IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 11, Issue 6, November-December 2024

Impact Factor: 7.394

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106045

Fraud Detection in Medical Insurance Claim System Using Machine Learning

Dr. R. Atul Kumar[1], Shaik Toufeeq Umar[2], Shubham Pal[3], Piyush Patla[4]

Assistant Professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[1]

Student, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[2,3,4]

ABSTRACT: Since the beginning of the insurance sector, the issue of false insurance claims has existed. The insurance sector loses billions of dollars every year as a result of these many illegal operations, the most of which go undiscovered. An estimated 600–600 million rupees are lost annually by the insurance sector in India as a result of the country's expanding economy, increased awareness, and improved distribution systems. Every year, false claims cause losses of 800 crores. India ranks 10th in terms of the gross premiums collected by life insurance companies and 15th in terms of the overall revenue generated by non-life. Thus, we are offering a framework for choosing features to be applied in machine learning, allowing for the reliable classification of insurance claims. Preventing financial losses and preserving the integrity of healthcare services depend on the detection of fraud in medical insurance claim systems. In order to efficiently identify fraudulent claims, this study explores the use of Support Vector Machines (SVM) in conjunction with GridSearchCV for hyperparameter optimization. In order to increase model accuracy, the study preprocesses a large dataset of medical insurance claims using rigorous feature selection and engineering. To find the ideal hyperparameters for the SVM model, GridSearchCV is used to conduct an exhaustive search over predetermined parameter ranges. The model's performance is evaluated using metrics including accuracy, precision, recall, and F1-score. The findings show that, in comparison to baseline models, the optimized SVM model greatly improves the detection of false claims.

I. INTRODUCTION

Data analytics is transforming industries, and healthcare is no exception. As one of the largest financial sectors in the U.S., healthcare generates an enormous amount of data—ranging from health records and clinical information to prescriptions, insurance claims, and patient demographics. Health insurance agencies process billions of claims every year, contributing to healthcare expenditures that exceed three trillion dollars annually. Unfortunately, fraudulent claims are a persistent issue, costing the industry billions.

To tackle this problem, we've developed a machine learning framework to detect fraudulent claims efficiently. Our approach uses Support Vector Machines (SVM), a powerful algorithm well-suited for analysing complex and high-dimensional datasets. To further improve accuracy, we leverage GridSearchCV, which systematically fine-tunes the model's parameters for optimal performance. This automated solution not only speeds up claim evaluations but also significantly enhances accuracy, reducing the reliance on manual investigations. By adopting this technology, insurance companies can better combat fraud, minimize financial losses, and boost operational efficiency.

II. LITERATURESURVEY

X. Wu et. al(2022)Infrared objects acquired from a long-distance have small sizes and are easily submerged by a complex and variable background. The existing deep network detection framework suffers greatly from the feature spatial resolution loss caused by the networks' depth and multiple down-sampling operations, which is extremely detrimental for small object detection. So, a crucial and urgent goal is, how to trade-off network depth and feature spatial resolution, while learning feature context representation and interaction to distinguish from the background. To this end, we propose a deep interactive U-Net architecture (short for DI-U-Net) with high feature learning and feature interaction ability. First, feature learning is first achieved through a multi-level and high-resolution network structure. This structure ensures feature resolution as the network depth increase, and also focus on the object's global context information. Then, the feature interactive is further achieved by the dense feature encoder (DFI) module to learn object local context information. The proposed method yields strong object context representation and well discriminability, as well as a good fit for infrared small object detection. Extensive experiments are conducted on the SISRT dataset and Synthetic dataset, demonstrating the superiority and effectiveness of the proposed deeper U-Net compared to previous state-of-the-art detection methods.

