Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Comparative Analysis and Mining of Unstructured Datasets from an Empirical Standpoint

Pragya Lekheshwar Balley[1], Dr. Shrikant V. Sonekar[2]

[1]PGTD of Computers, RTMNU Nagpur, pragyaballey26@gmail.com
[2]Professor, Department of CSE, J D College of Engineering and Management Nagpur

Unstructured data is stored in NoSQL type databases, and is used to represent real-world data samples. Because of its unstructured characteristics, the complexity of storage and retrieval is very low, which makes is useful for high scalability application scenarios. But as the scale of data increases, processing & analysis of this data becomes exponentially complex under real-time scenarios. Due to this, researchers have designed a wide variety of low-complexity data-processing methods, which can be applied to multiple NoSQL scenarios. Context-specific methods are needed for data cleaning, reduction, transformation, model building, and pattern mining processes. Due to such a wide variation in data processing models, it is difficult for system designers to identify optimal methods for their context-specific use cases. To overcome this ambiguity, and assist readers in identification of highly efficient data mining techniques, this text discusses design of recently proposed mining techniques that can be applied to unstructured data sets. Based on this discussion, readers will be able to evaluate the reviewed models in terms of their contextual nuances, functional advantages, application-specific drawbacks, and deployment-specific future scopes. This discussion is further extended by a comparison of the reviewed models in terms of processing efficiency, delay of operation, complexity of deployment, cost of deployment and scalability metrics. This will allow readers to identify optimally performing models for different performance-specific use cases. Based on these evaluations, a novel NoSQL Data Mining Rank Metric (NDMRM) is calculated, which combines the evaluation metrics in order to identify optimally performing NoSQL Data Mining Models under real-time scenarios.

Keywords: Data, Mining, NoSQL, Unstructured, Complexity, Delay, Efficiency, Cost, Complexity, Scalability, Scenarios.

1. Introduction

Enterprises produce and employ enormous amounts of unstructured data, including audio, video, and animation files. To get insights from this massive amount of unstructured data, analysis and handling are essential. Data mining is one such alternative, which makes use of a wide variety of tools and techniques to extract knowledge from both organised and unstructured data. Due to their pre-set format, structured data are simpler to analyse than

unstructured data. Thanks to technical advancements, a number of data mining platforms, like Talk-Walker Analytics, Orange, and RapidMiner, can now quickly analyse previously unstructured data. In the field of computer science called "data mining," enormous data sets are analysed to discover hidden links and patterns. Data mining is a method for extracting potentially useful information from vast volumes of unstructured data. It is feasible to combine data from several sources. The Internet, libraries, archives, databases, and data warehouses are a few examples.

Knowledge Discovery in Data is another name for data mining, which seeks to find hidden links in large datasets (KDD). It's conceivable that data mining might provide some insight on the reasons behind patterns seen in the target dataset. Large amounts of information are collected and stored by modern organisations from many different sources. Prior to starting the model building process, it is difficult to extract crucial data due to the massive number of big datasets. Data mining is a method used by businesses to sift through enormous volumes of data in search of useful patterns. Specialised personnel are often in-charge of cleaning data before analysis in firms. Time and money are needed when system issues appear or wrong inferences are made as a consequence of incomplete or missing information sets. Developers must thus use a variety of data cleansing strategies, each of which must take the organization's cost constraints into mind. Cleaning data involves a number of time-consuming processes that may be carried out sample by sample, such as adding missing data, removing duplicates, etc. Data reduction in action may be seen in the limitation of the amount of data shown.

Examples of data reduction techniques used to portray datasets more succinctly include dimensionality reduction and numerosity reduction. Both data minimization and improved security are achieved with the help of these solutions. Data minimization, however, has little impact on the outcomes of data mining. Engineers convert data formats as part of the data transformation process. The raw data is normalised to make it simpler to get when needed. Data mapping may incorporate extra approaches in addition to more traditional data science processes like aggregation, standardisation, and discretization. The creation of models for different forms of unstructured data may start after the transformation is finished.

To find patterns or trends in your data, utilise association rules, correlations, or other techniques. The classification of datasets based on similarity is made possible by the application of deep learning algorithms. Machine learning techniques like random forests and decision trees may be utilised for classification if the data has been labelled. You may use clustering techniques like K-means and DBSCAN as well as centroid-based, distribution-based, and density-based algorithms even if your data are unlabelled. As shown in figure 1, data must be processed and comprehended after collection from different sources. These conclusions must be transparent and unequivocal in order for businesses to implement new strategies and accomplish their goals.

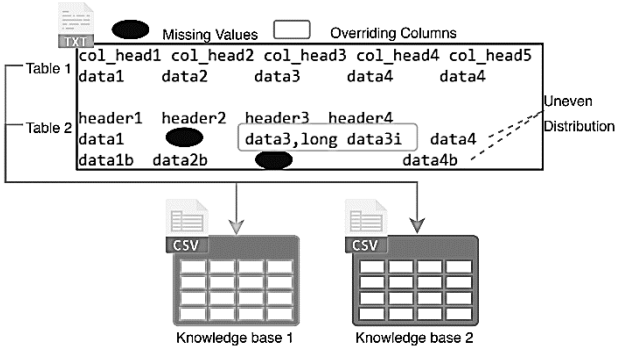

Figure 1. Design of a cross-domain knowledge data mining scenario for real-time use cases

The current system's rich logs and trace messages may be used to observe this behaviour during runtime. The intricacy and scalability of modern programmes result in a huge number of metadata being produced [1]. These unstructured log messages may be examined and processed for usage in applications including anomaly detection, cognitive management, autonomous issue validation, root-cause analysis, incident management, and management [2, 3, 4]. However, direct analysis may sometimes be challenging owing to a lack of language and/or hierarchical organisation. Print statements such as logs and traces may be beneficial in this situation. The process of extracting useful information from unstructured log messages has received a lot of attention in recent years. Due to the lack of automated frameworks for domain-independent knowledge base synthesis, human domain-specific knowledge base production predominates the literature [8].

It could be difficult to build a knowledge base for a particular topic in a large-scale system, however. There may be significant variation in the development of various telecom industry components even when logging standards like BSD1 are used. In a large-scale system, Radio Base Stations (RBS) produce a huge volume of log messages each hour. It is likely that the components' drastically differing designs may be attributed to the fact that they were separately created by several engineers using various programming languages. Explicit programming is required to create knowledge bases from these logs, and this is only possible if the programmer has a thorough understanding of how to recognise information-containing structures and remove unneeded material. Information extraction must be performed often to account for changes to rules or standards, even when it has been automated using bespoke parsers. The findings show that a supervised approach cannot be used to extract all significant data from a single large area. Several algorithms and methods for extracting unstructured data are covered, including ones that utilise semi-supervised or unsupervised learning to find correlations in template extraction use cases.

As a consequence, several researchers have developed fundamental data-processing methods that may be used in many NoSQL scenarios. Data cleansing, reduction, transformation, model building, and pattern mining are examples of tasks that need context-specific methodologies. It may be difficult for system designers to pinpoint the best practises for use cases that are unique to certain environments because of the diversity of data processing methodologies. By describing how data mining techniques have developed through time and how they may be used to analyse unstructured data sources, this article can dispel any misunderstandings. The

last section investigates possible comparison standards for these models. The models under examination are then given a short historical overview, followed by suggestions for enhancing their performance in real-world scenarios.

2. Literature Review

A wide variety of data mining & analysis models are proposed by researchers for processing unstructured data samples. For instance, According to research proposed in [1], the amount of data available and being collected by e-learning platforms, such as those used by massive open online course (MOOC) providers, is growing at a rate that has never been seen before. This means that there are many opportunities to use these data in a structured way to make decisions and improve education. Using data automation to figure out if a child is likely to drop out of school could be a useful way to keep students in school. Sequence mining looks like a good set of techniques to automatically pull out useful information from user activity data, since the data is based on time sequences. But there isn't a lot of information about how different strategies relate to each other or about how to use sequence mining in education in general. researchers are free from these limits because of two main things researchers have done. First, researchers show how sequence classification can be used to predict MOOC dropouts. The framework has two dropout definitions (TDD) that are based on data, as well as rules for how to properly format and prepare data for use and for training dropout prediction models at the best times during the course. Second, researchers do a benchmarking study of state-of-the-art sequence classification approaches. researchers test different parameterizations on 47 real-world datasets from MOOCs and compare and contrast over 18 thousand models. their research shows the main differences between the methods, so researchers can give good advice on how to set the hyperparameters that have the biggest effect on how accurate predictions are.

The results of a study reported in [2] show that data analysis tools are becoming more important as big data and cloud computing grow, which creates large market values. Some association rule mining tasks could be done in the cloud by people who don't have a lot of computing power. Owners of information should be aware that sensitive data could be lost or stolen during this process. Before sending their raw data, data owners can encrypt it to make sure that no sensitive information gets out during the outsourcing process. Researchers have been thinking about how to decipher encrypted data for a few years now. One way to get around this problem is to use homomorphic encryption, which is a type of cryptography. Data that has been encrypted can be used without first having to be decrypted. Academics are very interested in homomorphic encryption methods, which allow private data mining in a multikey context. In this paper, researchers describe a new homomorphic cryptosystem (HC) that lets a lot of cloud users each have their own private key. researchers also suggest a "twin-cloud" (TC) architecture for offloading data uploads that uses an "association rule mining" method that protects privacy. Based on real-world events, their method uses a way to show shopping mall transactions in databases. their tests with a real-world database of transactions show that their approach is pretty good.

Studies in [3] showed it more clearly. People use software all the time in their daily lives. As you get better, you'll want to know more about what makes people successful. researchers can

meet all of the Google Play Store's requirements because researchers use classifiers. For statistical analysis of things like hypothesis testing, correlation, and regression metric analysis, KNN and Stochastic Forest were used to run application regressions. Using KNN and Stochastic Forest (kNN SF) regression techniques, the goal of this research is to build inference engines that can predict how an application will be rated. The Stochastic Forest method gave better results than the KNN method.

