A Hybrid Machine-Learning Ensemble For Real-Time 4.0 Systems Anomaly Detection

[1]Dr Chandramouli VSA, [2]Mr. B Ganga Bhavani, [3]S.Divyateja, [4] P.Anitha, [5]NRDSSSrinivas, [6]PL kumar

[1,2,3,4,5,6] Department of CSE, BVC Engineering College, Odalarevu, A.P., India

[*1]chandramouli.ac@yahoo.com,[2] bhavanicse10@gmail.com,[3] S.Divyateja@gmail.com
[4] P.Anitha@gmail.com,[5] NRDSS.Srinivas@gmail.com,[6] PL.kumar@gmail.com

Abstract: Massive amounts of data are created and collected by data acquisition systems, including SCADA and embedded systems in industrial machines. The data may be analysed by AI algorithms to gain a better understanding of the process and identify any irregularities in the machines. This is a major benefit of Industry 4.0. This paper will examine predictive maintenance as a process in industry and demonstrate how it may make full use of technologies from Industry 4.0. Using a weighted average, our method integrates Three Machine Learning models: Auto encoder (AE), One-Class Support Vector Machine (OCSVM), and Local Outlier Factor (LOF). This allows us to discover anomalies in real-time. Anomaly detection will be enhanced as a result. Improve real-time anomaly detection using the suggested hybrid machine-learning ensemble pipeline by combining the outputs of three ML models: auto encoder (AE), one-class support vector machine (OCSVM), and local outlier factor (LOF).
Keywords: SCADA, Embedded Systems, AI Algorithms, Hybrid Machine-Learning, OCSVM, and Auto encoder (AE)

1. INTRODUCTION

The necessity to either enhance existing processes or create entirely new ones is one of the new obstacles that conventional industrial processes must overcome in order to take advantage of modern technologies. The new enabling technologies that form the basis of the Fourth Industrial Revolution, also known as "Industry 4.0," are a potential game-changer in the highly competitive business world. Robotics, nanotechnology, quantum computing, cyber-physical systems, artificial intelligence (AI), big data, the internet of things (IoT), and so on are all instances of such technologies. Industrial machinery generate and gather vast amounts of data using data acquisition systems like SCADA or embedded systems. The data may be analysed by AI algorithms to gain a better understanding of the process and identify any irregularities in the machines. This is a major benefit of Industry 4.0. In this article, we will look at predictive maintenance as an industrial process that might benefit greatly from the aforementioned Industry 4.0 technologies [2].Detecting anomalies is a popular area of research for many different fields. Machine upkeep, fraud prediction, intrusion protection, and illness diagnosis are only a few examples of applicable academic disciplines [4]. Anomaly detection finds states that don't match by comparing them to the normalcy data, which often represents the most common states.It is not an easy task to identify outliers. When dealing with streaming data, real-time detection becomes more complicated due to its limitations. Instead of utilising all available historical data without adding any new input to the models, stream learning differs from batch learning in that it involves consideration of five restrictions [5]. The system has to figure

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_128

out if the current sample is destroyed or archived, which is a big limitation when processing streaming data samples. This is due to the fact that samples can only be viewed once upon arrival on the internet. In order to retrieve data from the past, it is necessary to store samples in memory. Once that happens, the previous samples are erased by a forgetting process. Because of storage limitations, it is not possible to reverse decisions that were based on previous data. the processing time for each data sample should be short and consistent (iv). v) Just like a batch algorithm, the input-processing method should produce a model.
Due to these five limits, most anomaly detection algorithms that were created for batch processing cannot be employed for stream processing. While batch-learning algorithms are useful for creating a basic model, there are hybrid approaches that employ streaming anomaly-detection techniques.
This research evaluates and analyses various approaches to the challenge of real-time anomaly detection. The relative importance of the individual methods in the final model is defined by the performance-control criteria that were applied for each. This research takes an average of the projected output from three ML models, namely Autoencoder, One-Class Support Vector Machine (OCSVM), and Local Outlier Factor (LOF), and uses the F1-score value as a weight for each model. The integrated model's end aim is the real-time detection of anomalies in industrial environments. Information gathered from an actual industrial system of air-blowing equipment was used to construct the proposed hybrid model. This implies that the suggested hybrid anomaly detection approach is not limited to the Industry 4.0 domain, but can be used in many other industrial frameworks that have access to real-time data collecting systems.

