SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage: https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

"Precognito: The Emergence of Blockchain & Machine Learning-Based Student Record Authentication System

Sheikh Mohammad Tahsin Murtaj[1], Mohammed Mohammed[1], Abubakar Abubakar Yusif[1], Norzariyah Yahya[1,*], Rawad Abdulghafor[2]

[1] Department of Computer Science, Kulliyah of Information and Communication Technology, International Islamic University Malaysia, 53100 Jln Gombak, Malaysia
[2] Faculty of Computer Studies (FCS), Arab Open University–Oman, Muscat P.O. Box 1596, Oman

ABSTRACT

Education is advancing faster than ever before as a result of technological advancements. Two issues brought on by technological advancements are fake transcripts and fake degrees, both of which are mostly due to data security flaws. Despite technological advancement, many domains have not been sufficiently investigated, and there will always be room for improvement. This project aims to create a tool that combines two cutting-edge technologies, machine learning (ML) and blockchain, to combat problems like degree and transcript forgery. By storing student data on the Blockchain and leveraging machine learning techniques for precise analysis, the technology can prevent additional fraud and uncertainty in student achievements and can enable accurate prediction of future employment opportunities for graduates. Machine learning algorithms are used to train and make accurate predictions, and the requisite data are retrieved from a Blockchain ledger. PRECOGNITO will equip the institution with a decentralized and immutable alumni database that contains verified and transparent academic records. Furthermore, this system gives employers a place to check the legitimacy of their employees' academic credentials. Moreover, PRECOGNITO allows students to upload their academic credentials to social media and professional networking sites like LinkedIn. With this system, recruiters may easily locate verified student information.

Keywords:
Blockchain; machine learning; Degree Forgeries; Emergence of technology

1. Introduction

In today's automated society, the education industry is pioneering the adoption of innovative tools, specifically blockchain and machine learning, recognized as transformative forces with applications across sectors. Blockchain and machine learning, as two of the most disruptive 21st-century technologies, remain underutilized in education in which is a crucial sector grappling with challenges like degree fraud, cyber vulnerabilities, and the gap between academic qualifications and employability skills.

* *Corresponding author.*
E-mail address: norzariyah@iium.edu.my

https://doi.org/10.37934/araset.55.1.125140

To tackle these issues, we proposed and developed a system named PRECOGNITO that integrates blockchain and machine learning to create a secure, transparent platform for storing, managing, and analysing student records. Our hypothesis is that this platform can alleviate degree fraud, secure student data, offer precise performance assessments, and connect students with suitable employers. Employing a mixed-methods approach, we aim to test and evaluate the feasibility and effectiveness of our proposed solution.

2. Related Works

Blockchain technology in education has gained attention, especially in establishing a global assessment platform for securely storing and managing degree information as mentioned by Chen *et al.,* [1]. Schwardmann *et al.,* [2] stated that blockchain project inherent data immutability makes it ideal for safeguarding academic credentials with universities actively exploring blockchain to maintain students' records and facilitate sharing potential candidates' details with potential employers as suggested by Shah *et al.,* [3].

Machine learning (ML) is a distinct area within computer science, that utilizes recursive learning processes to glean insights from data [4]. However, ML faces challenges with tainted or unverified data, setting the stage for our proposed framework that integrates blockchain's secure storage with machine learning for precise predictions as stated by Abroyan and Hakobyan [5]. This section delves into systems that is similar to the one we proposed, touching on samples of proposed improvements to the educational system and healthcare. Their varied and ever-expanding use cases are detailed, along with the technologies that enable them (such as blockchain and machine learning).

Degree fraud is an increasing problem in educational institutions, as seen in cases like Chester Ludlow, a pug dog awarded an online master's in business administration (MBA) as reported by GetEducated [6], and the tragic consequences of a North Carolina "doctor" with degrees from bogus online institutions presented in Collier [7]. Moreover, a cryptographically signed alternative to paper certificates is proposed in Gräther *et al.,* [8] addressing the need for protecting the certificate registry and using an open digital signature standard to verify global digital certificates.

Blockchain technology addresses student data validity and security, with applications like secure academic certificate authentication as presented by Li and Wu [9], an electronic certificate infrastructure proposed by Gopal and Prakash [10], and healthcare record systems. Cheng *et al*., [11] introduce blockchain technology for secure and efficient healthcare record management, utilizing smart contracts to ensure data consistency and foster patient-physician coordination. The system prioritizes security, privacy, and real-time health data sharing. Operating on a blockchain, the system enables digital access to patient records and secure medication history sharing. Similarly, Han *et al.,* [12] offer a new way for individuals to manage their official transcripts and easily share them. The proposed approach uses blockchain technology, a distributed digital ledger of all cryptographic transactions, for secure and efficient record management beyond the university level.

Figure 1 shows the overall system architecture of the proposed system by Li and Wu [9], which consists of four components: The verification applications, the issuing application, the blockchain, and a local database.

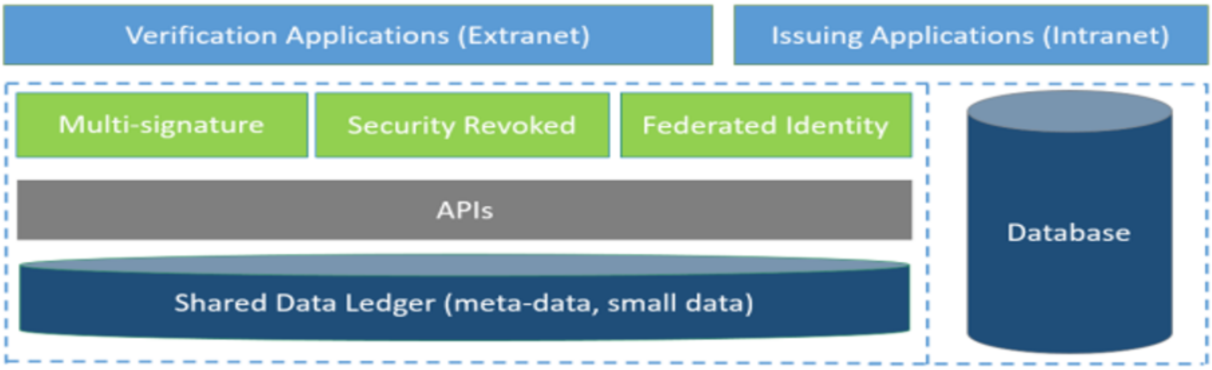

Fig. 1. System architecture overview [9]

Figure 2 illustrates the working process of a blockchain technology is producing e certificates for students in an educational institution. The interaction between the students, the institution, and the company that managing the e-certificates as the proposed system by Cheng *et al.*, [11]. The process started with the students applying for an e-certificate and the process involved the educational institution to review and approving the list of graduates there is a decentralized database blockchain that records the e-certificate to be issued to the students that requested the e-certificates. With the process in Figure 2 trailed the students will get a copy of e-certificates.