There is more information about this subject in [4]. Smart technologies are becoming more and more important in the field of education. The ever-growing amount of school-related data may make the usual ways of processing data ineffective or skewed. So, reconstructive data mining has become more popular in the field of education as a way to study. This research uses clustering, discriminating, and convolution neural network theories to analyse and predict how well students will do in school in the future. This is done to avoid unfair assessment results and keep track of how well students will do in the future. This research first suggests using a statistic that has never been used in the K-means approach before. This is so that the clustering number can be found as accurately as possible. Then, discriminant analysis is used to figure out how well the K-means clustering worked. It is possible to train and test convolutional neural networks with data that has been labelled. The model that was made can be used to make predictions. Lastly, the created model is tested with two metrics and two cross-validation strategies to make sure that the predictions are correct. The results of the experiment show that using the statistic makes the Improved K-Means Algorithm (IKMA) more predictable and makes it easier to estimate the number of clusters in a way that is both quantitative and objective.

In [5], a new framework is made for data mining that protects privacy. It is based on Multi Party Computation (MPC) and safe sums. In contrast to traditional MPC methods, which use a small number of aggregation peers instead of a central trusted entity, the current study proposes a distributed solution that uses a single server to store intermediate results and integrates all data sources into the aggregation process. In a large-scale situation, the possibility that data could become unavailable during the aggregation process is looked into. It is being closely looked at here because it could be caused by things like ad hoc network connections or sudden user logouts. For a more reliable system, researchers organise data sources into rings and have each ring member contribute to the process of aggregation. The probability that data mining results won't be precise enough to fulfil expectations is modelled analytically for two proposed protocol systems. Details are also given about how hard it is to process, how much it costs to communicate, and how anonymous each method is. The new protocols are then used in a small number of use cases to show how they work and what they can do.

Researchers have found a link between where hydrocarbon reservoir zones are and where it is safe to mine coal and where there are faults, cavities, and pinch-outs in the ground [6]. Discontinuities in the subsurface create diffractions that carry important information and make it easy to see these geological features. The low amplitude of a diffraction makes it easy for specular reflections to hide it. The low-rank method works well to get diffraction out of specular reflections. The results of traditional rank-reduction methods, on the other hand, are very sensitive to noise. This affects the next image because as noise gets worse, the strength of low-rank operators goes down. Together, the low-rank assumption of seismic recordings

and parameterized nonconvex penalty functions (LRA NCPF) are used to propose an improved diffraction-separation strategy that improves the quality of separation. arctangent penalty functions are used as regularisation factors in this unique method to get a good approximation of a low-rank matrix. The proposed method can be used to filter reflected energy out of data and pull diffracted energy out of noisy data. Synthetic and field samples are used to show that the new method can get high-quality diffractions, which helps find and reveal geological discontinuities below the surface.

Based on what was found in [7], it is now clear that the probabilistic c-means method (PCM) is an important fuzzy clustering method that has many uses in data mining, pattern recognition, image analysis, and knowledge discovery. But PCM has trouble clustering a lot of data, especially if the data are not all the same, because it was mostly made for small structured datasets. In this study, High-order Principal Component Analysis (HOPCA) is proposed as a solution to this problem by maximising the objective function in tensor space during the clustering of large datasets. On top of that, researchers build a MapReduce-based distributed HOPCM method for a lot of mixed-type data. researchers made a privacy-preserving HOPCM algorithm (PPHOPCM) by combining the BGV encryption method with HOPCM. This was done to protect private information in the cloud. PPHOPCM uses polynomial approximations of the functions for updating the membership matrix and clustering centres, which makes it safe to use the BGV method. Experiments show that PPHOPCM may be able to group together a lot of different types of data using cloud computing without revealing any private data samples.

In the cited study [8], a promising strategy is given for classifying the content of spaceborne radar images based on how they will be used. Compared to the most common methods for machine learning today, this one is an interesting change. In the sections that follow, we'll talk about how to explain how hard a data mining task is that doesn't require any prior knowledge: putting satellite images into groups without using any supervised features and based on a set of predefined classes. In particular, researchers talk about how important un-supervision, feature independence, and being able to explain things are. As more businesses make protecting customers' "right to explanation" a top priority, the need for machine learning models that can give explanations has gone up. The importance of feature-free classification comes from the fact that making basic features for complex pictures is hard and that using different features leads to different classification results. Researchers have suggested using unsupervised discovery methods to get around the problems of using features alone and not having enough data that has been labelled. This paper shows how one of the most popular unsupervised probabilistic methods, the efficient Latent Dirichlet allocation (eLDA) model, can be used to find the hidden structure in synthetic aperture radar data. Using LDA as a way to mine data that can be explained, the goal is to find meaningful semantic links that scientists can understand. The method's suitability as an explainable model is looked at, and interpretable subject representation maps are made to really show what "interpretability" means in an explainable machine learning paradigm. LDA looks at the data as a group of themes to find the hidden patterns. Then, researchers use these themes to make data visualisations that are easy to understand and figure out how common each theme is across different types of land cover. their research shows that each class has a certain mix of subjects that make it unique. Second, researchers could divide these groups into groups based on how similar their topics

are. Subject matter experts can describe the topics' contents and put them into groups.

Studies cited in [9] say that to get actionable insights from a large amount of data, it takes sophisticated methods and a complex architecture to analyse and process the data. Due to the semantics and classifications used by the processing models, it might be hard to see the collected data for real-time solutions. In this paper, researchers propose the fuzzy-optimized data management (FDM) strategy for classifying and improving coalitions based on information-based semantics and constraints. Data dependencies are put into different groups based on how the relationships between the attributes of the data are modelled. This method divides the attributes being evaluated into groups based on the limits of the similarity index. This makes it possible to process complex data quickly. A real-time weather forecast dataset made up of sensor data (what was seen) and image data is used to test how well the proposed FDM works (captured). Using this set of data, the FDM can do semantic analyses on the information it gets and put things into groups based on how similar they are. Classification, processing time, and similarity index are looked at for a variety of data sizes, classification instances, and dataset records. In different classification scenarios, the recommended FDM has been shown to cut processing time by 36.28 percent and increase similarity indices by 12.5 percent.

In [10], the authors suggest a system architecture for building a large knowledge network out of a lot of scholarly information, finding the meta routes that connect the entities, and ranking the entities by how important they are. The private and safe Amazon EC2 platform is the core of the system. YARN and Hadoop HDFS are used to manage computing resources efficiently. The open-source Spark framework is used to analyse data in parallel, and links between different types of entities are found in a decentralised way. On top of this framework, researchers offer four connection discovery tasks: suggesting venues for papers, finding possible collaborators, comparing venues, and suggesting sources to cite. researchers recommend the mixed and weighted meta path (MWMP) method for relationship mining jobs. This method can be used to look for possible connections between many different types of entities. Clusters made with Amazon EC2 platform sets were used to test how accurate the method was and how fast it could parallelize for different scenarios.

Studies in [11] say that mining for high-utility sequential patterns (HUSPs) is a growing problem in the field of finding information in databases. In particular, it involves finding what are called "HUSPs," which are parts of sequences that are very useful. HUSPs can be used in the real world for things like route planning, click-stream analysis, market basket analysis, and coming up with ideas for e-commerce. More than one method has been made for mining useful sequential patterns quickly and based on how useful they are. Due to the combinatorial growth of the search field for low utility thresholds and large amounts of data, these methods don't make the best use of time and memory. So, this study suggests that HUSP mining by UL-list is a good way to go about it (HUSP-ULL). It quickly finds HUSPs with the help of a lexicographic q-sequence (LQS) tree and a utility-linked (UL) list structure. HUSP-ULL also gives two pruning methods for getting accurate upper bounds on the values of the candidate sequences and for narrowing the search space by getting rid of unfavourable candidates early on in the search process. HUSP-ULL has been shown to be faster and more accurate than state-of-the-art algorithms on a number of benchmarks, such as the identification of real-world and artificial HUSPs.

As suggested in [12], people who are good at racquet sports like tennis and badminton use tactical analysis to figure out how their opponents play on the court. Most racket sports data are recorded as multivariate event sequences, and pattern mining is used by several data-driven methods to figure out what these sequences mean and find tactical advantages. But experts in the field may be confused by the tactics that are found this way, since they often go against the conclusions that experts draw from their experience in the field. This research suggests using an interactive method called RASIPAM (RAcket-Sports Interactive PAttern Mining) to help experts find new solutions by adding their knowledge to traditional data mining processes. RASIPAM is a constraint-based pattern mining method that was made to meet the analytical needs of professionals. Strategy identification is a text-based process in which experts give their ideas, which are then turned into algorithmic limits. RASIPAM also has a personalised visual interface that lets experts compare and contrast new methods with their predecessors and decide whether or not to make a certain change. This cycle of working together will keep going until all proposed solutions have been looked over by experts and accepted. researchers do a statistical test to show that their method lets people talk at the same time. With the help of two domain experts, two case studies are done, one with badminton and one with tennis, to show how efficient and flexible the system is.

The research in [13] says that the frequent itemsets mining and the association rules mining are two of the most popular ways to mine data. Data owners can keep their data mining tasks safe by sending them to the cloud. Because of mistrust in the cloud and among data owners, the old algorithms, which only work on plaintext, need to be rethought. For example, not all people who own data are happy to share it with other people during cooperative data mining. Both how well and how safe the old methods worked have been shown to be lacking. researchers propose the Secure and Efficient Data Mining Outsourcing (SecEDMO) scheme to make sure that frequent item sets mining and association rules mining across the joint database are safe (i.e., database aggregated from diverse data owners). SecEDMO uses a secure comparison method and their specialised lightweight symmetric homomorphic encryption technology to provide excellent privacy protection and reduce the amount of time it takes to mine data. Using virtual transaction insertion, data can also be hidden while still allowing data mining to be done on the cloud. The accuracy, security, and effectiveness of SecEDMO in real-time settings are shown by evaluating a numerical experiment and making theoretical comparisons.