I. Matloob et. al(2020)This article presents a novel methodology to detect insurance claim related frauds in the healthcare system using concepts of sequence mining and sequence prediction. Fraud detection in healthcare is a non-trivial task due to the heterogeneous nature of healthcare records. Fraudsters behave as normal patients and with the passage of time keep on changing their way of planting frauds; hence, there is a need to develop fraud detection models. The sequence generation is not the part of previous researches which mostly focus on amount-based analysis or medication versus diseases sequential analysis. The proposed methodology is able to generate sequences of services availed or prescribed by each specialty and analyze via two cascaded checks for the detection of insurance claim related frauds. The methodology addresses these challenges and self learns from historical medical records. It is based on two modules namely ‘‘Sequence rule engine and Prediction based engine’’. The sequence rule engine generates frequent sequences and probabilities of rare sequences for each specialty of the hospital. The comparison of such sequences with the actual patient sequences leads to the identification of anomalies as both sequences are not compliant to the sequences of the rule engine. The system performs further in detail analysis on all non-compliant sequences in the prediction based engine. The proposed methodology is validated by generating patient sequences from last five years transactional data of a local hospital and identifies patterns of service procedures administered to patients using Prefixspan algorithm and Compact prediction tree. Various experiments have been performed to validate the applicability of the developed methodology and the results demonstrate that the methodology is pertinent to detect healthcare frauds and provides on average 85% of accuracy. Thus can help in preventing fraudulent claims and provides better insight into how to improve patient management and treatment procedures

S. Wang et al (2010)To improve the accuracy of diagnosis and the effectiveness of treatment, a framework of parallel healthcare systems (PHSs) based on the artificial systems + computational experiments + parallel execution (ACP) approach is proposed in this paper. PHS uses artificial healthcare systems to model and represent patients’ conditions, diagnosis, and treatment process, then applies computational experiments to analyze and evaluate various therapeutic regimens, and implements parallel execution for decision-making support and real-time optimization in both actual and artificial healthcare processes. In addition, we combine the emerging blockchain technology with PHS, via constructing a consortium blockchain linking patients, hospitals, health bureaus, and healthcare communities for comprehensive healthcare data sharing, medical records review, and care auditability. Finally, a prototype named parallel gout diagnosis and treatment system is built and deployed to verify and demonstrate the effectiveness and efficiency of the blockchain-powered PHS framework.

S. Chakraborty et. al(2019)Blockchain, the technology of the future neutrally facilitated the financial transactions in crypt currencies by strictly eliminating the need for a governing authority or a management that was required to authorize the transactions based on trust and transparency. The Blockchain Network also follows the principle of absolute privacy and anonymity on the identification of the users associated in a transaction. Since the time of its inception, the Blockchain Technology has undergone research that has demonstrated some various kinds of methods to sort out the access control system of the conventional system. In recent years Blockchain has also shown optimum reliability in multiple sectors such as Smart Home, Healthcare, Banking, Information Storage Management, Security and etc. This work in terms is further concerned to the sector of Smart Healthcare, which has grown to a much affluence regarding the efficient technique of serving and dictating medical health care to the patients with the point of maintaining privacy of the patients' data and also the process of laying out real time accurate and trusted data to the medical practitioners. But in the scenario of Smart Healthcare, the primary concern arises in the fact of Privacy and Security of the data of the patients due to the interoperability of multiple stakeholders in the process. Also, there has been a fact of determining accurate and proper data to the doctors if the concerned subject is out of reach from the in hand medical service. Therefore, this Concern of privacy and also mitigation of the accurate data has been very much managed in the work by regulating, a monitoring and sensing paradigm with accordance to the IOT and the Blockchain as a transaction and access management system and also an appropriate medium for laying out accurate and trusted data for serving with deliberate medical care and benefits to the patients across.

N. Dhieb et. al(2019)The private insurance sector is recognized as one of the fastest-growing industries. This rapid growth has fueled incredible transformations over the past decade. Nowadays, there exist insurance products for most high-value assets such as vehicles, jewelry, health/life, and homes. Insurance companies are at the forefront in adopting cutting-edge operations, processes, and mathematical models to maximize profit whilst servicing their customers claims. Traditional methods that are exclusively based on human-in-the-loop models are very time-consuming and inaccurate. In this paper, we develop a secure and automated insurance system framework that reduces human interaction, secures the insurance activities, alerts and informs about risky customers, detects fraudulent claims, and reduces monetary loss for the insurance sector. After presenting the blockchain-based framework to enable secure transactions and data sharing among different interacting agents within the insurance network, we propose to employ the extreme gradient boosting (XGBoost) machine learning algorithm for the aforementioned insurance services and compare its performances with those of other state-of-the-art algorithms. The obtained results reveal that, when applied to an auto insurance dataset, the XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims. The obtained results reveal that, when applied to an auto insurance dataset, the XGboost achieves high performance gains compared to other existing learning algorithms. For instance, it reaches 7% higher accuracy compared to decision tree models when detecting fraudulent claims. Furthermore, we propose an online learning solution to automatically deal with real-time updates of the insurance network and we show that it outperforms another online state-of-the-art algorithm. Finally, we combine the developed machine.

