Leveraging Machine Learning for Enhanced Cyber Attack Detection and Defence in Big Data Management and Process Mining

Dr. Taviti Naidu Gongada[1], Dr. Amit Agnihotri[2], Kathari Santosh[3], Dr. Vijayalakshmi Ponnuswamy[4],
Narendran S[5], Dr. Tripti Sharma[6], Prof. Ts. Dr. Yousef A.Baker El-Ebiary[7]
Assistant Professor, Dept of Operations, GITAM School of Business, GITAM (Deemed to be) University, Visakhapatnam, India[1]
Assistant Professor (CS), Dept. of Computer and Information Sciences, Jrd State University, Chitrakoot (UP)- 210204[2]
Assistant Professor, Department of MBA, CMR Institute of Technology, Bengaluru, Bengaluru, India[3]
Professor, Department of Artificial Intelligence and Data Science, Koneru Lakshmiah Educational Foundation
(KL Deemed to be University), Green fields, Vaddeswaram, Guntur District, Andhra Pradesh, India, Pin code: 522302[4]
Assistant Professor, Dept of Nanotechnology, Institute of Electronics and Communication Engineering, SIMATS Engineering,
Saveetha Institute of Medical and Technical Sciences, Kanchipuram[5]
Department of Computer Science and Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]

Abstract—**The rapidly developing field of "Commercial Operation Divergence Analysis," this research seeks to identify and understand differences in commercial systems that exceed expected results. Approaches in this domain aim to identify the characteristics of process implementations that are associated with changes in process effectiveness. This entails identifying the features of procedural behaviours that result in unpleasant results and figuring out which behaviours have the biggest impact on increased efficiency. As the scale and complexity of big data management and process mining continue to expand, the threat of cyber-attacks poses a critical challenge. This research leverages machine learning techniques for the detection and defence against cyber threats within the realm of big data management and process mining. The study introduces novel metrics such as Skewness, Coefficient of Variation, Standard Deviation, Maximum, Minimum, and Mean for assessing the security state, utilizing variables like SPI, SPEI, and SSI. The research addresses prior issues in cyber-attack detection by integrating machine learning into the specific context of big data and process mining. The novelty lies in the application of Skewness and other statistical metrics to enhance the precision of threat detection. The results demonstrate the effectiveness of the proposed methodology, showcasing promising outcomes in identifying and mitigating cyber threats in the given dataset and which makes use of Support Vector Regression (SVR), has a standard deviation of 0.9, which is consistent with the variability shown in SVM. The results demonstrate a significant achievement, with a Mean Absolute Error (MAE) of 0.98, indicating the efficacy of the proposed approach in providing accurate and timely insights for cyberattack detection and defense, thereby enhancing the overall security posture in data-intensive systems. The results highlight how well the proposed method extracts significant insights from complicated event data, with important ramifications for real-world application and decision-making procedures.**

Keywords—Machine learning; data mining; cyber-attack detection; big data; support vector regression

I. INTRODUCTION

Effective extraction operations depend on the deep mine's ability to maintain a healthy and secure air atmosphere [1]. A crucial step in the analysis of data is outlier detection. Hawkins states that atypical is "a thing whether diverges sufficiently form other items as to be assumed that it has been produced by an alternate method" [2]. In 2008, the Global Financial Crisis (GFC) and the demise of the coal mining "super cycle" put a stop to a period of production-focused tactics during which operational costs increased faster than output [3]. Because it presents especially challenging compromises the extractive and material extraction sector is a desirable test case for the study of contamination. Individual plants may provide enormous value, up to millions of dollars annually [4]. Studies had lately claimed that the application of Process Mining (PM) might address these drawbacks through enabling auditors to efficiently and primarily automatically analyse all of the databases employing historic and/or present-day information [5]. Nevertheless, a number of issues brought on by the extraction and use of coal assets, including as sinkholes, erosion of soil, landslides, and the demolition of buildings, have had a significant detrimental impact on the daily lives and assets of local populations [6].

Mining processes is a field of study who tries to enhance process enhancements by offering based on reality observations on previous procedure implementations. The topic sits among system modelling and evaluation and intelligence computing as well as data mine on one's hand. Process variation assessment is described as a collection of methods that allow to contrast more than one event records belonging to various company procedure versions for the purpose to identify the differences between them [7]. A prime instance of contaminated soil includes the soils that make up anthracite mine dumps. The sedimentary layers that cover a coal seam are where the initial soil was formed. The excess soil is typically excavated using various excavators, then delivered into the spoil site via lorries or belt conveyors and deposited form different heights, either with or no choosing the material [8].

Multiple research studies indicate that these last class of computations, machine learning algorithms (MLAs), can be more accurate than statistical methods like discriminant evaluation or logistic regression, particularly if the feature space to be studied is complicated (i.e., once the dimension of the input feature time is believed to be quite large and the connection between the intended contributions along with the feedback transparent include is predicted to be non-linear) and the data sets being used are anticipated to include distinct characteristics [9]. One the contrary, machine learning is a branch of computing which seeks to give machines or different gadgets the capacity to understand sans needing directly controlled. It tries to provide methods and mathematical models for data-driven learning and forecasting. Upon accomplishment, machine learning techniques are used to simulate characteristics of the input in relation to anticipated result, predict production attributes in relation to past information, and characterise the behaviour within the data. A possible approach to predicting wind power using velocity data is machine learning techniques [10]. Machine learning has been immensely successful as information quantities and types have increased because of its ability to examine complex trends in seen information and generate predictive models or choices on fresh data. In the literature, a variety of machine learning methods and algorithms have been published [11].

Predicting how a business operation will behave in the years to come is an essential corporate competence. Procedure prediction, a form of statistical analysis used in management of business processes, uses information from previous process occurrences to forecast future ones [12]. Customer service representatives adjusting to requests about the amount of time left until an issue has been settled are a few examples of use instances. Other use cases include production managers forecasting the length of a manufacturing procedure for improved scheduling and higher utilisation or case supervisors determining probably violations of regulations to reduce business risk [13]. One kind of procedure mining work, called procedure learning, looks for a model that describes the behaviour of an organization's process using information about how it has previously been executed. The log of events is mapped onto a procedure model using a method known as a process identification procedure, which guarantees the model in question is a good representation of the behaviour shown in the event log [14].