For periodic pattern mining, algorithms look for repeating patterns in a time-series database, such as data from cell phones and Internet of Things sensors. A study that was published in [14] looked into this subject. This information can be used to evaluate risks, run systems, and come up with policies. This study shows a method for doing data analytics that has been shown to work, with a focus on how often it is done. So that it can work in a wider range of real-world settings and systems, it is based on the idea of flexible periodic patterns and doesn't take into account what happens in between. The proposed method also replaces the current data structure for mining periodic patterns with a new one that is based on symbols. Based on tests on real-world datasets about diabetes, oil prices, and bike sharing, their method is better than the current best methods for finding periodic patterns, which are called efficient periodic pattern mining (EPPM) and flexible periodic pattern mining (FPPM). The results of the tests show that the suggested method will take less time to run and use less memory than the current

methods. This is perfect for the vast majority of real-world data and needs.

According to the proposed study in [15], databases have improved their ways of finding information over the past few years, making it possible to find relevant and useful information. Frequent pattern mining and association rule mining have been studied a lot because they can be used in a lot of different ways in real life. Most methods of data mining use a central place to store the data being mined. This is especially true in the age of "big data," when the amount of data, bandwidth limitations, and energy limits make it hard to use these methods on databases that are spread out. Because of this, data mining in distributed settings has become an important area of study. Along with a fast data structure for storing items and counts to reduce the amount of data sent over the network, researchers present a set of FP growth-based algorithms for finding FPs that can provide fast and scalable service in distributed computing settings. In order to make sure that both DistEclat and BigFIM worked well, the teams thought about both of them during the experiment comparison. The results of the tests show that the proposed method works well in a variety of experimental settings and is a very cost-effective way to process large datasets. When the recommended method was used, the average time it took to run DistEclat was cut by 33% and its transmission cost was cut by 45%. Across a wide range of use cases, the proposed method only took 23.3% of the time it took BigFIM to run and cost 14.2% of what BigFIM did.

Studies cited in [16] say that topological data analysis (TDA) is a good way to reduce the number of dimensions in data, find hidden data links, and intuitively describe the structure of the data. The Mapper technique is one of these tools. It uses a filter function to project high-dimensional data into a one-dimensional space and then rebuilds the links between the underlying data structures. But the current TDA modelling frameworks don't take domain context knowledge or information into account. In this work, researchers develop and test a semi-supervised topological analysis (STA) framework that uses either discretely or continuously labelled data points to choose the best filter functions. researchers first tested the proposed STA framework with simulated data, and then researchers used it with real data from the genotype-tissue expression and ovarian cancer transcriptome datasets. The graphs that STA made from these two gene expression profile-based datasets match up with previous research, showing that the proposed frameworks are useful.

According to the research in [17], as big data mining technologies have become more popular, cross-domain themes driven by data predictive analysis have become important entry points for tackling old problems. The data are very time-dependent and unstable, and it's hard to explain how the process works because of how the trains work together and how the pressure sensor changes when the train slows down. In response to the growing trend of long groups of heavy-duty train commanders, train brake analysis that uses temporal data mining of small groups could lead to unique improvements in the whole train braking area. In this paper, researchers use cutting-edge technologies like machine learning, transfer learning, and lifelong learning to create the first predictive analytic research framework for railway braking systems. Using the idea of how a train brakes and temporal data from the intelligent experiment platform, a foundation has been set up to deal with both fixed-grouped and multi-grouped temporal prediction problems. The process ends with the creation of a predictive algorithm for the automated ongoing maintenance of model parameters in the service of lifelong learning. Lastly, transfer learning's (TL) parameter transfer is used to study the prediction of temporal

data across many groups. By comparing the training results of the "pre-trained" model on the general domain, the "tuned" model on both the general domain and the target domain, and the "target only" model on the target domain separately, the relevant scope and transfer requirements of multi-domain tuning are shown. So, the results of this research could make the semi-physical intelligent test platform for long-grouped heavy-duty trains even better.

Studies described in [18] say that matrix decomposition is one of the most important ways to get knowledge out of the huge amounts of data that modern applications create. Still, it would be hard, if not impossible, to use this method to manage huge amounts of data with a single system. Distributed storage is often used because large datasets are often collected and stored on many different devices. Because of this, it's not unusual for this kind of data to contain a lot of different kinds of noise. Distributed matrix decomposition is a useful technique that needs to be worked on for data analysis on a large scale. The problem of communication between systems that are spread out must also be solved, and the method must be scalable. For large-scale data mining and grouping, researchers recommend using a dispersed Bayesian matrix decomposition model (DBMD). researchers do distributed computing in particular by using these three methods: The first is called AGD, the second is called ADMM, and the third is called statistical inference. researchers look at how different algorithms tend to converge in theory. To deal with the different kinds of noise,researchers give a weighted average that fits well into existing statistical frameworks, lowering the estimation's variance. Compared to two popular distributed methods, Scalable-NMF and Scalable k-means++ related model techniques, their algorithms perform better or just as well in real-world investigations on huge data sets. Synthetic experiments back up the results of their theories.

Studies published in [19] say that improvements in remote sensing technology have led to Earth observation data with (very) high spatial resolution and a lot of hidden semantic information. Using standard methods for processing data, it is hard to get at the latent semantic information in this data. Methods based on data mining, such as latent Dirichlet allocation and bag of visual words models, can be used to find hidden pieces of knowledge. Even though semantic data mining is important for remote sensing, not much is known about it. The point of this article is to talk about this problem. Optical and synthetic aperture radar (SAR) data with different spatial resolutions is used in remote sensing applications. Three scenarios are used to test how well semantic information can be found in these datasets. First, researchers used the semantic discovery method to improve and fix the user-defined Ground Truth map that was part of the very high-resolution RGB data. In the second case study, which looks for areas damaged by wildfires in Sentinel-2 data, the potential of semantic discovery is looked at. In the third and final scenario, a Sentinel-1 SAR patch-based benchmark dataset is used to improve the robustness and accuracy of the annotation. This is done by using the semantic discovery technique (SDT) to find misclassified patches and patches with ambiguous or many semantic labels. In all three cases, researchers were able to show how data mining-based approaches to finding semantic information could be useful in remote sensing applications.

The [20] proposal for a study showed Artificial intelligence, cyber intelligence, and machine learning are just a few of the major technological advances that have made it possible for big data to be used widely not only in business and academia, but also in their everyday lives and in the many internet-connected apps that have come about as a result. In this study, researchers look at how a deep correlation mining method can be used with large amounts of different

data. A hierarchical hybrid network (HHN) model is built to describe how different types of entities interact with each other. It can measure both internal correlations within a layer and external correlations between layers. To make sure the HHN has the best routing operations, a deep reinforcement learning framework is used to build an intelligent router. The intelligent router then uses a restart-based algorithm to make a better random walk by taking into account the network's hierarchical structure and the many connections between its parts. The end goal of making and putting in place an intelligent recommendation system is to make it easier for users to work together in academic big data settings. Experiments done with data from ResearchGate and the Digital Science Learning Platform (DBLP) show that their strategy and methods work and can be used for different scenarios.

As described in [21], incremental mining may improve the quality of process mining and give relevant data for upgrading the reference model by comparing event logs to a reference model and analysing the differences. Existing incremental mining methods limit process mining projects based on domain knowledge, log completion, and the amount of time it takes for a business to finish. To fix these problems, it's best to look at live event streams to update the reference model using an incremental mining method based on the time between trustworthy behaviours. A clustering method is used to figure out the model's main structure and the trusted behaviour interval. This is done by starting with an existing reference model. Last, the behaviour and structure of the links between the online event streams and the reference model are looked at to come up with a candidate set that can be used. Using this set, an incremental update method (IUM) is made to improve the structure of the model and finish a dynamic online update of the reference model. The elbow method is used to figure out how many clusters are needed, and both simulated and real data are used to confirm that number. Then, researchers add the suggested method to the PM4PY and Scikit-learn frameworks. From what we've learned from their experiments, using this method improves the quality of the model with both complete and incomplete data sets. It also makes incremental mining more effective.

According to the research in [22], data pricing is the process of giving data a monetary value so that data sharing, exchange, and reuse can grow in a sustainable way. Yet, the unclear value index and lack of participation make the transaction process even more uneven than it already is. A good pricing plan for data and a well-organized data market can do a lot to increase data transactions. researchers use a three-agent model to better understand how the data market works. The data owner provides the data record, the model buyer wants to buy ML model instances, and the broker helps the data owner and the model buyer talk to each other. researchers know of two times when the two things interacted. In case 1, researchers use the Shapley value (SV) to objectively figure out how much each data record is worth, and researchers use the total SV at each data boundary to build a revenue optimization problem. By figuring out their derivatives, you can find the best answer. In case 2, researchers do some preliminary market research and create a problem called "revenue maximisation" (RM) to help us figure out how much the ML models will cost. Also, it is suggested that the RM problem be turned into a similar integer linear programming (ILP) problem and then solved with the Gurobi solver using the RM-ILP technique. Lastly, researchers do a lot of tests to show that the RM-ILP process could make the broker a lot of money and keep the model's target audience's costs low, compared to industry standards.

According to the research in [23], data visualisation is an important part of data mining because

it helps make sense of big data sets. The state-of-the-art and well-known t-distributed stochastic neighbour embedding method (TDSNEM) is one of the many ways to represent something that has been suggested. Yet, the most effective ways of visualising data have a major flaw: they can't explain how they get from the original properties of the dataset to the final way they show it. Several fields need to understand the data in terms of how it works, so there is a need for efficient ways to visualise data that use models that are easy to understand. In this study, researchers suggest using the genetic programming (GP) technique of truncated symmetric network expansion (GP-tSNE) to map the dataset to high-quality visualisations that can be understood. researchers come up with a multi-objective way to make visualisations that makes many different kinds of visualisations with different levels of visual quality and model complexity in a single run. Testing GP-tSNE against baseline methods on a variety of datasets shows that it has the potential to give deeper insights into data than standard data visualisation techniques. researchers also show how a multiobjective approach can be helpful by doing a detailed analysis of a potential front. This shows how many models can be evaluated at the same time to get a better understanding of the dataset samples.

Studies that were published in [24] show that traditional ways of sharing and updating educational resources lead to inefficient use. But if there is a lot of noise in the huge amount of data, the traditional way of filtering information won't be able to find the important data. Here, researchers use support vector machines to show how data fusion and conversion can be used to clean up abnormal data collected in the context of big data in education (SVMs). Also, the authors gave us a way to automatically put together a picture of school-related data. By filtering and mining the educational data, a neural network was trained to recover the classroom concepts, and plausible correlation rules were used to pull correlations between these concepts from the assessment data. The results show that their technique makes it much easier to find connections between ideas in education and to find good ideas.

