ERROR ANALYSIS IN IDENTIFYING FAULT DATA IN IOT DEVICES USING LASSO REGRESSION COMPARED OVER DECISION TREE MODELS

Hirangkar jyoti borah[1], P. V. Pramila[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The aim of this research is to perform an error analysis of fault data detection in IoT devices using lasso regression splines compared over decision tree model.
Materials and Methods: Lasso regression algorithm with sample size = 20 and decision tree algorithm were evaluated to predict the efficiency percentage. Lasso regression prediction updates its weights and configurations based on the input.
Results and Discussion: Lasso regression delivered significant results with 90.40% accuracy, compared to decision tree 85.80% accuracy. Lasso regression and decision tree statistical insignificance is p = 0.511 (p>0.05). Independent sample T-test value states that the results in the study are significantly not achieved with a 95% confidence level.
Conclusion: Lasso regression algorithm performed significantly better than the decision tree algorithm.

Keywords: Novel Lasso regression Algorithm, decision tree Algorithm, fault data Detection, IoT devices, IoT security

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105
[2*]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamil Nadu,India. Pincode:602105

1. Introduction

The decrease in expense of detectors and connection has led to the number of industrial items connected to the internet increased. As a result, Internet of Things (IoT) security has emerged as a critical component of Industry, as it permits the collection of huge volumes of data that are never used due to a lack of understanding or analysis (Teysseire et al. 2022). The best use of this data and real-time analysis can enhance productivity, increase machine health, improve daily production automated, and lead to genetic flaw assembly. Improved measuring equipment, which includes both hardware and technology, are some of the most essential aspects of the smart grid.It allows utilities and end users to communicate in two directions. Because an AMI has similar characteristics to a network connection, strategies used to resist information loss, malicious actions, and monetary gain in communications infrastructure can be used to power grids.(Lamont and Sayigh 2018) The danger of the system infrastructure outweighs the risk of the individual components.The system threat becomes more difficult and complicated to detect as the number of elements vulnerable to assault grows (Ustun 2019).

The most obvious concern is meter manipulation, which takes the form of altering the smart meter reading to give erroneous data to the utility (Vermesan and Friess 2014). This can lead to inflated invoices and inaccurate data for forecasting and monitoring, both of which can have serious consequences. Smart meters generate a vast amount of data that varies in time and pace. Computer vision approaches have the potential to make smart home devices more IoT secured. In today's smart grids, injecting fake data is a prevalent integrity attack that offers a malware concern.(Gunjan and Zurada 2020) Using a perimeter setting technique, an empirical equation to discover false data infusion was given in a data-centric paradigm. Analytical tools that make use of the smart grid's massive volumes of data can assist in detecting integrity attacks like fake data infusion (FDI). Various strategies for detecting outliers in sensor data have been used in the literature.

Our institution is keen on working on latest research trends and has extensive knowledge and research experience which resulted in quality publications (Rinesh et al. 2022; Sundararaman et al. 2022; Mohanavel et al. 2022; Ram et al. 2022; Dinesh Kumar et al. 2022; Vijayalakshmi et al. 2022; Sudhan et al. 2022; J. A. Kumar et al. 2022; Sathish et al. 2022; Mahesh et al. 2022; Yaashikaa et al. 2022). The main challenging problem is the recognition of a wide variety of spam files that are being recieved in IoT devices. Additionally, there are a variety of other factors that create differences in detecting those spam files, such as connected components, multi-oriented files, overlapping files, skewness of text lines, and pressure points among other factors (Pedir 2016)(N. Kumar and Makkar 2020), (Pedir 2016). Even a simple spam can be received differently. Thus, recognizing a particular spam is a challenging task. The aim of this research is to enhance the spam detection in internet of things devices using novel multivariate adaptive regression splines compared to XGBoost.

2. Materials and Methods

This research was carried out in the machine learning laboratory at the Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences. In this study there are two groups. A novel lasso regression is used for Group 1 and a decision tree is used for Group 2. With a sample size of 20, a confidence interval of 95 percent, and pretest power of 85 percent, the lasso regression and decision tree algorithms were examined for various iterations.

The softwares used machine learning models are Windows 11 with Python programming language version 3 and Pycharm IDE. Hardware configuration 8 Gb RAM in the system and the implementation is done in a jupyter notebook.

After dataset collection, the null values and unimportant content in the datasets were removed by preprocessing and data cleaning steps. After cleaning and preprocessing the data, an ideal input for the detection model is produced, which are processed into the detection model using opencv library and efficiency of both lasso regression algorithm and decision tree algorithm is calculated (Yin et al. 2019).