Our approach prioritizes adaptability to external influences by employing dynamic updating mechanisms. We continuously monitor cyber security policies, track advancements in attack techniques, and stay abreast of technological shifts. This proactive approach allows us to incorporate new knowledge into our models promptly, ensuring their relevance and effectiveness in evolving cyber security landscapes. Additionally, we leverage techniques such as transfer learning and ensemble methods to enhance model robustness and resilience to changing external factors. The proposed model exhibits robustness to changes in feature selection and extraction methods through rigorous validation and sensitivity analysis, ensuring consistent performance across varying feature sets. Additionally, automating feature engineering enhances efficiency and scalability while reducing the risk of human error, bolstering the reliability and adaptability of our models. Regular audits and oversight mechanisms further reinforce data privacy measures, mitigating potential privacy concerns and promoting responsible data stewardship in cyber security practices. Implementing the methodology may face challenges such as organizational resistance to change, integration with existing infrastructure, and compliance with regulations.

The following are the research Primary Contribution:

- The application of machine learning algorithms allows for improved accuracy and predictive power in identifying the characteristics of process behaviour that contribute to efficiency shifts.
- Machine learning algorithms provide a means to uncover the relevant factors that significantly affect process efficiency. By analysing the event logs and applying the proposed Declare-based coding, the research identifies the most influential aspects of a procedure, allowing organizations to focus on these factors for process optimization.
- The combination of machine learning algorithms and the proposed encoding technique constitutes an effective tool for the analysis of processes.
- The research compares the performance of different machine learning algorithms, such as Standardized Stream flow Index, Gene Expression Programming, Support Vector Regression, and M5 Model Tree. This comparative evaluation helps in understanding the strengths and weaknesses of each algorithm and provides guidance on selecting the most suitable approach for a given context.

Section I, the introduction, provides an overview of the research topic, establishing its relevance and context. Section II, related work, explores existing literature and research in the field to highlight gaps or connections with the current study. Section III, the problem statement, clearly defines the specific issue or gap that the research aims to address. Section IV, methodology, outlines the approach and techniques employed to conduct the study. Section V presents the results and engages in a discussion, while Section VI concludes the research, summarizing key findings and suggesting potential avenues for future exploration.

II. Related works

Richetti et al. [15] proposed to determine the aspects of a procedure which most affect its efficiency, they first use Treatment Learning as an original method in the realm of Deviation Mining. This is a novel encoding method enabling vector-based representations of process occurrences. The suggested encoding method may find more expressive solutions since it is built on declaring restriction framework fulfilment. Using publicly accessible logs of events from actual procedures, they do a number of tests that contrast our suggestion to the state-of-the-art activity decoding methods. The findings demonstrated that behavioural learning offered actionable and more descriptive insight from events logs when combined with our suggested Declare-based encoding, making it a useful tool for the analysis of processes.

Al-Shehari et al. [16] proposed the use of feature resizing and quick encoding strategies are used in the framework to alleviate the potential skew of identification outcomes that might emerge from an ineffective decoding procedure. The artificial minority sampling too much method (SMOTE) is additionally employed to alleviate the data set's balance problem. In order to discover a highly precise classification which can identify data leakage events carried out by malevolent outsiders throughout the crucial time when they depart an organisation, renowned machine learning methods are used. By applying our mathematical framework on the CMU-CERT Insider Threat Dataset and contrasting its results with the real world, we demonstrate the notion behind it. The results of the experiment demonstrate that our framework outperforms other methods which have been evaluated on the identical data in terms of detecting internal leakage of information events, with an AUC-ROC value of 0.99. The suggested framework offers practical approaches to deal with potential bias and class imbalance concerns in order to design a system that effectively detects insider data leaking.

Roldán et al. [17] proposed an approach that uses technologies like augmented reality and data mapping to teach workers in assembly operations. Firstly, skilled employees do assembly in accordance with their knowledge using a fully immersive environment. The next step is to use process mining methods to extract assemble model in the logs of events. Lastly, to understand the groups what the expert employees incorporated into the framework, learner employees utilise an improved immersion display with suggestions. Construction block experiments were designed as a toy example, and studies on a group of participants have been conducted. The outcomes demonstrate the suggested education system's competitiveness against more traditional options. It bases itself on procedure mining and mixed reality. In terms of mental effort, vision, learning, outcomes, and how they perform, user ratings are also superior.

Helm et al. [18] proposed 38 procedure mining instances related to health care reported from 2016 to 2018 that discussed the instruments, methods, and methodologies used as well as specifics on how the log data were found to have been medically significant. Utilising the common clinical coding schemes SNOMED CT and ICD-10, researchers then connected the diagnostic characteristics of the patient encounter setting, clinical speciality, and diagnosis of illness. The possible results of utilising a standardised method for categorising medical terms and events log data using common clinical codes are also highlighted.

Weinzierl et al. [19] proposed several prospective business process monitoring (PBPM) strategies that attempt to forecast potential process behaviours while the procedure is being executed. Methods for predicting subsequent event in particular have considerable promise for enhancing practical company processes. Many of these methods use deep neural networks (DNNs) and take into account data pertaining to the environment where the operation is occurring to provide recommendations that tend to be more reliable. Nevertheless, an in-depth analysis of such methods is lacking in the PBPM literature, making it difficult for academics and industry professionals to decide which approach is appropriate for a particular event log. To address this issue, they statistically assess the prediction performance among three potential DNN structures using five tried-and-true encoding methods and five context-rich real-world logs of events. They offer four conclusions that might aid researchers and practitioners in developing fresh PBPM methods for anticipating upcoming actions.