Two potential causes may contribute to the challenges of anomaly identification in a business environment. A proper definition and description of normal data behaviour is necessary for the accurate detection of anomalies

2. LITERATURE SURVEY

Methods for detecting outliers in sequencing data and their comparative assessment.(V. Kumar, V. Mithal, and V. Chandola),In this comprehensive review, we assess and contrast the performance of multiple anomaly detection algorithms on a range of synthetic and real-world datasets. Although a few of these methods are novel, the most are merely improvements or expansions of existing anomaly detection strategies for sequence data.we use semi-supervised anomaly detection to water analytics.J. Davis, G. Verbruggen, K. Maes, R. Baumer, and W. Meert are the authors.
A dashboard is a common tool in modern manufacturing for monitoring and displaying all process components in real-time. In order to monitor machines, track resource usage, and document actions, a variety of sensors are utilised. Discovering out-of-the-ordinary patterns in sensor-generated time series data is a prevalent challenge in this setting. The potential for automated methods to outperform the costly and time-consuming manual data analysis is substantial. Approaching anomaly detection as a supervised learning problem is usually not viable since there are either no tagged instances of abnormal behaviour or acquiring them would be too difficult, unpleasant, or both. Consequently, unsupervised methods are commonly used; these algorithms usually classify outliers as unusual occurrences rather than regular occurrences. But there are a lot of real-life settings where certain kinds of typical behaviour are more prevalent than others. Our proposed method is an innovative hybrid of semi-supervised and unsupervised restricted clustering-based anomaly detection. The approach can take an unlabeled data set as a starting point and enhance performance by incorporating expert opinions. The Colruyt Group is a large Belgian retailer, and we collaborated with them to gather time series data from supermarkets that are used for actual water monitoring in order to validate our methodology. Based on the data, our method is superior to other baselines, including the existing detection system. Currently, our system is being utilised by the organisation to track water usage at twenty separate sites every single day.A comprehensive literature study on methods for detecting, analysing, and predicting anomalies in an Internet of Things (IoT) setting.Along with A. Sillitti, M. Fahim

Due to advancements in sensor monitoring technologies, their affordability, and their substantial impact across multiple application domains, anomaly detection has recently attracted a lot of attention from scientists. Sensors generate massive volumes of data while monitoring real-world locations and objects. By analysing these streams of massively collected data, harmful behaviours can be uncovered. It could prevent the systems from breaking down, reduce functional hazards, and uncover hidden flaws. A plethora of research methodologies have been established and refined in the fields of security and risk analysis with the express purpose of detecting such anomalous actions. We present the results of a comprehensive literature review on anomaly detection methods, excluding these dominant areas of research. Smart things, industrial systems, transport systems, healthcare systems, and intelligent homes are the areas of application where our study is concentrated on papers published between 2000 and 2018. We found that there is a lack of published research on real-world anomalous behaviour prediction, processing constraints of statistical methods, data collection, and large-scale imbalanced dataset analysis. Academics and practitioners can learn about existing approaches, put them to use in real-world problems, and/or help develop new methods for identifying, predicting, and analysing anomalies with the help of our analysis.Spotting irregularities in discrete sequences: a review.“V. Kumar,” “A. Banerjee,” and “V. Chandola”The purpose of this study is to provide a thorough and organised summary of the research on anomaly detection in discrete/symbolic sequences. A comprehensive grasp of the sequence anomaly detection problem and the interplay between various approaches is the aim. This review's strongest point is the way it organises the literature: into three distinct parts, each addressing a different set of problems. The first is to locate sequences that do not conform to a normal sequence database; the second is to locate an extended sequence subsequence that does not add up; and the third is to locate an unexpectedly frequent pattern inside a sequence. We evaluate the usefulness of these problem formulations in different domains and demonstrate how they vary from one another. Each of these formulations is addressed along with methods from a broad range of unrelated and distinct application sectors. For each problem statement, we classify methods according to the algorithm type. We describe a fundamental method for detecting anomalies in each category and demonstrate how the current methods are actually just variations on this subject. This strategy is useful for comparing and contrasting methods that fall into the same category. Our classification uncovers hitherto unexplored varieties and combinations for anomaly identification. Also included are the pros and cons of certain approaches. We provide a wealth of fresh ways to address the various problem formulations by demonstrating how methods designed to handle one problem formulation may be modified to solve another. We further emphasise two more relevant fields that can benefit from discrete sequence handling methods: time series anomaly identification and live anomaly detection. A survey on data mining strategies for anomaly detection.[Agrawal brothers S. and J.]

