A Survey on Data Mining Techniques and Applications

Jin Woo Kim, **Bilal Ahmed** , **Kavya Suresh**, **Priyanka Vyas**
jkim80@kent.edu, **bahmed2@kent.edu** , **ksuresh@kent.edu** , **pvyas2@kent.edu**

Abstract
Digital technology in everyone's hands makes a vast amount of data available. Consequently, the last decade has seen a rise in interest in data mining for the purposes of discovering previously unknown correlations and predicting future trends. In fact, Knowledge extraction from massive data sets is a tough task. However, Data mining and Machine learning tools play an important role here and their applications are used in a wide variety of fields, including education, finance, Clinical research, healthcare, agriculture, defense, and marketing, amongst others. This paper presents a survey on Data Mining Techniques and Applications by showing recent work in this field and discusses them.

Introduction:
It has long been a practice to dig over data to find hidden relationships and forecast upcoming trends. In order to forecast outcomes, data mining is the act of identifying anomalies, patterns, and correlations within huge data sets. The new knowledge can be used to strengthen customer interactions, lower risks, raise profits, and more by employing a variety of strategies. The enormous proliferation of information data in today's time makes it challenging to extract usable information from it, which is why data mining came into being. Data mining has a wide range of uses in the modern world. Applying algorithms to a huge volume of data, interpreting the data, and mining the hidden information is known as data mining. Classification, clustering, regression, association rule mining, and correlation are the major methods of data mining and few examples of these techniques include neural networks, SVM support vector machines,etc. Data mining is a cutting-edge method of data analysis that can efficiently extract useful information from various forms of data.

Specifically, following operations are implemented during data mining techniques 1) Predicting trends and behavior automatically. 2) Hidden associations in the data can be found by association analysis. 3) Clustering can help people better grasp how items are similar. 4) When comparing the observations with the reference values, deviation detection might seek for significant deviations. The majority of data mining diagnosis techniques rely on a single algorithm model for implementation. Several industries have implemented data mining techniques to sort through the inconsistent and recurring noise in their data, recognize the pertinent facts, then apply it effectively to predict future outcomes and improve decision making ability. Data mining refers to a variety of methods or techniques used in various analytic capabilities that meet a range of organizational needs, ask various questions, and rely on differing degrees of human input or rules to reach a conclusion. Data Mining is a first step, to use data as a basis to build a model to predict future patterns.

Data modeling aims to leverage historical data to forecast future patterns. A step in the data modeling process is data mining. Data modeling methods are of two types (1) Descriptive Modeling Method, in order to understand the causes of success or failure, it uncovers commonalities or clusters in historical data. One example is classifying customers according to their attitudes about particular products. The descriptive modeling technique and its applications is more detailed in Section 1-1.1 (2) Prediction Modeling Method, this modeling method is used to categorize future events or anticipate unknowable outcomes, such as determining a person's likelihood of repaying a loan by utilizing credit score. The predictive modeling technique and its applications is more detailed in Section 1-1.2

Section 1: Data Mining Methods and its Applications:
Data mining is the technique of obtaining useful information from massive amounts of data by applying various approaches and procedures to huge datasets. These methods and algorithms are essentially the procedures, and they are used on data sets. Data mining techniques typically make use of

relational databases, transactional databases, and data warehouses. For more complicated data types, such as time series, symbolic sequences, and biological sequential data, there are, however, some advanced mining techniques.Following are few of the data mining methods or technique:

1.1 Description Method

1.1.1.Clustering

Clustering is a data mining technique that identifies similar clusters of objects in a given dataset. This approach uses unsupervised machine learning and works on unlabeled data. All the items in a cluster that is formed by a collection of data points would be members of the same group.Clustering analysis is generally used when no presumptions are made on the possible relationships present in the data.

Data mining and machine learning are used in multiple domains, likewise it is also used in modern energy management systems (EMS) for optimal load dispatch. The voltage-current variation during High Impedance Faults (HIFs) is like the sudden change in load. Vijay kale et al. proposed a method using ML and Density-based Clustering Algorithm with Noise (DBSCAN) to decrease the ambiguity and improve the fault discrimination **in their paper[2]**. Discrete Wavelet Transform (DWT) is used to process the input vector for the ML models such as Support Vector Machines (SVM), Decision Tree (DT), Multi-Layer Perceptron Neural Network (MLPNN), and Random Forest (RF).System state estimation is one of the key responsibilities of EMS as in most cases it is often undetected mainly due to high impedance in the faulty section which in turn decreases the current causing blinding of the overcurrent-principle-based system. The need for capturing uniqueness during HIF other than CTs monitoring became necessary to distinguish HIF from change in load. Either data mining or machine learning is generally used by modern EMS for this issue, however detection accuracy can be compromised in the events that can occur in the microgrids. They also proposed using both CT and VT errors together along with other events like low impedance faults, transients due to switching, and sudden load change to increase the accuracy. Here data mining helps in differentiating routine data and the event data prior as a preprocessing step. The first contains HIF samples and the second data contains low impedance faults. This preprocessing really increases the accuracy in HIF detection. The algorithm also can generalize all communication assisted microgrids as it only uses partial knowledge of system events for training purposes and can accurately detect HIF even for small amounts of change in load. The implementation of the proposed algorithm only calls for a few extra sensor installations and can improve the accuracy of HIF prediction in microgrids to a great extent.

Another approach using neural network for data clustering for prediction. In this paper, they used fuzzy neural network model to analyze the data prediction and created a prediction framework based on fuzzy C clustering. From their result, the clustering effect of IDWFCM altorithm is not affected by noise data, therefore, the convergence speed of the system is higher than the traditional clustering algorithm. The overall increase of 60% and the clustering accuracy to 94.2% from 88.4% which can be seen in **table 1**.

The algorithm name	The number of iterations	Clustering accuracy
FCM	21	88.4%
DWFCM	25	94.7%
IDWFCM	7	94.2%

Table1: Performance Comparison Table

1.1.2. Association Rule

In a given dataset, this data mining technique is used to search for recurring associations. The database's various items are analyzed for any relevant correlations and relationships in order to spot patterns.The patterns are characterized in the form of association rules. Association rules are created to calculate from itemsets, which are generated by two or more items.

If all of the potential item sets from the data were analyzed to build the rules, there would be so many rules that they would be meaningless. For this reason, association rules are frequently constructed using rules that are accurately reflected in the data.Apriori, Eclat, and FP-Growth are a few important algorithms, but as they are designed for mining frequent itemsets. Once the frequent item sets are identified, the next step is to create rules from frequently occurring item sets

identified in a database.

In paper**[32]** Tong Su et al proposes PSOFP growth algorithm which is an improved algorithm of association rule **Fig(1).** The search for the optimal support is made using the particle swarm optimization technique and a fitness function is proposed for association mining. It uses information entropy as a measure of association rule effectiveness and seeks to minimize the incidence of erroneous rules.

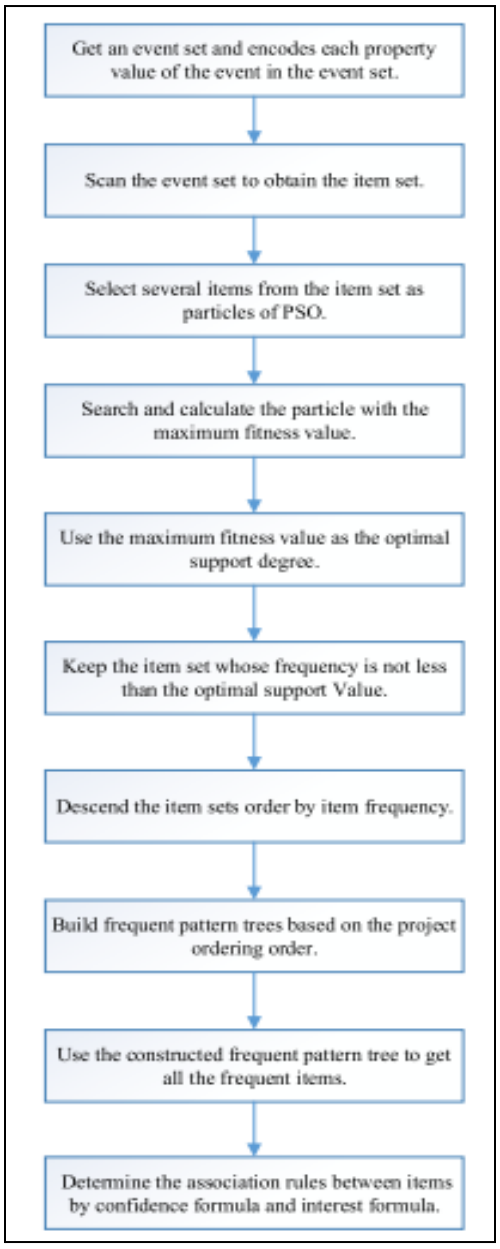

Fig 1: Flowchart of PSOFP Growth Algorithm

This research selects 5000 social security incidents with the following characteristics: time, location, attack target, and number of casualties. The association rules between events are mined using our improved algorithm, which also improves the validity of the rules.

The model is compared with traditional algorithms and the following improvements are observed that after determining the minimum support, the PSO algorithm is used to first determine the optimal support. When scanning the event set for the first time, filter out the important items, and then remove any applicable association rules. After obtaining the association rule in accordance with the confidence formula, calculate the information entropy to determine the degree of interest and then find the association rule that does so. Firstly 5000 social security events with attributes of time, place, attack target and casualty number are selected.

Any missing attributes and duplicate events are deleted. All of the death toll's attribute values were assigned to the same value within each interval after the attributes were divided into categories. additional characteristics, such as the time, location, attack target, etc. corresponding letters and numbers, respectively, stand in for various attribute aspects.The idea behind FP-tree building is that only items that are frequent along the same tree path may be frequent, but items that are frequent along other paths are not.

Support (count)	Apriori (count)	FP-growth (count)	PSOFP-growth (count)
20	2150	2030	1480
40	1385	1289	855
50	898	975	558
80	350	367	105

Table 2: The number of rules generated is compared with the traditional algorithm.