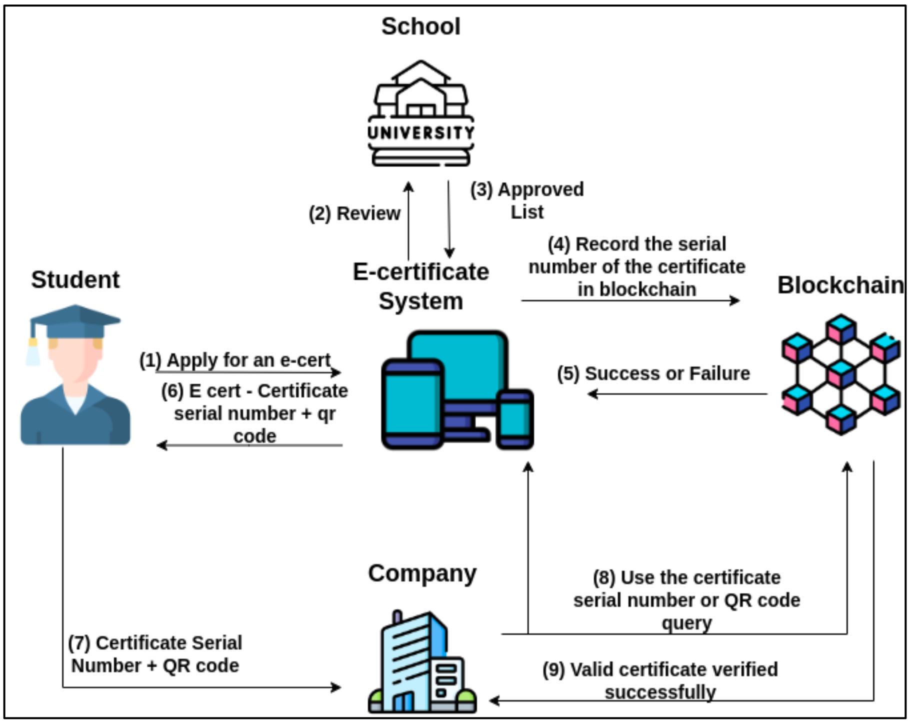

Fig. 2. System overview [11]

Additionally, Bhattacharya *et al.,* [13] explore into the transformative power of blockchain in healthcare, focusing on its impact on health record management, data security, and insurance billing. It presents innovative tools and an Access Control Policy Algorithm, highlighting blockchain's advantages over traditional Electronic Health Record (EHR) systems in terms of efficiency and security. The article introduces a blockchain-based Electronic Health Record sharing system, evaluating its performance and scalability using Hyperledger Fabric and Docker. Findings show that blockchain can enhance data collection, verification, and overall system security, heralding a potential revolution in healthcare systems. Chelladurai and Pandian [14] propose blockchain technology for secure and efficient healthcare record management. It utilizes smart contracts to ensure data consistency and foster patient-physician coordination. The system prioritizes security, privacy, and real-time health data sharing. Another example of using blockchain technology for healthcare data management is MedChain proposed by Shen *et al.,* [15], a system that separates mutable and immutable data to improve data integrity and security. MedChain allows patients to share their data securely with various stakeholders using cryptographic keys.

Machine learning predicts student performance by utilizing robotic process automation (RPA) as stated by Qazdar *et al.,* [16] and aids instructional design suggested by Kotsiantis [17], emphasizing the need for comprehensive ML systems tailored to learning centres in the twenty-first century. Moreover, Siddiqui [18] employs machine learning algorithms to provide foresight into student performance, utilizing data from the Scholar Management System MASSAR. Figure 3 represents the proposed framework for predicting students' performance which starts with collecting and preparing data from the SMS -Massar (School Management System "MASSAR").

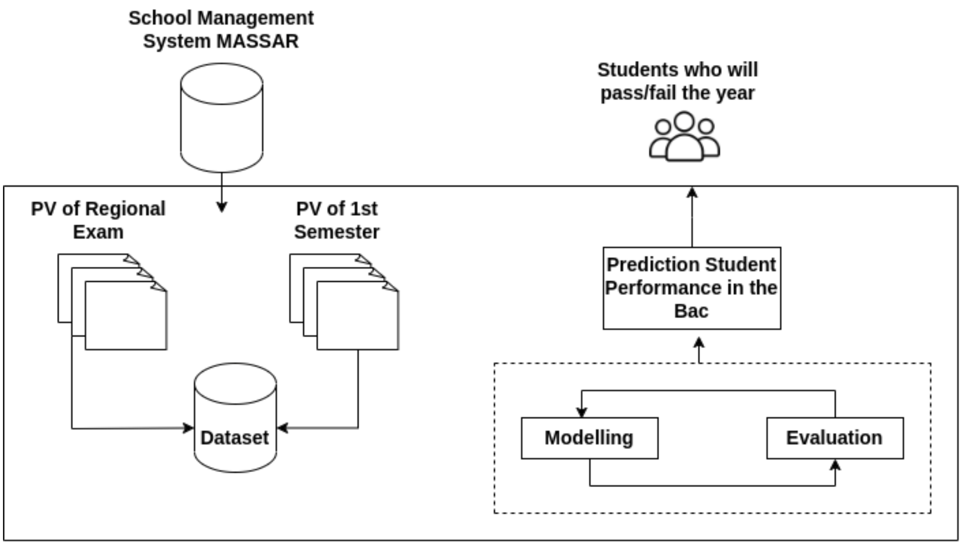

Fig. 3. The proposed framework of MASSAR [18]

While existing literature explores solutions in blockchain and machine learning individually, there is a clear gap for a comprehensive system that integrates both technologies.

