www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

Prediction of Silica Impurity Using Deep Learning Techniques for Mining Environment

Sanda Sri Harsha[1]; K. Venkata Prasad[2]; Katragadda Raghuveer[3]

[1,2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur (Dist), India.

[1]sriharsha.sanda@gmail.com

[2]prasad_kz@yahoo.co.in

[3]Department of Business Management, VR Siddhartha Engineering College, Vijayawada, India.

[3]katragadda.raghu@gmail.com

Abstract

This research paper proposes the percentage of Silica is measured in a lab experiment it takes at least one hour for the process engineers to have this value. As this impurity is measured every hour and it takes a lot of time for a day and causes delay in the mining process. The environment is polluting while reducing the number of ore that goes to tailings as you reduce silica in the ore concentrate. The overall goal is to predict impurity in the ore concentrate in mining process. In this case impurity is specifically Silica concentrate. Silica concentrate is a measured variable but takes time to report results, thus reducing efficiency in the mining process. Being able to predict the silica content without stopping to test is the extended goal of this project. This appears to be a continuous batch process, where raw material is fed into a flotation system, processed, removed, and the process repeated. The purpose is to evaluate the feasibility of using machine learning algorithms like Multiple Linear Regression, Random Forest and Decision tree to predict in real-time. And also, by using Deep Learning techniques like LSTM, we can predict the silica impurity in the ore in less time and help the engineers for early prediction and reduce the impurities. We also developed a web application to display the prediction. The web application is built by using flask framework and it is integrated with trained ML model and it help the engineers, giving them early information to take actions (empowering!).Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and also help the environment.

Keywords: Mining Process, Machine Learning, Deep Learning, LSTM.

1. Introduction

Data storage in various format such as records, files, documents, sound, movies, science and a lot of new information forms has been driven by the emergence of information technology in numerous domains for better decision-making, the data generated from diverse applications require an appropriate way of extracting information from big repositories. The purpose is to discover usable information from the wide collection of data in databases (KDD), often dubbed data extraction (Data Mining). In order to find and extract patterns of recorded data, the basic features of data mining are numerous approaches and algorithms. Data mining and application for knowledge discovery have become an important component in many organizations, as they play an important part in decision making. In the new sectors of statistics, databases, machine learning, model reorganization and artificial intelligence, computer capabilities etc., data mining technology has been incorporated.

In several industries, a model for quality prediction was established for the production of faultless products. However, in single-stage production, most quality prediction model is established. Previous research demonstrates that one-stage quality system in multi-stage production cannot effectively tackle the quality problem.

Linear regression is a statistical study that determines how a relationship between two types of variables is modeled, dependent and independent (predictor). Regression has major goal to investigate if independent factors have predicted the result variable well and which independent variables are important predictors of the result.

There has recently been increasing interest in investigating primordial factors for iron ore recovery at a froth flotation treatment plant. The Financial Times shows that ore iron is more of a raw material for steel manufacture than any other commodity, with the exception perhaps of land, in the world economy. Every tonne of iron ores produced has been shown to discharge around 2.5-3.0 tonnes. In addition, figures indicate that about 130 million tonnes of iron ore are produced annually. This suggests that if the mining reservoirs contain, for instance, an average of around 12% iron ore, over 1.52 million tonnes of iron would be waste per year. In the brotherhood of iron mining, stakeholders rely on standard laboratory testing techniques, which generally take more than two hours to attain their target quality in the froth flotation processing facility. Since environmental protection is highly important and good iron grade is needed which is mined from ore. Then we may forecast a single dependent variable with two or three separate versions by applying machine learning methods such as Multiple Linear Regression, and we can employ Random Forest Decision Trees further. Deep

learning techniques such as LSTM's neural network are utilized to predict silica impurity in mining process, which is well known for its time series prediction applications.

1.1 Data Mining

Data mining is used to use massive amounts of data to find hidden patterns and associations that are useful in decision-making. Alternatively, exploratory analysis, discovery by data and inferior learning were called. In this standard access, data mining access to a database differs in a number of ways: query, data and output. A data mining algorithm is a well-defined technique in which data is taken as input and generated as models or patterns. The phrase well-defined shows that it is precisely possible to encode the operation as a certain number of rules. In order to characterize the whole (most of the) data set, the structures found during the data mining process are called "models." There are also occasions where the identified structures have some local data characteristics and the term pattern is applied in this case.