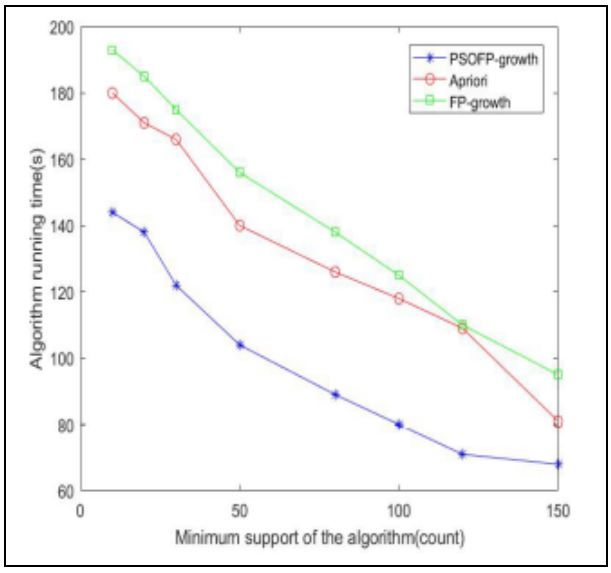

Fig 2: Algorithm running time

As noted in the **table 2** above, numerous invalid association rules are eliminated by changing the preset support value, which results in the PSOFP-growth algorithm producing over 30% less association rules than Apriori and FP-growth algorithm. **Figure(2)** shows the results of experiments on the running times of the classic algorithm Apriori and FP-growth and the new, improved method PSOFP-growth. The new algorithm runs substantially faster than the old approach, as shown in the figure. Further evidence that the value of support has a significant impact on the algorithm comes from the fact that the shorter the running duration of the algorithm is, the higher the support is set.

Fake news is another data mining application presented in **[30]** example Demand for methods to identify fake news stories has risen rapidly in response to their proliferation and the damage they do to institutions like democracy, justice, and public confidence. The multifaceted nature of the problem with false news necessitates the combined efforts of researchers from several fields, including but not limited to computer science, political science, journalism, sociology, psychology, and economics. The authors describe approaches to detecting false news from a number of angles, including data mining, machine learning, natural language processing, information retrieval, and social search, and focusing on both news content and social network information.

Due to the nature of fake news, spotting it is a difficult and multifaceted endeavor. A wide variety of news- and social-related data (such as feedback, propagation channels, and spreaders) are utilized by the detection strategies. Text, multimedia, network, etc. are all examples of different types of information, and each has its own set of methods for processing and sets of useful tools at one's disposal. This video looks at how to spot fake news from four different angles: content understanding, presentation, virality, and authority. To be more precise, when looking at fake news detection from a knowledge viewpoint, it is a "compare" between the relational knowledge retrieved from the unverified news articles and knowledge-bases representing facts/ground truth.

Xiaofeng Li et al **[40]** proposes a density-based approach for mining association rules from streams of medical data. This study builds a neural network using the compound neural network algorithm, trains it, extracts the distribution pattern association rules for the medical data stream, and completes the mining research using the pre-processing of medical data as the mining support. The following network structural parameters are as below **table(3)**:

Parameter	value
Dimension	[1-50]
Number of convolutions	2
Number of iterations	60
Weights	[-100,100]

Table 3: Structural parameters for model

where, the experiential indicators are :

1. Data redundancy probability
2. Histogram density estimation
3. Stability of medical data distribution pattern, and,
4. Mining result approximate root mean square error

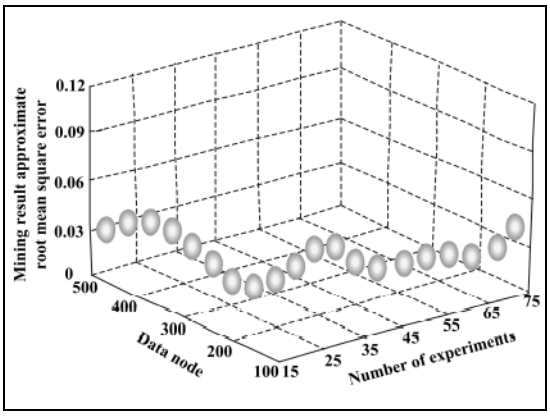

Fig 3: RMSE mining result of proposed algorithm

As shown in **Fig (3)** The medical data is more stable and can effectively identify duplicate data and eliminate data redundancy; the proposed mining technique obtains a more accurate histogram density estimation; With reduced mining time, the mining outcomes have an RMSEA of 0.03. The dispersion level of medical data is within a suitable range and the medical data has excellent stability; the proposed algorithm's contour curve is most similar to the true probability density curve; the likelihood of duplicate data is lower; RMSEA of the mining results is low; data mining takes less time.

Another example of data mining applications studied in **[28]** is Educational Data mining. In order to keep up with the rapid development of technology, numerous efforts have been made, such as educational data mining studies and projects that make use of artificial neural networks.In order to improve student retention, student progression, and cost savings, universities could benefit from utilizing the massive amounts of data they collect on their students to provide more personalized academic advising that supports adaptive learning. Predictive data mining tools and machine learning techniques, such as ANN, are necessary for optimally mining students' data. Educational data mining is a data-driven, technology-enhanced method of teaching that uses data science techniques like artificial intelligence (AI), data mining, and data warehouse to use data about learning to make decisions in the learning environment that are more informed. Artificial neural networks can be used to extract objective and non-discriminatory data about students to power personalized learning. It's a data-driven method for forecasting academic outcomes and categorizing students' approaches to learning by applying methods of function approximation (regression), pattern identification (classification), and predictive analytics to information gathered from classroom observations. As a result, ANN can be used to support innovative teaching methods like intelligent tutoring and academic advice.

A Data association rule mining method based on RBF neural network optimization algorithm is stated in **[12]**. To develop a real-time accurate data mining, Based on considering the constraint association rules, the data frequent itemsets which are reduced to get the corresponding candidate datasets that meet the rules. In their model, the candidate dataset is input of the RBF neural network and output is optimized by combining with rough set theory. In their experiments, they compared their method with other approaches where the redundant data capacity and the amount of available data related to the algorithm execution rate are stated in **figure(4)**. Their approach showed better accuracy than the others.

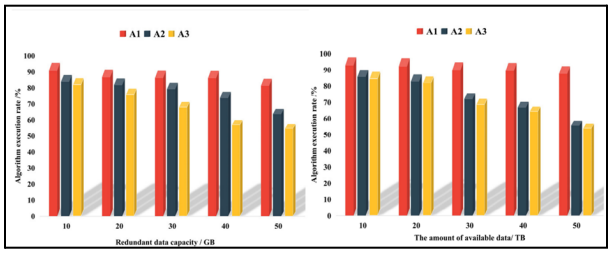

Fig 4: relation between algorithm efficiency and redundant data capacity.

1.1.3. Sequential Pattern

Sequential Pattern Analysis focuses on identifying statistically significant patterns between data samples where the values are presented sequentially. Sequence Pattern Mining differs slightly from Time-Series Mining in that the sequence does not always need to have a sense of time.The following steps are essential for

sequential pattern analysis
(1)Identifying recurring trends,
(2) Sequence comparison in the data,
(3)Locating missing sequence components, and
(4)Constructing effective sequence information indexes

Wang-Cheng Kang et al **[33]** proposes a self attention based sequential model (SASRec) that is, similar to an Recurrent neural network, allows us to capture long-term semantics, but bases its predictions on only a small number of activities (like an Markov Chain)

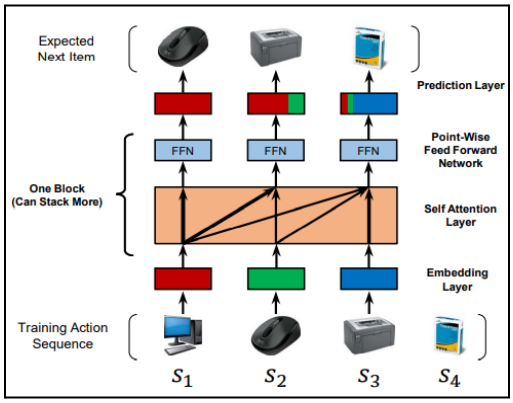

Fig 5: Training process of SASRec

In order to forecast the subsequent item at each time step, SASRec looks for "relevant" things from a user's action history. The study evaluates the method on four datasets Amazon, Steam, and MovieLens. The domains, platforms, and sparsity of the datasets differ significantly.

As in **Fig 5** this study uses two self-attention blocks (b = 2) and the learnt positional embedding for the architecture of SASRec's default implementation. The prediction layer and the embedding layer share item embeddings. TensorFlow is used to implement SASRec. The learning rate is set to 0.001 and the batch size is 128. The optimizer is the Adam optimizer . Due to their sparsity, MovieLens-1m has a dropout rate of turning off neurons of 0.2 while the other three datasets have a dropout rate of 0.5. For MovieLens-1m, the maximum sequence length is 200; for the other three datasets, it is 50

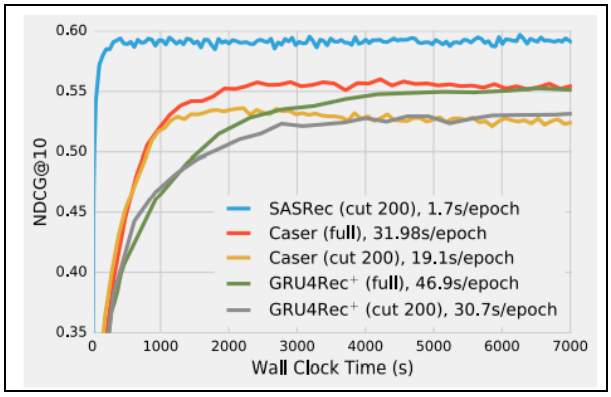

Fig 6: Training efficiency on ML-1M.

As in **fig 6** In terms of processing speed, SASRec is over 11 times quicker than Caser (19.1 s/epoch) and over 18 times faster than GRU4Rec+ (30.7 s/epoch) for one epoch of model changes. Additionally, it is observed that SASRec converges to optimal performance on ML-1M in about 350 seconds as opposed to substantially longer times for other models. Also, Caser and GRU4Rec+ perform better when full data is used.

Researchers in **[26]** propose a study about Motif Prediction. When mining unstructured data, higher-order network analysis is a useful tool. However, it does not have the tools and methodologies necessary to foresee how the linked datasets would develop. One of the most important challenges in graph mining is link prediction. Recent research, however, has shown the value of higher-order network analysis, in which motifs, which are complex structures, are treated as equals.

The authors also successfully extend the suggested architecture to the prediction of more random clusters and communities, showcasing the importance of using it for graph mining applications as well. Authors have devised a challenge of forecasting general complex graph structures termed motifs, such as cliques or stars. They have shown how it differs from simple link prediction and proposed heuristics for motif prediction that are size-independent and can detect connections among links that make up a motif. The study allows domain knowledge to be included, so it can serve as a basis for creating

motif prediction systems within certain domains, just like link prediction can. While heuristics are efficient, they have room to grow in terms of prediction accuracy. To solve this problem, researchers created a GNN structure for motif prediction. Their work shows superior performance than the state-of-the-art by up to 32% in area under the curve, providing exceptional accuracy that grows in tandem with the size and complexity of the projected motif. They have highlighted that they successfully utilized the proposed architecture to predict more arbitrarily organized clusters, demonstrating its greater potential in mining irregular data.