The literature review showcases several developments in machine learning and process mining applications across a range of industries. But there is a clear research vacuum when it comes to combining these technologies to improve cyber security—more especially, when it comes to insider threat defence and detection. Research on process mining, efficiency assessment, healthcare procedures, and prospective business process monitoring has been greatly aided by studies by Richetti et al. [15], Al-Shehari et al. [16], Roldán et al. [17], Helm et al. [18], and Weinzierl et al. [19]. However, none of these studies specifically address the crucial problem of using machine learning for cyber security in the context of Big Data Management and Process Mining. Novel encoding strategies, predictive modelling, or anomaly detection approaches specifically designed for cyber security in massively distributed data settings are not well explored in the literature. This gap in the literature highlights the necessity for a thorough investigation that carefully incorporates machine learning techniques into cyber security frameworks, with an emphasis on the particular difficulties presented by big data and process mining scenarios.

III. Problem Statement

The problem statement of this work is to address the limitations of existing techniques for business process deviance mining. These techniques are based on the extraction of patterns from event logs but have limited expressiveness, particularly in capturing complex relationships in highly flexible processes. The previous research is to apply Treatment Learning, a novel approach in the context of machine learning, to identify the characteristics of a process that have the most significant impact on its performance. The study aims to compare the proposed encoding technique with current process encoding techniques through a series of experiments using publicly available event logs from real-life processes [15]. By incorporating machine learning approaches to strengthen cyber-attack detection and defence mechanisms, particularly within the fields of Big Data Management and Process Mining, the research seeks to expand the breadth of cyber security while boosting the effectiveness and resilience of digital systems.

IV. Regarding Discovery and Declarative Process Modelling

Conventional urgent process diagrams are produced by the majority of mining process methods. These methods work effectively for organised processes since there aren't numerous additional ways an operation may be carried out. Declarative language modelling is suggested as a way to create an improved equilibrium amongst flexibility and guiding support for these types of models, despite the fact that many of these approaches are capable of handling event logs form flexible or unorganised models. Due to expressive modeling's relevance to log files from dynamic or unstructured processes, the potential of mining declarative models has also emerged. Potential bottlenecks may arise in resource-intensive tasks such as model training and feature extraction, requiring adequate computational resources and optimization strategies. To mitigate these challenges, we implement techniques such as data partitioning, caching, and resource allocation optimization to ensure efficient utilization of computational resources and maintain scalability as data volumes increase.

Declaring continues to be the most commonly employed languages for studies regarding declaratory modelling and mineral extraction, although having very little application in business. This is because it's versatile and particularly suited for use in extremely volatile procedures, which are characterised by extreme complexity and variety. The addition enables associations among actions taken upon KiPs to be described using domains limitations as opposed to sequential ordering. Additionally, it enables occurrences in a KiP to signal chronological ties, behavioural consistency restrictions, or choice-of-action relationships in its instances by using these extra notions. Compliance with laws and regulations controlling the application of machine learning for cyber security in various sectors and regions is given top priority in our approach. In addition, we keep clear records of all procedures, guaranteeing responsibility and traceability for our compliance initiatives. On the other hand, difficulties could emerge because regulations are dynamic and have different meanings in different places. The proposed models are designed to complement human-driven cyber security processes by providing automated support in threat detection and response. A collaborative approach where the proposed models serve as decision- support tools, aiding human analysts in identifying and prioritizing threats more efficiently. The models leverage time-series analysis methods to identify and respond to cyclical or recurring patterns in threat behaviour. Through this approach, the models demonstrate the ability to adapt to changes in threat behaviour over different time intervals, ensuring robust and effective threat detection capabilities in dynamic cyber security environments.

Deviance mining with machine learning and declare-based encoding of event logs in rapidly evolving environments like cyber security, where threats change constantly, machine learning models face challenges due to their assumption of stationary data distributions. This means they struggle to adapt to new patterns and trends. To overcome this, techniques like online learning algorithms and anomaly detection are crucial. Online learning allows models to update in real-time, while anomaly detection helps identify unusual behaviour. By employing these adaptive methods, machine learning models can better keep pace with evolving threats and enhance cyber security defenses. Machines are the simplest collection of rules that may be used in machine learning to discriminate between circumstances that include numerous highly weighted classes and scenarios with few strongly weighed categories. Machine learning, within contrast to association-rule mineral extraction, specifies a preferred type worth, that serves as a benchmark for weighing various class values and allows it to highlight machines with strong or poor performance as determined by a particular class characteristic in a dataset. Diverse datasets play a pivotal role in enhancing the generalizability of machine learning models. Models trained on diverse datasets are inherently more adaptable to variations across industries, company sizes, and geographic locations. This adaptability broadens the applicability of the models, ensuring they can effectively perform across a variety of contexts. By exposing the model to a wide range of scenarios and data distributions, diverse datasets enable the model to learn robust representations and patterns that transcend specific instances, thereby enhancing its ability to generalize and make accurate predictions in real-world scenario. Fig. 1 shows the steps to perform dataset encoding and machine learning analysis.

They introduce a unique rules-based technique to analyse company procedure footprints in the next section. By using a machine learner to find those intriguing regulations that have the greatest impact on the results of company procedure cases, our idea builds on previous methodologies centred around rule mining for associations and comparison items sets mine. Usually, indicators of success may be used to track system outputs. As a result, it can be seen trace-level indicators of success as trace-level characteristics that may be utilised as class variables in machine learning applications.

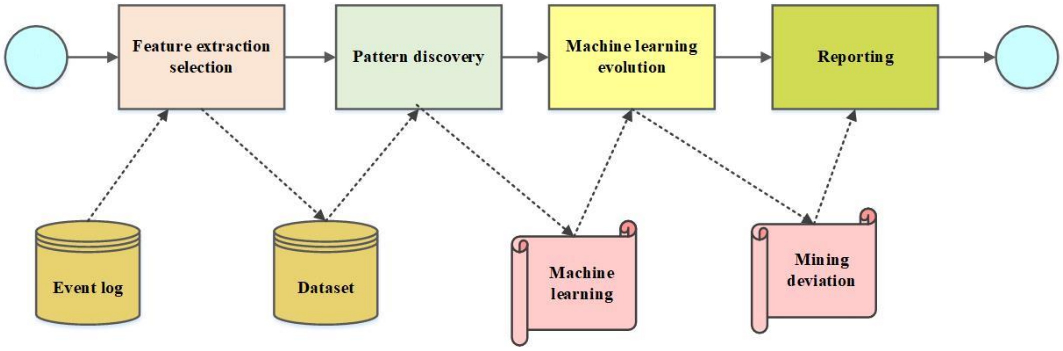

Fig. 1. Steps to perform dataset encoding and machine learning analysis.