1.2 Considerations for Applying Data Mining

In order to construct an effective data mining solution, the user has to investigate and articulate his objective. The problem objective leads the user to the correct learning algorithm. When hidden groupings in data may be detected or a connection between key data variables is established, users want to find information and select a technique for clustering or the association mining. Alternatively, a predictive model may be created that may divide samples into a category such as low air quality or a real world result such as an aviation quality score. There are a big, rising number of algorithms inside the prediction paradigm and the knowledge finding paradigm. The choice between the methods of any paradigm is a problem of his own accord. Domingo’s highlights some of the important issues in helping new practitioners in implementing algorithms for machine learning. When making this decision, the User should consider the intricacy and quantity of data presented. For example, a basic linear classifier is not suitable for a sophisticated, non-linear classification task. However it requires the employment of advanced study methods, such as deep artificial neural networks, in consideration of considerations regarding the storage, memory and durations of training.

1.2.1 Predicting Quality

Quality prediction involves the development of models in which quality input features are related to quality inputs and the use of models in order to forecast what the resulting quality property value will be of a collection of input parameters. For predication, regression approach can be adjusted. The regression analysis can be employed for modeling a link between one or more separate and dependent variables. Individual variables are already known attributes for data mining, and the answer variables are to be anticipated. Sadly, not just predictions are numerous real-world difficulties. Therefore, the prediction of future values can require more advanced algorithms (e.g. regression of logistics, decision trees or neural nets). For regression and classification, the same model types can often be utilized.

1.3 Neural Networks

The neural grid is a set of interconnected I/O modules with a connecting weight. The weight change of the network will enable the appropriate input to be anticipated during the learning phase. Neural networks can be used to uncover patterns and to find trends that human or computer technology can recognize far too difficult to draw significance from concatenated or inaccurate data. They are perfect for inputs and outputs that are valued continuously. Neural networks are ideally suited for prediction or forecasting requirements when determining data patterns or trends.

2. Literature Review

Marco Canaparo et. all (2019) In this study, the data mining strategies were initially comparable as far as software fault prognosis was concerned. The author employed existing literature to collect on-line data set, procedures and performance criteria in order to attain this objective. authors paid greater attention to open source and deep learning approaches than earlier studies. Data set linked to open source projects. By analyzing the findings, the author can find the best average accuracy of all data sets achieved by Bagging and Random Forest. Data mining can also serve to determine and predict software quality and can be used in conjunction with statistical analysis. [1]

Brijesh Kumar Baradwaj et. all (2011) This research uses classification tasks to forecast the division of students on the basis of an old database from a student database. Since there are numerous ways for the classification of data, the decision-tab method is being applied. Information like attendance, class testing, seminar and markings have been collected from the previous database of the

student to determine performance at the end of the semester. The students and teachers can raise the division of students through this research. This study will also highlight those students who have to pay special attention to lesser degrees of failure and take appropriate steps for the upcoming semester. [2]

Yunus Kologlu et. all (2012) The successful and invaluable players gather economically in their clubs and generally young talents such as Kylian Mbabpe and Paulo Dybala are worthwhile. 180 and 184 have proven beneficial, assuming that larger players do not have dribble skills and that shorter players lack air ball control. The premier league is also recognized as the most difficult league in Europe as another thumb rule, therefore there is no surprise that English players are more valuable. The fact that card numbers have not affected the player's market value is only an interesting feature in the study, which can be explained by the fact that valued players are more careful to avoid a charge. In general, a more reasonable data collection might be used to improve the study. [3]

Thuraiya Mohd et. all (2020) In this research, a qualitative and quantitative factors (dumb variables) were empirically experimented utilizing the property dataset in the Kuala Lumpur area, Malaysia. The results revealed that statistically significant contributions have been achieved by elements like the main floor area, Green Certification, Tenure and Number of Bedrooms. In other words, all these factors played key roles in the prices of transactions. Main floor area characteristics, green certificate and tenure are related to the transaction price positively. The main floor area has provided the most contribution to the model based on the standardized beta coefficient. [4]

Rajat Chaudhari et. all (2020) After an analysis of a variety of documents, author find that the soil fertility forecast will help to decrease farmers' troubles and to offer farmers with effective information to achieve high yields and hence maximize earnings, thereby reducing suicide rates and reducing their problems. To predict soil fertility, a model is implemented. The system uses supervised and uncontrolled algorithms to learn the machines and provides the highest possible precision results. It compares results from the four algorithms and selects the one that gives the best and most precise output. [5]

Turóczy Zsuzsannaa et. all (2012) The major objective is to enhance the competitiveness, flexibility, adaptability and reactivity of ceramic companies. Since the ceramic sector is an essential part of manufacturing, authors concentrated on this area in order to assess the progress of companies in the sector. The value of the research lies in its novelty and efficacy, as in the instance of a company producing advanced ceramic items, the performance indicators have been analyzed by multiple regression analyses. This analysis is frequently a multivariate and explanatory approach of analysis.