Carson K. Leung et al **[36]** proposes an item-centric privacy-preserving method (PP-UV-Eclat) for mining common patterns from large-scale uncertain data in the Spark environment. While some algorithms (such as Apriori and FP-growth) see the transaction database TDB "horizontally," others (such as Eclat) view it "vertically" as a collection of items (i.e., TDB = {tidset(x) | x ∈ domain items}) where, tidset is a set of transaction IDs. Here, it's important to record the existential probability P(x, t_i) connected to each item x in each transaction t_i in the uncertain database for mining uncertain data (UDB). In this study, the mapped record's key changes to "item|transactionID." The algorithm initially transforms the original data into the <item|transactionID,value> format. The data is gathered and multiplied in the reduction phase using the same key, "item|transactionID."In order to speed up the search, items are stored in a trie structure. The support value is kept in the mapped key and kept on the associated item and particular transaction. The key is formatted as itemKey|transactionID|supportValue. PP-UV-Eclat records the existential probability value of the uncertain data related to each item. By adding the existential probability values across all transactions, it then calculates the predicted support of every singleton frequent pattern. By intersecting the tidsets of the related (k − 1)-itemset and 1-itemset, the frequent k-itemset is identified.By multiplying their respective expected support values, the corresponding expected support can be calculated.

PP-UV-Eclat algorithm performs a post-processing phase to assure privacy preservation after identifying all frequent patterns. To hide the true numbers, the algorithm more specifically introduces statistical noise and probability.

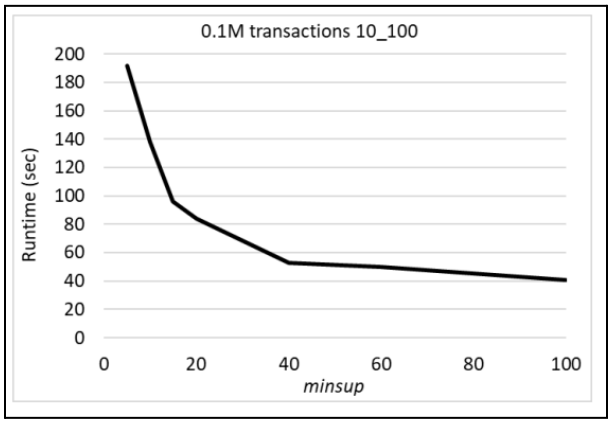

Fig 7: Evaluation of PP-UV-Eclat: runtime vs. minsup.

In **fig 7** the outcome demonstrates that for two datasets, the runtime dropped as minsup increased.

Opinion Mining is another example studied in **[27].** An organization's success can be attributed in large part to its ability to identify and connect with its desired customer. As a result, feedback from customers on social media platforms is a gold mine of information for businesses' long-term marketing strategies. Recent developments in technology, particularly in the field of social networks, have made it feasible to disseminate knowledge to a broader audience at a tremendous speed. Users of the Internet can record and share their thoughts and perspectives through a number of channels, including message boards, weblogs, Twitter, and other social media sites. Since this approach to gathering feedback is more cost-effective, time-efficient, and adaptable, it has received a great deal of focus from many different types of businesses. Opinion mining, also known as sentiment analysis, is the study of extracting and analyzing people's subjective thoughts and feelings from online content. One of the most

explored topics in data, web page, and text mining, it is also one of the most active study areas in natural language processing. Whether the text is a post, tweet, review, or remark, opinion mining can be applied to relevant portions of it.

Attempts should be made to classify opinions using both conventional text mining methods and cutting-edge deep learning technologies. Researchers have analyzed the quality and efficiency of deep learning and non-deep learning algorithms applied in opinion mining of hotel reviews posted on TripAdvisor. Algorithms used in the study are Latent Dirichlet Allocation, Naive Bayes, Logistic Regression, and Long Short-Term Memory Units. Authors examined classifiers' accuracy, precision, recall, F1-measure, and training/classification time. LSTM produced the best quality metrics as a result. However, it did not give results that were satisfactory in reference to the processing time.

Online Study Behavior Modeling is proposed in **[29].** As the Internet and education continue to become more intertwined, we have seen the rise of online tutoring platforms like Coursera and Udemy in recent years. With the COVID-19 pandemic forcing schools to relocate their classrooms, there has been a surge of interest in distance and online learning. The study quality is compromised by the fact that students learn by viewing videos, which lacks face to face interaction from teachers and fellow students, however online tutoring services give students access to huge learning materials and tools. In this research, authors propose investigating the challenge of modeling student online study activity patterns for evaluating and predicting online study quality as a means of mitigating this issue. The authors of [29] define a study behavior sequence that they use to characterize various types of online studying activities. Each time stamp represents a different study behavior, such as "viewing," "dragging ahead," or "dragging back," when watching a video lecture. Researchers have created a neural hawkes process framework for modeling digital research habits. Authors claim that EduHawkes **(Fig 8)** is an innovative hierarchical encode-decode architecture that can optimize both the study behavior prediction task (event level) and the study quality prediction task in tandem (course-level). To show how well the suggested method performs in predicting online study activities, the authors conducted tests using EduHawkes for the purposes of study quality prediction and identifying careless students. The amount of time spent watching lecture videos by students as a percentage of all time spent on the online tutoring platform is how the authors define the term "online study engagement." The authors obtained the data from a website that provides educational services, such as tutoring. A student's details for a single class are recorded in the student log.

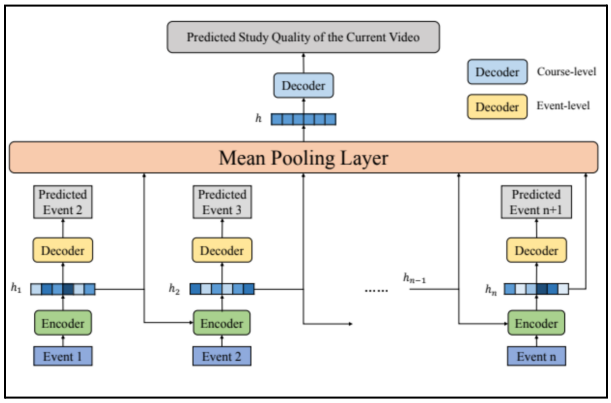

Fig 8: Eduhawkes Model

Qian Guo et al **[38]** proposes a learning model that is capable of directly mining and reasoning the relationship between two input images and one output image, without presetting any reasoning patterns. This study uses Fashion MNIST data to mine fashion logic task. There are 50,000 training samples, 5,000 validation samples, and 5,000 testing samples in each data set for a visual logic relation. Additionally, arithmetic procedures use decimals and binary numbers for bitwise operations. The size of the images is set to 28 × (28 × 14), for the Bitwise And and Bitwise Or data

sets, meaning that the maximum number that may be encoded in one image is a 14-digit integer. The size of the images is set to 28 × (28 × 7), for the Addition and Subtraction data sets, so the maximum number that can be embedded in one image is a seven-digit number. The model takes two images as input and outputs one image for comparison with the original two images. Through mean-square error (mse) loss optimization and the use of the ADAM or SGD optimiser, the model is trained to generate a single correct output image.

Model	Operations			
	Bitwise And	Bitwise Or	Addition	Subtraction
LSTM	100%	100%	73.46%	74.62%
CNN-LSTM	100%	100%	57.78%	55.10%
MLP	100%	100%	99.44%	99.38%
CNN-MLP	100%	100%	99.82%	99.8%
Autoencoder	100%	100%	97.60%	97.52%
ResNet18	100%	100%	99.38%	99.08%
ResNet50	100%	100%	99.54%	99.80%
ResNet152	100%	100%	100%	100%

Table 4: Model test accuracies

This study shows that all models outperform the random guess method on all tasks, it is possible to directly mine logic relations from data using a data-driven approach. All models outperform Addition and Subtraction tasks in terms of accuracy on Bitwise And and Bitwise Or tasks. LSTM and CNN-LSTM models perform less accurately on addition and subtraction tasks when compared to other models as shown in **table 4.**

In another example authors in **[24]** propose a Web Data Extraction model Using CNN +LSTM. It's early days for the use of deep learning to filter web data for useful information. Web scraping methods (including open source and commercial crawlers and APIs) are used to extract data. Traditional end-to-end automated online data extraction systems based on ML/DL and DL have something called the "Extraction Engine" as one of its fundamental sub-components.

According to the authors, the Yolo model was conceptualized in 2015. The first step is to classify an image, and the second is to pinpoint where in that image an object is located. A single category, such as "electronics" or "book," is allocated to each image in the process of image classification. Using the region of interest pooling mechanism, this strategy rapidly developed into fast R-CNN. Following the initial fully differentiable R-CNN and Mask R-CNN models, faster variants have since appeared. In addition, Yolo was launched to facilitate Detection of Images or Objects in Real Time.

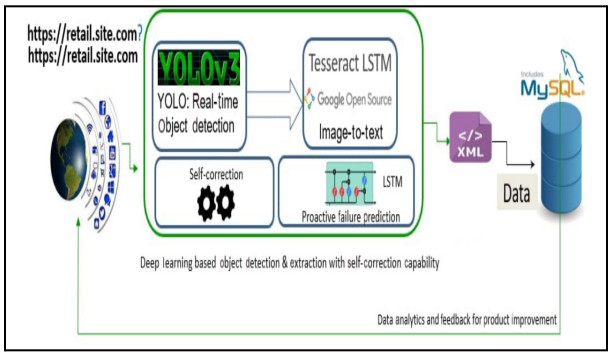

Fig 9: Web data extraction system architecture using Yolo and Tesseract

Using deep learning object detection technologies like Yolo **(fig 9)**, the suggested approach in this study fills this void in modern and machine learning techniques by doing away with DOM or wrapper methods for data extraction. This is because the suggested system's innovation relies on an object identification technique, which can identify the object anywhere it may be on a webpage. Once a website's item or product page is identified, LSTM is used to parse the object for information about the product. The proposed approach is not only responsive to shifting website design but also performs data extraction mechanically. The execution time of Yolo and LSTM architectures is greater during the initial run, but reduces in successive cycles, hence decreasing the overall execution time. As graphic image processing is computationally costly, the training time of a model is dependent on the system GPU. Additionally, it automatically extracts data. The suggested approach eliminates the subcomponents of the extraction engine that are fundamental to traditional and machine learning strategies for successful data extraction.

Therefore, the method has the potential to revolutionize automated data extraction from several websites in the future.

It is also important to mention that the capacity to handle these enormous data sets in real time using Big Data Analytics (BDA) tools and Machine Learning (ML) algorithms has numerous advantages as shown in the survey. However, It might be difficult to determine which of the many available free BDA tools, platforms, and data mining tools is best suited for a certain project. Authors in **[25]** discuss relevant areas of application, difficulties, and, most crucially, research gaps. On December 16, 2020, the daily volume of data created worldwide was 59 Zettabytes. As we go into a future that is increasingly more data-driven, its size is projected to grow to 149 zettabytes by 2024. Consequently, it is crucial for businesses around the world to extract useful insights and competitive benefits from these massive data sets. The research, however, reveals that it is difficult to efficiently and correctly obtain useful insights from Big Data in a rapid and easy manner. Therefore, Big Data Analytics is now a must-have for any company that wants to enhance their bottom line and expand their market share. While the majority of AI/ML algorithms and the necessary platforms for implementing BDA are open source and free to use, doing so effectively requires a new set of skills that is not typically found among practitioners.