According to [25], analysing the slope stability of open pit mine operations to predict the risk of mine landslides is an important part of the digital mining system's safety management. This analysis of stability looks at a wide range of environmental and human factors in the area. Using these factors to train a learning model and figure out how they affect slope stability is a natural way to figure out how likely a mine landslide is to happen. The main problem is figuring out how to study the complex, nonlinear interactions between these parts using only a few high-dimensional historical slope data. Traditional factor-based solutions to the problem only look at the effects of the different parts of a landslide. They don't look at the correlations between historical slope data, which have more important information. In this paper, researchers show a new way to use knowledge graphs to estimate the likelihood of mine landslides. After a gradient-boosting decision tree (GBDT) has been run, the crossover features in the historical data are used to build a landslide semantic network using a knowledge graph and the correlation information between historical slope data. The model is good for dealing with the small set of high-dimensional data because it takes into account both the features of the landslide component values and the correlations between past slope data. Data from real open pit mining slopes are used in the tests. The experiment's results show that the proposed model for finding landslide risks from a small set of high-dimensional historical data samples works well and efficiently.

For sequence classification, it's important to find features that are different from each other.

Sequential pattern mining (SPM) is often used to find common patterns in sequences that can be used as features, as described in the proposed research project [26]. Data scientists use a method called "contrast pattern mining" to better group data and make patterns easier to understand. This method looks for patterns with high contrast rates across many categories. At the moment, contrast SPM methods have a lot of problems, such as not being able to count occurrences well and needing very specific parameter values. This study suggests a top-k self-adaptive contrast SPM that can find top-k SCPs in both positive and negative sequences by changing the gap limits on the fly. One of the main goals of the mining problem is to figure out how many times a pattern shows up in each sequence. To make counting faster and more accurate, researchers store all instances of a pattern in a single array inside a Net Tree, which is a big tree with many roots and many parents. researchers use the array to figure out where all of its super-patterns show up using one-way scanning so researchers don't have to do the same calculations twice. researchers also suggest the Zero and Less and Contrast-First mining (ZLCF) processes to get rid of candidates with the highest contrast rates and figure out the contrast rates of all their super-patterns, since the contrast SPM problem doesn't meet the Apriori requirement. The data show that the proposed strategy works and that contrast patterns are much better at classifying sequences than common patterns.

As the body of work beginning with [27] has gone on, the number of people interested in time series forecasting has grown. But the ability of unusual times to predict the future is lessened when time series aren't balanced and there are clear patterns between unusual and normal times. The main goal of this study is to come up with a model that can be used to fix the difference and make better predictions for certain time periods. There are two main things that make this work so hard: The trade-off between finding similar patterns and finding different distributions over time, and how the series depends on time. To deal with these problems, researchers suggest a self-attention-based prediction model that changes over time and is trained in two phases. To find common patterns in a time series, researchers first use an encoder-decoder module along with the multi-head self-attention (ED MHSA) technique. Lastly, researchers propose an optimization module that changes over time to improve the results of some periods and bring things back into balance. researchers also suggest using reverse distance attention instead of the more common dot attention to show how important similar past values are for predicting what will happen in the future. Lastly, a lot of testing shows that their model has less mean absolute error and mean absolute percentage error than other baselines.

Based on mining production data from commercial dyeing processes and a modular system architecture, the study cited in [28] suggests a unique way to suggest fabric dyeing recipes. Traditional ways of coming up with dyeing recipes sometimes require time-consuming calibration tests between the amount of dye and the colour it makes. their method, on the other hand, doesn't need such tests. It's common to change systems to fit the needs of a certain dyeing job. researchers talk about the ideas behind their method and how to put it into practise. It is a powerful tool thanks to its modular design and set of gradient-boosting regression tree models (GBRT). Every GBRT can make predictions about the dye concentrations in a DCS that are made for a specific type of fabric. The study also talks about realistic ways to train models and how models usually do.

Studies published in [29] say that the current algorithms for figuring out what gear to put the

car in are limited by numerical model variation and the designer's personal preferences. This makes it hard for the car to adapt to the driver's intentions and the driving situation. Good drivers of cars with manual transmissions know how to choose the right gear to get where they're going safely and quickly adjust to things that didn't go as planned. So, data mining is used to come up with a strategy and design for a vehicle's automatic transmission system that can handle a wide range of driving intentions and tough road conditions. Top drivers are hired to drive cars with manual transmissions so that a lot of driving data can be collected. After driving data is collected, it must be pre-processed, cleaned, and outliers must be removed to get the shift boundary points needed to make the shift rule surface (SRS) for each gear. The proposed gear decision design method can use a lot of driving data to figure out how expert drivers change gears. The resulting strategy can adapt better to the driver's intentions, get better gas mileage, and avoid gearshifts that aren't necessary than the default strategies used by automatic transmissions. Using the uphill scenario, the comparisons show that this statement is true for real-time use cases.

Researchers say in [30] that utilities may be able to change their supply in a smart grid when smart metres report their power use in real time. Yet, attackers may be able to get personal information about users from data that is available to the public about how much power they use in real time. Data aggregation methods are used to keep user data safe. There are still big problems that haven't been fixed with dynamic membership and metre failure. To solve these problems, researchers came up with a way to combine different data sets based on a group whose members change over time. As a first step towards fixing metre failure, a way to set up cooperative keys is made. The data from each metre is encrypted with keys made by people in the same group. The damage to the metres of one group won't affect the other groups. The dynamic join, dynamic leave, and metre replacement methods (DJD LMR) are then talked about. All of these are possible because the dynamic membership lets metres update their keys. The simulation results show that the proposed approach is better for the IoT scenario, since the cost of calculating and communicating with a metre is the lowest of all the linked processes. Along with the methods for encoding and retrieving data, researchers also came up with multiple attack scenarios.

The research project proposed in [31] made it clear that the high computational cost of machine learning makes it hard to use in many big data mining situations with a large number of samples. Machine learning must iteratively compute by going through the whole dataset without taking into account the roles of different samples in training computation. researchers argue, though, that most of the samples that use most of the computer resources don't add much to the gradient-based model update, especially when the model is almost done. In the field of machine learning, this is called the Sample Contribution Pattern (SCP). This research suggests two ways to use SCP. The first is to find gradient features. The second is to get people to use old gradients again. This paper specifically reports research findings on (1) the definition and description of SCP to reveal an intrinsic gradient contribution pattern of different samples, (2) a novel SCP-based optimising algorithm (SCPOA) that outperforms competing tested algorithms in terms of computation overhead, (3) a variant of SCPOA that includes discarding-recovery mechanisms to carefully trade-off between model accuracy and computation cost, and (4) the implementation and evaluation of SCPOA. their tests show that the suggested methods cut calculation costs by a lot while keeping competitive levels of accuracy.

Work presented in [32] said that the unique features and ways that microgrids work create protection issues because there are more distributed energy resources. The proposed plan uses data mining and the Hilbert transform to protect the microgrid from these issues. First, the Hilbert Transform (HT) is used to separate the sensitive fault characteristics from the voltage and current signals that are wrong. The extracted feature data set is then made, and the logistic regression classifier is used to find errors. Then, the AdaBoost classifier (AC) is taught to find mistakes. In the suggested method, Python is used to train and test data mining models, and the results of feature extraction simulations are checked on a standard medium-voltage microgrid from the International Electrotechnical Commission (IEC) that is compatible with the MATLAB/SIMULINK software environment. By modelling different fault and no-fault scenarios, the results are evaluated for both looping and radial designs in grid-connected and islanded modes. The results show that the suggested logistic regression and AdaBoost classifier methods are more accurate than the decision tree, support vector machine, and random forest methods. The results show that the suggested method can handle high levels of measurement noise.

The study referred to in [33] explained One of the biggest problems with itemset mining is that every time a user wants to get a new type of itemsets, they have to make a new data structure or method. Generic Itemset Mining based on Reinforcement Learning (GIM-RL) is a method researchers suggest to solve this problem. GIM-RL gives you a single way to train an agent to get any kind of itemsets. In the GIM-RL environment, iterative steps are used to pull a target type of item sets out of a dataset. Every time the agent does something to add or remove an item from the current itemset, the environment gives the agent a reward that shows how relevant the new itemset is to the target type. Through a series of trial-and-error phases in which different rewards are earned by different actions, the agent learns how to get the most rewards overall. As a result, it figures out the best way to act to make as many item sets of the goal type as possible. This architecture can be used to train an agent to get any kind of item set, as long as a reward that fits the type can be given. The detailed tests on mining frequent item sets, association rules, and high-utility items show GIM-performance RL's as a whole, as well as one particularly interesting possibility (agent transfer). researchers think that GIM-RL will open the door to more research on learning-based processes for mining itemsets.

According to research published in [34], mining frequent weighted patterns (FWPs), which take into account the different semantic importance (weight) of items, is better for practise than mining frequent patterns. Because of this, it is very important in the real world. There are a few things you can't do when you use techniques for mining FWPs that were made for static data on growing datasets, especially data streams. In this paper, a method is proposed for mining FWPs across different data streams. At first, researchers suggest using a sliding window model to find FWPs across multiple data streams. Next, we'll talk about the SWN-tree, which is a variation of the weighted node tree (WN-tree) that can keep the data even when data streams change. The paper then shows how to use a sliding window model based on SWN-tree to pull FWPs out of data streams. The name of this method is the Frequent Weighted Patterns Over Data Stream (FWPODS) algorithm. Last but not least, researchers do real-world tests to compare the results of their strategy to the most recent method (NFWI) for mining FWPs across data streams. The results of the test show that their method works better than the NFWI algorithm when it is used in batch mode with sliding window sets.

