IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

USING HYBRID DATA MINING TO DISCOVER KNOWLEDGE

[1]Deepak Saraswat

[1]Research Scholar
[1]Departement of Physics and Computer Science,
[1]Dayalbagh Educational Institute (Deemed University), Agra- 282005, Uttar Pradesh, India

Abstract: A hybrid intelligent system is a set of artificial intelligence (AI) tools which can be used to handle difficult medical challenges in healthcare. Materials informatics can be advantageous for industrial engineering and manufacturing applications when analyzed from a data mining approach. Knowledge Discovery in Databases (KDD) is the goal of the data mining process. KDD uses processes such as data selection, data mining, data preprocessing, pattern assessment, and data transformation to extract knowledge from raw datasets. To extract knowledge from massive datasets, many data mining approaches like association, clustering, and classification are utilized. This work reviewed the literature on KDD applications in the wide area of manufacturing by a particular focus on the types of data functions to be performed. Description, characterization and association, classification, prediction, clustering, and evolution analysis are the primary data mining functions to be done. In addition, the Knowledge Discovery using Hybrid Data Mining Approach module's settings and outcomes are assessed.

Index Terms **- Data mining, Knowledge Discovery, hybrid data mining, Naïve Bayes, Recurrent Neural Network.**

I. INTRODUCTION

The vast number of massive data sets and materials informatics have been affected by rapid breakthroughs in materials science and information technology correspondingly. Materials informatics is a branch of research that uses informatics ideas to engineering and materials science in order to realize how materials are used, developed and discovered. Because many of the traditional analytic methods used for materials structural-properties analysis are no longer effective in some situations, researchers in the manufacturing industries and other areas of industrial engineering are confronted with original research matters in systematic analysis of materials data sets. As a result, materials informatics has emerged as a new study topic in material science and technology, changing experimental methodologies and thinking in materials research, and posing even more obstacles in multidisciplinary research. Data mining is an interdisciplinary area that brings together concepts from statistics, machine learning (ML), information science, visualization, and different fields (Hemanth, Vastrad & Nagaraju, 2011).

The highly important method for integrating information and theory for knowledge discovery in any informatics field, like as Bioinformatics, Cheminformatics, Materials informatics, Nano informatics and many more. Many effective research trial outcomes have demonstrated the significance of Data Mining on knowledge discovery. As a result, data mining can be utilized to remove non-trivial, hidden, formerly unknown, potentially beneficial, and eventually intelligible knowledge from large databases of materials. Data understanding, business understanding, modeling, data preparation, assessment of the model and deployment. The removal of useful and earlier undiscovered information or trends from data sources is known as data mining. It can be used in a variety of fields, including biological data analysis, financial data analysis, and weather forecasting. The core point of the KDD process is data mining, which corresponds to the modelling stage in the KDD process (Hemanth, Vastrad & Nagaraju, 2011). It entails the use of intelligent technologies to identify new and valuable patterns from enormous amounts of data. There have been several models developed for the KDD process but the extremely well-known is the industrial model-CRISP-DM (Hemanth, Vastrad & Nagaraju, 2011). KDD is an iterative and participatory process, according to this model, that consists of 6 steps: data understanding, business understanding, data preparation, modelling, model assessment, and deployment. Raw data is transformed into valuable information or expertise using data mining procedures (DMT). Data is highly helpful and intriguing. Many advanced technologies make clever use of data as useful information. KDD is the method of extracting desired output in various formats from raw data. KDD may alternatively be characterized as a method for identifying meaningful patterns in data (Dutt, Ismail & Herawan, 2017).

DMT is most often used in EDM. It is utilized to create databases that help students and administrators make decisions more quickly (Dutt, Ismail & Herawan, 2017). It is described as a new educational programmer that investigates various sorts of data generated by educational institutions. It examines data created by the educational system in order to enhance learning and educational outcomes (Dutt, Ismail & Herawan, 2017). It is part of a body of work on data mining, visualization, ML and computing. Furthermore, Nave Bayes, Neural Networks (NN), K-Nearest Neighbor (KNN), Decision Trees, and several other approaches are employed in ML (Nagy, Aly & Hegazy, 2013). Other fields are also combined with data mining. Such as, various potential techniques that combine data mining with semantic web have been developed. Correspondingly, data mining and ML methods are employed in a variety of applications (Buczak &Guven , 2015), (Madni, Anwar, & Shah, 2017).

II. DATA MINING

The term data mining indicates to the process of extracting information from massive amounts of data in order to create useful tools for a range of educational purposes. Research in the subject of education is rapidly developing due to the vast amount of student data that can be used to uncover crucial patterns connected to student learning behavior.