The authors conclude that there is a dearth of work on the integration of ML technology and big data analytics across a wide range of fields. Big data is causing waves in every industry, yet no previous research have done enough to address this issue. Therefore, it is not appropriate to confine the big data analytics assessment to just one or two sectors. The authors provide a quantitative synopsis of the findings from 66 papers. This literature review identifies five major themes that can be used to classify the state of the art in Big Data Analytics: I the central BD domain for dealing with scale; (ii) dealing with data noise and fuzziness; (iii) considering privacy and security concerns; (iv) data engineering; and (v) meeting the needs of both BD and data science. Performance of a machine learning model can be measured in a number of ways; the appropriate metric(s) will vary from task to task. The most popular ones include MSE, RMSE, RMSE, MAE, accuracy, precision, recall, AUC, and F-score.

1.2.Prediction Method

1.2.1. Classification:

The data mining technique of classification divides up the objects in a collection into specific groups or classifications. Identifying the correct target class for each occurrence in the data is the aim of classification. According to the WHO report, over 17 million people died of heart attacks in the last few decades.

Aushtmi Deb et al. [1] has done research on the various data. mining models like Naive Bayes, Random Forest Classification, Decision tree, K-Nearest Neighbor, Logistic Regression, and Support Vector Machine that can be utilized in effectively predicting heart disease and compared their efficiency to eventually build an intelligent Heart Disease prediction system that can give accurate diagnosis based on historical data. As machine learning lets the machine learn. Gradually improving the accuracy and data mining is extraction of previously unknown and potentially useful information from data, these two approaches can be combinedly utilized. In their approach, they divided 20% data into a testing data set and the 80% as the training data to train the **model[10**]. The main focus was given on the factors like predictive accuracy, precision, recall and F1 score. The flow of the model is given as the **flowchart below[11]**. The accuracy of prediction is best in the case of RF model and the lowest accuracy was given by the logistic regression approach.

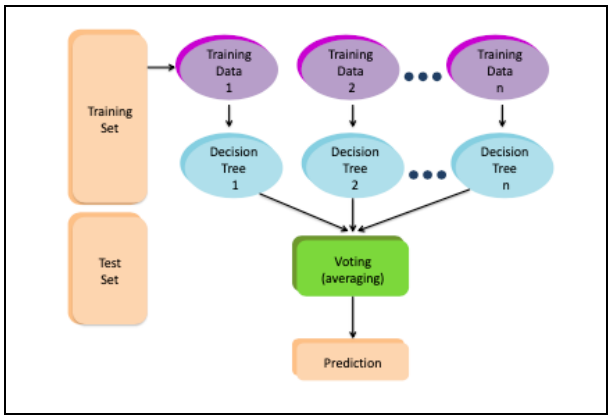

Fig 10: Working process of Random Forest classifier

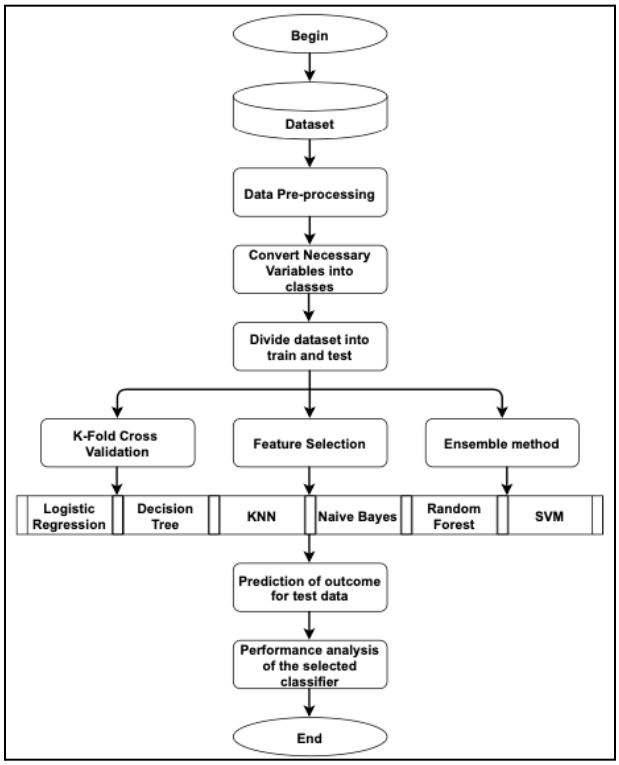

Fig 11 : The flow diagram of the proposed data mining and machine learning approach

Credit card usage has become very common in current society. The increase in the credit card usage also increased in the credit card frauds in different forms. The credit card fraudulent activities can happen from anywhere like via internet service, mobile phone, e-banking platforms etc. What makes it even worse is when the end user fails to recognize that such activities are happening in their account. Most common cyber-attacks that are carried to hack into user and credit card information are Phishing, Vishing, Mishing and Smishing. Mainly for detecting the credit card fraud activities, the financial institutions mainly use different data mining techniques, ML and AI tools are also used along with other mathematical tools where they try to do pattern recognition on features like sensitivity, accuracy, efficiency, precision, specificity to validate a transaction. Avinash Sharma et al. discussed on various data mining techniques like C4.5, CART algorithms, J48, Naïve Bayes algorithm, EM algorithm, Apriori algorithm, SVM etc which can be used in the analysis of credit card fraud detection in **their paper [4]** and the results of these supervised and unsupervised methods are verified. The use of these techniques along with increasing the awareness by alerting the user via notification can be helpful in decreasing the credit card fraud activities.

The data accumulation in various departments has exploded as the data in the society is increasing exponentially. This call for automation in various data mining operations to retrieve information. In case of this kind of massive data, the data mining process is overwhelmed by its own data processing, Machine learning algorithms can be made use in this situation. **Yongxu Li, Chanchun et al. [5]** proposed a Three-stage positioning method which combines the SVM-based positioning and K- nearest neighbor method which can be used in problems like positioning mobile terminals so that the call traffic can be controlled efficiently. This new method seems to improve the accuracy of positioning and decrease the computational complexity. The Support Vector Machine (SVM) Algorithm is a novel classification algorithm which can be used in problems with nonlinear sample and high dimensional pattern recognition needs. SVM positioning algorithm makes data into small grids which are then further abstracted into categories and then from the positioning area, a large amount of information is collected. K- nearest Neighbor Method (KNN) is an algorithm majorly used in data mining and machine learning problems. In the proposed method mainly make use of the support vector from positioning and later the position is rasterized. Large amount of terminal information is collected from the positioning position, which can be later analyzed to create the measurement report. For a single user, first the

longitude and the latitude are located and then it is further divided into grids generally of one kilometer. Multiple road tests are conducted to find the average value of the longitude and latitude of the measured position. This helps in improving the precision. This information is then given to supervised learning algorithms like K-Means. With the help of SVM, the supporting vector and the decision function are calculated. This method of positioning has a higher accuracy and lesser computational complexity as compared to the traditional methods.

Nowadays, cardiovascular diseases contribute a lot to the mortality rate. It is not possible to monitor the patients every time. But we do have a repository of information from all the previous patients. This can be utilized wisely to give better care for the patient. Different heart attributes can be analyzed to find the most important features that contribute to cardiovascular health. Mostly these can be seen as costly and not always productively ready in a health care system. As coronary disease can even compromise the lifespan of a person, it is crucial to foresee the patient's coronary illness. **Simran Verma et al. [6]** utilized the coronary illness dataset at UCI repository for performing data mining to retrieve useful information regarding the key attributes which are responsible for the diseases.

There is an immense quantity of data available in the healthcare domain which are stored in servers. Generally, the users can access these data with the help of a search tool by using questionnaire methods. Even now the conventional approach is followed for disease prevention in the clinics and applied on an individual basis. **[8] Use of data mining** in this domain can combinedly make use of various decision fragments and past cases experiences can help in creating meaningful observations. The main categories of data mining utilization in heath domain are health informatics, client association supervision, effectiveness therapy, testing for waste and deception, medicinal investigation. The therapy quality can be monitored in multiple patients to find the most effective practice, also the data analysis can help in detecting diseases which causes the deterioration in health and monitor its correlation with other diseases and hence helping them by giving the patient a better chance to recover. The challenges that can be faced while trying to implement data mining in the medical domain are many. The first and foremost is the collection of the data. The data can be misused in many forms like identity theft or privacy breaching etc. and the process can be costly. Another issue is in the variation in the type of data like the data can be either numerical or sometimes even pictorial as in cases of scan reports. Data mining models that are discussed in **[8]** are the neural network model and decision trees. Artificial Neural Networks (ANNs) consist of several interconnected computer units and its brain circuits. Grouping, sampling and creation of prediction model can be carried out in the ANNs which often uses linear data mining methods and fuzzy logic is used for reasoning. A decision -making tree is a statistical framework used in forest, clarification, regression and grouping the health data. It can be used for visualization to aid the pragmatic and graphical interpretation of the information. The use of decision trees in the health domain can minimize the dynamic connections by division in major subsets of the input factors. Since it can be visualized, it will be easier to understand and provides an anti-parametric, distribution-free solution and is simple to manage.

Weather prediction has a vital place in human civilization. Application of different data mining techniques in weather prediction is discussed by **Mohammad Zeyad et al. [10]**. The data for this study was mainly atmospheric study data collected by the government institutions.Data mining in weather prediction examines the available meteorological data to find previously unknown patterns. The different approaches that are used for weather prediction are the synoptic method where several weather patterns are observed over a short span of time to create synoptic charts and are documented on a daily basis, the next method is numerical weather prediction where a computer program is made use for the weather prediction. If the early stages of the weather is not understood properly, this method fails in accurately forecasting the weather. The third method is by using statistical equations, but even this method also cannot be dynamically used to predict the weather accurately. In utilizing data mining techniques and ML approach in real time on

different factors like humidity, temperature, wind gust etc K-means algorithm and decision trees are the best approaches for the prediction of weather phenomena such as temperature, thunderstorms, rainfall, and precipitation.

To boost sales forecasting accuracy, the authors in **[22]** used a ML model as shown in **fig (12**) to examine sequential data from marketing datasets. They emphasize the need for model analysis that accounts for consumer heterogeneity in order to correctly assess client attributes and conduct precisely designed marketing approaches. Their work centers on artificial intelligence and includes client diversity into machine learning model settings. The authors use convolutional neural networks and attention mechanisms to help extract specific trends from clients' purchase histories and use this information for sales forecasting.
To better serve their customers, online retailers like Amazon collect detailed records of when, how often, and what their customers buy. Even while services like Google BigQuery and Amazon Web Services (AWS) have established a foundation for data analytics and expanded the types of data that can be accessed, there are still inefficiencies between companies. If customer segmentation is implemented based on the model's expected purchase amount, it is possible to narrow down the clients who will make more purchases in the following months. This improves marketing accuracy.