Huge amounts of data are being stored and transmitted by people all over the world these days. Data becomes more vulnerable to intrusion as it is transferred or stored. Despite all the precautions and tools you use, data protection is not 100% effective. Data mining technologies have made it easier to analyse data and identify different types of assaults, which has reduced its vulnerability to these risks. A data breach or attack is more likely to take place if anomaly detection employs these data mining techniques to reveal the concealed anomalous behaviour in the data. Several hybrid methods have been developed to further enhance the accuracy of assault detection. To further educate readers on the methods currently employed and to support future research in this subject, this paper examines a range of data mining methodologies for anomaly discovery.

3. PROPOSED SYSTEM

To further enhance its anomaly detection rate, our proposed system now incorporates a Hybrid Ensemble Machine Learning technique. Prior to merging their output predictions, we are comparing the F-SCORE of the Hybrid algorithm with those of three independent methods: Auto Encoder (AE), One Class Support

Vector Machine (OCS), and Local Outlier Factor (LOF). The rate of anomaly identification is enhanced by retraining using an auto-encoder following the combining of three different algorithms' predictions.

Dataset Pre-processing: The author uses this module to normalise the features using the MIN-MAX technique after deleting missing data.
Features Selection: After the characteristics have been normalised, a training set will be constructed.
Dimension reduction: This module will be used to use principal component analysis (PCA) in order to decrease the dataset's dimensions by selecting important characteristics and removing unrelated ones.
KMEANS Clustering: In this module, we will use KMEANS to group similar data values into two clusters, Normal and Anomaly, using a dataset that has been analysed using PCA.
Train LOF with Grid Search: The Grid Search LOF model will be built using clustered data. Then, it will be applied to TRAIN and TEST data to calculate the prediction F-SCORE.
Train One Class SVM with Grid Search: The Grid Search One Class Support Vector Machine (OCS) model will be constructed using clustered data. It will thereafter be used to both TRAIN and TEST data sets in order to determine the prediction F-SCORE.
Train Auto Encoder with different epochs: In order to calculate the prediction F-SCORE, the auto encoder will be fed clustered data. Then, the AE model will be applied to the TRAIN and TEST datasets.
Hybrid Ensemble Algorithm: We will use the three algorithms to forecast values, retrain Auto-Encoder to create a hybrid model, and then apply the model to both the training and testing sets of data to determine the F-SCORE.
CNN Extension Algorithm: Prior to applying the CNN model to the TRAIN and TEST datasets to determine the prediction F-SCORE, the clustered data will be used to train the model.
Comparison Graph and Tables: This module will be used to display, graphically and tabularly, the F-SCORE comparison of all algorithms using TRAIN and TEST data.

4 ALGORITHMS

An unsupervised approach to anomaly detection, the Local Outlier Factor (LOF) algorithm compares a single data point's relative local density deviation to that of its neighbours. When it comes to density, outliers are samples that don't fit the pattern. To do single-class classification using clustered data, use Grid Search to enter the data into One Class SVM. Applying the OCS model (which is built with a single class SVM utilising TRAIN and TEST data) allows one to obtain the prediction F-SCORE.Auto Encoder: After training the model using clustered data, it is possible to calculate the prediction F-SCORE by applying it to both the TRAIN and TEST sets of data.

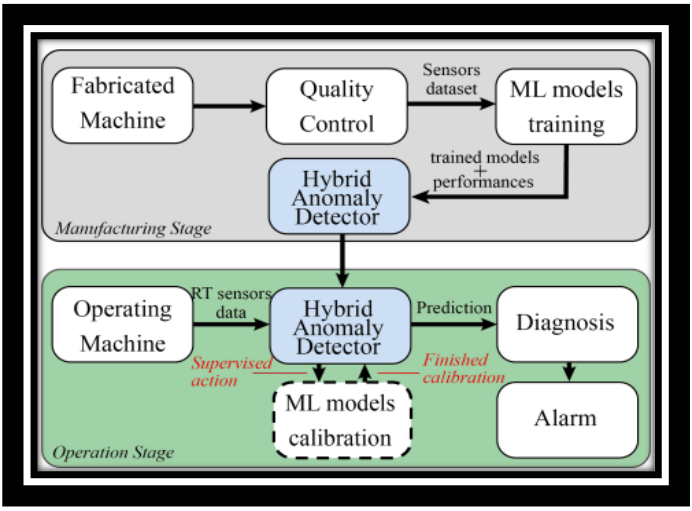

Fig1. Architecture

5. RESULTS AND DISCUSSION

Launch JUPYTER Notebook by doubling-clicking the "run.bat" file; the following output will be shown.
The amount of normal and anomalous records in the dataset in the graph up there. We used the principal component analysis (PCA) algorithm to reduce the number of dimensions in the dataset, which went from five features before PCA to four features after PCA.