It is crucial to bear in mind which (Process Efficiency Indicator) PPIs can be present at additional levels for abstractness in relation to business procedures, such as at the stage of activity, in which a particular task might be tracked from a PPI without consideration of the outcome of every other task carried throughout the same procedure. At the leadership threshold, it can be more important to keep track of a business' overall efficacy, which is often accomplished by combining the findings of a trace-level PPI. For instance, management is interested in monitoring a service level agreement that requires at least 95% of problems to be resolved within 24 hours, thus they would like to measure the amount of incidents resolved in less than 24 hours. This PPI identifies every procedure trail according to its finish time. It is an incorporation of a tracing-level PPI. Trace-level PPIs are of importance for the purposes of this work.

The idea behind machine learning is this, given a realisable choice, demonstrating the disparities among possibilities could prove more obvious than presenting every single event. As opposed to just listing the details of the present-day scenario, a machine learner quickly determines the critical aspects that most affect that circumstance.

A company's record of events might be transformed into a set of data for the purposes of machine learning. Following that, they describe a compression strategy that mines an archive of processes traces for features using declarative mining of processes. Declare's expressiveness is sufficient to record both basic count of each task and complex related to time interactions between pairs of activity. The idea is to use only one, condensed syntax in this manner to record both basic and complicated themes that could be present in an event log. As far that we comprehend, research has not yet investigated a declarative-language oriented encoding strategy to build vector illustrations of process occurrences.

TABLE I. SEQUENCE ENCODING

h_{id}	σ	boa			bigram		mrs		mra	
		x	*y*	*z*	*xy*	*xz*	*xy*	*xyz*	*x,y*	*X,y,z*
h_1	xyzxy	2	1	2	1	1	1	1	1	2
h_2	xyzx	2	1	1	0	1	0	0	1	2
h_3	xyzyzx	3	1	2	1	1	1	1	1	3

Table I is a non-exhaustive illustration of characteristics that may be retrieved given the occurrences of events inside the context of multiple activity traces that together make up a log of events P′. The illustration used known coding methods including bag-of-activities (boa), bigram, maximum repeat sequence (mrs), and maximal repeat alphabet (mra). Such encoding techniques track the frequency with which each encoding pattern is present in the process traces. The choice containing directly extracting tracing-level characteristics off the set h_{attr} of an operation trace and adding those into the collection of instance properties j_{attr} is also taken into account by our methodology. It is feasible to convert the incident log P′ to a datasets after extraction properties from a set of activity traces H ′ by transferring each h to H ′ to a dataset instance q, so that each q = $\left(h_{id}, h_{attr}, c_{name}\right)$, with $q_{1.....n} \in X^{'}$.

TABLE II. EVENT –LOG USING BOA ENCODING

h_{id}	**x**	**y**	**z**	**et**	**Pc (** c_{name} **)**
h_1	2	1	2	4.50	false
h_2	2	1	1	3.20	true
h_3	3	1	2	7.30	false

The incident log change example's information is shown in the Table II. Imagine the identical examples traced from before that additionally have additional trace-level characteristics: processing duration (et), in days, as well as effective process conclusion (pc), containing an integer categorization value of "True, False." This provides an illustration of how an event log may be completely transformed into a dataset. A finite a number of occurrences within each trace $h_{1..3}$ may then be encoded using an encoding approach, such as bag-of-activities. Four unique qualities (characteristics) were identified utilising the BOA technique taking into account the peculiarities of the aforementioned activity traces: a, b, and c. It is therefore feasible to create a dataset that includes the gathering of each of the event-driven & trace-level characteristics by taking into account both of the current trace-level characteristics, et and pc. In this manner, the procedure's control-flow and information properties may be examined to one another. In order to connect to the element which serves as the foundation for verifying deviant behaviour, the given name for an attribute of a class (c_{name}) has to be identified in the dataset. The term "c_{name}" must be used to identify a trace-level performance marker that is relevant for examination. False-valued (unsuccessful) footprints are regarded as aberrant instances in our scenario since the effective completion characteristic is specified as a class variable, c_{name} = pc.

A. Mining Declare Constraints as Trace-level Attributes

Compared to the currently used series encoding methods, they also suggest a fresh method employing logical process mining in order for extracting traits from periods of happenings. They took into account the Declaration programming syntax and its restriction examples, which offer the primary relationship and presence restrictions forms. They took into account the meaning of Declare restrictions using standard patterns included in both Unrestricted Miner and MINERFul++ declaratory mining algorithms with the goal to execute the discovery of limitations at the track levels. They must stay away from vacuously fulfilled restrictions since pattern fulfilments are the things that we want to engage in. To eliminate simply met restrictions, a different labelling collection of support automaton for vacuity detecting is suggested. In our search process, the comparable routine expressions used by the vacuity detecting support automata have been taken into account. Declarative syntax mining methods now in use seek to identify a collection of restriction patterns to describe the behaviour of a whole event record as one procedure paradigm. To determine if a restriction template is valid and meaningful, these techniques may take into account several threshold characteristics at the event log level, such as support, confidence, and interest factor. Through examining the achievement of a set Announce specifications for every step in the trace, they hope to employ Declare constraints as features at the trace level in this study. Similar to the previously discussed current encoding methodologies, those Declare-based attributes for each process trace may be used to create a collection of examples.

They use declaratory procedure mined approaches to identify whether Declaration requirements was satisfied in every programme tracing $h \in H$, provided an events log P. A number of Predefined limitations have to be established before mine can be done correctly. A Declaration restriction generators collection may represent all of it or a portion of it in this case. It then needs to be paired to a collection of unique occurrences that are recorded on the occurrence log. This occurrence set includes the parameters that Declaring requirement patterns require in order to function, while this mixture produces the collection of characteristics produced by this encode technique. By creating unique ordinary expressions, the list of default requirement templates may be expanded to include additional restrictions as appropriate. The label of the restriction example, that symbolises an abstraction of a restriction (at first used stated in LTL or via an ordinary expressions), plus a group of parameters are combined to form a Declare restriction d, where d = name ({args}). The total amount of parameters differs based on the pattern; for instance, a init restriction theme only requires a single query since it applies to the occurrence who initiates the trace's execution, but the coexisting restraint pattern requires two inputs because it applies whenever two occurrences occur in the same processes trail.