Multifaceted challenges in education, including fraudulent degrees, credential forgery, and the lack of a standardized, secure, and decentralized database, underscore the need for a reliable information delivery system. Leveraging the insights from Gräther *et al.,* [8], our proposed solution aims to address these challenges comprehensively. Our project, PRECOGNITO, addresses the global challenge of degree fraud by leveraging blockchain and machine learning, building on the foundation laid by Lu [19] in enhanced data security. The significance lies in enhancing data security, streamlining verification, empowering students with predictive insights, and fostering collaborative stakeholder engagement. PRECOGNITO represents a transformative solution, reshaping educational record management and contributing to transparency, security, and informed decision-making in education and employment verification.

3. Methodology

The proposed PRECOGNITO system aims to steer the educational sector into the Industry Revolution 4.0 (IR 4.0) direction of physical security and blockchain technology by combining both technologies' strengths (Blockchain and Machine Learning). To ensure that the proposed system meets the needs of parties involved in the education institutional management and accessibility to academic records (students, firm employers, colleges, and authorities), the researchers have considered their concerns about privacy, convenience, and availability. Following two sections are discussing the components of PRECOGNITO as illustrated in Figure 4 also the PRECOGNITO flow as demonstrated in Figure 5.

3.1 The Precognito Components

For clarity, the system is depicted hierarchically, with each level including a distinct group of stakeholders and the duties that pertain to them. Figure 4 presents a diagrammatic representation of the hierarchical structure of the proposed system.

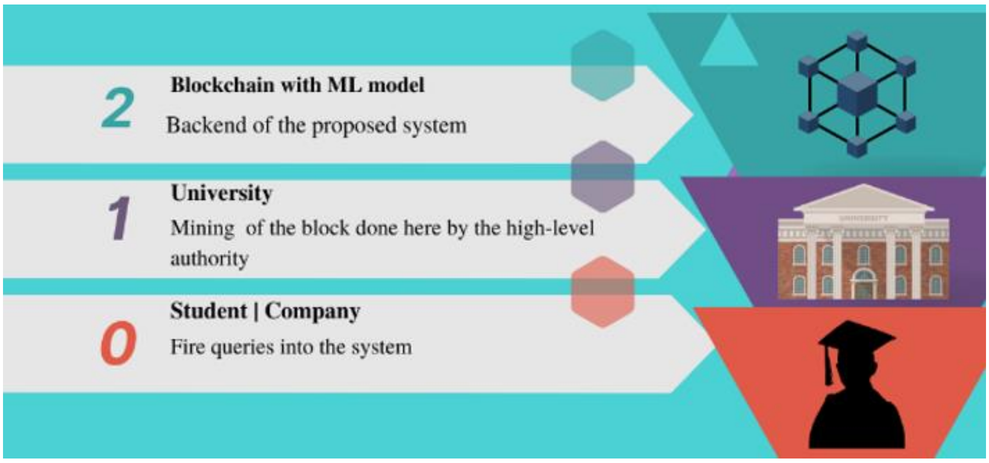

Fig. 4. Level hierarchy of the proposed system, the PRECOGNITO

3.1.1 Student or company

At the foundational level of the framework are students and companies, representing end-users who interact with the PRECOGNITO platform. Students can access their academic records, while companies can query the blockchain for student data, ensuring data accessibility and security. Students play a pivotal role in the PRECOGNITO framework as the primary beneficiaries. Their active participation involves the initial entry of their academic records into the blockchain. By providing accurate and reliable data, students ensure the integrity of their academic achievements. Furthermore, they can utilize the insights generated by the machine learning model to make informed decisions about their career paths. Companies (Employers) constitute another essential stakeholder group in the PRECOGNITO ecosystem. They benefit from the system's transparent and verified academic records, streamlining their hiring processes. Employers have direct access to the blockchain to verify the educational credentials of potential hires, ensuring the authenticity of applicants' qualifications.

3.1.2 University

The university functions as the administrative authority within the framework. It serves as an intermediary entity responsible for maintaining the integrity of the system. This role includes providing the data for the blockchain and verifying student data before it is stored in the blockchain. Additionally, the university oversees the system's smooth operation and resolves potential issues. Moreover, Educational institutions, represented by universities and colleges, serve as the custodians of student records. Their role is vital in maintaining data accuracy and security. Institutions are responsible for verifying and validating student data transactions on the blockchain. They also oversee the smooth operation of the system and address any technical challenges that may arise.

3.1.3 Blockchain with ML model

The highest level of the framework is the decentralized database that combines blockchain and machine learning. This critical component is responsible for processing user-initiated queries, safeguarding data integrity, and facilitating data retrieval. Should this component fail, the entire system's functionality could be compromised. Thus, meticulous care is taken in its management and maintenance.

Integrating Blockchain and ML technologies and the roles and interactions of the many stakeholders are all depicted in Figure 5 as part of the overall system flow. The process flow of this project started with the data collection on students' transcripts and degree awarded details. Then, the collected data will be stored into a blockchain decentralized database, the distributed ledger technology which consists of smart contracts to store the students' transcript details, and the degree awarded to the students. Following are the details of the process as shown in Figure 5.