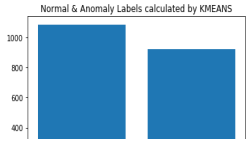

Fig. 2. Normal and Anomaly Labels Calculated by KMEANS

We can observe that 1086 records are classified as Normal and 923 records as Anomaly in the blue text of the preceding image, which is also visible in the graph, when we apply K-MEANS clustering.
The total dataset comprises 2009 samples, with 80% (1808 samples) used for training and 20% (201 samples) reserved for testing.
The LOF Training F1 Score is 0.4670780846957522 and the LOF Validation F1 Score is 0.417910447761194.

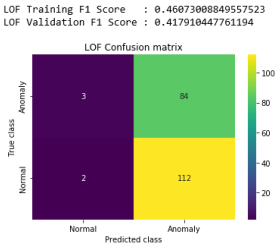

Fig. 3. LOF Confusion Matrix

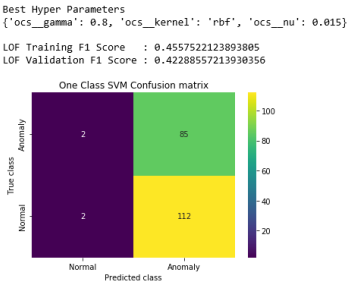

Fig. 4 . One Class SVM Confusion Matrix

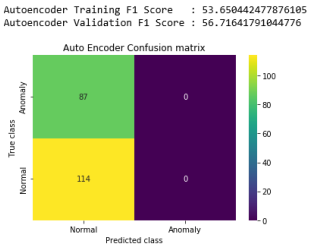

Fig. 5. Auto-Encoder confusion matrix.

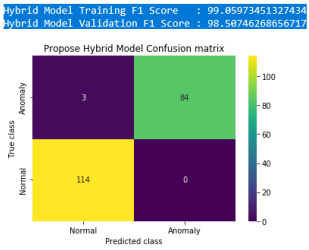

Fig. 6. Proposed Hybrid Model Confusion Matrix

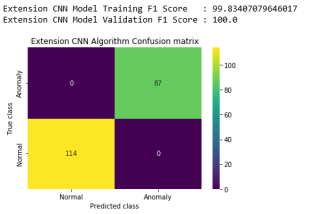

Fig. 7. Extension CNN Algorithm Confusion Matrix

	Algorithm Name	Training FScore	Validation FScore
0	LOF	0.460730	0.417910
1	One Class SVM	0.455752	0.422886
2	Auto Encoder	53.650442	56.716418
3	Hybrid Ensemble Model	99.059735	98.507463
4	Extension CNN Model	99.834071	100.000000

In [ ]:

Fig. 8. Performance Metrics

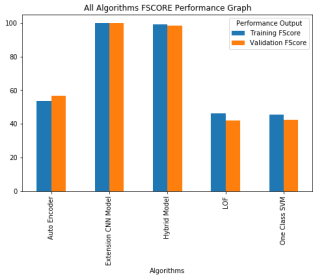

Fig. 9. All Algorithms FScore Performance Graph

CONCLUSION

This work presents the results of developing and testing a hybrid machine-learning ensemble for use in real-time industry 4.0 anomaly detection systems. This ensemble consists of two stages, the manufacturing and operating stages, which are based on traditional industrial models. We are not aware of any other ML methods that consider these industrial stages at this moment.

BIBLIOGRAPHY

[1] M. Xu, J. M. David, and S. H. Kim, ‘‘The fourth industrial revolution: Opportunities and challenges,’’ Int. J. Financial Res., vol. 9, no. 2, pp. 92–95, 2018.

[2] M. Reis and G. Gins, ‘‘Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis,’’ Processes, vol. 5, p. 35, Jun. 2017. [Online]. Available: http://www. mdpi.com/2227-9717/5/3/35

[3] S. H. An, G. Heo, and S. H. Chang, ‘‘Detection of process anomalies using an improved statistical learning framework,’’ Expert Syst. Appl., vol. 38, no. 3, pp. 1356–1363, Mar. 2011.