Lasso regression Algorithm

On challenging problems, Novel Lasso Regression is a multi-linear technique. The method requires determining a set of fundamental linear functions that, when integrated, yield the best prediction results. It's thus a type of basic linear function ensemble that excels at difficult regression problems including a large number of input variables and complex non-linear interactions.

Pseudocode for lasso regression algorithm.

1) Start program
2) Input data set
3) Give the path to lasso regression weight and configs and store it in a variable
4) Start with loop to loop frame one by one
5) Use the modules to install the library
6) Implement the dataset
7) Use the lasso regression algorithm to get efficiency values
8) After compare the dataset
9) Run the program
10) If all modules will run then the graph will generate
11) End the program

Decision tree Algorithm

Decision tree is a decentralized support vector toolkit optimized for speed, customization, and adaptability. It creates Machine Learning algorithms using the Vector Support framework. It employs parallel tree boosting to efficiently and precisely solve a wide range of data science problems. Fig**.** 2 shows the algorithm for decision from dataset processing to output generation.

Pseudocode for decision tree algorithm.

1) Start program
2) Import the data set as home appliance.csv
3) Analyze the information stored in images.
4) After loading the dataset, it was splitted into training and testing samples.
5) Preparing the data for training and reshaping it.
6) Imported relevant libraries and modules. Next, we splitted the data into half training and half testing using the function train_test_split.
7) The model is trained and tested to ensure accuracy.
8) Import train_test_split into the coding.
9) Run the program
10) If all modules will run then the graph will generate
11) End the program

S
tatistical Analysis

The control variable is pixel length, and the outcome variable is bounding box color. IBM SPSS is used for statistical analysis. The testing variable is specified as GroupID. GroupID for lasso regression is Group 1 and Group 2 for decision tree model. Group Statistics is used for the Statistical analysis for the home appliances (SPSS) dataset. By performing the statistical analysis of group statistics, By using lasso regression and decision tree model, we analyze the error of identifying fault data in IoT devices.("Two-Group Multivariate Analysis of Variance" 2015)

3. Results

Table 1 represents the simulated efficiency analysis of lasso regression and decision tree algorithms.

Table 2 represents group statistical analysis with the mean value of 90.40 and 85.80 , standard deviation of 1.517 and 3.564 for lasso regression and decision tree respectively.

Table 3 represents the independent T test analysis of both the groups with tailed insignificance values of 0.511 ($p>0.05$). The p-value shows that there is no insignificant difference amongst the group of algorithms used in this analysis.

Fig. 1 shows the architecture for Error analysis of bug detection in IoT security devices using lasso regression compared over decision tree model.

Fig. 2 shows the bar graph analysis based on efficiencies of two algorithms. The mean

efficiencies of novel lasso regression and decision 90.40 and 85.80 respectively. From the results obtained it is inferred that the lasso regression object detection algorithm is more efficient than the decision tree algorithm.

4. Discussion

In this research work, novel lasso regression and decision tree were evaluated for the efficiency of scam analysis. It was discovered that the novel lasso regression approach outperforms the decision tree algorithm after assessing the two models using the identical datasets (National Research Council et al. 2012). It will detect the scam user and the post which is posted by the id. it will analyze and show in graphs like comments, share etc. The datasets from different ranges help to improve the efficiency percentage (National Research Council et al. 1988).

(Bae, Kim, and Kim 2016) used AI networks and they obtained an analysis based on fault data entry. But their network model usually required more time to train data. (Gao et al. 2013) specifically proposed a localized zonal technique to identify fault data, which has been found to significantly improve detection accuracy levels. Based on the home appliances dataset, this method accounts for detection placement by using two independent convolutional IoT networks with test accuracies of 97.2% and 76.8%, respectively. (Zambom and Akritas 2015) presented fault data identification approaches as well as fault data modifiers. It is determined by the fault data and bug combination and has an identification rate of 75% on average. ((Dixit et al. 2020; Cleff 2019)proposed a technique that uses an AI model of decision tree characteristics to identify the fault data.

In future scope, the comparison study should be significant now that it is insignificant. Error analysis is found in lasso regression, but it is minimal when compared to the decision tree model.