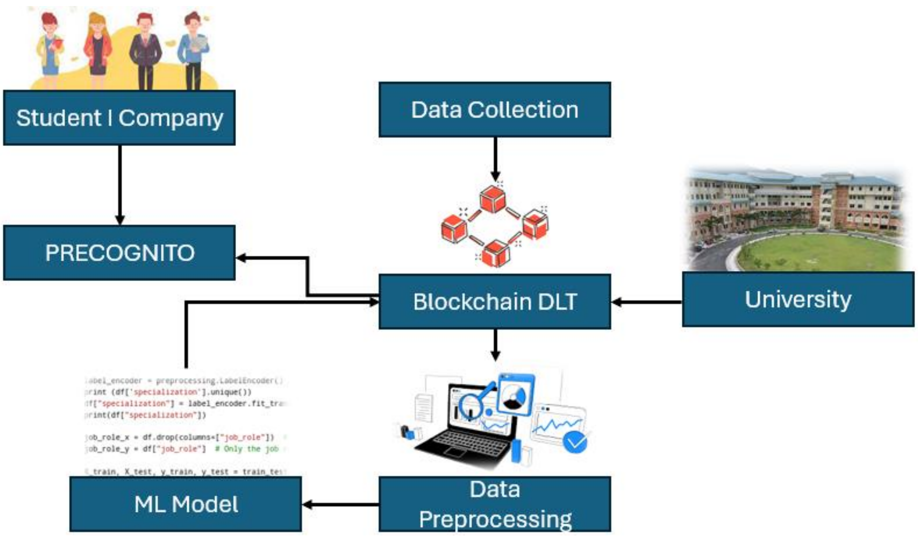

Fig. 5. Flow of the PRECOGNITO proposed system

3.2 Data Collection

The first step to take in any machine learning project is the collection of data that will be used to train the models. In this project, the survey is used as the data collection instrument, and the population sample is students who have already graduated from the International Islamic University Malaysia (IIUM) sample dataset can be found at /dataset on GitHub repo Murtaj and Yusif [24].

3.3 Blockchain

In this project, we use Python as the platform to create the smart contract to store and authenticate the students' and graduates' data. Two important data that are stored are the students' grades of courses taken and the degree awarded to the graduates. As for this blockchain module of this project, all the data are gathered by distributing a survey for students to fill in their grades for courses taken, as well as degrees for those who have graduated. Take note, PRECOGNITO is an initiative of an independent project to establish the emergence of the blockchain and machine learning to securely store and manage degree information to avoid forgery of a degree.

In addition, Python is used as it is not a problem-specific language. Therefore, it may be used to create a wide variety of programs. Besides the main objective of the smart contract which is to store and authenticate the students' and graduates' data, the blockchain modules in PRECOGNITO consist of classes that contain numerous methods, each of which will do a specific action, such as registering a new block, authenticating the chain, adding a new node, adding additional transactions, etc. It is also possible to use the secure hash algorithm 256 to determine a block's hash (SHA256). Code for generating a proof of work, authenticating the chain, and adding a block to the chain is displayed in Figure 6. In this scenario, the transactions are the student records. The process accepts a Python dictionary as input and produces a newly constructed block that can be added to the existing chain.

```
def add_block(self, transactions):
    previous_hash = (self.chain[len(self.chain) - 1]).hash
    new_block = Block(transactions, previous_hash, str(datetime.datetime.now()))
    # calculate nonce
    proof = self.proof_of_work(new_block)
    # new_block.hash = proof
    self.chain.append(new_block)
    return proof, new_block
```

Fig. 6. Code to add a new block to the chain

The type of blockchain used in this project is a private permissioned blockchain based on the Hyperledger fabric framework. This blockchain suits the projects' goal of securely, efficiently and reliably processing and storing student records [20]. It also enables faster and cheaper transactions, as it does not need a resource-intensive consensus algorithm like proof of work according to Seth [21]. Moreover, it supports smart contracts, which automate the systems' logic and rules records [20]. This blockchain has the advantage over other blockchains in security, efficiency, and scalability as stated by Seth [21]. It prevents unauthorized access, data tampering, and attacks suggested by MacDonald [22]. It reduces transaction time, cost, and complexity. It handles more transactions and data without affecting performance and quality [23].

3.4 Data Preparation and Preprocessing

The dataset in Murtaj and Yusif [24] used in this work contains 15 independent features and 1 dependent feature, a categorical value. Three more categorical variables in the dataset must be converted to numeric values by factorization and categorical encoding to apply a specific machine learning technique to the dataset. The future job role of students is the key variable of interest in this dataset. Hence, the student's job is the target variable, dependent on the remaining variables.

The dataset is collected through a Google Form survey of current and former IIUM students. The survey also takes into account information found in a student's permanent records at the university. Therefore, many of the responses were not parallel. To fix the issues, such as unwanted spaces were replaced using regular expressions. Then, rather than removing rows containing duplicate values acting as unique values, those values were replaced with actual values. For example, the "loc" method of the pandas DataFrame is used to find rows in the "JOB ROLE" column where the value is "Network engineer." Once the rows are chosen, the value "Network engineer" in the "JOB ROLE" column is changed to "Network engineer." This is a case correction operation, where all instances of "Network engineer" in the specified column of the DataFrame "df" are changed to "Network Engineer."

3.5 Machine Learning: Model Training, Selection and Evaluation

In this section, we describe the process of model training and the rationale behind the selection of the random forest (RF) algorithm as our primary machine-learning approach. We also provide an overview of the other models considered and their respective performance metrics.

3.5.1 Data splitting

Before delving into the details of model training and selection, it's essential to clarify the division of our dataset into training and testing subsets. To ensure an unbiased evaluation of model performance, we randomly split our dataset into two portions: a training set and a testing set. The training set comprised 70% of the total dataset, allowing our models to learn from a substantial portion of the data. This larger training set size contributed to robust model training and feature learning. The remaining 30% of the dataset was dedicated to the testing set. This separate testing set was utilized to assess the generalization capability of our models, measuring their performance on unseen data.

3.5.2 Model selection

The choice of an appropriate machine learning algorithm plays a pivotal role in achieving accurate and reliable results. After careful consideration and a review of the literature, we opted for the Random Forest (RF) algorithm as our primary modelling approach. RF is a widely acknowledged ensemble learning technique known for its robustness and effectiveness in diverse domains as stated by Breiman [25]. The decision to use RF is supported by numerous studies in the literature that have demonstrated its superiority over other algorithms in classification tasks as presented in Breiman [25].

In particular, RF excels in handling high-dimensional data and is less prone to overfitting compared to single decision trees. It combines the predictions of multiple decision trees, each trained on a different subset of the data, resulting in an ensemble model that generalizes well to unseen data. Additionally, RF provides important insights into feature importance, which can aid in understanding the underlying factors contributing to the classification results as presented in Breiman [25]. The model's performance metrics, including accuracy, precision, recall, and F1-score, showcase the effectiveness of RF in our dataset. With an accuracy of (0.936 – Table 1) and high precision and recall values, RF demonstrates its capability to accurately classify instances, which is essential in our application.