Considering the occurrence logging instance P ′ from earlier, that includes a collection of three separate occurrences (a, b, and c). Three Declaration requirements init (a), init (b), and init (c) can be produced from a Declaration restriction generator of class init. Every limitations, represented by "1" as a fulfilment or "0" alternatively, makes up as a trace-level attribute-value pairing in the sake of decoding by obtaining an amount matching to the Declaring condition's fulfilment. Common attribute-value pairings associated with the $init$ model, for instance, are as follows: $h'_{attr=((init(a),1),(init(b),0),(init(c),0))}$. The exactly_n model, which counts an exact n of instances of events within the entire track, corresponds to the lone alternative. Activity tracing containing Declare-based attribute-value pairings can then be converted into database objects in a manner similar to that shown in the table for boa coding. A typical dataset is shown in Table III and is made up of objects with characteristics that correspond to an example of Declaration restrictions obtained from the event log P ′. Declare-based characteristics may represent timing connections among actions in a manner that sequence- based set-based encoding methods can't, in contrast with other current encoding methods. For instance, the boa, bigram, mra, and mrs methods do not have an equivalent for the answer (b,c) restriction. Customised constraints for incident sequencing representations of features may nevertheless be defined. Declaring also offers a number of predefined templates that can handle a variety of timing connections between procedure incidents, which is a further advantage. Concerning methodology, each of the four rules may be represented with the current Announce limitation components.

TABLE III. EVENT LOG USING DECLARE ENCODING

h_{id}	Init(x)	Last(x)	Exactly(x)	Response(x,y)	et	pc
h_1	1	0	2	1	4.5	false
h_2	1	1	2	0	3.2	true
h_3	1	1	3	1	7.3	false

B. *Machine Learning Evaluation*

1) Standardized stream flow index (SSI): Similarly to indicators of severe weather, the majority of investigations used standardised criteria for assessing hydrologic dryness. Two significant standardised indices are flow indices and standardised runoff indices, both which have an analogous theoretical foundation. The sole difference between SSI computations and other computations is that run-off from the surface data are utilised in place of precipitation data. For example, this index displays correct beta dispersion. As a result, for each month, the total flow values are separately estimated before the SSI is computed.

2) Gene Expression Programming (GEP): Genetics can be made using genetic algorithms in the Gene Expression Programming (GEP) algorithm, which uses communities of people and selects these according to fitness. The GEP method's initial step is to create a main collection of answers. This level can be finished by an unintentional procedure or by using some knowledge about the issue. The chromosomal structures were then visualised as a tree expression and evaluated using a fitting method. In general, processing a number of target issues, also known as fitting problems, allows for the evaluation of the appropriate function. The research process ends and the most effective resolution is determined once the answer has an appropriate standard or if a certain number of iterations have passed. The most suitable response form the latest generation is maintained if the most favourable scenario cannot be discovered, and the remaining options are left to be chosen from. The best people are more likely to have children, based on the decision. For many generations to come, every step has been repeated, and it is anticipated the group in question quality will generally increase as new generations are born. GEP chooses the candidates using the renowned roulette wheel approach. In contrast to genetic algorithms and genetic programming, GEP uses a number of genetic operatives to reproduce modified people. Replica is a procedure designed for preserving a few of the most talented members of this era into the following one. A mutation operator's objective is to insert arbitrary changes into an individual chromosome. To avoid producing people deemed rule-deficient, this operator conducts some of the perfect procedures. GEP employs a one-point and two-point combination, similar to a biological algorithm. The genomic equivalent problem (GEP) employs a single-point and two-point combinations. The kind of two-point combo is a little more intriguing because it can largely switch on and off the chromosomal regions that are not encoded. Additionally, the GEP also performs a different kind of combining known as gene combination, in which genes are entirely combined. To create two new children, this operator randomly chooses genes on both-parent chromosomes that are located in the same location.

C. Support Vector Regression

Over the following decades, Support Vector Machine (SVM) evolved into a linear classification algorithm using optimum hyper plane concept. Utilising statistical learning theory, this approach is used. Additionally, they utilised kernel algorithms to create nonlinear classifications. SVM's classification algorithm serves to categorise problems associated with data into multiple classes, while its regression technique is applied to solve prediction issues. Regression on fit data produces a hyper plane. A given location's deviation from its hyper plane revealed the inaccuracy of that location. The most effective technique for regression analysis is advised is the least squares approach. But it can happen that using a least-square estimation for analysis issues in the form of outliers may not be entirely rational, which would lead to the analysis performing poorly. In order to avoid bad performance that is not responsive to minute modifications to the model, a robust estimator should be created. As mentioned, the SVM is built upon the principle of minimising risk, a hierarchy generated by the theory associated with statistical training. a distance from real values termed an error function to employ SVM in regression issues that overlook mistakes in a -insensitive manner. This function's definition translates as follows in Eq. (1) and Eq. (2):

$$P(a, f(d,y) = \left|a - f(d,y)\right|_{\varepsilon} \tag{1}$$

$$= \begin{cases} 0 \;\; for \left|a - f(d,y)\right| \le \varepsilon \\ \left|a - f(d,y)\right| - \varepsilon \; if \; \left|a - f(d,y)\right| > \varepsilon \end{cases} \tag{2}$$

Below, this mistake function does not take into account errors.