According to research presented in [35], information about high-speed trains' vehicles can not only tell if the train's different parts are working well, but can also predict how the train will work in the future. It's hard to figure out how to get useful information from a huge amount of vehicle data. Then, researchers split the data from a high-speed train into 13 subsystem datasets based on the roles of the parts that collected the data. Based on the grey theory and the Granger causality test, the Gray-Granger Causality (GGC) model is then proposed. This model can build a network of information about vehicles based on how well the different parts of the collection fit together. By using the complex network theory to mine the networks, researchers find that the vehicle information network and its subsystem networks have the characteristics of a scale-free network. Also, the subsystem network is well connected and not easy to attack, but the vehicle information network is.

One of the hardest things about association rule mining, according to a research study [36], is that when a new incremental database is added to an original database, some frequent item sets can change from being frequent to being infrequent, and vice versa. This could make some old rules for associations no longer valid and make room for new ones. researchers came up with a new, more effective way to find incremental association rules by using a Fast Incremental Updating Frequent Pattern growth algorithm (FIUFP-Growth), a new Incremental Conditional Pattern tree (ICPT), and a compact sub-tree that is good for incremental mining of frequent item sets. This technique takes the frequently mined items from the original database and adds their support counts. This makes it easier to find frequent item sets in the updated database and ICP-tree and reduces the number of times the original database has to be rescanned. When researchers compared their method to the stand-alone FP-Growth, FUFP-tree maintenance, Pre-FUFP, and FCFPIM algorithms, researchers found that their method used less time and resources to create unnecessary sub-trees. Based on the results, their method is 46% better than FP-Growth, FUFP-tree, Pre-FUFP, and FCFPIM in terms of the average time it takes to execute pattern growth mining at a minimum support threshold of 3%. their approach to incremental association rule mining and the results of their experiments may help the people who make business intelligence systems for computers.

The work suggested in [37] made this clear. Because sensor networks and smart devices that constantly collect data are everywhere, it is hard to analyse the growing stream of data in real time. In the past few years, it has become much more important to get information by looking at event sequence data step by step. Even though episode pattern mining methods have been around for a while, it's only recently become clear that they can be used to solve problems in the real world, such as with factory records, financial markets, and weather forecasts. More and more, it's important to use episode pattern mining techniques to look at complicated event data and solve problems in the real world in a wide range of fields. Few studies, on the other hand, have focused on making a framework based on mining large event sequences for episode patterns that can be used in many different fields. In this work, researchers introduce a new framework called SAAF (Scalable Analytical Application Framework). It is based on complex event episode mining approaches, such as batch episode mining, delta episode mining, incremental episode mining, and pattern merging. Also, to make the system easier to scale, researchers use the lambda architecture with Apache Spark and Apache Spark Streaming as the system development framework. Using three real datasets from different domains and two benchmark datasets, the final results showed that the proposed SAAF framework works very

well in terms of efficiency, accuracy, and scalability.

The study referred to in [38] explained from a power grid point of view, there is no literature that can be used to figure out the fault level of sensitive equipment (FLSE) caused by voltage sag. In practise, the voltage sag of the node is the most important part of FLSE. However, the voltage sag of the node is related to the whole power grid, including the voltage grade and location of the node, the distance from the concerned node to the location of fault, and the weather and date at the time of fault. All of these things are called voltage sag properties (VSP). In order to fill this gap, this work uses data mining techniques based on multidimensional matrix simplification and enhanced grey target theory to measure the FLSE using long-term monitoring data of VSP of certain regional power grids. The data mining process has three steps: With long-term monitoring data, you can: 1) make a database with VSP and FLSE data; 2) use multidimensional matrix simplification to mine the association rules between VSP and FLSE; and 3) use the enhanced grey target theory to match some real-world situation with the mined association rules. Using simulations and real-world examples, the performance and effectiveness of the suggested method are confirmed.

Research published in [39] shows that a nonnegative latent factorization of tensors (NLFT) model accurately catches the hidden temporal patterns in multichannel data from different applications. It usually uses the SLF-NMUT method, which stands for single latent factor-dependent, nonnegative, and multiplicative update on tensor. The fixed learning depth of the algorithm, on the other hand, leads to frequent training changes or poor model convergence due to overshooting. This work uses SLF-NMUT to look closely at the connections between an NLFT model's performance and its learning depth in order to come up with a method for adjusting both. Based on this method, a unique depth-adjusted nonnegative latent-factorization-of-tensors (DNL) model is made. This model is called a Depth-adjusted Multiplicative Updating on tensor model. Experiments with two industrial data sets show that a DNL model can estimate missing data on a high-dimensional and incomplete tensor very quickly and with a much higher level of accuracy than the best NLFT methods. Note to Professionals — Multichannel data is often used in a number of big data applications. For effective knowledge acquisition and representation, a data analyst must accurately find the temporal patterns that are hidden in the data. The main topic of this paper is the analysis of temporal QoS data, which is a kind of representative multichannel data. To get the right temporal patterns from them, an analyst must explain their non-negativity correctly. To reach this goal, the single latent factor-dependent, nonnegative, and multiplicative update on tensor (SLF-NMUT) method can be used to build a nonnegative latent factorization of tensors (NLFT) model. Since the learning depth of an NLFT model is fixed, its training error tends to swing a lot or even fail to converge. This paper looks at the learning rules for an NLFT model's decision parameters using an SLF-NMUT to solve this problem, and it suggests a combined learning-depth-adjusting scheme. This method changes the learning depth by changing the multiplicative parts of the SLF-NMUT-based learning rules in a linear and exponential way. It is the basis for the unique depth-adjusted nonnegative latent-factorization-of-tensors (DNL) model developed in this work. A DNL model is a better reflection of multichannel data than the current NLFT models. It meets business needs well and can be used to do well in data analysis tasks, such as estimating missing data that takes into account time.

According to research [40], all schools put a high priority on improving students' grades in

order to improve the quality of education as a whole. In this way, Educational Data Mining (EDM) is a rapidly growing field of research that uses the basic ideas of Data Mining (DM) to help academic institutions find useful data on the Student Satisfaction Level (SSL) with the Online Learning process (OL) during COVID-19 lock-down. Several strategies have been tried out using EDM to predict how students will act in order to find the best learning environments. So, Feature Selection (FS) is often used to find the subset of characteristics that are the most important and have the lowest cardinality. This research looks closely at how well the SSL model works with FS approaches, since the FS process has a big effect on how well a satisfaction model predicts the truth. In this way, a dataset was first put together through an online survey of student reviews of OL courses. With this dataset, fitness values were used to measure how well wrapper FS approaches in DM and classification algorithms worked. In the end, 11 wrapper-based FS algorithms that use the NN and Support Vector Machine (SVM) as baseline classifiers are evaluated based on how well they predict and how many features they use. The studies showed the best way to do things and how many dimensions the feature subset should have. The results of this study clearly show that the well-known connection between fewer characteristics and better ability to predict is true. It's amazing how useful FS is for high-accuracy SSL prediction, since the right mix of attributes can really help develop good ways to teach. Using the real-time dataset samples, researchers chose; their research leads to a reduction in feature size of up to 80% and a classification accuracy of up to 100% for different scenarios.

The study in [41] suggests that reliability could be improved by learning more about why power systems break down. In this study, researchers look at data about distribution power network failures to find the most common reasons for blackouts that are caused by nature or machines. After integration, real-world data like weather, load, and outages are used to pull out the most important features. This article uses visualisation techniques to show how different factors affect the frequency of outages. It then uses association rule mining to find attributes that are linked to each type of outage and to each other. Two important metrics for Apriori association rule mining are the chi-square and the lift index (ARMCL) for different inputs. Each piece of equipment has its own outage study to figure out what restrictions go with it. The results that show how well and how valid the proposed method is for figuring out why outages happen can be used to plan for the future and figure out when distribution power networks will be working under real-time scenarios.

Research in [42] shows that frequent itemsets mining is a popular topic of study in the data mining and knowledge discovery communities. Due to the rise of database systems and the exponential growth of the amount of data that needs to be stored, there is an urgent need for fast, efficient ways to find the most important information. Mining for frequent itemsets (FIs) looks at a transactional database to find groups of items that happen together more often than a certain threshold. Since it usually takes a lot of database searches to find these FIs, FI mining needs to use efficient methods. As part of this research, researchers come up with a way to describe a whole transactional database using graphs. The proposed graph-based method (GBM) makes it possible to save all of the important (for FI extraction) data in the database in a single step. The FIs are then taken out of the graph-based structure that was made. Using six benchmark datasets, the results of experiments are shown that compare the suggested method to 17 others similar FIs mining approaches. In a number of different situations, the results

show that the proposed method is faster than other options.

Research cited in [43] says that sensing errors can cause data to be different from what it really is in a number of situations, especially for the Internet of Things (IoT). During the last 10 years, there has been a lot of focus on data mining that protects privacy, but not much research has been done on data values that are wrong. Differential privacy is the standard way to keep sensitive information safe by hiding it with random noise. If the goal value already has some mistake in it, there's no reason to add more. In this paper, researchers propose a brand-new privacy model called true-value-based differential privacy (TDP). This model applies normal differential privacy to the "real value," which the data owner or anonymizer doesn't know, but not to the "measured value," which has higher error levels. TDP's calculations show that their method could cut the noise from differential privacy methods by about 20%. Because of this, the mean square error metric and the Jensen-Shannon divergence metric can be used to make histograms that are up to 40.4% more accurate and 29.6% more accurate, respectively. researchers use five real data sets and one fake data set to back up this result. More importantly, researchers showed that the level of privacy protection doesn't go down when it's measured in real time, as long as the measurement error isn't too big.

Literature cited in [44] says that traditional Probabilistic Integration Method (PIM) parameter inversion cannot be used with Line of Sight (LOS) distortion based on Differential Interferometric Synthetic Aperture Radar (DInSAR) approaches. It was suggested that parameter inversion might be more accurate if it was based on a 3D deformation model. Since it uses PIM directly, the model is able to show how something changes in 3D. The square root of the sum of squared errors (SSE) between the PIM results and the deformation results that were seen was used to figure out the GA fitting function. This GA model can be used to figure out accurate PIM parameters. Six Global Navigation Satellite System (GNSS) monitor sites were set up above the 011207 and 011809 working panels to find out how the surface of the Jinfeng coal mine moves. Due to the small number of points and the large distances between them, GNSS alone is not enough to get accurate PIM parameters. Also, the LOS direction deformation was found using a small baseline subset (SBAS) DInSAR analysis of 83 Sentinel-1A images. With a 3D inversion model, LOS direction deformation, and GNSS-monitored deformation, the exact PIM parameters were found. Then, these traits were used to estimate how much panels 011207 and 011809 bend in coal mines, and the results were the same as what was expected. Integrating DInSAR and GNSS into the model could reduce the need for GNSS points and reduce the effects of radar decoherence, allowing the model to be used to study the rules that govern mining subsidence. This is a way to look at the law of surface movement in mining subsidence data sets using cutting-edge technology.