Regression analysis describes the connection between a dependent variable and several distinct factors. [6]

Fahmi Arif et. all (2013) In this study, algorithms of decision tree are utilized to disclose the relationship between factors of product and the final product quality level. The consistent number of values for all product characteristics as the attribute of the decision tree is very low, low, low medium, upper medium, high and very high, all of which are extremely low. It is also possible to summarize that ID3 performs better than C4.5 and CHAID in the case of imbalanced datasets with a uniform number of attribute values, whereas DS, RT and RF fail in the classification of minorities. [7]

E. V. Ramana et. all (2017) Neural net and rule induction models were surpassed by a 95 percent prediction accuracy on the test data set over Naive Bayes model (80 percent). The rule induction model shows that sink marks are created by high temperature of moulding, low injection velocity, dust temperature and injection duration. [8]

T B Chistyakova et. all (2019) A computer data mining approach is described to forecast quality in multi-sortiment, large-scale and multinational polymer films. The following paper provides a library of statistical and data mining methods that enable the testing of normal distribution data, the predicting of the quality of film polymers for various line configurations and film types using statistical tests. The methods include recurring neural networks, a long-term memory neural network and a convolutions network. [9]

Umesh Kumar Pandey et. all (2011) The student database uses the Bayesian classification method to forecast the division of students on the basis of the preceding year record. This research helps students and teachers increase the student's division. The study will also identify pupils who needed additional attention to reduce failures and take suitable action in due course. [10]

Amjad Abu Saa et. all (2016) Multiple data mining tasks were performed in this research article to develop qualitative prediction models that could efficiently and efficiently predict student grades from a dataset of collected training. The first was a survey that targeted and collected a number of personal, social and academic information about university students. Secondly, the obtained data set has been pre-processed and examined so that data mining jobs are suitable. Thirdly, data mining operations were carried out on the dataset in hand to create and test categorization models. [11]

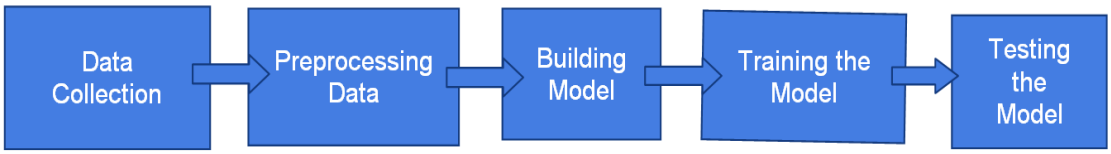
Colin Bellinger et. all (2017) The number of articles reporting on the use of data mining tools to monitor air pollution has increased considerably from our survey. This is because massive data groups and computer power are being made more available and because authors are becoming aware

of the potential advantages of data mining. Despite this tendency and the possible advantages within the sector, a study of the current state-of-the art has not been carried out to the best of our knowledge. [12]

3. Methodology

Gathering the data is the first step to build a model and the process of gathering data depends on the type of project and how we are collected or taken the data. Kaggle is one of the most popular websites where we can download free dataset which is related to the project and we downloaded the dataset that contains 24 features like %iron feed, %silica feed, starch flow, Amina flow, Ore Pulp pH, Ore Pulp Density, Flotation Column 01- 07 Air Flow, Flotation Column 01-07, % Iron Concentrate, % Silica Concentrate and737453 samples which include data and time. Data Preprocessing is done for cleaning the data and removing the out layers which are present in the dataset[13]. In data preprocessing step outliers are removed by IQR score method. After that by using matplotlib we can visualization the data and analyze the data. For building a regression model this are the five steps in the following figure 3.1.