As a consequence, data mining should be called knowledge mining, because it is concerned with extracting knowledge from large amounts of data (Shukla, Sharma, Samaiya, & Kherajani, 2020). It is a computer technique for identifying patterns in large data sets that combines approaches from ML, AI, and database systems. The purpose of the data mining approach is to extract information from a collection of data and convert it into a usable structure.

Some of the most essential aspects of "data mining" are listed below (Shukla, Sharma, Samaiya, & Kherajani, 2020).

- Creation of actionable information
- Concentrate on databases and large datasets
- Forecast of results
- Automatic discovery of patterns.

Tasks of Data Mining

It consists of following tasks

- **Anomaly detection**

Looks for correlations among variables. For example, a supermarket can collect information on client purchase behaviour. The store can utilize association rule learning to discover which goods are usually acquired together and using that knowledge for marketing reasons.

- **Association rule learning**

The explorations for connections among variables. Applying association rule learning, the supermarket can expose goods that are commonly purchased simultaneously and utilize this data for marketing objectives.

- **Clustering**

Clustering is the finding of structures and groups in the data that are in several way or other “related” exclude of utilizing identified formations in the data.

- **Classification**

This is the task of simplifying structures to be relevant to the latest data. Such as: an electronic-mail program can make an effort to categorize an electronic mail as legitimate or as spam.

- **Regression**

This discover a function which shows the data with the least error.

Following are some of the most important techniques which are commonly used in data mining process (Omar, Alzahrani, & Zohdy, 2020).

i. K means Clustering

The k-means clustering technique is a data mining and ML tool that used clusters observations into groups of similar observations without knowing the links between them. A variety of attributes are used to define a data sample in real-life datasets. In order to analyze the datasets, it is necessary to categorize these samples based on their similarity in characteristics. The collection of data samples is an essential part of data analysis procedures.

It is an unsupervised approach of segmenting a large amount of data items into clusters. Clustering differs from supervised classification in that the goal is to arrange a set of unlabeled patterns into logical groups. For the same collection of data samples, different clustering techniques might provide distinct groups (Omar, Alzahrani, & Zohdy, 2020).

Partition-based and hierarchical clustering are two types of clustering.

K-means and k medoids are 2 instances of partition-based clustering methods. The k-means algorithm is used in the classification technique. Initially, the mean value of the points inside a cluster is determined as k number of centroids (Omar, Alzahrani, & Zohdy, 2020).

The data samples are then assigned to these clusters depending on the distance parameter among the sample location and cluster's centroid. The centroids values for the clusters are upgraded repeatedly, and data items are reallocated to all clusters based on the new centroids values. It improves in achieving a local best solution (Omar, Alzahrani, & Zohdy, 2020).

K-Means grouping is an unsupervised iterative clustering method.

- The given data set into k predefined separate clusters.
- A cluster is described as a collection of data points exhibiting certain comparisons.

Following are the advantages of k-means Clustering:

- This is comparatively effective with time complexity. Figure 1 shows k-means clustering process.

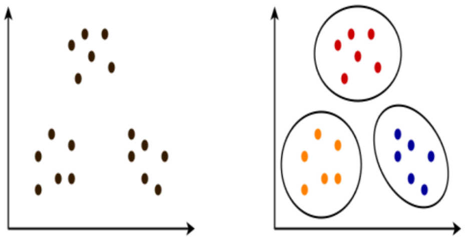

Figure 1: K means Clustering (Ghoussaini, et al. (2012))

- Simulated Annealing or Genetic Algorithms can be utilized to discover the global optimum.
- It usually concludes at local optimum (Omar, Alzahrani, & Zohdy, 2020).

It has the following disadvantages-

- It can’t hold noisy data and outliers.
- It is not appropriate to detect clusters with non-convex shapes.
- It needs to identify the number of clusters (k) in advance.

On the same amount of data, the elbow technique can yield the similar amount of clusters K. Based on the case study, the outcome of identifying the optimal quantity of clusters with the elbow technique be the avoidance for the characteristic procedure (Omar, Alzahrani, & Zohdy, 2020).

- **Elbow Criterion**

K-Means is a graph of cluster connection with decreasing error, rising value of K, then progressively reducing graph until consequence of value of K is fixed. For instance, from K = 2 to K = 3, then from K = 3 to K = 4, the value of the cluster K = 2 to K = 3, then from K = 3 to K = 4, it exhibits a significant reduction to create the elbow at point K = 3, therefore the optimum cluster k is K = 3 (Ghoussaini et al. (2012).

The value of K at the best cluster can be determined using the mixed Elbow and K-Means Methods. Affect the number of clusters created by multiplying the number of clusters by k.