D. M5 Model Tree

This technique is an amalgam of machine learning and data mining techniques. Data mining techniques identify several, suitable frameworks before extracting data from a pool of set values. Because data mining techniques differ from statistical approaches because they were established for huge datasets with multiple variables, they were created for smaller datasets with fewer variables. Among the most popular data mining approaches, decision tree-based methods use input data to forecast or categorise target qualities as an output in the shape of an equation having a structure of trees. The M5 modelling trees are a structure of choices that may be utilised for forecasting continuous quantitative qualities. Its branches are representations of regression operate, and it has lately sparked a substantial development in classifications and predictions. When contrasted to other theories, the tree algorithm's data has higher precision and is simpler to replicate and comprehend. A tree of choices is composed of four components: the root, the branch, the nodes, and the leaves. The rectangular shape denoted each node, while the connections between them were shown as branches. The tree of choices usually goes from left to right or from top to bottom, with the base (first node) on the very top to make it easier to create. The leaf denotes the conclusion of a series of events. For the reason of minimising the total of the squared variances from the average information for each node, splitting is carried out by one of the predictive variables. Utilising the splitting criterion is the first step in creating a tree model. The M5 algorithm's dividing criteria relies on the accuracy of the usual variation of the numbers acquired in every node that correspond to every class or subcategory. In a consequence of checking every characteristic at that node, dividing criteria determines the amount of erroneous for that component and determines the smallest predicted error type. In most circumstances, the predictive inaccuracy is determined by assessing how well the desired outcomes for hypothetical cases are predicted. SDR, or standard deviation reduction, is given in Eq. (3).

$$SDR = sd(H) - \sum \frac{\left|H_i\right|}{\left|T\right|} sd(H_i) \tag{3}$$

The total number of specimens approaching all nodes is shown by H, and H_i is the portion of examples which correspond to the nth outcome of a possible test. sd stands for standard deviation. Up till reaching the final cluster (the leaf), the method of division is repeated multiple times at every node. So when it reaches the leaves, the total of the squared differences above the average information is virtually zero. The consequence is going to be the growth of a huge tree. Using numerous limbs and nodes, it is going to difficult to operate using this large tree; as a result, undesirable branches must be removed to create an ideal and effective tree. There are a total of two ways to prune: (1) while the plant forms its full potential, (2) trimming following the peak of shrub development. The second strategy begins by forming the largest possible tree before beginning the trimming manipulate, unlike the initial method, which prevents the tree from growing further branching. Choosing the best branch is dependent on reducing errors in prediction.

E. Evaluation parameters

The root mean square error (RMSE) (4), relative absolute error (RAE) (7), mean absolute error (MAE) (5), and correlation coefficient (CC (6)) were used to analyse the error values between the anticipated and observed data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - a_i)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|d_i - a_i\right| \tag{5}$$

$$CC = \frac{(\sum_{i=1}^{n} d_i a_i - \frac{1}{n}\sum_{i=1}^{n} d_i \sum_{i=1}^{n} a_i)}{(\sum_{i=1}^{n} d_i^2 - \frac{1}{n}(\sum_{i=1}^{n} d_i)^2)\,(\sum_{i=1}^{n} a_i^2 - \frac{1}{n}(\sum_{i=1}^{n} a_i)^2)} \tag{6}$$

$$RAE = \frac{\sum_{i=1}^{n}\left|a_i - d_i\right|}{\sum_{i=1}^{n}\left|d_i - \bar{d}\right|} \tag{7}$$

When n represents the total amount on assessments, and xi, yi are the anticipated & observed results of the SSI. Complete correlation (CC) among measured and anticipated numbers. Correlation that is direct is shown by values that are positive, and the opposite relationship is indicated by negative values. Additionally, the RMSE and MAE values are errors, therefore smaller values suggest lesser modelling mistakes.

V. Results and Discussion

The effectiveness of the three models—SVM, GEP, and M5—in projecting the Standardised It Index utilising the SPI and SPEI indices at Navrood station throughout six-time delays (a one-month to six-month) is examined in the current work. A 48-month grade was chosen for investigation in this study out of the several scales for predicting SSI since it had a stronger correlation and was predicted by the mathematical models that were provided. The statistical characteristics of the drought indices used in the research region are shown in Table IV.

TABLE IV. Statistical Characteristics of the Utilized Data

SSI	skewness	coefficient of variation	standerd deviation	maximum	minimum	mean	variable
0.69	0.13	958.7	0.98	1.98	-2.09	0.0011	SPI
0.69	-0.69	19.098	0.99	1.45	-2.023	-0.054	SPEI
1	0.08	530	0.99	1.98	-1.67	0.003	SSI

The Fig. 2 shows the Root Mean Square Error (RMSE) values for three machine learning algorithms: GP, M5, and SVR. Each row represents a different evaluation scenario or experiment. The values indicate the accuracy of the algorithms, with lower RMSE values indicating better accuracy. Based on the table, GP consistently has the lowest RMSE values across different scenarios, suggesting it performs better than M5 and SVR in terms of accuracy.

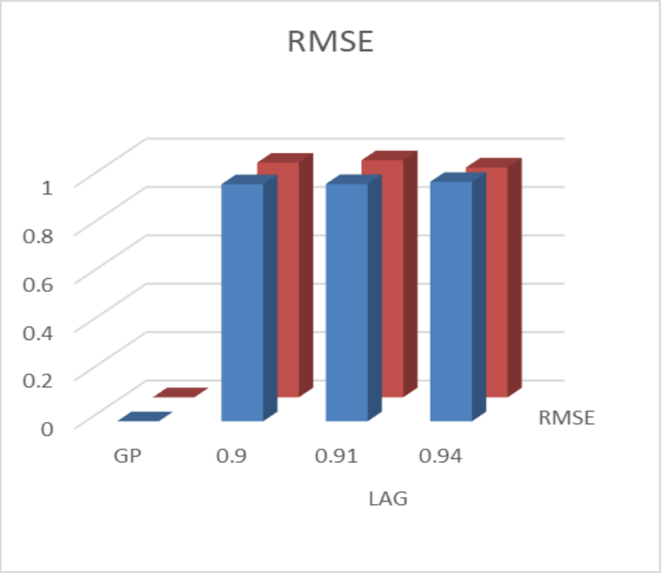

Fig. 2. RMSE model.

The Fig. 3 represents the Mean Absolute Error (MAE) values for three machine learning algorithms: GP, M5, and SVR. Each row corresponds to a different evaluation scenario. MAE is a metric used to measure the average absolute difference between the predicted and actual values, where lower values indicate better accuracy. Based on the table, GP consistently has the lowest MAE values across different scenarios, indicating it performs better in terms of accuracy compared to M5 and SVR.