III. EXISTING SYSTEM

Fraud poses a major challenge to the healthcare system, leading to significant financial and operational losses. To tackle this issue, advanced machine learning methods like Long Short-Term Memory (LSTM) networks have been utilized. LSTMs, a type of recurrent neural network (RNN), are particularly effective at analyzing sequential data, making them well-suited for identifying fraudulent activities in medical insurance claims where the sequence of events is critical. This system processes historical claim data by organizing it into sequences that reflect event order and trains the LSTM model to recognize patterns unique to both legitimate and fraudulent claims. Feature engineering further enhances the model's ability to differentiate between normal and suspicious activities, improving fraud detection accuracy.

DISADVANTAGES

- **High Computational Demands:** LSTM networks require substantial processing power, resulting in extended training times, especially for large datasets.
- **Scalability Challenges:** The system struggles to maintain efficiency as dataset sizes grow, leading to slower performance with increased data volume.
- **Intensive Memory Requirements:** LSTMs consume significant memory, necessitating high-capacity hardware, which can drive up infrastructure costs.
- **Risk of Overfitting:** The complexity of LSTMs increases the likelihood of overfitting, which limits their ability to detect new or unseen fraud patterns.
- **Difficulty with Non-linear Relationships:** LSTMs may fail to capture complex non-linear patterns that are not temporal, impacting detection accuracy.
- **Complex Hyperparameter Optimization:** Tuning LSTM hyperparameters is a time-consuming process that requires expertise and extensive experimentation.
- **Latency in Real-Time Detection:** LSTMs process data sequentially, which can lead to delays and limit their effectiveness in real-time fraud detection systems.
- **Lack of Interpretability:** As "black-box" models, LSTMs offer limited insight into why a claim is flagged, making it harder to build user trust in the system.

IV. PROPOSED SYSTEM

The proposed system improves fraud detection in medical insurance claims by integrating Support Vector Machines (SVM) with GridSearchCV to achieve optimal performance. SVM's capability to manage high-dimensional data, combined with GridSearchCV's parameter tuning, ensures efficient preprocessing, model development, and seamless integration into existing workflows. Through meticulous feature engineering, the system accurately differentiates between valid and fraudulent claims. Performance is validated using metrics like accuracy, precision, and F1-score, with cross-validation ensuring the model's reliability. A user-friendly interface facilitates real-time claim reviews, offering detailed analysis and visualization tools to support informed decision-making.

ADVANTAGES

- **High Accuracy and Reliability:** The system combines SVM with GridSearchCV to deliver exceptional accuracy in detecting fraudulent claims. Its robustness against diverse fraud patterns ensures consistent and reliable performance.
- **Effective Non-linear Data Handling:** Leveraging SVM’s strength in managing non-linear relationships, the model can identify complex fraud patterns that deviate from simple or sequential trends.
- **Scalable and Efficient:** Built for scalability, the system handles large datasets effortlessly while maintaining fast processing speeds. This makes it well-suited for real-time fraud detection in high-volume environments.

V. SYSTEM ARCHITECTURE

The System architecture for training and deploying a machine learning model is structured into key stages. It starts with Data Collection, where raw data is divided into Training and Testing Datasets. The Algorithm Layer processes the Training Dataset, enabling the model to identify patterns and relationships. Next, the Evaluation Layer tests the model's performance using the Testing Dataset and metrics like accuracy and precision, allowing for iterative refinements. Once validated, the Trained Model is moved to the Prediction Layer, where it processes new Production Data and generates real-time predictions, such as fraud detection alerts. This architecture provides a systematic and efficient approach to building, testing, and deploying a model capable of delivering accurate, real-world results.