The elbow criterion technique can be applied to choose the number of k clusters to be applied for grouping data using the K-Means algorithm in this study. Sum of Squared Error describes the elbow method (Ghoussaini et al. (2012). At the start, choose a random centre point for the cluster.

$$\mathrm{SSE} = \sum_{\mathrm{K}=1}^{\mathrm{k}} \sum \left|\left|\mathrm{X_i} - \mathrm{C_k}\right|\right| 2^2 \qquad (1)$$

The early centroid is determined at random from the accessible objects up to cluster k, then the next i-cluster centroid is calculated using the formula:

$$v = \frac{\sum_{i=1}^{n} x_i}{n};\ i = 1,2,3, \ldots\ldots n \qquad (2)$$

Using the Euclidian Distance, calculate the distance among each item and each centroid.

$$d(x,y) = \left|\left|x - y\right|\right| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2; i = 1,2,3, \ldots\ldots\ldots.., n} \qquad (3)$$

With x_i: Variable on object x to-i and y_i: Variables output y.

The number of items is indicated by the n. Place each object in the centroid that is closest to it. Using k-means, allocate items to each cluster at each iteration the closeness distance between each cluster member item and the cluster's Centre point has been measured. Iterate the procedure, then use equation to find the position of the new centroid. If the new centroid position differs from the previous centroid, repeat step 3 (Agrawal & Gupta, 2013).

ii. Decision tree classification

To create classification and regression models in data mining process, Decision Tree is employed. It is used to build data models that forecast class labels or values in order to aid in decision-making. The models are created using the training data that has been given into the system (supervised learning). It is an algorithm that determines the tree structure using a top-down recursive method. This algorithm's goal is to create a decision tree from a data collection in order to display categorization rules. The entropy or information levels for each characteristic are used to select class labels for categorization. The C4.5 algorithm formulas are utilized to compute the value of information gain for every attribute test. Let freq (Ci, S) be the quantity of samples in S that correspond to class Ci. The number of samples in the set is represented by S. The set S's entropy is therefore defined as:

$$\text{Info (D)} = \sum\left(\frac{\text{freq}\ (C_i,\ D)}{|D|}\log\left(\frac{\text{freq}\ (C_i,\ D)}{|D|}\right)\right) \qquad (4)$$

Considering the partitioning of set S based on the n results of one attribute test x,

$$\text{Inf ox (D)} = \sum\frac{|Di|}{|D|}\ \text{Inf o (Di)} \qquad (5)$$

By equation (1) and (2):

$$\text{Gain (X)} = \text{Inf 0 (D)} = \text{Inf ox (D)} \qquad (6)$$

The decision tree's root node is selected as the element with the maximum gain value. For each property X, the data samples are separated in reducing order based on models. The minimal number of samples for the leaf nodes is the blocking criteria applied to limit the repetitions of iterations required to create the tree. The decision tree that is created is utilized to assess developments or patterns and make decisions in order to attain the intended goals (Agrawal & Gupta, 2013).

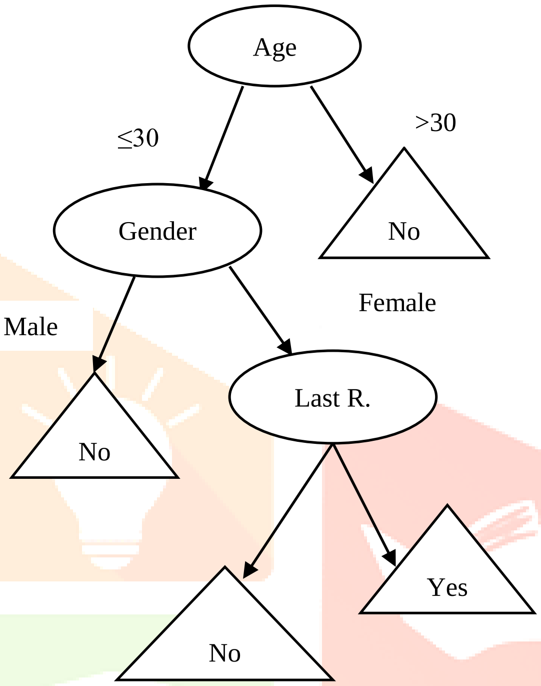

Figure 2: Decision Tree

Figure 2 shows a decision tree for determining if a prospective customer can reply to a direct mailing. Interior nodes are shown as squares though leaves are shown as triangles. It's worth noting that this decision tree has both insignificant and numeric properties. With this classifier, the predictor can anticipate a customer’s reaction and consider the behavioral patterns of the entire population of potential customers when it comes to direct mailing. Any node is labelled with the attribute it is evaluating and all divisions are labelled with the values associated with that attribute (Agrawal & Gupta, 2013).

iii. Naive Bayes

It is simple and quick to forecast the test data set's class. The Naive Bayes classifier does classification by assuming conditional freedom which decreases the amount of constraints to be calculated when modelling P(X|Y) from 2(2n-1) to only 2n [14].