5. Conclusion

In this study, the fault data score is used to assess the reliability of IoT devices in a smart home network. The lasso regression algorithm is used to calculate the defect data score for all IoT devices. Lasso regression and decision tree model techniques were used to examine improvements and the time information produced by smart meters during extensive tests and experiments (Panigrahi et al. 2018). The findings suggest that the device failure data score based on decision tree model aids in improving the basis for successful IoT device operation in the smart home.

Declaration

Conflict of Interest

The author declares no conflict of interest.

Authors Contribution

Author HJB was involved in data collection, data analysis, and manuscript writing. Author PVP was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organization for providing financial support that enabled us to complete the research.

1. Codedion technologies Pvt.Ltd.,Chennai, Tamil nadu - 600044
2. Saveetha University.
3. Saveetha Institute of Medical And Technical Sciences.
4. Saveetha School of Engineering.

6. References

Chapter 19: Admission Control-Based Load Protection in the Smart Grid—Security and Privacy in Cyber-Physical Systems. Available online: https://learning.oreilly.com/library/view/security-and-privacy/9781119226048/c19.xhtml (accessed on 30 April 2020).

Smart Meters—Threats and Attacks to PRIME Meters—Tarlogic Security—Cyber Security and Ethical Hacking. Available online: https://www.tarlogic.com/en/blog/smart-meters-threats-and-attacks-to-prime-meters/ (accessed on 5 May 2020).

Makkar, A.; Garg, S.; Kumar, N.; Hossain, M.S.; Ghoneim, A.; Alrashoud, M. An Efficient Spam Detection Technique for IoT Devices using Machine Learning. IEEE Trans. Ind. Inform. 2020. [Google Scholar] [CrossRef]

Choi, J.; Jeoung, H.; Kim, J.; Ko, Y.; Jung, W.; Kim, H.; Kim, J. Detecting and identifying faulty IoT devices in smart homes with context extraction. In Proceedings of the

48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2018, Luxembourg, 25–28 June 2018; pp. 610–621. [Google Scholar] [CrossRef]

Tang, S.; Gu, Z.; Yang, Q.; Fu, S. Smart Home IoT Anomaly Detection based on Ensemble Model Learning from Heterogeneous Data. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 4185–4190. [Google Scholar] [CrossRef]

Wang, Y.; Amin, M.M.; Fu, J.; Moussa, H.B. A novel data analytical approach for false data injection cyber-physical attack mitigation in smart grids. IEEE Access 2017, 5, 26022–26033. [Google Scholar] [CrossRef]

Alagha, A.; Singh, S.; Mizouni, R.; Ouali, A.; Otrok, H. Data-Driven Dynamic Active Node Selection for Event Localization in IoT Applications—A Case Study of Radiation Localization. IEEE Access 2019, 7, 16168–16183. [Google Scholar] [CrossRef]

Mishra, P.; Gudla, S.K.; ShanBhag, A.D.; Bose, J. Enhanced Alternate Action Recommender System Using Recurrent Patterns and Fault Detection System for Smart Home Users. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 651–656. [Google Scholar] [CrossRef]

Bae, Mungyu, Kangho Kim, and Hwangnam Kim. 2016. "Preserving Privacy and Efficiency in Data Communication and Aggregation for AMI Network." *Journal of Network and Computer Applications*. https://doi.org/10.1016/j.jnca.2015.07.005.

Cleff, Thomas. 2019. "Regression Analysis." *Applied Statistics and Multivariate Data Analysis for Business and Economics*. https://doi.org/10.1007/978-3-030-17767-6_10.

Dinesh Kumar, M., V. Godvin Sharmila, Gopalakrishnan Kumar, Jeong-Hoon Park, Siham Yousuf Al-Qaradawi, and J. Rajesh Banu. 2022. "Surfactant Induced Microwave Disintegration for Enhanced Biohydrogen Production from Macroalgae Biomass: Thermodynamics and Energetics." *Bioresource Technology* 350 (April): 126904.

Dixit, Umesh D., Rahul Hiraskar, Raghavendra Purohit, and Sagar Shivanagutti. 2020. "Recognition of Handwritten Word Images." In *2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC)*. IEEE. https://doi.org/10.1109/b-htc50970.2020.9297853.

Gao, Xunbing, G. A. O. Xunbing, M. A. Chunguang, Ping Zhao, and Liang Xiao. 2013. "Fine-Grained Access Control Scheme for Social Network with Transitivity." *Journal of Computer Applications*. https://doi.org/10.3724/sp.j.1087.2013.00008.

Gunjan, Vinit Kumar, and Jacek M. Zurada. 2020. *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2020*. Springer Nature.