Research in [45] shows that it is a common task in databases to find the top k most-used items. Finding the k objects that are used the most is a new challenge that has grown in popularity over the past few years. Users want to know what the common and reliable parts are. These are two problems that no solution has been able to solve at the same time. Also, the state of the art can't guarantee accuracy for fast-moving data streams when there isn't enough memory. In this study, researchers present a new challenge for finding interesting information and suggest a state-of-the-art way to deal with it. researchers call this approach LTC. Long-tail Restoration and Resetting are two of its most important features. It also has three improvements. LTC also has a way to find significant items based on a threshold. researchers

use theory to figure out the right rate and error limit, and then researchers do a lot of testing on three real-world datasets to see how well LTC works. Based on the results of their tests, LTC can multiply with a rate of error that is an order of magnitude lower than that of other algorithms that do the same thing. Lastly, LTC is used to find DDoS attacks, which proves once and for all that finding rare events is less useful than finding common ones.

In response to growing safety concerns and an ageing workforce in the traditional mining industry, a new trend is emerging: smart mining, and especially autonomous mining, which has the potential to solve the safety problem while also increasing productivity [46]. Even so, there are no set rules for smart mining. This has led to a lot of different machines and development costs that are too high to be practical. So, researchers will talk about the steps that Chinese academic and industrial mining organisations have taken to set benchmarks for automation and autonomy. researchers also talk about how autonomous trucks are used in mines around the world now, as well as some of the most important cognitive technologies for autonomous mining transportation. With the adoption of new standards, these important intelligent technologies should improve a lot, making them more widely used in the mining industries.

According to the research shown in [47], the growth of information and communication technologies in the real world has brought a lot of attention to computer programming. Meeting the growing need for skilled programmers in the information and communications technology sector is a big problem. With this in mind, it's helpful to have resources outside of the classroom for learning and practising programming, and online judge (OJ) systems give you just that. Because of this, OJ systems have collected a lot of information about how people solve problems. This information comes in the form of solution codes, logs, and scores, and it could be used to learn more about how to teach programming. In this study, researchers use unsupervised algorithms to create a framework for educational data mining. This will help advance the study of computer science. The following steps are part of the framework: First, the data needed to solve the problem is collected (OJ logs and scores are used). Then, the data is clustered in Euclidean space using the MK-means clustering algorithm. Next, statistical features are extracted from each cluster. Finally, the frequent pattern (FP)-growth algorithm is used on each cluster to mine data patterns and association rules. Lastly, a set of recommendations are made based on the patterns and rules found in the data. To find the best setting for clustering and association rule mining, several parameters are changed. Five hundred thirty-seven students in an introductory programming course (Algorithms and Data Structures) were asked to gather information on how to solve 70,000 real-world problems. But in order to show how well the MK-means algorithm works, it has been necessary to use data that has been made up. The results of the tests show that the proposed framework is able to pull out useful features, patterns, and rules from data about how to solve problems. These boiled-down features, patterns, and norms not only show how programming education is right now, but also how it could be made better in many ways.

According to research reported in [48], urban flow pattern analysis is an important field of study for the development of smart cities because it looks at how cities flow all the time. It may be hard to find, store, and use traffic patterns from urban multi-source, heterogeneous, and huge data sets. In this study, a network for mining and storing information about urban flow patterns is suggested. The network would be used to mine information about regional

flow patterns. The model that is being suggested has two main parts. In the first module, the flow pattern of the area and its unique properties are taken as the entity and the relationship. Module 2 is all about modelling POI attributes to improve embedding of relations and entities. The translation distance method is used to find triplets of information about how things flow in a region. Lastly, the Chengdu Didi order and POI datasets are used to compare the proposed model to different benchmark methods. Experiments show that the proposed paradigm is a good idea. Along with showing the knowledge triplets, researchers also give a few examples of how they can be used in real life.

E-commerce businesses need access to data, which can be gathered with sophisticated web scraping tools, in order to compete in today's market of hyper-personalized customer experiences. But the basic data extraction technologies don't work well because they can't adapt to how online content changes all the time. This study looks at a smart and flexible web data extraction system that uses convolutional and Long Short-Term Memory (LSTM) networks [49] to automatically recognise web pages using the Yolo algorithm and Tesseract LSTM to get product information from online sites that are labelled as photos. This state-of-the-art solution doesn't need a centralised data extraction engine, so it can easily adapt to new website designs. Studies of real-world retail situations have shown that character extraction can be done with 97% accuracy and image recognition can be done with 99% accuracy. In different situations, whether 45 items or photos are used as the input dataset, a mean average accuracy of 74% is also reached.

Studies cited in [50] show that there is a big need in the material science community for a central database with the compositions of materials and the analyses of metal strength that can be made from them. Many research groups have their own Excel spreadsheets that they made by tabulating experimental data from scholarly articles and then manually analysing the data with formulas to figure out how strong the material is. researchers propose a big data storage for material science data and its processing parameters, data mining techniques to get the information from databases to do big data analytics, and a machine learning prediction model to get insights into material strength, all to make the time-consuming process of tabulating data from scientific articles easier. The three models that are suggested are the Stochastic Forest, the Support Vector Machine, and the Logistic Regression. Each of these models is based on a different machine learning technique. The 10-fold cross validation method is used to train these models and check how accurate they are. For the independent dataset, Stochastic Forest was more accurate than Logistic regression and SVM (87% vs. 72%) for practical use cases. Thus, a wide variety of models are proposed for analysis of unstructured data samples. A comparative analysis of these models is discussed in the next section of this text.

3. Statistical Analysis

As per the detailed review on unstructured data mining techniques, it can be observed that deep learning & augmented classification techniques showcase better performance for real-time scenarios. A comparative analysis of these models is discussed in terms of processing efficiency (E), delay of operation (D), complexity of deployment (CC), cost of deployment (C) and scalability (S) metrics. All these evaluation metrics were referred from the reviewed texts. Out of these metrics, efficiency levels were directly inferred from Accuracy, Precision

& Recall levels, while other metrics were converted into quantized levels of Low (L=1), Medium (M=2), High (H=3), and Very High (VH=4), which will assist in comparison of the models under an augmented unified set of scales. Based on this strategy, these parameters were tabulated in table 1 as follows,

Table 1. Comparative analysis of different data mining techniques

Model	E	D	CC	C	S
TDD [1]	85.4	M	H	M	H
HC TC [2]	90.5	H	H	H	M
kNN SF [3]	83.1	M	H	H	H
IK MA [4]	97.5	M	H	H	H
MPC [5]	93.4	H	H	M	VH
LRA NCPF [6]	92.8	H	H	H	M
PPHOPCM [7]	93.1	H	M	H	H
eLDA [8]	90.4	H	H	M	H
FDM [9]	91.2	VH	H	H	VH
MWMP [10]	93.1	H	H	VH	H
HUSP ULL [11]	90.5	H	H	VH	H
RASI PAM [12]	97.4	H	H	VH	H
SecE DMO [13]	90.5	VH	H	VH	L
EF PPM [14]	91.4	H	M	H	L
Big FIM [15]	90.2	H	VH	M	H
STA [16]	91.8	VH	H	H	L
TL [17]	98.5	M	H	H	H
DBMD [18]	95.2	M	H	H	M
SDT [19]	75.8	H	VH	H	H
HHN [20]	85.2	M	H	VH	H
IUM [21]	97.5	L	M	L	H
SV RM [22]	90.8	M	H	H	L
TD SNEM [23]	91.5	H	H	H	H
SVM [24]	90.4	M	H	H	H
GBDT [25]	98.3	L	M	M	H
ZLCF SPM [26]	99.1	H	H	H	M
ED MHSA [27]	99.2	L	VH	H	H
GBRT [28]	98.5	M	VH	H	H
SRS [29]	90.4	H	VH	M	H
DJD LMR [30]	91.3	H	H	H	M
SCP [31]	83.5	H	H	H	M
GIMRL [33]	85.5	H	H	H	H
FWP SWN WT [34]	94.2	M	VH	H	H
GGC [35]	94.8	H	H	H	H
ICPT [36]	95.5	H	H	M	VH
SAAF [37]	90.4	H	VH	H	VH
FLSE [38]	90.5	M	H	H	H
SLFNMUT [39]	95.4	H	H	L	H
NN SVM [40]	97.9	H	VH	H	H
ARM CL [41]	95.1	H	M	L	H
GBM [42]	98.9	M	H	L	VH
TDP [43]	98.4	H	H	M	H
GA SBAS [44]	98.1	H	H	H	VH
LTC [45]	95.5	M	VH	M	H
FP OJ [47]	97.2	H	VH	M	H

POI [48]	93.5	H	H	H	H
YoLO LSTM [49]	99.2	L	H	L	L
SF SVM [50]	89.5	H	M	H	H

As per this evaluation, it can be observed that ED MHSA [27], YoLO LSTM [49], ZLCF SPM [26], GBM [42], TL [17], GBRT [28], TDP [43], GBDT [25], and GA SBAS [44] are capable of achieving higher efficiency of mining when deployed on multimodal data samples. Work in IUM [21], GBDT [25], ED MHSA [27], and YoLO LSTM [49] showcased lower computational delays, thus can be deployed for high-speed scenarios.