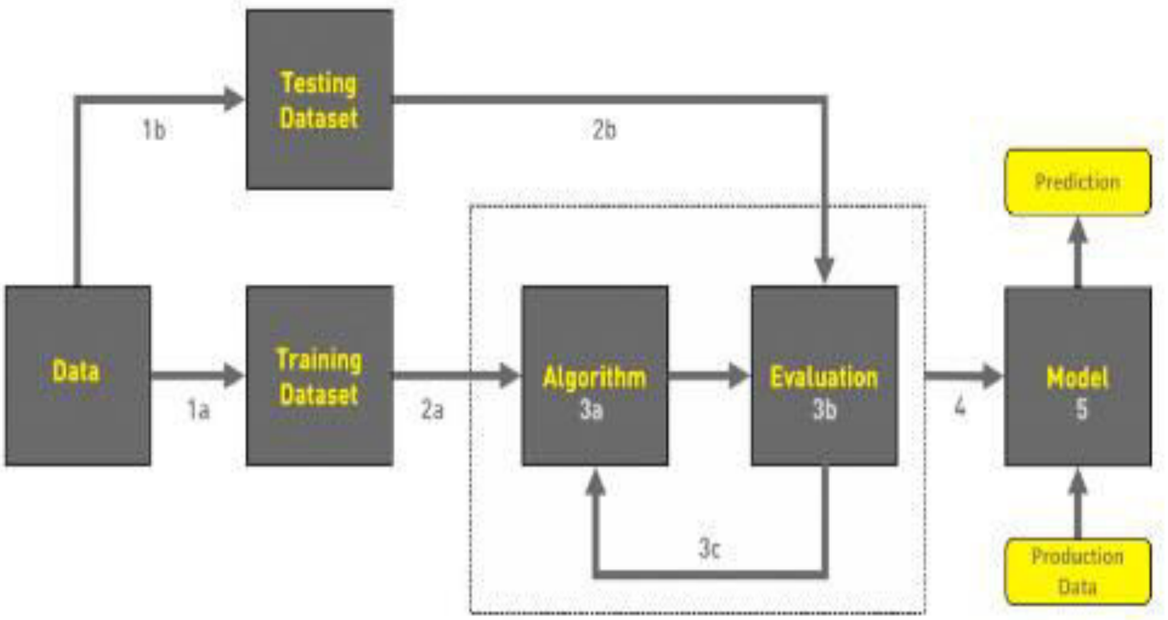

Fig:SYSTEM ARCHITECTURE

VI. METHODOLOGY

The proposed methodology involves a systematic approach to building, evaluating, and deploying a machine learning model for fraud detection in medical insurance claims. It is designed to ensure accuracy, scalability, and real-world applicability.

MODULES:

1. **Data Collection and Preprocessing:**

Data Sources: Gather data from insurance companies, public datasets, or simulated sources containing details about medical insurance claims.
Data Cleaning: Address missing values, outliers, and inconsistencies to ensure data quality.
Data Transformation: Convert categorical variables into numerical formats using methods like one-hot encoding.

2. **Feature Engineering:**

Feature Selection: Identify relevant features that indicate potential fraud, such as claim amounts, claim frequency, claimant demographics, and medical provider information.
Feature Creation: Generate additional features as needed, like the ratio of claim amount to policy limit or the time interval between consecutive claims.

3. **Model Development:**

SVM Implementation: Build an SVM model using a library like scikit-learn in Python.
Kernel Selection: Experiment with various kernels (linear, polynomial, RBF) to identify the one that works best for the data.

4. **Hyperparameter Tuning with GridSearchCV:**

Parameter Grid: Define a range of hyperparameters to test, including C (regularization), gamma (kernel coefficient), and kernel type.

Cross-Validation: Use GridSearchCV to perform cross-validation and identify the optimal hyperparameter values.

5. **Model Evaluation:**

Evaluation Metrics: Evaluate model performance using accuracy, precision, recall, F1-score, and AUC-ROC.

Confusion Matrix: Examine the confusion matrix to assess true positives, false positives, true negatives, and false negatives.

6. **Implementation and Deployment:**

User Interface: Create a simple and intuitive interface to integrate the model into the insurance adjuster's workflow, allowing seamless access to predictions and insights.

VII. IMPLEMENTATION

ALGORITHM:-

1. Initialize the Flask app and set a secret key for session management.
2. Import necessary libraries, including Flask and pickle, for building the web application.
3. Load the pre-trained machine learning model using pickle. Handle exceptions to ensure the application remains functional even if the model fails to load.
4. Configure the upload folder and define allowed file extensions for handling user-uploaded files.
5. Define a route for the home page (/) that renders the index.html template.
6. Implement a login functionality with the /login route:

- For a POST request, validate the username and password from a predefined dictionary of users. If valid, log the user in by setting a session variable and redirect to the services page. If invalid, display an error message on the login form.
- For a GET request, render the login form.

7. Set up a user registration system with the /register route:

- For a POST request, collect user details from the form, check for existing usernames, and validate passwords. Add valid user entries to the dictionary and redirect to the login page. If validation fails, display appropriate error messages.
- For a GET request, render the registration form.