3.5.3 Model evaluation

To evaluate the models, we employed a range of performance metrics, including accuracy, precision, recall, and the F1-score. These metrics provide a comprehensive view of each model's ability to classify instances accurately and deal with imbalanced data, which is often the case in classification problems. Figure 7 illustrates the connection between the two variables as measured by the Pearson correlation coefficient as described by Benesty *et al.,* [26]; Linear dependence measures the degree to which two variables are linked. Its value can be anywhere from -1 to 1, based on the following factors: A -1 correlation indicates a negative linear relationship when comparing two variables. If both values are 0, there is no linear relationship between them. If the correlation coefficient between two variables is 1, it indicates a perfect linear link between them. The distance that a correlation coefficient for a set of variables traverses away from zero indicates the intensity of the relationship between those variables. Furthermore, those who spend more time-solving coding problems on Code force (a platform to operate, organize, and discuss programming contests) are more likely to spend time participating in programming contests, as shown in the figure, which displays a correlation between "P_contest" and "Code_Force_PRACTICE" of 0.83.

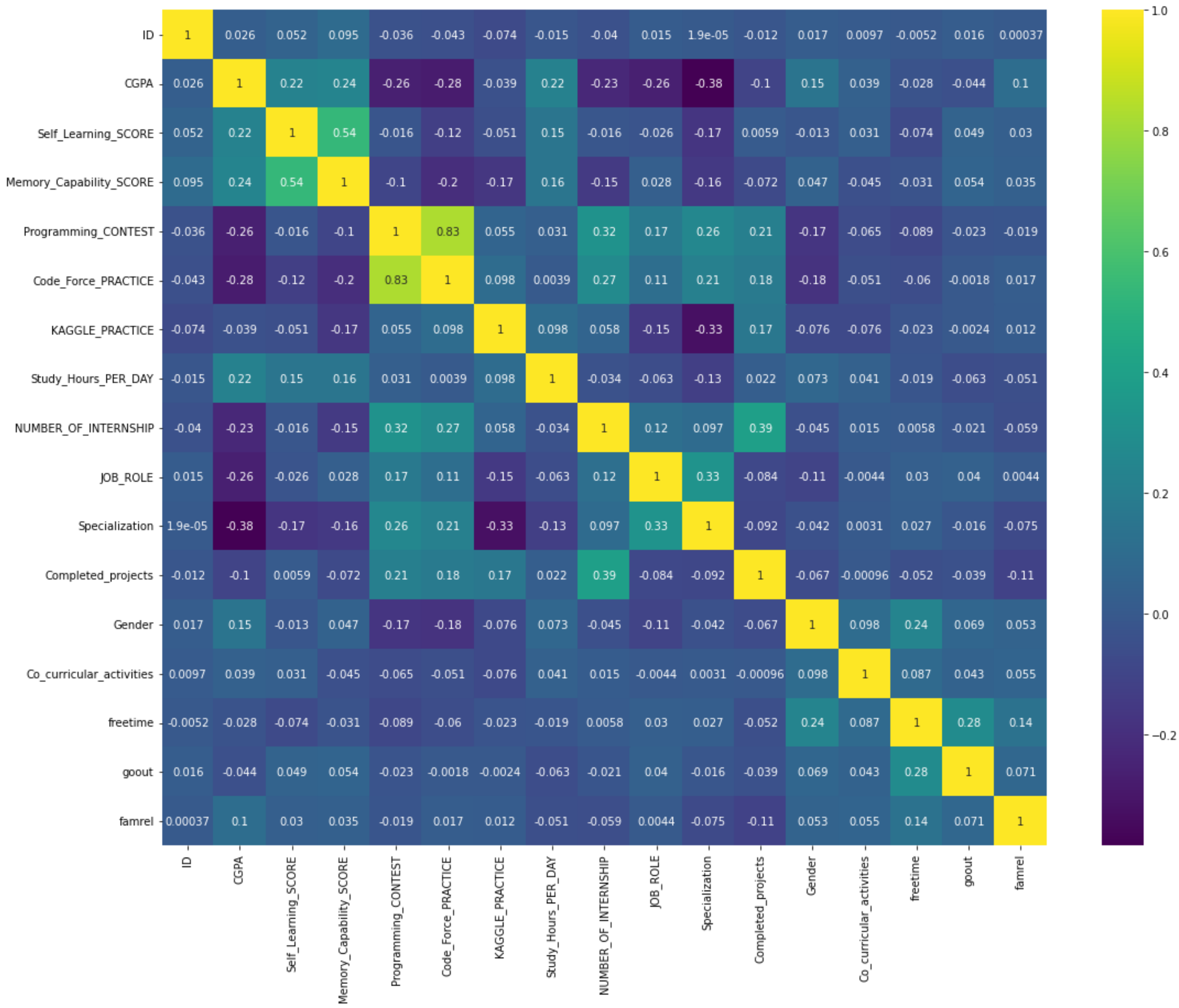

Fig. 7. Correlation matrix of the dataset

Figure 8 represents the ML model's top 10 most important features. "Specialization" feature has the highest importance in the ML model, indicating that a student's area of specialization significantly impacts the target variable "JOB_ROLE."

3.6 Integration of Machine Learning and Blockchain

The fusion of machine learning (ML) and Blockchain technologies through a flask-based Application Programming Interface (API) signifies a pivotal step toward creating innovative, secure, and efficient data-driven applications. This integration bridges two transformative technologies: Blockchain, known for its data integrity and security, and ML, renowned for its data analysis and prediction capabilities. The Flask API serves as a vital conduit for seamless communication between these technologies, the key objectives include using ML to enhance data analysis within the Blockchain ecosystem. This integration enables real-time fraud detection, market trend prediction, and the execution of smart contracts powered by ML models.

In essence, this Flask-based API opens doors to a new era of intelligent, data-driven applications, leveraging the strengths of both ML and Blockchain while maintaining efficiency and security.