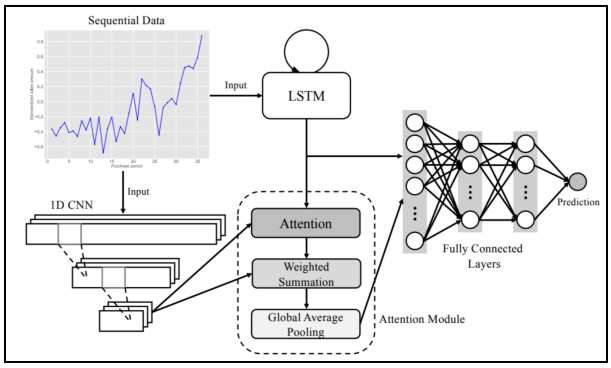

Fig 12: Flow chart of the Sales forecasting algorithm

In paper **[31]** Yuanpan zheng et al presents two key models for sensor data modeling:Long Short Term Memory (LSTM) prediction model and Support Vector Machine (SVM) model based on Internet of Things (IoT) data. The two data analysis models are designed on a cloud platform of IoT systems based on Hadoop framework so as to improve the efficiency of algorithms.

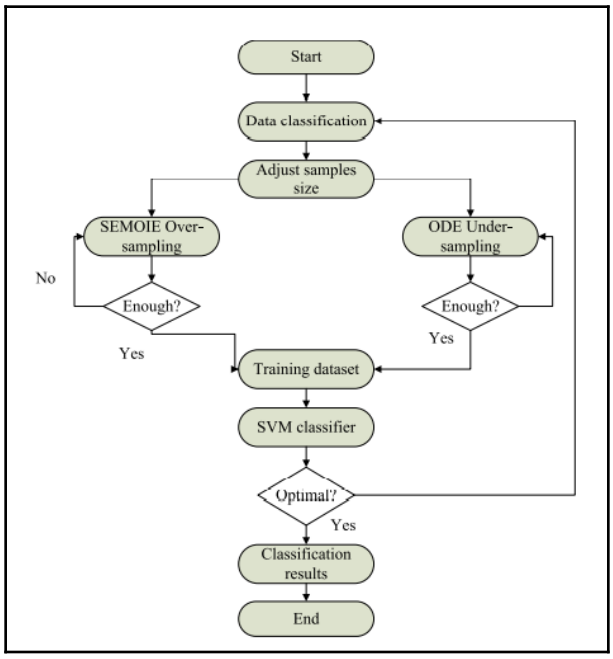

Fig 13: Flow chart of SVM algorithm

The technique has high requirements for the initial trajectory of the particle filter and high data noise. SVM's weight is optimized using PSO, which also increases the algorithm's accuracy and speed of convergence. It consistently increases the algorithm's effectiveness, lowers its energy usage, and increases its viability. The flowchart of SVM Algorithm in **Fig 13** shows that the data sets of each sub-site are first subjected to SVM local mining; following this, a multi-tree building technique is used to map the support vectors recovered locally into local feature multi-trees, and support vectors and data are then loaded into the following site via mobile agents. Then, the fresh samples—shell vectors from the first few sites—and the old samples—samples from the following site—are combined to mine. The mining of SVM in a distributed environment is eventually realized as a result of the accumulation of sample sets (the movement of each site). After that, the numbers are tallied. Finally, data definition, data mining procedures, and data valid information are

developed. In SVM, it is relatively simple to regulate the number and degree of support for the rules, and the resulting rules are of higher quality.

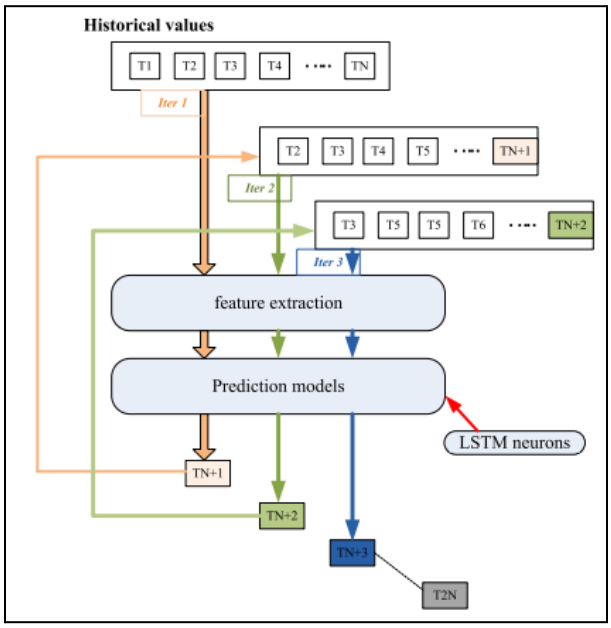

Fig 14: Prediction model based on LSTM

In simulation the main analysis data points are the wind sensor data, air pressure sensor data, temperature data and relative humidity. For LSTM prediction the data **fig 14** is divided into a training set and test set. The first data point represents training data, and the last data point represents test data. A comparative study of prediction achieved by the respective model shows that LSTM has a greater descent gradient and improved training efficiency compared to SVM based model.

There is another approach that used neural network data mining system for the sales analysis. **[11]**. Especially, they used neural network data mining for the sales prediction. There neural network system consists of BP neural network, Bayesian network, radial basis neural network and time series of combining decompoisition and neural network. To verify their function of neural network data mining algorithm in the sales forecast analysis of network marketing product, they used decision tree and Bayesian network algorithm in the experiment. The figure [15] shows the Bayesian network algorithm initialization training and the profit forecast using the neural network model. They did not provide any evaluation method of similarity but as I see in the figure, the predicted result is very similar to the real data.

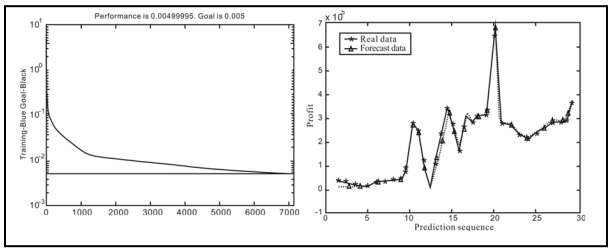

Fig 15: Bayesian network algorithm initialization training and profit forecast for the first quarter of 2013.

There is another classifier method using multi-classifier fusion approach **[16]**. They used student credit classification data to make a classification data mining method using BP neural network fusion classification algorithm based on AdaBoost. Then they combined the algorithm with student credit classification. The data mining algorithm based on multi-classifier fusion can be seen in the **figure[16]**.

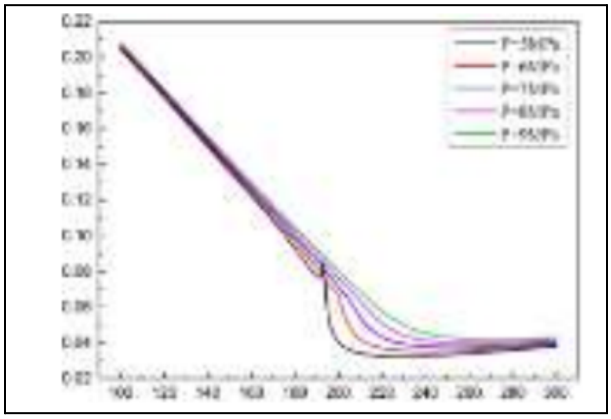

Fig 16: data mining algorithm visualization

Another example of classification is stated in the paper **[17]**. In this paper, the aquaculture water quality indicator were analyzed by a feedforward error back propagation algorithm which is a BP neural network approach with strong nonlinear mapping capacity. Then the pa reduce distributed programming model was used to perform parallel design of the neural network algorithm to cover the needs of massive data processing in aquaculture platform. The **fig [17]** shows the execution process of map reduce programming

framework.

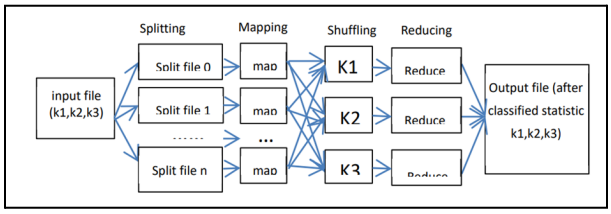

Fig 17: execution process of MapReduce programming framework.

There is an approach of prediction using neural network **[18]**. In this paper, they are focused on the prediction of time series data stream. Because of the major features of time series data which are non stationary and non linear characteristics, current solutions don't show good accuracy and time consumption. For this reason, they used memory wavelet neural network algorithm which is focused on the higher accuracy and minimize time consumption. They used 5 different models which are MWNN, WNN, GM, ARIMA, LSTM using 4 different CPU. And they figured out the MWNN showed the smallest root mean error among 5 models which is shown in the **figure [18]**

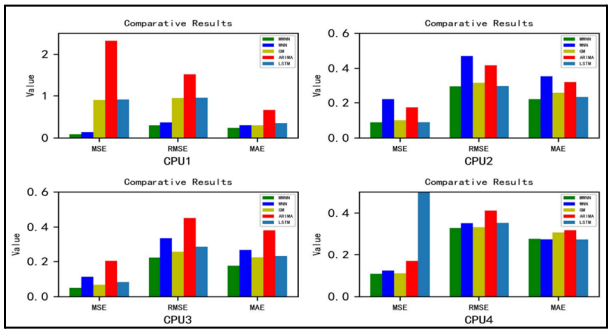

Fig18: Comparison of prediction accuracy of four models.

There is a data assimilation model error correlation using neural networks in the paper **[20]**. They implemented the neural network with the physical models by data assimilation algorithms to improve the assimilation process and forecasting result. In their experiments, they compared the data assimilation method and the neural network combined with an assimilation method which is their own method that can be seen in the **fig [19]**. The result showed more accurate prediction with the neural network model.

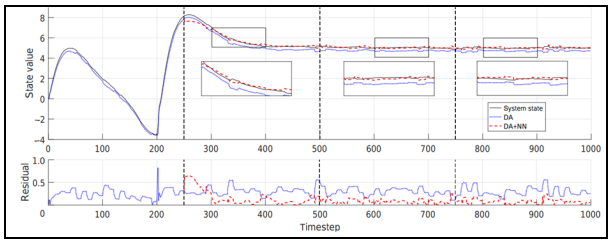

Fig 19: parallel composition error. Simulation result of DA+NN on a double-integral system.