8. Define static routes (/about, /blog, /client, /contact) to render corresponding HTML pages.
9. Implement the fraud detection service on the /services route:

- Ensure the user is authenticated by checking session variables. If not authenticated, redirect to the login page.
- For a POST request:
- Extract input features from the form, validate the data, and prepare it for model prediction.
- Pass the features to the pre-loaded machine learning model and generate a prediction. Interpret the prediction result as either "Fraud" or "No Fraud" and display it to the user.
- If a file is uploaded, validate its extension, save it to the server, and notify the user of the upload status.
- For a GET request, render the services.html form to collect inputs.

10. Provide a logout mechanism using the /logout route to clear the session and redirect to the home page.
11. Add error handling to manage model-loading issues and exceptions during form processing or prediction. Display user-friendly error messages when needed.
12. Start the Flask application in debug mode to enable detailed error logs during development.

VIII. EXPERIMENTAL OUTCOME

Objective:

This report evaluates the performance of the model trained on a given dataset, where the test set constitutes 30% of the total dataset.

Model Metrics:

• **Test Set Split:** 30% of the total dataset was allocated for testing the model, ensuring a sufficient amount of data was used for evaluating its performance.

• **Accuracy Score:** The model achieved an accuracy score of 0.9872 (98.72%). This indicates that approximately 98.72% of the model's predictions were correct, suggesting a highly accurate model that performs well on unseen data.

- **F1 Score:** The F1 score of 0.9874 (98.74%) is a measure of the model's ability to balance precision and recall. This score is particularly useful when dealing with imbalanced datasets, as it takes both false positives and false negatives into account. The high F1 score further emphasizes the model's effectiveness in classifying both the positive and negative classes with minimal errors.

CONFUSION MATRIX:
The confusion matrix for the model on the test set is as follows:
[[80706 2121]
[349 82484]]
Where:

- True Positives (TP): 82,484
- True Negatives (TN): 80,706
- False Positives (FP): 2,121
- False Negatives (FN): 349

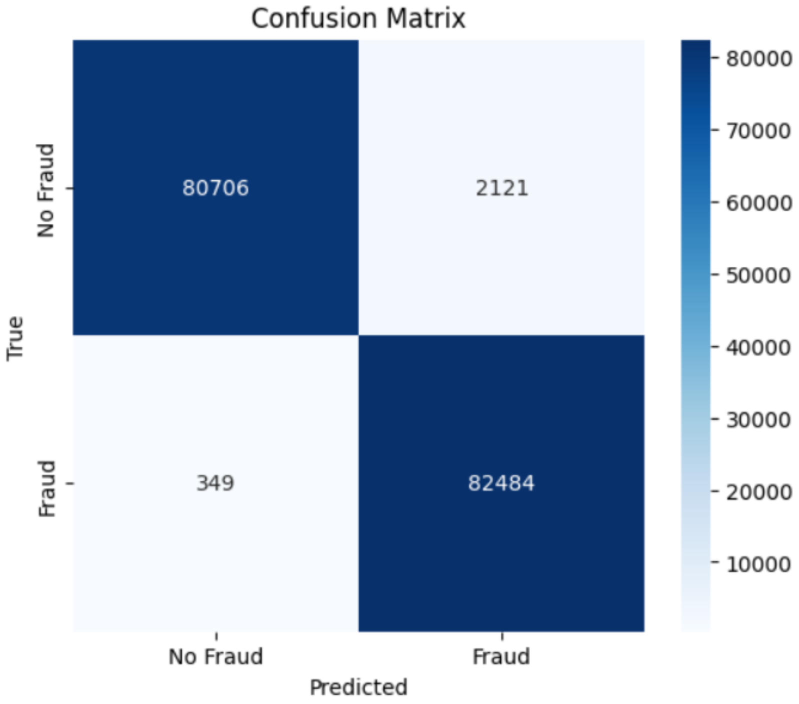

Fig: Confusion Matrix Graph

IX. CONCLUSION

To conclude, combating healthcare insurance fraud is vital for protecting financial resources and ensuring patient care. The current use of Support Vector Machines (SVM) along with Grid Search Cross-Validation (CV) lays a strong foundation for identifying fraudulent activities. Looking forward, incorporating more advanced machine learning techniques has the potential to create a more sophisticated and effective fraud detection system. By refining a detailed classification of fraud patterns and incorporating additional machine learning methods, this enhanced system aims to boost detection accuracy, improve security, and increase efficiency. This approach not only minimizes financial losses but also reinforces the trust and integrity of the healthcare sector, ultimately benefiting patients, providers, and insurers alike.