Figure 1 - Building ML Model



Data Collection

Data collection is defined as the process of accurate research insights using standard verified methodologies to be collected, measured and analyzed. On the basis of acquired data, a researcher might evaluate his theory. In most circumstances, data gathering, irrespective of the subject of research, is the main and most significant step in research.[14]

Preprocessing Data: Data Preprocessing is the step of any machine learning process in which the data is modified or encoded to make it so easy to comprehend. In other words, the data's characteristics may now easily be read using the algorithm.[15]

Building Model: Construction of the model Regression analysis involves the development of a probabilistic model, in which the link between the dependent and the independent variables is better described.[16] The multiple linear regression attempts by fitting a linear equation into the observe data to

modeled the association between the two or more explicatory variable and the response variable. Each value of the independent variable x is linked to the dependent variable y value.

$$Y = b0 + b1 * x1 + b2 * x2 + b3 * x3 + \ldots\ldots bn * xn;$$

Dependent variable = Y

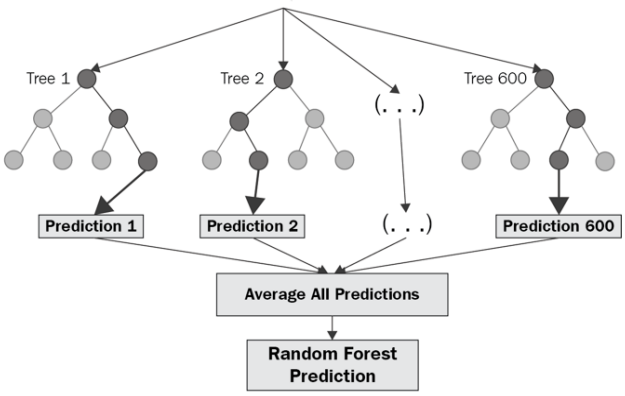
Independent variables=x1,x2,x3,…..xn

For building Multiple Linear Regression model we need to import the data to the operating environment like jupyter notebook. For loading the data sets and to preprocess them we need to import the libraries such as pandas, numpy and matplotlib for visualization.[17] Check if the dataset consists of any missing values and remove the poorly correlated independent variables by using the correlation heatmap. And then we have split the dataset as 80% of the data to train the model and 20% to test the model. After splitting of data, we can train and test the model.

The following step is to analyze model performance when a machine learning model is built and to comprehend the model that is best. Root Mean Square is the measure of how well a regression line fits the data points.[18] And Adjusted R-squared is used to determine the goodness of fit in regression analysis.

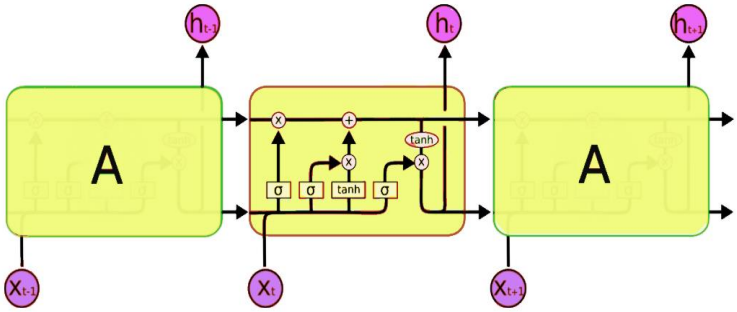
Random Forest is a popular learning system that is frequent in numerous academic publications, competition for the Kaggle and blog posts for categorization purposes. Random forests may also be utilized for regression tasks in addition to classification.[19] The non-linear character of a Random Forest is a fantastic option for it over linear methods. Tree decision also has an excellent regression problem approach and is utilized to forecast silica impurity in an ear in these controlled Machine Learning techniques.[20]

Figure 2 - Random Forest Algorithm



The method of research is based on an examination of regression. This form of analysis is utilized for several variables to be modeled and analyzed. The multiple regression analysis includes a description of the connection between a dependent variable and numerous independent variables.[21]Then, we can broaden this model by employing profound learning approaches and various ways of using predictive analysis systems, which are a sort of recurring neural network capable of learning orders in sequence prediction issues using Long Short term Memory.[22]