Gabriela Czibula et al **[39]** proposes IntelliDaM, a machine learning-based framework helpful for increasing decision-making processes by enhancing the performance of data mining tasks. Education data mining has been the subject of extensive research and the discovery of hidden knowledge from data is made easier by IntelliDaM. The proposed framework effectively mines educational data with the aim of analyzing student performance.

The research makes use of real data sets from an undergraduate course called "Logic and Functional Programming" (LFP). Since students are divided according to how well they perform, it is highly likely that they would also learn in ways that are similar. The two data sets used as case studies are:

•D2018−2020- data gathered for two academic years (2018–2019 and 2019–2020), during which all teaching and evaluation activities were carried out in-person.

•D2020−2021 - information gathered for the academic year 2020–2021 during which all activities were shifted online because of the COVID–19 pandemic.

Data set	Regressor	Classification	
		7 categories	5 categories
$D_{2018-2020}$	**Tweedie**	10	10
	SGD	10	10
	Poly	11	11
$D_{2020-2021}$	**Tweedie**	9	7
	SGD	10	10
	Poly	9	10

Table 5: Optimal number of features from the Relief-based feature set

As shown in **table(5)**, in this study a Relief -based set of features for a data set. A grid search is used

for various k to find the Relief-based set of features for a given data set and Poly regressor. A 10-fold cross-validation is used to test the performance of the classification (into 5 and 7 categories) for a specific set of data (the first k features produced by the ReliefF algorithm) .The k-value that maximizes performance over the course of a 10-fold validation procedure is considered to be the optimum value.

In terms of all evaluation metrics (Acc, Prec, Recall, and F1), classification into 5 categories performs better than classification into 7 categories. The classification task successfully completed with a very good performance (an F1 score of about 92%), and the findings also showed that utilizing a technique for selecting discriminative data features improved performance statistically significantly.

Researchers in paper **[23]** show how to use two DNN-based methods—ADL(Autonomous Deep Learning) and DEVDAN (Deep Evolving Denoising Autoencoder)—to sidestep dynamic data streams. The potential of DNNs to solve issues with continuously-increasing data streams is still largely untested.

Authors demonstrated that, when dealing with high-dimensional issues, their suggested adaptive synapses pruning is capable of retaining the computational and spatial complexities without reducing the generalization power. Initial findings of the work on the semi-supervised learning challenge, DEVDAN+ may be helpful in solving the above mentioned issue.
The direct deployment of DNNs to manage data streams is frequently unfeasible, however, because the majority of them are based on an iterative training process that is computationally and memory-intensive. Technically, data streams should be processed using an online algorithm that employs single-pass learning over the data to prevent the catastrophic forgetting problem and to accommodate non-stationary situations.

In paper **[21]** authors present another example of Pattern classification. It is defined to be a general problem that is an important part of both Data Mining and Data Stream Mining. There are many ways to solve this problem now. When it comes to classifying how we react in response to graphics like photos or videos, we run into problems since object classes tend to mix together, their boundaries are fuzzy, and there are no established standards for what constitutes high quality. In contrast, the cross-entropy objective function enables the development of a learning criterion that provides a high rate of learning.

Reaction data is increasingly used in software development. Analyzing this data can lead to a number of improvements, such as new versions of the interface, the addition of personalized features and services, and the creation of policy decisions about the user (wake a driver who has fallen asleep). A person's reaction can be gauged by observing their breathing rate, heart rate, speech pattern, gestures, or facial expressions.

Deep neural networks are today's most accurate classification methods (DNNs). DNNs have a short learning process, which makes them difficult to apply in Data Stream Mining, when training set data enters consecutively online. However, Probabilistic neural networks (PNN) with "lazy learning" are fast. The authors proposed a neo-fuzzy unit-based pattern recognition system as a solution to the issue. Accurate recognition is made possible by the hybrid nature of these neurons, which assures a high approximation accuracy. In order to complete the work in real-time, a learning algorithm that is tailored to speed must be used. The studies demonstrated the suggested system's rapid pace of learning as well as its accuracy in tackling recognition-related issues.

1.2.2. Regression

Regression Analysis is a type of supervised machine learning technique that is used to forecast any attribute with a continuous value. It is a very important tool for analyzing data that may be utilized for time series modeling and financial

forecasting.
A regression analysis is used to forecast the value of a continuous dependent variable from a number of independent variables. It is the technique of fitting a line or a curve to a large number of data points such that the distance between the data points and the solution turns out to be the smallest.

The effective utilization of data mining and machine learning in the healthcare domain can indeed increase the precision of diagnosis. **Syed Javeed Pasha et al. [7]** proposed a novel algorithm to enhance the performance of health prediction by the practitioner by decreasing the error rate. It is the Ensemble Gain ratio Feature Selection (EGFS) algorithm. It helps in extracting the most important features which are highly contributing. Many metric evaluations like Area Under Curve (AUC), accuracy, precision, recall and f-measure are used to be extra cautious in avoiding misdiagnosis. They performed the EGFS on texting the iodine level in patients as it is one of the major elements which can lead to cerebral diseases. Thyroid and cancer patients are growing widely in the world and iodine can be seen as one of the major deficiencies in them. The use of machine learning along with data mining has proven to enhance diagnosis in the medical domain especially in early diagnosis of serious health conditions which helps in improved patient care. The dataset used is the thyroid dataset, which is highly imbalanced, so the data is first processed by replacing the missing values with the mean of corresponding column, oversampling is done by adding extra data which are similar to the positive tuples. Synthetic Minority Over-sampling Technique (SMOTE) is used for this. The AUC, precision, recall and f-measure etc is calculated as

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

below

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F-Measure} = \frac{2 * TP}{(2 * TP+FP+FN)}$$

The EGFS algorithm then tries to figure out the strongest features that contribute to the highest accurate prediction. The Random Forest algorithm (RF) produces Classification and regression trees (CART). After effective feature extraction by RF algorithm, this feature credibility is verified by gain ratio algorithm. This returns the average of all the test information. classifiers like KNN, LR and NB are then applied to the features to refine key features. As EGFS is a combination of several ML approaches, it performs much better and precise than the traditional methods

There was another neural network approach to help the data flow management system at a virtual machine in the paper **[19]**. They created a neural network on data flow paradigm which can be seen in the **figure [20]**. They created this model to solve a simple XOR function with two inputs and one output using sigmoid command which is required to make a neuron activation function of the model.

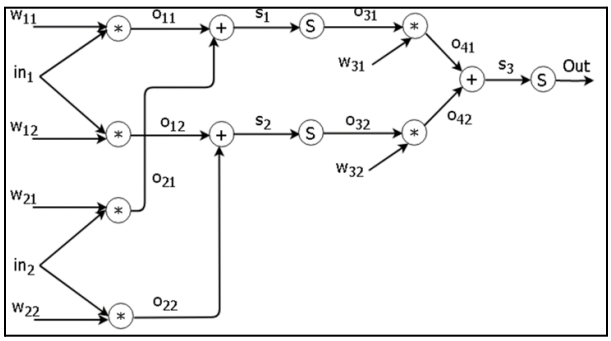

Fig 20: Neural network on data flow paradigm.

There have been a lot of studies to increase the efficiency of data analysis. Huge data sets are generally preferred in case of building a training set. But in many cases, it is not a feasible option as it increases the labor and computational speed and can be often costly. Also, most of the machine learning algorithms behave like a black box where the users will get the result accurately despite not

having much grasp on how the algorithm works. Data visualization along with the ML algorithm can give the user a better perspective in this kind of scenario. Users can make use of the interactive selection via visualization to guide the ML training product by selecting relevant features and hence, improving the efficiency of the training process. **Suvendu Kumar Jena et al. [3]** has researched on the collaboration of machine learning platforms with visualization and how this approach will enhance the efficiency and efficacy of the model. In their model, they used the SVM ML approach on preselected quantities of data which are smaller and better qualified for the training purpose which can save the collection expense. The flow chart fig **[21]** illustrates the flow of the proposed model. They used triangle-based interpolation to generate an interpolation graph which produces a smoother curve. Classification was done with various approaches like decision tree, Naïve Bayes, Neural Network etc. This research further affirmed the fact that the use of ML and data mining in diagnosis of cardiac disease along with the human cognitive interpretation is more effective and there is need for higher research on hybrid data mining techniques in the same. As here the research with strong sample data was proven to be more effective than big data interpretation with brute force approach.

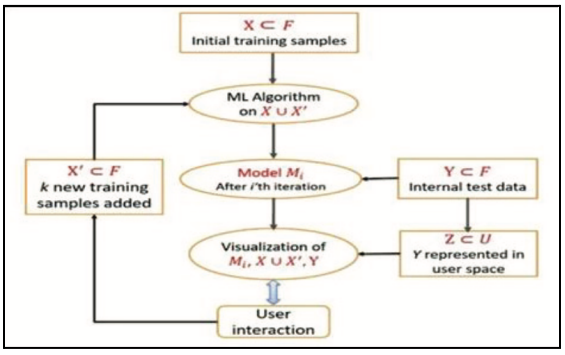

Fig 21:The interactive machine learning system's structure flowchart

1.2.3. Deviation/Anomaly

Anomaly detection or Outlier detection is a very important domain of unsupervised learning and has applications in a wide range of fields. The process of locating data points that deviate from the norm is known as anomaly detection. While outlier detection can increase the accuracy of the current model by bringing attention to the handling of outliers, anomaly detection can result in the discovery of an entirely new model when the anomaly is proven to be a component of a distinct dataset.

There is another paper that presents a new optimization model based on the constructed seismic anomaly database **[14]**. In their paper, a neural network was used for the earthquake prediction. Using the earthquake anomaly database, they created a model that predicts the earthquake level with 4 different approaches which are the decision tree, Bayes theorem, Association rules and Machine learning which is shown in the **table 6**.

Algorithm	Error rate (%)
The decision tree	17.25
Bayes theory	22.32
Association rules	16.53
Machine learning	5.26

Table 6: Data Mining Error Rate of Various Algorithms

Guansong Pang et al **[37]** in the paper outlines and reviews the problems with detecting anomalies and various deep learning methods for anomaly detection that focus on using neural networks to learn feature representations or anomaly scores for the purpose of anomaly detection. The major problems experienced while detecting anomalies are like anomalies associated with unknownness eg. network intrusion, heterogeneous anomaly classes eg. traffic accidents, rarity and class imbalance eg, misclassification of anomalies and diverse types of anomalies eg, point anomalies, conditional anomalies or group anomalies. The study explains main challenges while tackling deep anomaly detection like CH1: Low anomaly detection recall rate. CH2: Anomaly detection in high-dimensional and/or not-independent data. CH3: Data-efficient learning of normality/abnormality. CH4: Noise-resilient anomaly detection. CH5: Detection of complex anomalies and CH6: Anomaly explanation.