Kumar, J. Aravind, J. Aravind Kumar, S. Sathish, T. Krithiga, T. R. Praveenkumar, S. Lokesh, D. Prabu, A. Annam Renita, P. Prakash, and M. Rajasimman. 2022. "A Comprehensive Review on Bio-Hydrogen Production from Brewery Industrial Wastewater and Its Treatment Methodologies." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123594.

Kumar, Neeraj, and Aaisha Makkar. 2020. *Machine Learning in Cognitive IoT*. CRC Press.

Lamont, Lisa, and Ali Sayigh. 2018. *Application of Smart Grid Technologies: Case Studies in Saving Electricity in Different Parts of the World*. Academic Press.

Mahesh, Narayanan, Srinivasan Balakumar, Uthaman Danya, Shanmugasundaram Shyamalagowri, Palanisamy Suresh Babu, Jeyaseelan Aravind, Murugesan Kamaraj, and Muthusamy Govarthanan. 2022. "A Review on Mitigation of Emerging Contaminants in an Aqueous Environment Using Microbial Bio-Machines as Sustainable Tools: Progress and Limitations." *Journal of Water Process Engineering*. https://doi.org/10.1016/j.jwpe.2022.102712.

Mohanavel, Vinayagam, K. Ravi Kumar, T. Sathish, Palanivel Velmurugan, Alagar Karthick, M. Ravichandran, Saleh Alfarraj, Hesham S. Almoallim, Shanmugam Sureshkumar, and J. Isaac JoshuaRamesh Lalvani. 2022. "Investigation on Inorganic Salts K2TiF6 and KBF4 to Develop Nanoparticles Based TiB2 Reinforcement Aluminium Composites." *Bioinorganic Chemistry and Applications* 2022 (January): 8559402.

National Research Council, Division of Behavioral and Social Sciences and Education, Commission on Behavioral and Social Sciences and Education, and Committee on Basic Research in the Behavioral and Social Sciences. 1988. *The Behavioral and Social Sciences: Achievements and Opportunities*. National Academies Press.

National Research Council, Division on

Engineering and Physical Sciences, Board on Mathematical Sciences and Their Applications, and Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification. 2012. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. National Academies Press.

Panigrahi, Bijaya Ketan, Munesh C. Trivedi, Krishn K. Mishra, Shailesh Tiwari, and Pradeep Kumar Singh. 2018. *Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS 2017*. Springer.

Pedir, Samah Samir. 2016. "Evaluation of the Factors and Treatment Options of Separated Endodontic Files Among Dentists and Undergraduate Students in Riyadh Area." *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*. https://doi.org/10.7860/jcdr/2016/16785.7353.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102102.

Rinesh, S., K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, and Yosef Asrat Waji. 2022. "Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms." *Journal of Healthcare Engineering* 2022 (February): 2761847.

Sathish, T., V. Mohanavel, M. Arunkumar, K. Rajan, Manzoore Elahi M. Soudagar, M. A. Mujtaba, Saleh H. Salmen, Sami Al Obaid, H. Fayaz, and S. Sivakumar. 2022. "Utilization of Azadirachta Indica Biodiesel, Ethanol and Diesel Blends for Diesel Engine Applications with Engine Emission Profile." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123798.

Sudhan, M. B., M. Sinthuja, S. Pravinth Raja, J. Amutharaj, G. Charlyn Pushpa Latha, S. Sheeba Rachel, T. Anitha, T. Rajendran, and Yosef Asrat Waji. 2022. "Segmentation and Classification of Glaucoma Using U-Net with Deep Learning Model." *Journal of Healthcare Engineering* 2022 (February): 1601354.

Sundararaman, Sathish, J. Aravind Kumar, Prabu Deivasigamani, and Yuvarajan Devarajan. 2022. "Emerging Pharma Residue Contaminants: Occurrence, Monitoring, Risk and Fate Assessment – A Challenge to Water Resource Management." *Science of The Total Environment*. https://doi.org/10.1016/j.scitotenv.2022.153897.

Teysseire, Fabienne, Valentine Bordier, Aleksandra Budzinska, Nathalie Weltens, Jens F. Rehfeld, Jens J. Holst, Bolette Hartmann, et al. 2022. "The Role of D-Allulose and Erythritol on the Activity of the Gut Sweet Taste Receptor and Gastrointestinal Satiation Hormone Release in Humans: A Randomized, Controlled Trial." *The Journal of Nutrition*, February. https://doi.org/10.1093/jn/nxac026.