[4] A. Boukerche, L. Zheng, and O. Alfandi, ‘‘Outlier detection: Methods, models, and classification,’’ ACM Comput. Surv., vol. 53, no. 3, pp. 1–37, May 2021.

[5] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. D. Carvalho, and J. Gama, ‘‘Data stream clustering: A survey,’’ ACM Comput. Surv., vol. 46, no. 1, pp. 1–31, 2013.

[6] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, ‘‘Unsupervised realtime anomaly detection for streaming data,’’ Neurocomputing, vol. 262, pp. 134–147, Nov. 2017.

[7] V. Chandola, V. Mithal, and V. Kumar, ‘‘Comparative evaluation of anomaly detection techniques for sequence data,’’ in Proc. 8th IEEE Int. Conf. Data Mining, Dec. 2008, pp. 743–748.

[8] J. Rabatel, S. Bringay, and P. Poncelet, ‘‘Anomaly detection in monitoring sensor data for preventive maintenance,’’ Expert Syst. Appl., vol. 38, no. 6, pp. 7003–7015, Jun. 2011.

[9] V. Vercruyssen, W. Meert, G. Verbruggen, K. Maes, R. Baumer, and J. Davis, ‘‘Semi-supervised anomaly detection with an application to water analytics,’’ in Proc. IEEE Int. Conf. Data Mining (ICDM), Nov. 2018, pp. 527–536.

[10] M. Fahim and A. Sillitti, ‘‘Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review,’’ IEEE Access, vol. 7, pp. 81664–81681, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8733806, doi: 10.1109/ ACCESS.2019.2921912.

[11] B. R. Priyanga and D. Kumari, ‘‘A survey on anomaly detection using unsupervised learning techniques,’’ Int. J. Creative Res. Thoughts (IJCRT), vol. 6, no. 2, pp. 2320–2882, 2018. [Online]. Available: http://www.ijcrt.org/papers/IJCRT1812118.pdf

[12] V. Chandola, A. Banerjee, and V. Kumar, ‘‘Anomaly detection for discrete sequences: A survey,’’ IEEE Trans. Knowl. Data Eng., vol. 24, no. 5, pp. 823–839, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/5645624, doi: 10.1109/TKDE.2010.235.

[13] Ramesh, P. S., Vanteru, M. K., Rajinikanth, E., Ramesh, J. V. N., Bhasker, B., & Begum, A. Y. (2023). Design and optimization of feedback controllers for motion control in the manufacturing system for digital twin. SN Computer Science, 4(6), 782.

[14] M. Hubert and E. Vandervieren, ‘‘An adjusted boxplot for skewed distributions,’’ Comput. Statist. Data Anal., vol. 52, no. 12, pp. 5186–5201, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167947307004434

[15] S. Agrawal and J. Agrawal, ‘‘Survey on anomaly detection using data mining techniques,’’ Proc. Comput. Sci., vol. 60, pp. 708–713, Jan. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915023479

[16] S. Ferreiro, B. Sierra, I. Irigoien, and E. Gorritxategi, ‘‘A Bayesian network for burr detection in the drilling process,’’ J. Intell. Manuf., vol. 23, no. 5, pp. 1463–1475, Oct. 2012, doi: 10.1007/s10845-011-0502-z.

[17] B. Sierra, E. Lazkano, E. Jauregi, and I. Irigoien, ‘‘Histogram distancebased Bayesian network structure learning: A supervised classification specific approach,’’ Decis. Support Syst., vol. 48, no. 1, pp. 180–190, Dec. 2009.

[18] Y. Yuan, S. Li, X. Zhang, and J. Sun, ‘‘A comparative analysis of SVM, naive Bayes and GBDT for data faults detection in WSNs,’’ in Proc. IEEE Int. Conf. Softw. Qual., Rel. Secur. Companion (QRS-C), Jul. 2018, pp. 394–399.

[19] Morasa, B., Anwaraly, Y.B., Chennuru, J.R., Devara, C., Gavvalla, M. (2022). Smart Monitoring System for Waste Management Using IoT. In: Sugumaran, V., Upadhyay, D., Sharma, S. (eds) Advancements in Interdisciplinary Research. AIR 2022. Communications in Computer and Information Science, vol 1738. Springer, Cham.

[20] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, ‘‘LOF: Identifying density-based local outliers,’’ ACM SIGMOD Rec., vol. 29, no. 2, pp. 93–104, Jun. 2000.