 (ncr.com). Credit card companies must vet applicants to mitigate the risk of non-payment and identity fraud. A machine learning tool that detects fraudulent credit applications and transactions can provide a competitive advantage. If companies collaborate to develop such a tool, both the industry and the economy would benefit. This paper explores fraudulent credit applications and the use of machine learning in this context.

A. Data Set and Processing

The credit card fraud detection dataset from Kaggle has 121 predictor variables and one target variable for fraud ('TARGET'), with 8.07% of the 307,511 instances being fraudulent. Irrelevant columns, including gender (to avoid bias), are dropped, leaving 75 columns. Columns with many null values, like OWN_CAR_AGE, EXT_SOURCE_1, and OCUPATION_TYPE, are also excluded. For numerical consistency, object data types are converted to dummy variables, adding 102 uint8 columns. A correlation matrix helps identify and exclude highly correlated attributes, improving model performance.

B. Machine Learning in Fraud Detection

A decision tree, neural network, and random forest classifier were used for training and testing. The dataset was split 80/20 for training/testing, and SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the fraud instances. SMOTE increased the number of fraud cases from 24,755 to 281,464.

C. Model Evaluation

(1) Decision Tree: Best performer with SMOTE, achieving a recall of 91% and accuracy of 90%. Without SMOTE, recall

was 15% and accuracy 85%. (2) Neural Network: Accuracy dropped from 92% to 75% with SMOTE, but recall improved from 0% to 77%. (3) Random Forest: Similar performance to the decision tree without SMOTE; recall improved but accuracy decreased with SMOTE.

SMOTE effectively improves recall across all models. The decision tree combined with SMOTE is the most effective model, balancing high recall and accuracy. Creditors should use similar data preparation techniques and apply SMOTE for imbalanced fraud data. Further information should be gathered if a model identifies an application as fraudulent.

TABLE III. AVERAGE CLASSIFICATION PERFRONCE (WITH SMOTE)

Classifier	Precision	Recall	F1-score	Accuracy
Decision Tree	0.90	0.90	0.90	0.90
Neural Network	0.75	0.75	0.75	0.75
Random Forest	0.75	0.64	0.59	0.64

VI.

VII. LIMINATIONS AND FUTURE RESEARCH

In the current study, we proposed a methodological approach applying both machine learning and statistical analyses. The open dataset, such as the credit card fraud detection dataset extracted from Kaggle, can provide large-scale and easily applicable datasets to researchers. However, the results of the classification and prediction of performance can be biased or not generalized to high school students with diverse linguistic, cultural, and academic backgrounds. It should be noted that the follow-up study should verify the currently proposed approach so that it can compare the effectiveness of the suggested approach across high school students from various geographical and cultural backgrounds. Additionally, each school district has a different digital environment and infrastructure available for students, teachers, and families. Fraud detection can be related to or impacted by various factors around personal, school, district, and higher-level dimensions. Understanding these hierarchical characteristics of variables and clustered data is needed when detecting and investigating fraud data and incidences. In future research, researchers should consider these various levels of factors and data in detecting fraud information.

VIII. CONCLUSION

During the pilot study of the course, student comments and survey scores will be recorded. Based on feedback from these surveys and interviews, assignments and readings should be adjusted to better suit the learning characteristics and needs of high school students. The course can be offered in either synchronous or asynchronous online formats, allowing students to easily participate and enabling instructors to act as instructional designers to create a meaningful learning experience in an informal setting. Future studies will be conducted to gather additional data and expand the scope of this research, delving deeper into the teaching and learning experiences of cyber forensics and fraud investigation at the high school level. With the increasing need for valid measurement tools in school environments to employ data-based decision-making, the proposed techniques can provide methodological demonstrations that both researchers and practitioners can use in evaluating learners' behaviors of growth over time throughout the school year related to cyber forensics and fraud investigation. Through the use of open and user-friendly online tools such as Weka, ongoing investigation of approaches to implement cases across students, schools, and districts is needed in future research.

ACKNOWLEDGMENT

This work has been partially supported by the Office of Naval Research (ONR), USA, Award Number N00014-23-1-2454.

REFERENCES

[1] Dixon, P.D., 2005. An overview of computer forensics. IEEE Potentials, 24(5), pp.7-10.
[2] M. K. Rogers and K. Seigfried, "The future of computer forensics: a needs analysis survey," *Computers & Security*, vol. 23, no. 1, pp. 12-16, 2004.
[3] Forensics science; https://en.wikipedia.org/wiki/Forensic_science
[4] Elrick, D., et al. "Design and evaluate a forensic science online course for high school students of color." In Society for Information Technology & Teacher Education International Conference, pp. 678-683. AACE, 2018.
[5] Nelson, B., et al. "Guide to computer forensics and investigations.", Course Technology Cengage Learning, 2010.
[6] An Overview of the Forensic Investigation Process; https://www.exterro.com/
[7] Hall, M., et al. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.
[8] K. Bissadu and G. Hossain, "Designing a High School Course on Machine Learning for Cyberthreat Analytics," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference*, IEEE, 2024.
[9] Cherif, A., et ala. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, *35*(1), 145-174.
[10] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, *17*(3), 235-255.
[11] Raj, S. B. E., et ala. (2011, March). Analysis on credit card fraud detection methods. In *2011 Inte Conference on Computer, Comm and Electrical Technology (ICCCET)* (pp. 152-156). IEEE.
[12] Alpan, K., & İlgi, G. S. (2020, October). Classification of diabetes dataset with data mining techniques by using WEKA approach. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-7). IEEE.
[13] O. Hazzan and Y. Dubinsky, "Teaching a software development methodology: The case of Extreme Programming," in Proc. 16th Int. Conf. on SE Education and Training, Piscataway, NJ: IEEE Publishing, 2003, pp. 176–184.
[14] Dimov, R., et al. "Weka: Practical machine learning tools and techniques with java implementations." AI Tools SeminarUniversity of Saarland, WS 6, no. 07 (2007).
[15] S. P. Mohammed, G. Hossain, and S. Y. Q. Ameen, "Cybersecurity Data Visualization: Designing a Course for Future High School Students," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, 2024.
[16] G. Hossain and D. Yarra, "Learning Blockchain Technology in High School: Towards Cybersecurity Education," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference*, IEEE, 2024.
[17] S. Grover, X et al. "Strengthening early STEM learning by integrating CT into science and math activities at home," in Comp Thinking in PreK-5: Empirical Evid for Integration and Future Directions, pp. 72-84, 2022.
[18] Heritage, M. (2018). Supporting Teachers' Successful Implementation of Formative Assessment. Presentation for Assessment Learning Network, Michigan Assessment Consortium
[19] Bentivegna, M. (2021, January 12). *Precision vs. recall - evaluating model performance in credit card fraud detection*. Medium. Retrieved November 21, 2021
[20] *Credit card transaction fraud continues to climb to new heights*. NCR. (2021, April 5). Retrieved November 21, 2021.
[21] *Equal credit opportunity (regulation B); discrimination on ...* (n.d.). Retrieved November 21, 2021.
[22] Mishra5001. (2019, July 15). *Credit card fraud detection*. Kaggle. Retrieved November 21, 2021.