"Two-Group Multivariate Analysis of Variance." 2015. *Applied Multivariate Statistics for the Social Sciences*. https://doi.org/10.4324/9781315814919-9.

Ustun, Taha Selim. 2019. *Advanced Communication and Control Methods for Future Smartgrids*. BoD – Books on Demand.

Vermesan, Ovidiu, and Peter Friess. 2014. *Internet of Things Applications: From Research and Innovation to Market Deployment*.

Vijayalakshmi, V. J., Prakash Arumugam, A. Ananthi Christy, and R. Brindha. 2022. "Simultaneous Allocation of EV Charging Stations and Renewable Energy Sources: An Elite RERNN-m2MPA Approach." *International Journal of Energy Research*. https://doi.org/10.1002/er.7780.

Yaashikaa, P. R., P. Senthil Kumar, S. Jeevanantham, and R. Saravanan. 2022. "A Review on Bioremediation Approach for Heavy Metal Detoxification and Accumulation in Plants." *Environmental Pollution* 301 (May): 119035.

Yin, Hujun, David Camacho, Peter Tino, Antonio J. Tallón-Ballesteros, Ronaldo Menezes, and Richard Allmendinger. 2019. *Intelligent Data Engineering and Automated Learning – IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part II*. Springer Nature.

Zambom, Adriano Zanin, and Michael G. Akritas. 2015. "Nonparametric Significance Testing and Group Variable Selection." *Journal of Multivariate Analysis*. https://doi.org/10.1016/j.jmva.2014.08.014.

Preethi, P. S., Hariharan, N. M., Vickram, S., Manian, R., Manikandan, S., Subbaiya, R., ... & Awasthi, M. K. (2022). Advances in bioremediation of emerging contaminants from industrial wastewater by oxidoreductase enzymes. Bioresource Technology, 127444.

Tables and Figures

Table 1. Analysis of Lasso regression and decision tree. The novel lasso regression algorithm is 15% more efficient than the decision tree algorithm.

ITERATION NO	Lasso regression	Decision tree
1	96.1	94.2
2	95.2	92.4
3	94.4	90.3
4	90.1	89.5
5	87.6	87.3

Table 2. Group Statistics of Lasso regression and decision tree algorithm with the mean value of 90.40% and 85.80%

GROUP	N	Mean(%)	Std.deviation	Std.Error Mean
Lasso regression	5	90.40	1.517	.678
Decision tree	5	85.80	3.564	1.594

Table 3. The insignificance and standard error of the two groups are determined using an independent sample T-test. One tailed insignificance value is 0.511 and it is statistically insignificant

		Levene's test for equality of variances.		**T- test for equality of means**						
		F	**Sig.**	t	**df**	**Sig.(2-tailed)**	**Mean difference**	**Std. error difference**	**95% confidence interval of the difference**	
									Lower	**Upper**
ACCURACY	Equal variance assumed	6.826	.511	2.656	8	.029	4.600	1.732	0.606	8.594
	Equal variances			2.656	5.403	.042	4.600	1.732	0.246	8.954