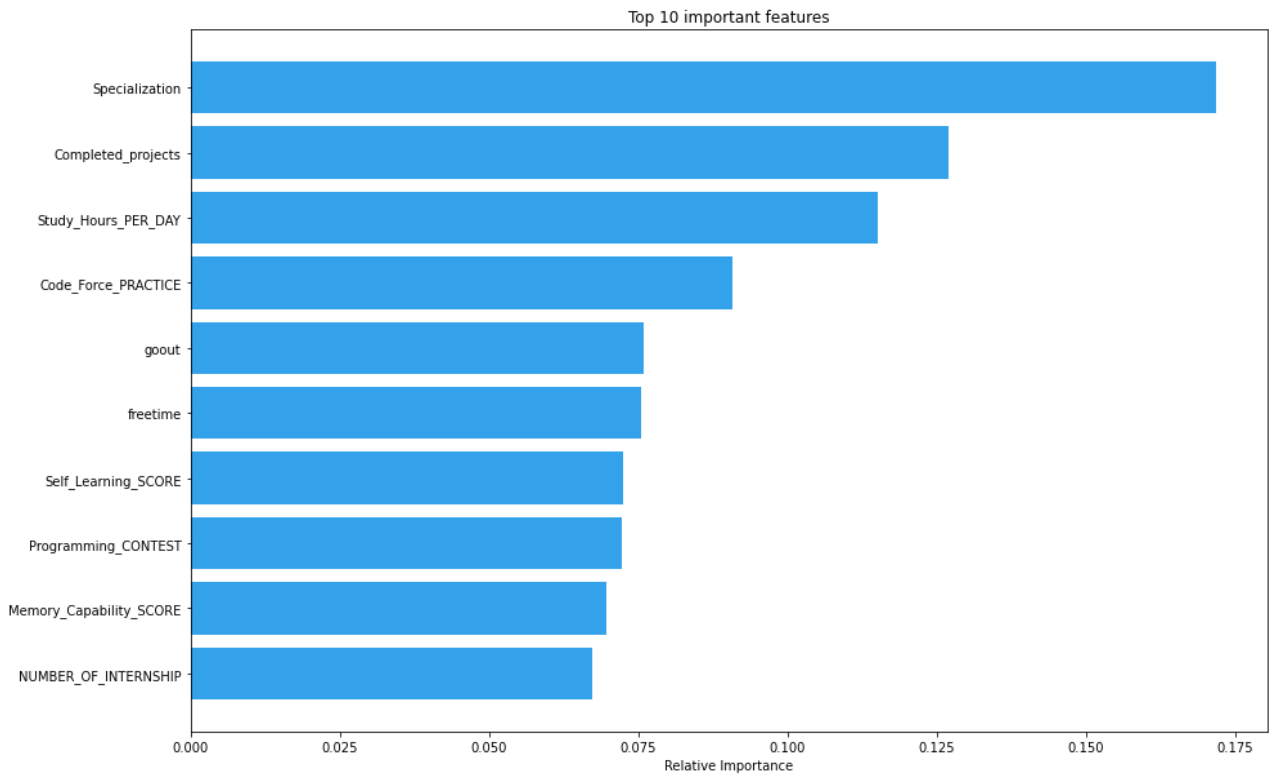

Fig. 8. Top 10 most important features

3.7 Testing

The testing phase would begin after the implementation and involve all functional testing. The goal of the test is to check that the requirements have been met and that the system is functioning correctly. For blockchain-based components, these are the features tested:

i. Functional testing: It is a comprehensive process that assesses the performance of the Blockchain's functional components.

ii. Application Programming Interface (API) testing examines how applications interact in the blockchain environment. It verifies that API requests and responses are appropriately structured and handled.
iii. Performance testing: detects performance bottlenecks, proposes ways to fine-tune the system, and determines whether the programmer is ready to go live.
iv. Node testing: To achieve smooth cooperation, all heterogeneous nodes on the network must be checked independently.

4. Results and Discussion

The PRECOGNITO system stands as a transformative and comprehensive solution that emerges two technologies blockchain and machine learning in a new era in educational record management. Within this section, we delve into the primary discoveries and profound insights uncovered through our rigorous examination of the system's multifaceted components, intricate stakeholder dynamics, and cutting-edge technology implementations. This comprehensive analysis underscores the system's pivotal role in reshaping conventional paradigms of educational record management. By merging blockchain and machine learning technologies, PRECOGNITO introduces a groundbreaking approach to address the longstanding challenges in this domain.

The blockchain is a component of PRECOGNITO that runs the smart contract to store students' and graduates' data. Blockchain is to ensure that the data is immutable and decentralized, and the distributed ledger technology allows the transactions to be transparent however with certain restrictions in access due to the confidentially of the data stores in the blockchain module; thus, in PRECOGNITO the blockchain modules being developed as a permissioned blockchain.

In the development of the machine learning (ML) model, a robust validation process was employed to ensure its reliability and accuracy. We utilized a k-fold cross-validation technique, specifically a 5-fold validation, to assess the model's performance. This method involves dividing the dataset into five subsets, training the model on four subsets and validating on the remaining one in each iteration. The process was repeated five times, and the average performance metrics were computed. The application of k-fold cross-validation enhances the model's robustness by providing a more comprehensive evaluation across different subsets of the data, minimizing the risk of overfitting.

To justify the selection of the Random Forest classifier for our machine learning model, we conducted a comparative analysis with other models, including Stacking Classifier, Logistic Regression, and Gradient Boost. The rationale behind choosing the Random Forest classifier lies in its ability to handle complex relationships in the data, mitigate overfitting, and deliver high accuracy across diverse datasets. This careful consideration of alternative models supports our decision to employ Random Forest as the most suitable choice among the models evaluated. The comparative results demonstrated that the Random Forest classifier consistently outperformed the alternative models in terms of accuracy. This superiority can be attributed to its ensemble learning approach, which aggregates predictions from multiple decision trees, resulting in a more robust and accurate predictive model.

Table 1 summarizes the accuracy, precision, recall, and F1-score of our machine learning models prior to incorporating cross-validation techniques.

Table 1
Accuracy, precision, recall, F1-score of machine learning models

Model Name	Accuracy	Precision	Recall	F1-score
Random Forest	0.936	0.973	0.960	0.960
Stacking Classifier	0.852	0.731	0.776	0.842
Logistic	0.404	0.169	0.190	0.334
Gradient Boost	0.904	0.925	0.948	0.924

We applied Stratified K-Fold Cross-Validation to assess the models' robustness. The results demonstrated a mean accuracy of 0.975 with a low standard deviation of 0.011, indicating consistent and reliable performance across varied data subsets.