Method	End-to-end Optimization	Tailored Representation Learning	Intricate Relation Learning	Heterogeneity Handling
Traditional	×	×	Weak	Weak
Deep	✓	✓	Strong	Strong
Challenges	CH1-6	CH1-6	CH1, CH2, CH3, CH5	CH3, CH5

Table 7: Deep Learning Methods vs. Traditional Methods in Anomaly Detection

Deep techniques allow it to optimize the entire process for anomaly detection from beginning to end and to learn representations that are designed with anomaly detection in mind. Table(7) Deep approaches, which frequently use black-box models instead of traditional ones, provide ways to combine anomaly detection and explanation into a unified framework, leading to a more accurate justification of the abnormalities that a particular model has identified. The study further focuses on few diverse modeling approaches on harnessing deep learning techniques for anomaly detection.

1.End-to-end One-class Classification: This class of techniques aims to end-to-end train a one-class classifier that can distinguish between a given instance being normal or not. It has an end-to-end adversarially optimized anomaly categorization model. The advanced techniques and concepts of adversarial learning and one-class classification can be used to design and support it.

2. Softmax likelihood model: With this method, anomaly scores are learned by increasing the chance of events in the training set. The benefit of this technique is that the learning process for anomaly scores can take into account various interactions and the anomaly scores are accurately optimized in relation to the particular anomalous interactions that we want to detect.

3.Prior driven models: In this method, the anomaly score learning is encoded and driven by a previous distribution. The prior may be imposed on either the internal module or the learning output (i.e., anomaly scores) of the score learning function τ because the anomaly scores are learned in an end-to-end manner. The benefit of this model is it is possible to directly optimize the anomaly scores given a prior. Also, It offers an adaptable framework for including various previous distributions in the learning of anomaly scores. For anomaly detection, many Bayesian deep learning approaches may be adopted and In comparison to other methods, the prior can also produce anomaly scores that are easier to understand.

4. Ranking models: In order to sort data instances based on an observable ordinal variable linked to the absolute/relative ordering relation of the anomaly, this class of approaches aims to directly learn a ranking model. The observable ordinal variable drives the neural network for anomaly scoring. The advantages of using this model is by using adapted loss functions, the anomaly scores can be directly optimized and by applying a weak assumption regarding the ordinal order between anomaly and normal occurrences, they are typically exempt from the definitions of anomalies.

The aim of this area of research paper is to learn scalar anomaly scores from beginning to end. This kind of approach has a neural network that directly learns the anomaly scores rather than relying on existing anomaly measures, which is different from anomaly measure-dependent feature learning. The anomaly scoring network frequently requires the use of novel loss functions. Formally, the goal of this method is to train an entire anomaly score learning network: $\tau(\cdot; \Theta) : X \rightarrow R$)

Section 2: Hybrid approach

There are few applications that are implemented using a combination of descriptive and predictive methods of data mining technique. With the increasing availability of Spatio-temporal datasets such as maps, virtual globes, remote-sensing images, decennial census and GPS trajectories, the Spatio-temporal data mining has become so important in the big data realm. The ST data are generally highly self-correlated **[9]** and are embedded in a continuous space and can show both spatial and temporal properties. Event data, trajectory data, point reference data, raster data, video data are all different available ST data types. Deep learning with S data is done in various fields like transportation, on-demand service, climate & weather, human mobility, location-based social networks (LBSNs), crime analysis, and neuroscience.

There is a data mining algorithm with a hybrid approach which uses genetic algorithms and the neural network **[13]**. In their algorithm, the neural network is formulated by probabilistic approach and the genetic algorithm is generalized by discrete distribution of variables. This algorithm is

developed to predict academic information however, they did not provide any results.

Yunliang Wang et al **[34]** proposes a hybrid data mining approach based on clustering, association rules, and stochastic gradient descent for the classification and prediction of power system faults.

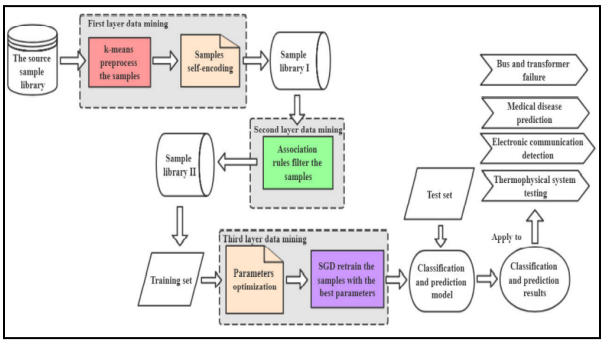

Fig 22: Three layer data mining process

Fig(22) The research proposes a three layer data mining model technique: The first layer proposes using self-encoding to streamline the data form after preprocessing the original fault data source with the K-means clustering technique where K at randomly K is selected as the initial point for the classification clusters, then after the criterion function converges, the contour coefficient approach is used to determine the optimal K value for the samples on each node.

The second layer uses association rules (Apriori) to effectively remove the data that have minimal bearing on the outcomes of the prediction, and the highly associated data are mined to produce the regression training data. Apriori algorithm is used to mine the association rules for the data samples. **table (8)** The minimal confidence is set to 0.7, while the minimum support is set to 0.3. The common itemsets are the sample sets from sample library I with a support degree greater than 0.3. Then, every frequent itemset that satisfies the requirements of having a lift larger than 1.0 and a confidence greater than 0.7 is discarded as a sample of the strong association rules.

Group	Sample group	Support	Confidence	Lift
1	{905}→{1}	0.35	0.749	2.14
2	{306}→{1}	0.42	0.749	1.78
3	{205,407}→{1}	0.33	0.850	2.56
4	{306,905}→{1}	0.41	0.846	2.06
5	{205,606,701}→{2}	0.55	0.884	1.61
6	{606}→{2}	0.44	0.738	1.74
7	{403,505}→{2}	0.45	0.749	1.68
8	{101,206,906}→{3}	0.53	0.874	1.64
9	{305,406}→{3}	0.57	0.765	1.34
10	{105,406,502}→{3}	0.55	0.889	1.62
11	{405,506,901}→{3}	0.5	0.889	1.61
12	{205,506}→{3}	0.45	0.714	1.59
13	{105,604,908}→{4}	0.36	0.659	1.83
14	{108,303}→{407}	0.45	0.683	1.52
15	{402,503,607}→{105}	0.48	0.702	1.46

Table 8:Partial rules mined from the frequent itemsets.

The third layer uses stochastic gradient descent (SGD) for data regression training to first find the optimal parameters for each fault model, followed by cross-validation to provide a classification and prediction model for each fault type.A short circuit fault classification and prediction test is carried out using a power grid corporation project in a specific urban area distribution network to prove the scalability of the proposed technique. There are 56 data collection points, 12 transformers, and 8 generators in the distribution network. The distribution network has more than 20,000 sets of voltage, together, the voltage data from the 56 collection points in the distribution network have an impact on the fault types.

Systems	Accuracy	Runtime(s)	Computational cost(B)
WSCC 9 bus system	**93.8%**	**3.7**	**0.014**
Practical application system	**89.2%**	**11.6**	**0.235**

Table 9: Comparison of the various systems' fault classification accuracy, runtimes, and computing costs

Since the algorithm model has processed the data at each collection point 210 times, the computational cost of the classification and prediction model in the distribution network is 56 × 20000 × 210 = 0.235B. Additionally, the classification and prediction model in the WSCC 9 bus system has a computing cost of 9×20000×78 ≈ 0.014B, where B stands for billion which is the unit of the number of the times the computer runs.

From the **table(9)** it can be concluded since runtime and computational cost are correlated with the complexity of the actual system, such as the number of nodes, the number of line branches, etc., it implies that as the number of system nodes increases, so will the runtime and computational cost. As a result, it is possible to predict that the computing cost will increase in an actual system that is more complex.

One another research conducted by Nawaf Alsrehin **[35]** aims to study data mining and machine learning techniques hybrid approaches for Intelligent Transportation and Control Systems. The study focuses mainly on traffic management approaches using data mining techniques and machine learning models to detect and predict traffic parameters. Following approaches are studied in this research for traffic parameters prediction:

1.Estimate and predict real-time traffic flow.

This mechanism integrates a number of scalable data mining methods, including neural networks, association rules, and decision trees. These procedures include historical data as well as a few traffic parameters. The Artificial Neural Network (ANN) was utilized to anticipate the short-term traffic flow using historical traffic data. Traffic volume, speed, density, time, day of the week, and each category's speed are used as input factors in the model.

2.Predict short-term traffic flow in heterogeneous conditions.

A lane path is followed by identical vehicles in homogeneous traffic. With no lanes, heterogeneous traffic includes both motorized and non-motorized vehicles, including two- and three-wheelers, as well as numerous more cars and trucks. The absence of lane discipline in this heterogeneous traffic leads to complex traffic behavior, making traffic flow prediction more difficult than in homogeneous traffic.

A deep learning model that combines a linear model with a series of tanh layers, which are used to represent nonlinear relationships and find spatiotemporal relationships between predictors. For short-term traffic flow projections, a deep learning model delivers accurate results. The authors also found the empirical discovery that predictions based on data from the most recent 40 minutes of traffic data produce better results than predictions based on values from the most recent 24 hours. This suggests that rather than using features derived from earlier days of observations, a powerful model might be built on features derived from recent observations (i.e., within the last few minutes).

3.Estimate and predict travel time in real-time.

Travel time is the amount of time it takes for drivers to get from one place to another. Timely travel time estimation reduces traffic and improves the efficiency of the overall transportation system. The study used historical data and a classification system like k-NN to forecast when the next bus would arrive. This real-time prediction makes use of a Kalman filtering-based model-based recursive estimation approach. The estimated trip time can be shown at bus stops, within buses, or on websites as the amount of time left to get to the destination. The evaluation's results showed that, in comparison to existing methods based on static inputs, the proposed method increased prediction accuracy.

4.Estimate and predict the real-time traffic density.

The main metrics for measuring traffic congestion on roads other than signalized intersections include throughput, travel time, safety, fuel consumption, emission, reliability, and traffic density. Aerial photography with a loop detector, a data-driven approach using linear models, linear regression, ANN, k-NN, pattern matching, PCA, nearest neighbor approach, Kalman filtering, clustered neural networks, wavelet neural networks, k-NN and linearly sewing principle components, and image processing techniques can all be used to predict real-time traffic density. The kNearest Neighbor (kNN) and Artificial Neural Network (ANN) machine learning techniques is used in this study to evaluate trip time and traffic density and it is observed that a better performance for ANN is achieved provided the training dataset used is huge.