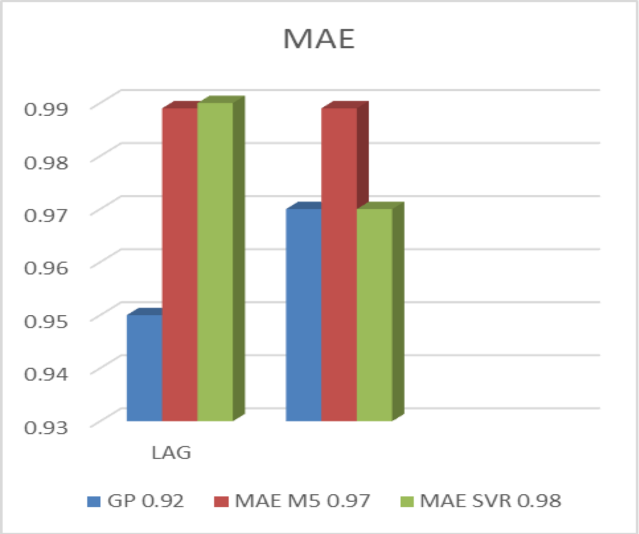

Fig. 3. MAE model.

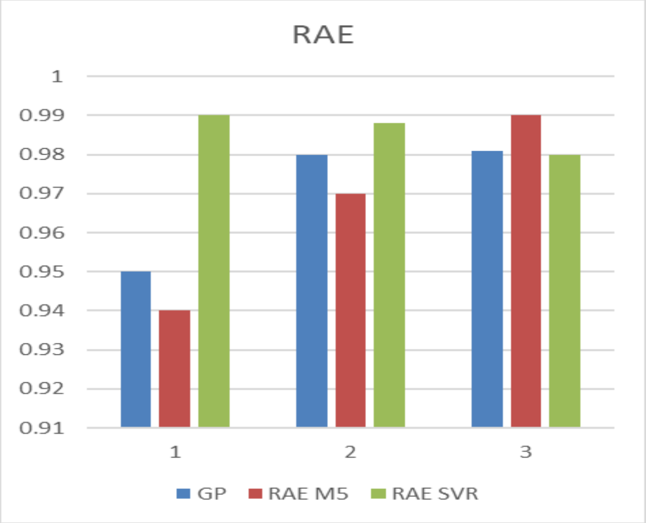

Fig. 4. RAE model.

The Fig. 4 shows the Relative Absolute Error (RAE) values for three machine learning algorithms: GP, M5, and SVR. Each row represents a different evaluation scenario. RAE is a metric used to measure the relative difference between the predicted and actual values, indicating the performance of the algorithms in relation to the magnitude of the target variable. Lower RAE values indicate better accuracy. Based on the table, GP generally has lower RAE values across different scenarios, suggesting it performs better in terms of accuracy compared to M5 and SVR in relation to the magnitude of the target variable.

Additionally, determined by Pearson's correlation but cross-correlation, it was determined that the USDA hydrological dryness index was better and predicted with a smaller error even though the drought index is better and more dependent on climatic circumstances.

TABLE V. PERFORMANCE METRICS COMPARISON

Method	Standard Deviation
SVM [20]	0.9
MLP[21]	0.6
MLP[22]	0.6
Proposed Method	0.9

A comparison of performance measures, namely standard deviations, for various approaches is shown in Table V. The Support Vector Machine (SVM) shows performance variability with a standard deviation of 0.9. Two presentations of the Multilayer Perceptron (MLP) technique are made; in both cases, the standard deviation is 0.6, indicating higher consistency in performance as compared to SVM. Interestingly, the suggested technique, which makes use of Support Vector Regression (SVR), has a standard deviation of 0.9, which is consistent with the variability shown in SVM. These measures provide insights into the stability and reliability of each method, with lower standard deviations often reflecting more consistent performance.

VI. CONCLUSION

The research contributes significantly to the domain of commercial operation divergence analysis, offering valuable insights into the identification of variations in commercial systems beyond anticipated outcomes. By delving into the characteristics of procedure executions, the study illuminates' behaviours impacting process efficiency, encompassing both detrimental and optimal aspects. Success in this context is gauged through domain-specific efficiency metrics, encompassing cost-effectiveness, time optimization, and resource utilization. Users may have concerns regarding the dependability and efficacy of machine learning models in detecting and mitigating threats in applications like cyber security or automated threat detection. It's critical to fully assess the models' performance using real-world data and stringent testing protocols in order to address this. Moreover, adding human supervision to the machine learning procedure might offer still another level of security. Clearly defined procedures for human evaluation and intervention, particularly in crucial decision-making roles can guarantee responsibility and reduce the hazards connected with automated systems. The paper introduces an innovative decoding strategy that utilizes Declare constraint templates, enabling more expressive treatments through vector-based representations of procedure scenarios. Additionally, the research pioneers the application of Machine Learning, incorporating algorithms like Standardized Stream flow Index, Gene Expression Programming, Support Vector Regression, and M5 Model Tree within the realm of Deviance Mining. This approach effectively identifies the aspects of a procedure significantly influencing its efficiency, surpassing traditional trend mining methods to handle intricate linkages within highly variable systems. The experimental outcomes underscore the efficacy of machine learning when integrated with the proposed Declare-based coding. Analysing event logs through this approach yields pertinent and insightful conclusions, offering a comprehensive understanding of process behaviour and performance. While acknowledging these contributions, it is crucial to recognize the limitations of the current study, such as the specific contextual constraints and the need for further validation across diverse industry scenarios. Future work in this field by iteratively validating the model's performance across various data partitions, cross-validation provides a more robust estimate of its generalization ability, helping to identify and address over fitting issues before deployment in real-world scenarios. This research sets the stage for practical and effective tools in process analysis, empowering organizations to make informed, data-driven decisions for optimizing efficiency, reducing costs, and enhancing overall performance.

REFERENCES

[1] M. A. Semin and L. Yu. Levin, "Stability of air flows in mine ventilation networks," Process Saf. Environ. Prot., vol. 124, pp. 167–171, Apr. 2019, doi: 10.1016/j.psep.2019.02.006.