Comparing model performance metrics before and after cross-validation revealed a marginal increase in mean accuracy and a reduction in standard deviation. This enhancement signifies improved stability and reliability of our models.

Figure 9 showcases the results of an API request, exemplifying the system's functionality. The /predict route processes a post request, yielding a JSON response predicting 'Web Developer' for the 'id' '1891881'. This amalgamation of blockchain and machine learning technologies establishes a secure and innovative alternative to centralized database systems, emphasizing the significance of authentic data for model training.

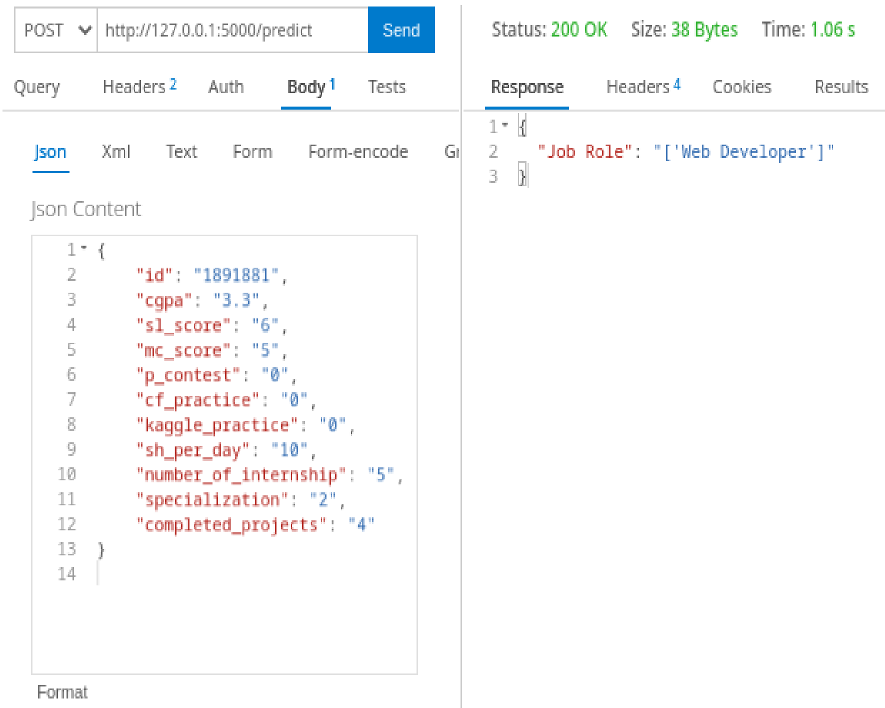

Fig. 9. POST request sent from thunder client to the flask server's predict route

Within the ML testing domain, our analysis utilized a dataset of 1177 students, employing 11 independent features and 1 dependent feature. The Random Forest classifier demonstrated notable accuracy, as detailed in Table 2.

Table 2
ML model testing

Input	Expected result	Status
Student Data	Predicting Jobs	93% accuracy

Table 3 presents the results of our performance testing scenarios, which aimed to evaluate the scalability and efficiency of our system under different workloads and database sizes. We used various parameters, tools, and metrics to simulate and measure the system's behavior in realistic situations. The table shows that our system can handle low to high user activity, sudden bursts of load, peak usage, and extended duration without significant degradation in response time or query time. The table also demonstrates that our system can process large volumes of data efficiently, as the query time does not vary much with the database size. These results indicate that our system is scalable and performant and can meet the users' expectations.

Table 3
Scalability and performance analysis

Scenario	Objective	Parameters	Tools	Metrics
1. Low Workload	Evaluate system behavior under minimal user activity.	Simulated a low number of concurrent users (e.g., 10).	JMeter	• Average Response Time: 20 ms • Minimum Response Time: 15 ms • Maximum Response Time: 25 ms
2. Burst Load	Assess how the system handles sudden spikes in user activity.	Rapidly increased the number of concurrent users (e.g., from 10 to 1000).	LoadRunner	• Average Response Time: 150 ms • Minimum Response Time: 120 ms • Maximum Response Time: 180 ms
3. Peak Usage	Mimic scenarios where the system experiences maximum usage.	Set workload patterns to simulate peak hours (e.g., 1000 users for 1 hour).	Gatling.	• Average Response Time: 180 ms • Minimum Response Time: 160 ms • Maximum Response Time: 200 ms
4. Extended Duration	Measure the system's performance over an extended period.	Run tests for an extended duration (e.g., 24 hours) and monitor for resource exhaustion or degradation.	JMeter	• Average Response Time: 220 ms • Minimum Response Time: 200 ms • Maximum Response Time: 240 ms
5. Database Size Impact	Evaluate the impact-of different database sizes on system performance.	Used datasets of varying sizes (e.g., 10 MB, 100 MB, 1 GB, 10 GB, etc.).	Gatling	• Average Response Time: 30 ms • Minimum Query Time: 25 ms • Maximum Query Time: 35 ms

Incorporating these insights into our academic discourse enriches our understanding of the validation process, model performance, and the system's scalability, contributing to the broader discourse on educational record management systems.

5. Conclusion

PRECOGNITO offers a comprehensive system that emerges blockchain and machine learning to address critical challenges in the education sector. The primary contribution lies in the holistic

PRECOGNITO system, where the emergence of two remarkable technologies sets the stage for a paradigm shift in educational management. By integrating blockchain as a secure foundation and leveraging machine learning for predictive analytics, PRECOGNITO provides a multifaceted solution. Blockchain enhances academic record security, fortifying the credibility of certifications while simplifying the verification process. Simultaneously, machine learning empowers students with predictive insights into future career opportunities based on academic performance.

Central to our vision is the creation of a decentralized student data management system, enabled by blockchain technology, offering efficiency and transparency for all stakeholders not only to students, and academia but also the employers and industries in getting accurate data of graduates. In essence, PRECOGNITO is evidence of the profound integration between blockchain and machine learning, offering tangible benefits to students, educational institutions, and employers alike. It encapsulates a transformative step towards a more secure, data-driven, and future-oriented educational landscape.

Our research contributes significantly to the ongoing discourse on the advancement of educational technologies. As we look ahead, we foresee PRECOGNITO as an indispensable tool in shaping the future of education, instilling trust, and empowering students to make well-informed decisions about their career paths.