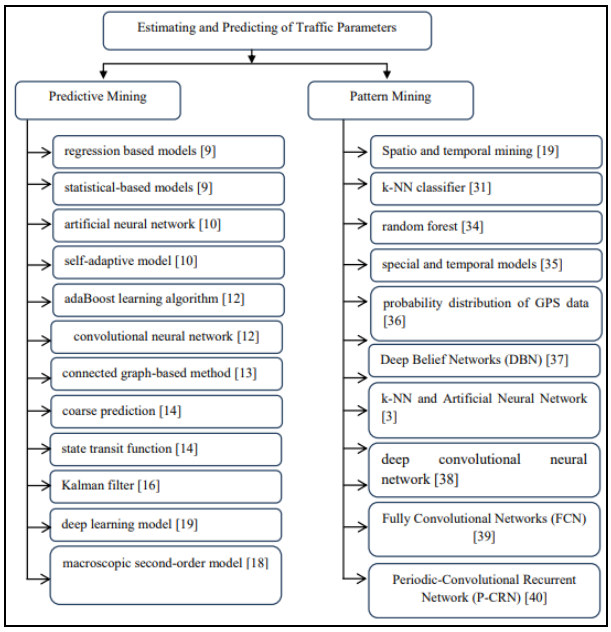

Fig 23: General structure of the methods used to estimate and predict traffic parameter

Fig (23) The research also provides some insights on various other data mining used in the research field that has been implemented for estimating and predicting traffic parameters and to improve the efficiency and accuracy of the overall system.

Conclusion

From this survey, we found out the neural network is widely used in all fields of data mining technology which would be a great aspect to research with. For the future plan, we decided to use a neural network approach to solve the problem which is not solved from the papers with the better efficiency or accuracy.

Reference:

[1] *A. Deb, M. S. Akter Koli, S. B. Akter and A. A. Chowdhury, "An Outcome Based Analysis on Heart Disease Prediction using Machine Learning Algorithms and Data Mining Approaches," 2022 IEEE World AI IoT Congress (AIIoT), 2022, pp. 01-07, doi: 10.1109/AIIoT54504.2022.9817194.*

[2]*S. Khond, V. Kale and M. S. Ballal, "A combined Data Mining and Machine Learning approach for High Impedance Fault Detection in Microgrids," 2022 IEEE International Conference on Power Electronics, Smart Grid, and Renewable Energy (PESGRE), 2022, pp. 1-7, doi: 10.1109/PESGRE52268.2022.9715823.*

[3]*S. K. Jena, P. Sahu and S. Mishra, "Dynamic Data Mining for Multidimensional Data Based On Machine Learning Algorithms," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-7, doi: 10.1109 /ISCON52037.2021.9702355.*

[4]*P. Kaur, A. Sharma, J. K. Chahal, T. Sharma and V. K. Sharma, "Analysis on Credit Card Fraud Detection and Prevention using Data Mining and Machine Learning Techniques," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), 2021, pp. 1-4, doi: 10.1109/ICCICA52458.2021.9697172.*

[5]*Y. Li, "Practice of Machine Learning Algorithm in Data Mining Field," 2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI), 2020, pp. 56-59, doi: 10.1109/ICAACI50733.2020.00016.*

[6]*S. Verma and A. Gupta, "Effective Prediction of Heart Disease Using Data Mining and Machine Learning: A Review," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 249-253, doi: 10.1109/ICAIS50930. 2021.9395963.*

[7]*S. J. Pasha and E. S. Mohamed, "Ensemble Gain Ratio Feature Selection (EGFS) Model with Machine Learning and Data Mining Algorithms for Disease Risk Prediction," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 590-596, doi: 10.1109/ICICT48043.2020.9112406.*

[8]*N. A. Shemu, M. Z. Hossain, S. M. Saleh and K. A. A. Pavel, "A Machine Learning View for Health Data Mining Emphasizes on the Decision Trees," 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2020, pp. 130-133, doi: 10.1109/ICCAKM46823.2020.9051467.*

[9]*S. Wang, J. Cao and P. S. Yu, "Deep Learning for Spatio-Temporal Data Mining: A Survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 8, pp. 3681-3700, 1 Aug. 2022, doi: 10.1109/TKDE.2020.3025580.*

[10]*M. Zeyad and M. S. Hossain, "A Comparative Analysis of Data Mining Methods for Weather Prediction," 2021 International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 167-172, doi: 10.1109/ComPE53109.2021.9752344.*

[11]*Quan, X., & Xu, L. (2020, October). Sales Analysis of Network Marketing Products Based on Neural Network Data Mining Algorithm. In 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA) (pp. 219-222). IEEE.*

[12]*Xia, T., Ye, Z., Wu, H., & Liu, Y. (2020, July). Data Association Rules Mining Method Based on RBF Neural Network Optimization Algorithm. In 2020 International Conference on Communications, Information System and Computer Engineering (CISCE) (pp. 433-436). IEEE.*

[13]*Tiwari, A. K., Ramakrishna, G., Sharma, L. K., & Kashyap, S. K. (2019, October). Neural Network and Genetic Algorithm based Hybrid Data Mining Algorithm (Hybrid Data Mining Algorithm). In 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 95-99). IEEE.*

[14]*Long, Y., & Rong, J. (2021, September). Research on Model of Seismic Anomaly Data Mining Based on Neural Network. In 2021 IEEE 4th International Conference on Information Systems and Computer Aided Education (ICISCAE) (pp. 518-522). IEEE.*

[15]*Long, Y., & Rong, J. (2022, February). Analysis of Data Mining and Dynamic Neural Network for Data Prediction. In 2022 11th International Conference of Information and Communication Technology (ICTech)) (pp. 312-315). IEEE.*

[16]*Li, L. (2022, April). Research and Application of Data Mining Classification Algorithm Based on Multi-classifier Fusion. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 651-654). IEEE.*

[17]*Wen, Y., Li, M., & Ye, Y. (2020, April). MapReduce-based BP neural network classification of aquaculture water quality. In 2020 International Conference on Computer Information and Big Data Applications (CIBDA) (pp. 132-135). IEEE.*

[18]*Chen, L., Wang, W., & Yang, Y. (2020, December). An efficient dynamic neural network for predicting time series data stream. In 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) (pp. 625-632). IEEE.*

[19]*Kharchenko, K., Beznosyk, O., & Romanov, V. (2018, August). Implementation of neural networks with help of a data flow virtual machine. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (pp. 407-410). IEEE.*

[20]*Zhu, J., Hu, S., Arcucci, R., Xu, C., Zhu, J., & Guo, Y. K. (2019). Model error correction in data assimilation by integrating neural networks. Big Data Mining and Analytics, 2(2), 83-91.*

[21] *Kulishova, N., Bodyanskiy, Y., & Timofeyev, V. (2020, August). The Fast Image Recognition System Based on Neuro-Fuzzy Units and its Online Learning for Data Stream Mining Tasks. In 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) (pp. 147-151). IEEE.*

[22]*Kaneko, Y. (2019, November). Customer-Base sequential data analysis: An application of attentive Neural Networks to sales forecasting. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 349-355). IEEE.*

[23]*Ashfahani, A. (2019, November). Autonomous Deep Learning: Incremental Learning of Deep Neural Networks for Evolving Data Streams. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 83-90). IEEE.*

[24] *Patnaik, S. K., Babu, C. N., & Bhave, M. (2021). Intelligent and adaptive web data extraction system using convolutional and long short-term memory deep learning networks. Big Data Mining and Analytics, 4(4), 279-297.*

[25] *Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics, 5(2), 81-97.*

[26] *Besta, M., Grob, R., Miglioli, C., Bernold, N., Kwasniewski, G., Gjini, G., ... & Hoefler, T. (2022, August). Motif prediction with graph neural networks. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 35-45).*

[27] *de Oliveira Lima, T., Colaço, M., Prado, K. H. D. J., & de Oliveira, F. R. (2021, December). A Big Data Experiment to Evaluate the Effectiveness of Traditional Machine Learning Techniques Against LSTM Neural Networks in the Hotels Clients Opinion Mining. In 2021 IEEE*

International Conference on Big Data (Big Data) (pp. 5199-5208). IEEE.

[28] *Okewu, E., Adewole, P., Misra, S., Maskeliunas, R., & Damasevicius, R. (2021). Artificial neural networks for educational data mining in higher education: A systematic literature review. Applied Artificial Intelligence, 35(13), 983-1021.*

[29] *Jiang, L., Wang, P., Cheng, K., Liu, K., Yin, M., Jin, B., & Fu, Y. (2021). Eduhawkes: A neural hawkes process approach for online study behavior modeling. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM) (pp. 567-575). Society for Industrial and Applied Mathematics.*

[30] *Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019, January). Fake news: Fundamental theories, detection strategies and challenges. In Proceedings of the twelfth ACM international conference on web search and data mining (pp. 836-837).*

[31] *Y. Zheng and G. Chen, "Energy Analysis and Application of Data Mining Algorithms for Internet of Things Based on Hadoop Cloud Platform," in IEEE Access, vol. 7, pp. 183195-183206, 2019, doi: 10.1109/ACCESS.2019.2958377.*

[32] *T. Su, H. Xu and X. Zhou, "Particle Swarm Optimization-Based Association Rule Mining in Big Data Environment," in IEEE Access, vol. 7, pp. 161008-161016, 2019, doi: 10.1109/ACCESS.2019.2951195.*

[33] *W. -C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 197-206, doi: 10.1109/ICDM.2018.00035.*

[34] *Y. Wang, X. Wang, Y. Wu and Y. Guo, "Power System Fault Classification and Prediction Based on a Three-Layer Data Mining Structure," in IEEE Access, vol. 8, pp. 200897-200914, 2020, doi: 10.1109/ACCESS.2020.3034365.*

[35] *N. O. Alsrehin, A. F. Klaib and A. Magableh, "Intelligent Transportation and Control Systems Using Data Mining and Machine Learning Techniques: A Comprehensive Study," in IEEE Access, vol. 7, pp. 49830-49857, 2019, doi: 10.1109/ACCESS.2019.2909114.*

[36] *C. K. Leung, C. S. H. Hoi, A. G. M. Pazdor, B. H. Wodi and A. Cuzzocrea, "Privacy-Preserving Frequent Pattern Mining from Big Uncertain Data," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 5101-5110, doi: 10.1109/BigData.2018.8622260.*

[37] *Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. ACM Comput. Surv. 54, 2, Article 38 (March 2022), . https://doi.org/10.1145/3439950*

[38] *Q. Guo, Y. Qian and X. Liang, "Mining Logic Patterns from Visual Data," 2019 International Conference on Data Mining Workshops (ICDMW), 2019, pp. 620-627, doi: 10.1109/ICDMW.2019.00094.*

[39] *G. Czibula, G. Ciubotariu, M. -I. Maier and H. Lisei, "IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining," in IEEE Access, vol. 10, pp. 80651-80666, 2022, doi: 10.1109/ACCESS.2022.3195531.*

[40] *X. Li, Y. Wang and D. Li, "Medical Data Stream Distribution Pattern Association Rule Mining Algorithm Based on Density Estimation," in IEEE Access, vol. 7, pp. 141319-141329, 2019, doi: 10.1109/ACCESS.2019.2943817.*