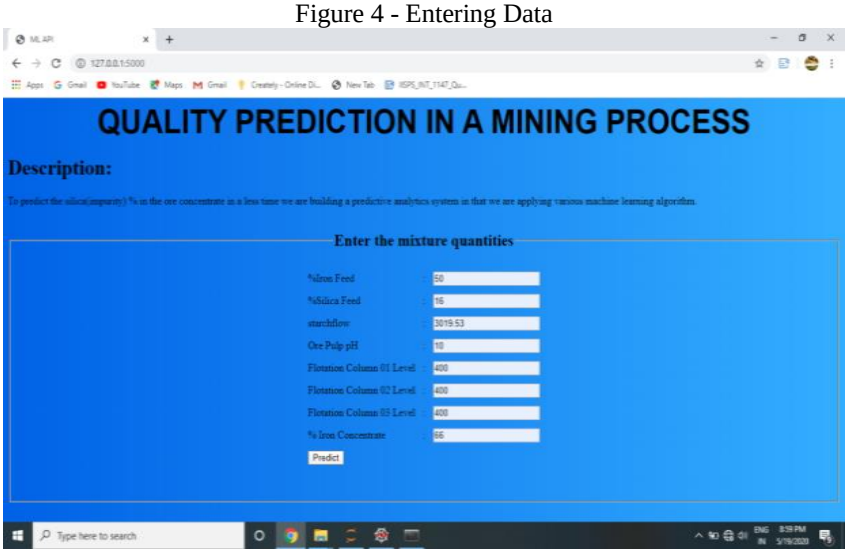
Figure 3 - LSTM Algorithms



4. Results

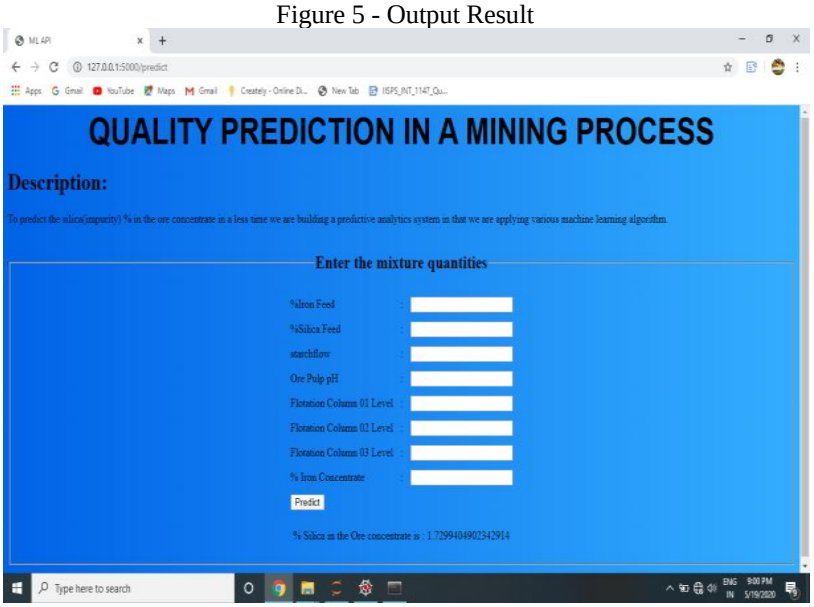
After successful trained and tested the models which are build by different algorithms like Multiple Linear Regression, Random Forest, Decision Tress and all are given good results. The results of Multiple Linear Regression are acceptable and we used Multiple Linear Regression to build flask web application and predict the results as shown in the figure and also, we further implemented a LSTM model which gives better results.

Figure 4 - Entering Data



Received: 22.04.2021 – Accepted: 13.05.2021

Figure 5 - Output Result



5. Conclusion

The findings from this study suggest that machine learning algorithms have the predictive power to predict percentage of silica concentrate in iron ore froth flotation processing plant in real-time as opposed to 2 hours laboratory analysis. However, after the 3 months' observations of iron ore froth flotation processing plant dataset was analyzed, on average, the silica concentrate predictions will be off by 0.38% with a standard deviation of approximately 0.12%, which is significant considering the fact that silica concentrate ranges from 0.77% to 5.53%.This result should be interpreted with caution because the silica concentrate variable used in the analysis was lagged 2 hours and could be further explored with diverse residence time.

However, it is worth noting that each observation in the froth flotation plant can be estimated with respect to silica concentrate as fast as possible. This further connotes that not only the execution time is significant but also the precision of the predictive task.Thus, when both the prediction accuracy and execution time are significant features of an automating the froth flotation plant system, the best option is artificial neural network. This provides in effective predictive in real-time.

6. Future Enhancement

The dataset analyzed in this study was small and we have scope to deal with large datasets by using different techniques in Machine Learning and Deep Learning. And also, we can collect different datasets across the world and deal with them and know the best results to predict different

impurities in ore concentration. On the other hand, we can extend the application of the methodology for different froth flotation processing plants preferably paper mills industry and mineral processing.

References

Marco Canaparo and Elisabetta Ronchieri "Data Mining Techniques for Software Quality Prediction in Open Source Software" *EPJ Web of Conferences* 2019.

Brijesh Kumar Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students‟ Performance" IJACSA 2011.

Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz, *"A Multiple Linear Regression Approach for Estimating the Market Value of Football Players in Forward Position"* Abdullah Gül University Industrial Engineering Department 2012.

Thuraiya Mohd, Syafiqah Jamil and Suraya Masrom "Multiple Linear Regression on Building Price Prediction with Green Building Determinant" *International Journal of Advanced Science and Technology* 2020.

Rajat Chaudhari, Saurabh Chaudhari, Atik Shaikh, Ragini Chiloba, Prof. T.D. Khadtare "Soil Fertility Prediction Using Data Mining Techniques" *International Journal of Future Generation Communication and Networking* 2020.

Turóczy Zsuzsannaa, Liviu Marian "Multiple regression analysis of performance indicators in the ceramic industry" *Emerging Markets Queries in Finance and Business* 2012.

Fahmi Arif, Nanna Suryana, Burairah Hussin "A Data Mining Approach for Developing Quality Prediction Model in Multi-Stage Manufacturing" *International Journal of Computer Applications* 2013.

E. V. Ramana, S. Sapthagiri and P. Srinivas "Data Mining Approach for Quality Prediction and Control of Injection Molding Process" *Indian Journal of Science and Technology* 2017.

T B Chistyakova and M ATeterin "Data mining system for predicting quality of polymeric films" *AMCSM* 2020.

Umesh Kumar Pandey S. Pal "Data Mining: A prediction of performer or underperformer using classification" *IJCSIT* 2011.

Amjad Abu Saa "Educational Data Mining & Students' Performance Prediction" *IJACSA* 2016.

Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane and Alvaro Osornio-Vargas "A systematic review of data mining and machine learning for air pollution epidemiology" *BMC Public Health* 2017.

Harsha, S.S., Simhadri, H., Raghu, K., Prasad, K.V. "Distinctly trained multi-source cnn for multi-camera based vehicle tracking system *"Published in International Journal of Recent Technology and Engineering,* 2019, 8(2), pp. 624–634

Atmakuri, K.C., Prasad, K.V. "A COMPARATIVE STUDY on PREDICTION of INDIAN AIR QUALITY INDEX USING MACHINE LEARNING ALGORITHMS" *Published in Journal of Critical Reviews,* 2020, 7(13), pp. 41–46

Vidya Sagar, P., Harsha, S.S., Prasad, K.V., Moparthi, N.R. “Transferable deep learning assisted radar signal processing model for sea-target detection and classification” *Published in Journal of Green Engineering,* 2020, 10(10), pp. 7661–7671

Pradeep, Ch Nikhil; Rao, M. Kameswara; Vikas, B. Sai “Quantum Cryptography Protocols for IOE Security: A Perspective”, *ADVANCED INFORMATICS FOR COMPUTING RESEARCH, ICAICR* 2019, PT II

Kumar, M. Tanooj; Katragadda, Revanth Kumar; Kolli, Vishnu Srujan; Rahiman, Shaik Lahir, “A Hybrid Approach For Enhancing Security In Internet Of Things (IoT)”, *PROCEEDINGS OF THE 2019 INTERNATIONAL CONFERENCE ON INTELLIGENT SUSTAINABLE SYSTEMS (ICISS 2019).*

Rao, M. Kameswara; Santhi, S. G.; Hussain, Md Ali, “Multi Factor User Authentication Mechanism using Internet of things”, *PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE ON ADVANCED INFORMATICS FOR COMPUTING RESEARCH (ICAICR '19).*

Sahu, Aditya Kumar; Swain, Gandharba, “Data hiding using adaptive LSB and PVD technique resisting PDH and RS analysis”, *INTERNATIONAL JOURNAL OF ELECTRONIC SECURITY AND DIGITAL FORENSICS, 2019.*

Rajesh, L.; Satyanarayana, Penke, “Vulnerability Analysis and Enhancement of Security of Communication Protocol in Industrial Control Systems”, *Published in HELIX in 2019.*

Babukarthik, Raju Govindaraj; Monica, John; Sambasivam, Gnanasekaran; Amudhavel, J. “Intelligent Decision Making System encompassing Security Framework for VM scheduling”, *Published in BIOSCIENCE BIOTECHNOLOGY RESEARCH COMMUNICATIONS 2018.*

Swain, Gandharba “High Capacity Image Steganography Using Modified LSB Substitution and PVD against Pixel Difference Histogram Analysis”, *Published in SECURITY AND COMMUNICATION NETWORKS – 2018.*