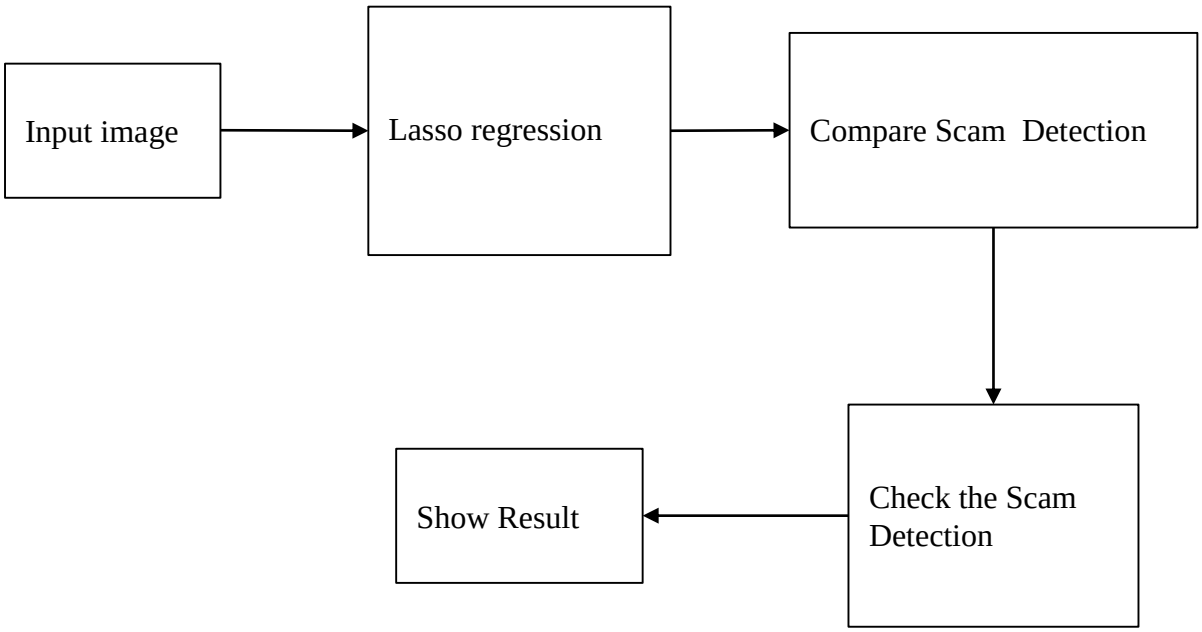

Fig. 1. Architecture for Error analysis of fault data detection in IoT devices using lasso regression compared over decision tree.

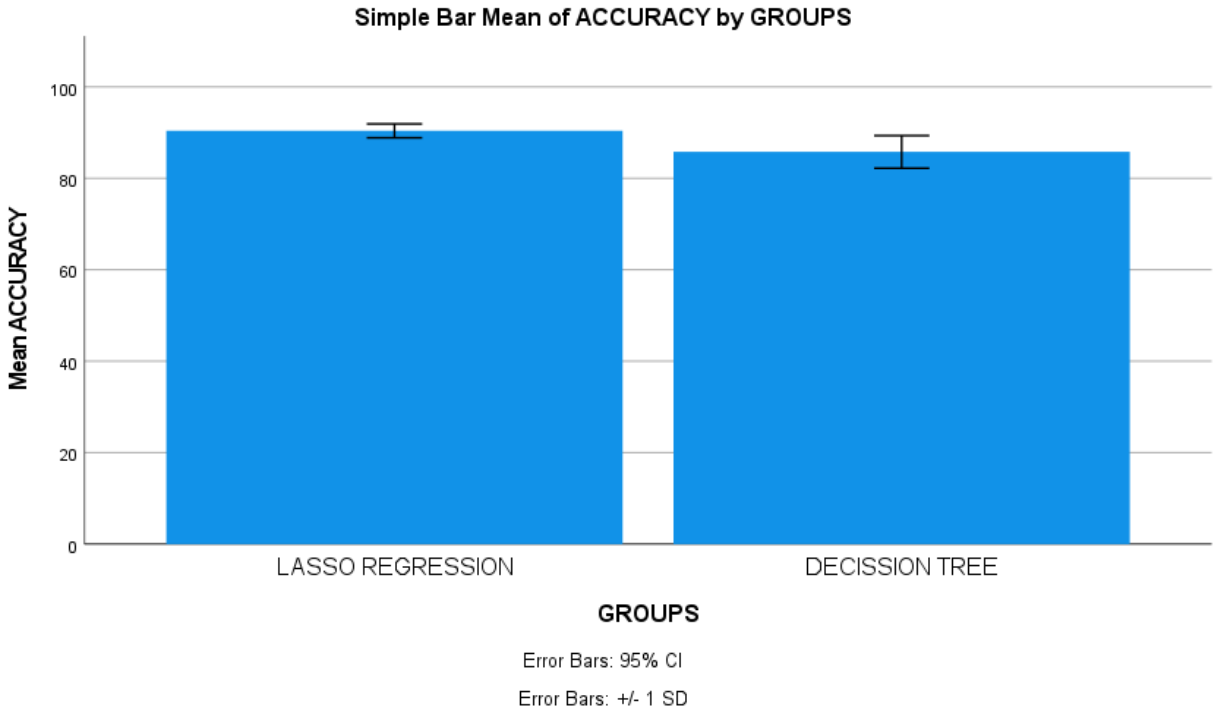

Fig. 2. Bar graph analysis of lasso regression algorithm and decision tree algorithm. Graphical representation shows the mean efficiency of 90.40% and 85.80% for the proposed lasso regression algorithm and decision tree respectively. X-axis : Lasso regression vs decision tree. Y-axis : Mean precision ± 1 SD.