In terms of computational complexity, work in PPHO PCM [7], EF PPM [14], IUM [21], GBDT [25], ARM CL [41], SF SVM [50], and Big FIM [15] showcase better performance, while in terms of deployment cost, work in IUM [21], SLF NMUT [39], ARM CL [41], GBM [42], and YoLO LSTM [49] are identified to b applicable for cost-aware scenarios. In terms of scalability levels, work in SecE DMO [13], EF PPM [14], STA [16], SV RM [22], and YoLO LSTM [49] showcase applicability for larger number of use cases.All these metrics were combined to form a novel NoSQL Data Mining Rank Metric (NDMRM) via equation 1,

$$\text{NDMRM} = \frac{\text{E}}{100} + \frac{\text{S}}{5} + \frac{1}{\text{CC}} + \frac{1}{\text{C}} + \frac{1}{\text{D}} \ldots (1)$$

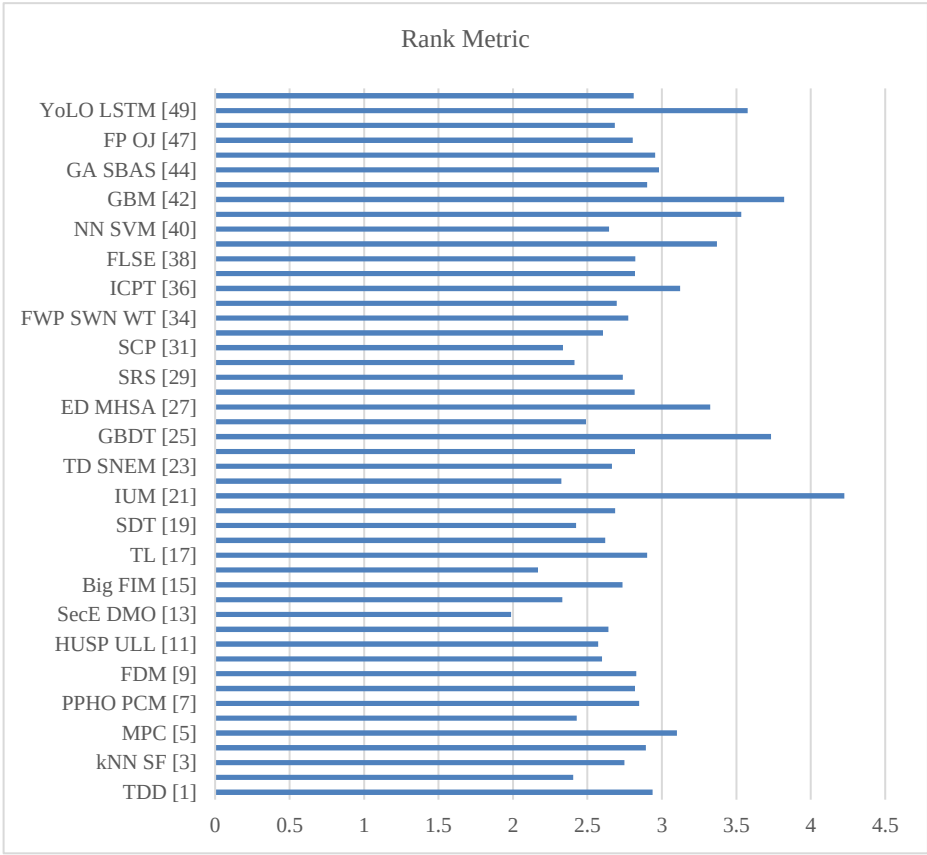

Figure 1. Evaluation of NDMRM for different models

As per this evaluation, and figure 2, it can be observed that IUM [21], GBM [42], GBDT [25], YoLO LSTM [49], ARM CL [41], SLF NMUT [39], ED MHSA [27], ICPT [36], MPC [5], and GA SBAS [44] are capable of high efficiency, high scalability, low complexity, low cost and high-speed use cases. Thus, these models must be used for deployments, and can be extended for multiple scenarios.

4. Conclusion and future scope

This article provides a discussion on the design of recently proposed mining techniques that can be used on unstructured data sets. Readers will be able to evaluate the reviewed models based on their contextual nuances, functional advantages, application-specific drawbacks, and deployment-specific future scopes after reading this discussion. A comparison of the models that have been reviewed in terms of their processing efficiency, delay of operation, complexity of deployment, cost of deployment, and scalability metrics is presented as a further extension of this discussion. This will make it possible for readers to determine which models have the best performance for a variety of performance-specific use cases. A novel NoSQL Data Mining Rank Metric (NDMRM) is calculated based on these evaluations. This metric combines the evaluation metrics in order to identify NoSQL Data Mining Models that perform optimally under real-time scenarios.

According to the results of this analysis, it was found that ED MHSA [27], YoLO LSTM [49], ZLCF SPM [26], GBM [42], TL [17], GBRT [28], TDP [43], GBDT [25], and GA SBAS [44] are all capable of achieving higher levels of mining efficiency when they are applied to multimodal data samples. The work that was done in IUM [21], GBDT [25], ED MHSA [27], and YoLO LSTM [49] demonstrated lower computational delays, and as a result, it is suitable for use in high-speed scenarios. Work in PPHO PCM [7], EF PPM [14], IUM [21], GBDT [25], ARM CL [41], SF SVM [50], and Big FIM [15] demonstrate better performance in terms of computational complexity. On the other hand, in terms of deployment cost, work in IUM [21], SLF NMUT [39], ARM CL [41], GBM [42], and YoLO LSTM [49] are identified to be applicable for cost-aware In terms of the levels of scalability, the work that was done in SecE DMO [13], EF PPM [14], STA [16], SV RM [22], and YoLO LSTM [49] demonstrates applicability for a greater number of different use cases. After combining these metrics, it was discovered that IUM [21], GBM [42], GBDT [25], YoLO LSTM [49], ARM CL [41], SLF NMUT [39], ED MHSA [27], ICPT [36], MPC [5], and GA SBAS [44] are all capable of high efficiency, high scalability, low complexity, low cost, and high-speed application scenarios. Consequently, these models need to be used for deployments, and they can be expanded to cover a variety of use cases.

In future, performance of these models must be validated on other applications, and can be improved via integration of Q-Learning, Bioinspired computing, and other incremental learning operations. Moreover, this performance can also be improved via integration of Generative Adversarial Networks (GANs), and Transformer Models for high-efficiency use cases under heterogenous data sample scenarios.

References

1. G. Deeva, J. De Smedt and J. De Weerdt, "Educational Sequence Mining for Dropout Prediction in MOOCs: Model Building, Evaluation, and Benchmarking," in IEEE Transactions on Learning Technologies, vol. 15, no. 6, pp. 720-735, 1 Dec. 2022, doi: 10.1109/TLT.2022.3215598.
2. H. Pang and B. Wang, "Privacy-Preserving Association Rule Mining Using Homomorphic Encryption in a Multikey Environment," in IEEE Systems Journal, vol. 15, no. 2, pp. 3131-3141, June 2021, doi: 10.1109/JSYST.2020.3001316.
3. R. Gomes da Silva, J. de Oliveira LiberatoMagalhães, I. R. Rodrigues Silva, R. Fagundes, E. Lima and A. Maciel, "Rating Prediction of Google Play Store apps with application of data mining techniques," in IEEE Latin America Transactions, vol. 19, no. 01, pp. 26-32, January 2021, doi: 10.1109/TLA.2021.9423823.
4. G. Feng, M. Fan and Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," in IEEE Access, vol. 10, pp. 19558-19571, 2022, doi: 10.1109/ACCESS.2022.3151652.
5. M. L. Merani, D. Croce and I. Tinnirello, "Rings for Privacy: An Architecture for Large Scale Privacy-Preserving Data Mining," in IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 6, pp. 1340-1352, 1 June 2021, doi: 10.1109/TPDS.2021.3049286.
6. P. Lin, C. Li and S. Peng, "Diffraction Extraction Using a Low-Rank Matrix Approximation Method," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 3007505, doi: 10.1109/LGRS.2022.3187045.
7. Q. Zhang, L. T. Yang, Z. Chen and P. Li, "PPHOPCM: Privacy-Preserving High-Order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing," in IEEE Transactions on Big Data, vol. 8, no. 1, pp. 25-34, 1 Feb. 2022, doi: 10.1109/TBDATA.2017.2701816.
8. C. Karmakar, C. O. Dumitru, G. Schwarz and M. Datcu, "Feature-Free Explainable Data Mining in SAR Images Using Latent Dirichlet Allocation," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 676-689, 2021, doi: 10.1109/JSTARS.2020.3039012.
9. G. Manogaran et al., "FDM: Fuzzy-Optimized Data Management Technique for Improving Big Data Analytics," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 1, pp. 177-185, Jan. 2021, doi: 10.1109/TFUZZ.2020.3016346.
10. D. Zhang and M. R. Kabuka, "Distributed Relationship Mining over Big Scholar Data," in IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 1, pp. 354-365, 1 Jan.-March 2021, doi: 10.1109/TETC.2018.2829772.
11. W. Gan, J. C. -W. Lin, J. Zhang, P. Fournier-Viger, H. -C. Chao and P. S. Yu, "Fast Utility Mining on Sequence Data," in IEEE Transactions on Cybernetics, vol. 51, no. 2, pp. 487-500, Feb. 2021, doi: 10.1109/TCYB.2020.2970176.
12. J. Wu, D. Liu, Z. Guo and Y. Wu, "RASIPAM: Interactive Pattern Mining of Multivariate Event Sequences in Racket Sports," in IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 1, pp. 940-950, Jan. 2023, doi: 10.1109/TVCG.2022.3209452.
13. J. Wu, N. Mu, X. Lei, J. Le, D. Zhang and X. Liao, "SecEDMO: Enabling Efficient Data Mining with Strong Privacy Protection in Cloud Computing," in IEEE Transactions on Cloud Computing, vol. 10, no. 1, pp. 691-705, 1 Jan.-March 2022, doi: 10.1109/TCC.2019.2932065.
14. H. Kim, U. Yun, B. Vo, J. C. -W. Lin and W. Pedrycz, "Periodicity-Oriented Data Analytics on Time-Series Data for Intelligence System," in IEEE Systems Journal, vol. 15, no. 4, pp. 4958-4969, Dec. 2021, doi: 10.1109/JSYST.2020.3022640.
15. P. -Y. Huang, W. -S. Cheng, J. -C. Chen, W. -Y. Chung, Y. -L. Chen and K. W. Lin, "A Distributed Method for Fast Mining Frequent Patterns From Big Data," in IEEE Access, vol.