However, our project is not without limitations and challenges. We acknowledge that our current model can be improved by employing new algorithms such as SVM, neural networks, ensemble methods (AdaBoost), etc., leading to better results or accuracy. We also recognize that our system requires further testing and evaluation from stakeholders, such as students, educators, and employers, to ensure its usability and effectiveness. Therefore, we plan to build a high-fidelity system prototype and send it out to the stakeholders for feedback and validation. Moreover, we intend to develop a mobile app version of our system, which can provide more convenience and accessibility to the users. These are some of the potential future enhancements and directions for our system development.

Acknowledgments

This research was not funded by any grant.

References

[1] Chen, Guang, Bing Xu, Manli Lu, and Nian-Shing Chen. "Exploring blockchain technology and its potential applications for education." *Smart Learning Environments* 5, no. 1 (2018): 1-10. https://doi.org/10.1186/s40561-017-0050-x

[2] Schwardmann, Ulrich, Christian Boehme, Dora B. Heras, Valeria Cardellini, Emmanuel Jeannot, Antonio Salis, Claudio Schifanella et al. "Correction to: Euro-Par 2019: Parallel Processing Workshops." In *Euro-Par 2019: Parallel Processing Workshops: Euro-Par 2019 International Workshops, Göttingen, Germany, August 26–30, 2019, Revised Selected Papers 25*, pp. C1-C1. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-48340-1_64

[3] Shah, Dhruvil, Devarsh Patel, Jainish Adesara, Pruthvi Hingu, and Manan Shah. "Exploiting the capabilities of blockchain and machine learning in education." *Augmented Human Research* 6 (2021): 1-14. https://doi.org/10.1007/s41133-020-00039-7

[4] UCI. "A Primer on Machine Learning." *UCI Division of Continuing Education* (2024).

[5] Abroyan, Narek Hovhannes, and Robert Grigor Hakobyan. "A review of the usage of machine learning in real-time systems." *Вестник Национального политехнического университета Армении. Информационные технологии, электроника, радиотехника* 1 (2016): 46-54..

[6] GetEducated. "Pug Dog Earns Online MBA in Rochville University Scam." (2020).

[7] Collier, Lorna. "Top 10 Instant Degree Scams." *GetEducated* (2020).

[8] Gräther, Wolfgang, Sabine Kolvenbach, Rudolf Ruland, Julian Schütte, Christof Torres, and Florian Wendland. "Blockchain for education: lifelong learning passport." In *Proceedings of 1st ERCIM Blockchain workshop 2018*. European Society for Socially Embedded Technologies (EUSSET), 2018.
[9] Li, Rujia, and Yifan Wu. "Blockchain based academic certificate authentication system overview." *IT Innov. Centre, Univ. Birmingham* 8 (2018).
[10] Gopal, Neethu, and Vani V. Prakash. "Survey on blockchain based digital certificate system." *International Research Journal of Engineering and Technology (IRJET)* 5, no. 11 (2018).
[11] Cheng, Jiin-Chiou, Narn-Yih Lee, Chien Chi, and Yi-Hua Chen. "Blockchain and smart contract for digital certificate." In *2018 IEEE international conference on applied system invention (ICASI)*, pp. 1046-1051. IEEE, 2018. https://doi.org/10.1109/ICASI.2018.8394455
[12] Han, Meng, Zhigang Li, Jing He, Dalei Wu, Ying Xie, and Asif Baba. "A novel blockchain-based education records verification solution." In *Proceedings of the 19th annual SIG conference on information technology education*, pp. 178-183. 2018. https://doi.org/10.1145/3241815.3241870
[13] Bhattacharya, Pronaya, Sudeep Tanwar, Umesh Bodkhe, Sudhanshu Tyagi, and Neeraj Kumar. "Bindaas: Blockchain-based deep-learning as-a-service in healthcare 4.0 applications." *IEEE transactions on network science and engineering* 8, no. 2 (2019): 1242-1255. https://doi.org/10.1109/TNSE.2019.2961932
[14] Chelladurai, Usharani, and Seethalakshmi Pandian. "A novel blockchain based electronic health record automation system for healthcare." *Journal of Ambient Intelligence and Humanized Computing* 13, no. 1 (2022): 693-703. https://doi.org/10.1007/s12652-021-03163-3
[15] Shen, Bingqing, Jingzhi Guo, and Yilong Yang. "MedChain: Efficient healthcare data sharing via blockchain." *Applied sciences* 9, no. 6 (2019): 1207. https://doi.org/10.3390/app9061207
[16] Qazdar, Aimad, Brahim Er-Raha, Chihab Cherkaoui, and Driss Mammass. "A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco." *Education and Information Technologies* 24 (2019): 3577-3589. https://doi.org/10.1007/s10639-019-09946-8
[17] Kotsiantis, Sotiris B. "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades." *Artificial Intelligence Review* 37 (2012): 331-344. https://doi.org/10.1007/s10462-011-9234-x
[18] Siddiqui, Ahsan Ali. "The Epidemiology of COVID-19 Novel Corona Virus in Incidence and the Distribution of the Disease across the World." (2020). https://doi.org/10.34297/AJBSR.2020.09.001407
[19] Lu, Yang. "Blockchain: A survey on functions, applications and open issues." *Journal of Industrial Integration and Management* 3, no. 04 (2018): 1850015. https://doi.org/10.1142/S242486221850015X
[20] Phemex Academy. n.d. "What Are Public, Private, and Permissioned Blockchains."
[21] Seth, Shobhit. "Public, Private, Permissioned Blockchains Compared." *Investopedia*. 2023.
[22] MacDonald, Ross. "Permissionless vs. Permissioned Blockchains: Pros & Cons." *1Kosmos.* 2023.
[23] Binance Academy. "What Are Permissioned and Permissionless Blockchains?" 2023.
[24] Murtaj, Sheikh Mohammad Tahsin, and Abubakar Abubakar Yusif. "PreCogNito GitHub Repository." *GitHub*. 2022.
[25] Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32. https://doi.org/10.1023/A:1010933404324
[26] Benesty, Jacob, Jingdong Chen, Yiteng Huang, and Israel Cohen. "Noise Reduction in Speech Processing." (2009). https://doi.org/10.1007/978-3-642-00296-0_5