[2] Y. Yuan, H. Cao, Y. Zhang, Q. Xie, and R. Yao, "Outlier Mining Based on Neighbor-Density-Deviation with Minimum Hyper-Sphere," Inf. Technol. Control, vol. 45, no. 3, pp. 267–277, Sep. 2016, doi: 10.5755/j01.itc.45.3.13164.

[3] J. A. Botín and M. A. Vergara, "A cost management model for economic sustainability and continuos improvement of mining

operations," Resour. Policy, vol. 46, pp. 212–218, Dec. 2015, doi: 10.1016/j.resourpol.2015.10.004.

[4] J. Von Der Goltz and P. Barnwal, "Mines: The local wealth and health effects of mineral mining in developing countries," J. Dev. Econ., vol. 139, pp. 1–16, Jun. 2019, doi: 10.1016/j.jdeveco.2018.05.005.

[5] P. Zerbino, D. Aloini, R. Dulmin, and V. Mininno, "Process-mining-enabled audit of information systems: Methodology and an application," Expert Syst. Appl., vol. 110, pp. 80–92, Nov. 2018, doi: 10.1016/j.eswa.2018.05.030.

[6] Y. Xu, T. Li, X. Tang, X. Zhang, H. Fan, and Y. Wang, "Research on the Applicability of DInSAR, Stacking-InSAR and SBAS-InSAR for Mining Region Subsidence Detection in the Datong Coalfield," Remote Sens., vol. 14, no. 14, p. 3314, Jul. 2022, doi: 10.3390/rs14143314.

[7] F. Taymouri, M. L. Rosa, M. Dumas, and F. M. Maggi, "Business process variant analysis: Survey and classification," Knowl.-Based Syst., vol. 211, p. 106557, Jan. 2021, doi: 10.1016/j.knosys.2020.106557.

[8] I. Bagińska, M. Kawa, and W. Janecki, "Estimation of spatial variability of lignite mine dumping ground soil properties using CPTu results," Stud. Geotech. Mech., vol. 38, no. 1, pp. 3–13, Mar. 2016, doi: 10.1515/sgem-2016-0001.

[9] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," Ore Geol. Rev., vol. 71, pp. 804–818, Dec. 2015, doi: 10.1016/j.oregeorev.2015.01.001.

[10] H. Demolli, A. S. Dokuz, A. Ecemis, and M. Gokcek, "Wind power forecasting based on daily wind speed data using machine learning algorithms," Energy Convers. Manag., vol. 198, p. 111823, Oct. 2019, doi: 10.1016/j.enconman.2019.111823.

[11] Z. Zhu, N. Anwer, Q. Huang, and L. Mathieu, "Machine learning in tolerancing for additive manufacturing," CIRP Ann., vol. 67, no. 1, pp. 157–160, 2018, doi: 10.1016/j.cirp.2018.04.119.

[12] J. Evermann, J.-R. Rehse, and P. Fettke, "Predicting process behaviour using deep learning," Decis. Support Syst., vol. 100, pp. 129–140, Aug. 2017, doi: 10.1016/j.dss.2017.04.003.

[13] J. Evermann, J.-R. Rehse, and P. Fettke, "A Deep Learning Approach for Predicting Process Behaviour at Runtime," in Business Process Management Workshops, vol. 281, M. Dumas and M. Fantinato, Eds., in Lecture Notes in Business Information Processing, vol. 281. , Cham: Springer International Publishing, 2017, pp. 327–338. doi: 10.1007/978-3-319-58457-7_24.

[14] C. D. S. Garcia et al., "Process mining techniques and applications – A systematic mapping study," Expert Syst. Appl., vol. 133, pp. 260–295, Nov. 2019, doi: 10.1016/j.eswa.2019.05.003.

[15] P. H. P. Richetti, L. S. Jazbik, F. A. Baião, and M. L. M. Campos, "Deviance mining with treatment learning and declare-based encoding of event logs," Expert Syst. Appl., vol. 187, p. 115962, Jan. 2022, doi: 10.1016/j.eswa.2021.115962.

[16] T. Al-Shehari and R. A. Alsowail, "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques," Entropy, vol. 23, no. 10, p. 1258, Sep. 2021, doi: 10.3390/e23101258.

[17] J. J. Roldán, E. Crespo, A. Martín-Barrio, E. Peña-Tapia, and A. Barrientos, "A training system for Industry 4.0 operators in complex assemblies based on virtual reality and process mining," Robot. Comput.-Integr. Manuf., vol. 59, pp. 305–316, Oct. 2019, doi: 10.1016/j.rcim.2019.05.004.

[18] E. Helm, A. M. Lin, D. Baumgartner, A. C. Lin, and J. Küng, "Towards the Use of Standardized Terms in Clinical Case Studies for Process Mining in Healthcare," Int. J. Environ. Res. Public. Health, vol. 17, no. 4, p. 1348, Feb. 2020, doi: 10.3390/ijerph17041348.

[19] S. Weinzierl et al., "An empirical comparison of deep-neural-network architectures for next activity prediction using context-enriched process event logs," 2020, doi: 10.48550/ARXIV.2005.01194.

[20] K. Ghanem, F. J. Aparicio-Navarro, K. G. Kyriakopoulos, S. Lambotharan, and J. A. Chambers, "Support Vector Machine for Network Intrusion and Cyber-Attack Detection," in 2017 Sensor Signal Processing for Defence Conference (SSPD), London: IEEE, Dec. 2017, pp. 1–5. doi: 10.1109/SSPD.2017.8233268.

[21] A. Nusret Özalp and Z. Albayrak, "Detecting Cyber Attacks with High-Frequency Features using Machine Learning Algorithms," Acta Polytech. Hung., vol. 19, no. 7, pp. 213–233, 2022, doi: 10.12700/APH.19.7.2022.7.12.

[22] T. T. Teoh, G. Chiew, E. J. Franco, P. C. Ng, M. P. Benjamin, and Y. J. Goh, "Anomaly detection in cyber security attacks on networks using MLP deep learning," in 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Jul. 2018, pp. 1–5. doi: 10.1109/ICSCEE.2018.8538395.