9, pp. 135144-135159, 2021, doi: 10.1109/ACCESS.2021.3115514.
16. T. Feng, J. I. Davila, Y. Liu, S. Lin, S. Huang and C. Wang, "Semi-Supervised Topological Analysis for Elucidating Hidden Structures in High-Dimensional Transcriptome Datasets," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 4, pp. 1620-1631, 1 July-Aug. 2021, doi: 10.1109/TCBB.2019.2950657.
17. W. J. Liu, G. C. Wan and M. S. Tong, "A Hybrid Temporal Data Mining Method for Intelligent Train Braking Systems," in IEEE Access, vol. 10, pp. 28739-28749, 2022, doi: 10.1109/ACCESS.2022.3157598.
18. C. Zhang, Y. Yang, W. Zhou and S. Zhang, "Distributed Bayesian Matrix Decomposition for Big Data Mining and Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3701-3713, 1 Aug. 2022, doi: 10.1109/TKDE.2020.3029582.
19. R. M. Asiyabi and M. Datcu, "Earth Observation Semantic Data Mining: Latent Dirichlet Allocation-Based Approach," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 2607-2620, 2022, doi: 10.1109/JSTARS.2022.3159277.
20. X. Zhou, W. Liang, K. I. -K. Wang and L. T. Yang, "Deep Correlation Mining Based on Hierarchical Hybrid Networks for Heterogeneous Big Data Recommendations," in IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 171-178, Feb. 2021, doi: 10.1109/TCSS.2020.2987846.
21. N. Fang, X. Fang, K. Lu and E. Asare, "Online Incremental Mining Based on Trusted Behavior Interval," in IEEE Access, vol. 9, pp. 158562-158573, 2021, doi: 10.1109/ACCESS.2021.3130758.
22. Y. Tian, Y. Ding, S. Fu and D. Liu, "Data Boundary and Data Pricing Based on the Shapley Value," in IEEE Access, vol. 10, pp. 14288-14300, 2022, doi: 10.1109/ACCESS.2022.3147799.
23. A. Lensen, B. Xue and M. Zhang, "Genetic Programming for Evolving a Front of Interpretable Models for Data Visualization," in IEEE Transactions on Cybernetics, vol. 51, no. 11, pp. 5468-5482, Nov. 2021, doi: 10.1109/TCYB.2020.2970198.
24. Y. Xie, P. Wen, W. Hou and Y. Liu, "A Knowledge Image Construction Method for Effective Information Filtering and Mining From Education Big Data," in IEEE Access, vol. 9, pp. 77341-77348, 2021, doi: 10.1109/ACCESS.2021.3074383.
25. L. Ma, J. Wang, J. Cheng, X. Wang and W. Zhu, "MLRP-KG: Mine Landslide Risk Prediction Based on Knowledge Graph," in IEEE Transactions on Artificial Intelligence, vol. 3, no. 1, pp. 78-87, Feb. 2022, doi: 10.1109/TAI.2021.3114652.
26. Y. Wu, Y. Wang, Y. Li, X. Zhu and X. Wu, "Top-k Self-Adaptive Contrast Sequential Pattern Mining," in IEEE Transactions on Cybernetics, vol. 52, no. 11, pp. 11819-11833, Nov. 2022, doi: 10.1109/TCYB.2021.3082114.
27. C. Hou, J. Wu, B. Cao and J. Fan, "A deep-learning prediction model for imbalanced time series data forecasting," in Big Data Mining and Analytics, vol. 4, no. 4, pp. 266-278, Dec. 2021, doi: 10.26599/BDMA.2021.9020011.
28. X. Qin and X. J. Zhang, "An Industrial Dyeing Recipe Recommendation System for Textile Fabrics Based on Data-Mining and Modular Architecture Design," in IEEE Access, vol. 9, pp. 136105-136110, 2021, doi: 10.1109/ACCESS.2021.3117261.
29. K. Cheng, D. Sun, C. Chen, D. Qin and K. Wang, "Intelligent Gear Decision Method for Automatic Vehicles Based on Data Mining Under Uphill Conditions," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 12, pp. 24235-24247, Dec. 2022, doi: 10.1109/TITS.2022.3193966.
30. Y. Chen, J. -F. Martínez-Ortega, L. López, H. Yu and Z. Yang, "A Dynamic Membership Group-Based Multiple-Data Aggregation Scheme for Smart Grid," in IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12360-12374, 1 Aug.1, 2021, doi:

10.1109/JIOT.2021.3063412.

31. X. Shi and Y. Liu, "Sample Contribution Pattern Based Big Data Mining Optimization Algorithms," in IEEE Access, vol. 9, pp. 32734-32746, 2021, doi: 10.1109/ACCESS.2021.3060785.
32. S. Baloch and M. S. Muhammad, "An Intelligent Data Mining-Based Fault Detection and Classification Strategy for Microgrid," in IEEE Access, vol. 9, pp. 22470-22479, 2021, doi: 10.1109/ACCESS.2021.3056534.
33. K. Fujioka and K. Shirahama, "Generic Itemset Mining Based on Reinforcement Learning," in IEEE Access, vol. 10, pp. 5824-5841, 2022, doi: 10.1109/ACCESS.2022.3141806.
34. H. Bui, T. -A. Nguyen-Hoang, B. Vo, H. Nguyen and T. Le, "A Sliding Window-Based Approach for Mining Frequent Weighted Patterns Over Data Streams," in IEEE Access, vol. 9, pp. 56318-56329, 2021, doi: 10.1109/ACCESS.2021.3070132.
35. J. Man, H. Dong, L. Jia and Y. Qin, "GGC: Gray-granger causality method for sensor correlation network structure mining on high-speed train," in Tsinghua Science and Technology, vol. 27, no. 1, pp. 207-222, Feb. 2022, doi: 10.26599/TST.2021.9010034.
36. W. Thurachon and W. Kreesuradej, "Incremental Association Rule Mining With a Fast Incremental Updating Frequent Pattern Growth Algorithm," in IEEE Access, vol. 9, pp. 55726-55741, 2021, doi: 10.1109/ACCESS.2021.3071777.
37. J. C. C. Tseng, S. -Y. Hsieh and V. S. Tseng, "A Scalable Analytical Framework for Complex Event Episode Mining With Various Domains Applications," in IEEE Access, vol. 10, pp. 130672-130685, 2022, doi: 10.1109/ACCESS.2022.3228962.
38. F. Xu et al., "Evaluation of Fault Level of Sensitive Equipment Caused by Voltage Sag via Data Mining," in IEEE Transactions on Power Delivery, vol. 36, no. 5, pp. 2625-2633, Oct. 2021, doi: 10.1109/TPWRD.2020.3024761.
39. X. Luo, M. Chen, H. Wu, Z. Liu, H. Yuan and M. Zhou, "Adjusting Learning Depth in Nonnegative Latent Factorization of Tensors for Accurately Modeling Temporal Patterns in Dynamic QoS Data," in IEEE Transactions on Automation Science and Engineering, vol. 18, no. 4, pp. 2142-2155, Oct. 2021, doi: 10.1109/TASE.2020.3040400.
40. H. E. Abdelkader, A. G. Gad, A. A. Abohany and S. E. Sorour, "An Efficient Data Mining Technique for Assessing Satisfaction Level With Online Learning for Higher Education Students During the COVID-19," in IEEE Access, vol. 10, pp. 6286-6303, 2022, doi: 10.1109/ACCESS.2022.3143035.
41. M. S. Bashkari, A. Sami and M. Rastegar, "Outage Cause Detection in Power Distribution Systems Based on Data Mining," in IEEE Transactions on Industrial Informatics, vol. 17, no. 1, pp. 640-649, Jan. 2021, doi: 10.1109/TII.2020.2966505.
42. Z. Halim, O. Ali and M. Ghufran Khan, "On the Efficient Representation of Datasets as Graphs to Mine Maximal Frequent Itemsets," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1674-1691, 1 April 2021, doi: 10.1109/TKDE.2019.2945573.
43. Y. Sei and A. Ohsuga, "Private True Data Mining: Differential Privacy Featuring Errors to Manage Internet-of-Things Data," in IEEE Access, vol. 10, pp. 8738-8757, 2022, doi: 10.1109/ACCESS.2022.3143813.
44. G. Wang et al., "Mining Subsidence Prediction Parameter Inversion by Combining GNSS and DInSAR Deformation Measurements," in IEEE Access, vol. 9, pp. 89043-89054, 2021, doi: 10.1109/ACCESS.2021.3089820.
45. S. Cheng, D. Yang, T. Yang, H. Zhang and B. Cui, "LTC: A Fast Algorithm to Accurately Find Significant Items in Data Streams," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 9, pp. 4342-4356, 1 Sept. 2022, doi: 10.1109/TKDE.2020.3038911.
46. S. Ge et al., "Making Standards for Smart Mining Operations: Intelligent Vehicles for Autonomous Mining Transportation," in IEEE Transactions on Intelligent Vehicles, vol. 7, no. 3, pp. 413-416, Sept. 2022, doi: 10.1109/TIV.2022.3197820.

47. M. M. Rahman, Y. Watanobe, T. Matsumoto, R. U. Kiran and K. Nakamura, "Educational Data Mining to Support Programming Learning Using Problem-Solving Data," in IEEE Access, vol. 10, pp. 26186-26202, 2022, doi: 10.1109/ACCESS.2022.3157288.
48. J. Liu et al., "Urban Flow Pattern Mining Based on Multi-Source Heterogeneous Data Fusion and Knowledge Graph Embedding," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 2, pp. 2133-2146, 1 Feb. 2023, doi: 10.1109/TKDE.2021.3098612.
49. S. K. Patnaik, C. N. Babu and M. Bhave, "Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks," in Big Data Mining and Analytics, vol. 4, no. 4, pp. 279-297, Dec. 2021, doi: 10.26599/BDMA.2021.9020012.
50. Chittam S, Gokaraju B, Xu Z, Sankar J, Roy K. Big Data Mining and Classification of Intelligent Material Science Data Using Machine Learning. Applied Sciences. 2021; 11(18):8596. https://doi.org/10.3390/app11188596