X. FUTURE ENHANCEMENT

Future research will concentrate on integrating blockchain technology with machine learning to provide a more effective and safe insurance fraud detection system. This will entail creating a thorough taxonomy of the various forms of fraud, taking into account situations such as upcoding, duplicate claims, and charging for services that were never rendered. Additionally, we will improve feature engineering, apply anomaly detection techniques, and investigate more machine learning algorithms. Smart contracts will enable automated claim verification while upholding strict data privacy and security standards, utilizing blockchain's transparency and immutability.

REFERENCES

1. N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, “A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement,” IEEE Access, vol. 8, pp. 58546–58558, 2020, doi: 10.1109/ACCESS.2020.2983300.
2. S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P. K. Singh, and W. C. Hong, “Machine Learning Adoption in Blockchain-Based Smart Applications: The Challenges, and a Way Forward,” IEEE Access, vol. 8, pp. 474–448, 2020, doi: 10.1109/ACCESS.2019.2961372.
3. M. Bärtl and S. Krummaker, “Prediction of claims in export credit finance: a comparison of four machine learning techniques,” Risks, vol. 8, no. 1, 2020, doi: 10.3390/risks8010022.
4. L. Ismail and S. Zeadally, “Healthcare Insurance Frauds: Taxonomy and Blockchain-Based Detection Framework (Block-HI),” IT Prof., vol. 23, no. 4, pp. 36–43, 2021, doi: 10.1109/MITP.2021.3071534.
5. I. Matloob, S. A. Khan, and H. U. Rahman, “Sequence Mining and Prediction-Based Healthcare Fraud Detection Methodology,” IEEE Access, vol. 8, pp. 143256–143273, 2020, doi: 10.1109/ACCESS.2020.3013962.
6. G. Kowshalya and M. Nandhini, “Predicting Fraudulent Claims in Automobile Insurance,” Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018, no. Icicct, pp. 1338–1343, 2018, doi: 10.1109/ICICCT.2018.8473034.
7. S. Wang et al., “Blockchain-Powered Parallel Healthcare Systems Based on the ACP Approach,” IEEE Trans. Comput. Soc. Syst., vol. 5, no. 4, pp. 942–950, 2018, doi: 10.1109/TCSS.2018.2865526.
8. S. Chakraborty, S. Aich, and H. C. Kim, “A Secure Healthcare System Design Framework using Blockchain Technology,” Int. Conf. Adv. Commun. Technol. ICACT, vol. 2019-February, pp. 260–264, 2019, doi: 10.23919/ICACT.2019.8701983.
9. T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi, and J. Wang, “Untangling Blockchain: A Data Processing View of Blockchain Systems,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 7, pp. 1366–1385, 2018, doi: 10.1109/TKDE.2017.2781227. 55
10. W. Kozlow, M. J. Demeure, L. M. Welniak, and J. L. Shaker, “Acute extracapsular parathyroid hemorrhage: Case report and review of the literature,” Endocr. Pract., vol. 7, no. 1, pp. 32 36, 2001, doi: 10.4158/ep.7.1.32.
11. M. Raikwar, S. Mazumdar, S. Ruj, S. Sen Gupta, A. Chattopadhyay, and K. Lam, “2018 9th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2018 - Proceedings,” 2018 9th IFIP Int. Conf. New Technol. Mobil. Secur. NTMS 2018 - Proc., vol. 2018-January, 2018.
12. R. Roy and K. T. George, “Detecting insurance claims fraud using machine learning techniques,” Proc. IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT 2017, 2017, doi: 10.1109/ICCPCT.2017.8074258.
13. X. Liang, J. Zhao, S. Shetty, J. Liu, and D. Li, “Integrating blockchain for data sharing and collaboration in mobile healthcare applications,” IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC, vol. 2017-October, pp. 1–5, 2018, doi: 10.1109/PIMRC.2017.8292361.
14. F. Tang, S. Ma, Y. Xiang, and C. Lin, “An Efficient Authentication Scheme for Blockchain Based Electronic Health Records,” IEEE Access, vol. 7, pp. 41678–41689, 10.1109/ACCESS.2019.2904300