- **Conditional Independence**

Assumed L, M, N are 3 sets of random variables. If and only if the probability distribution leading L is independent of the value of M provided N, then L is conditionally independent of M given N [14].

$$(\forall i, j, k)P(L = li\ |M = mj, N = nk) = P(L = li|\ N = nk) \qquad (7)$$

Consider the following 3 Boolean random variables to characterize weather like: Rain, Lightning and Thunder. The importance of Rain contains no new knowledge of Thunder until realize that whether or not there is Lightning. Of course, thunder is contingent on rain in general, but if one knows the importance of lightning there is no conditional dependency. But in this case, L, M, N all are specific random variables, the term can also be applied to groups of random variables.

- **Derivation of Naive Bayes Algorithm**

This algorithm utilizes the Bayes rule and a series of conditional independence supposition to classify data. The Naive Bayes algorithm assumes that all Li is provisionally independent of each of the different Xks given Y, as well as independent of every subset of the other Lks given M with the objective of learning P(M|L) where X = (L1...., Ln) [14].

This statement is useful because it significantly simplifies both the representation of P(L|M) and the issue of approximating it from training results. For example- the situation where L = (L1, L2).

$$P(L|M) = P(L1, L2|M) = P(L1|L2|, M)P(L2|M) \qquad (8)$$

where the 2nd line is derived by the common feature of probability and the 3rd line is derived explicitly by provisional independence concept. If it comprises n attributes that fulfil the provisional independence statement [14].

$$P(L1 \dots Ln \mid M) = \Pi P(Li|M)\ (3)\ ni = 1 \qquad (9)$$

When Xi and Y are boolean variables and just want 2n factors to define P (Li = li k |M = m j) for the i, j, and k that are needed. As related to the 2(2n-1) factors wanted to define P(L|M) if provisional independence is not assumed this is a substantial reduction (Wakefield, 2013).

Now, let defining the Naive Bayes algorithm considering that M is a discrete-valued vector and that the attributes L1. Ln are any discrete or real-valued attributes in general. For every new instance L to classify the task is to train a classifier can generate the probability distribution over every likely value of M. Corresponding to the Bayes theorem the likelihood that M can take on its kth potential formula is (Wakefield, 2013).

$$P(M = mk\ |L1 \dots Ln) = P(M = Mk)P(L1 \dots .\ Ln\ |M = mk)\Sigma P(M = Mj)P(L1 \dots .\ LN\ |M = mj)j \qquad (10)$$

where the amount is taken every probable value mj of M. If considering the Xi are provisionally independent given M and also equation (1) re-written as:

$$P(M = Mk\ |L1 \dots Ln) = P(M = yk)\Pi p(li\ |m = mk)i\Sigma P(M = mj)\Pi P(ijli \mid m = mj) \qquad (11)$$

The Naïve Bayes classifier’s fundamental equation is Equation (10). This equation illustrates how to quantify the likelihood when Y can take on any given value offered the pragmatic attribute values of X_{new} and the supplies P (Li| M) check and P(M) calculated from the training results, given a new instance Lnew = (L1...Ln). If just think about the most possible value of Y, it may use the Naïve Bayes classification rule:

$$M \leftarrow \text{argmax}\ L(M = Mk)\Pi P(Li\ |M = mk)i\Pi P(i(Li \mid M = mj) \qquad (12)$$

that make simpler to the successive equation-

$$M \leftarrow \text{arg maxmk}\ P(M = mk)\Pi P((Li \mid M = mk) \qquad (13)$$

In educational settings, it is critical to be able to anticipate a student’s performance. Academic success is influenced by a variety of elements, including personal, social, psychological, and environmental influences. A strategy for using a Recurrent Neural Network (RNN) to predict students’ final grades using log data kept in educational systems is also addressed (Wakefield, 2013).

iv. Recurrent Neural Network (RNN)

RNN attempts to produce the input patterns in the output which is important in data mining approach. Time series data is handled by using a Recurrent Neural Network (RNN). An RNN features a recursive loop, unlike a conventional Neural Network (Okubo, Yamashita, Shimada, & Ogata, 2017).

The RNN generates core information from a prior time to the present time and calculates the output value created on the current and previous information. Hence, it is feasible to output in matter of the past state. Backpropagation Through Time (BPTT) is used to train the RNN’s parameters over time. The BPTT traces back to time t-1 to propagate the error among the ground truth and the output at time t (Okubo, Yamashita, Shimada, & Ogata, 2017). In the same way, a mistake at time t-1 is propagated to time t-2, and training is done backwards. Though the RNN can theoretically output with all previous information taken into account, the mistake cannot spread far into the past. As a result, it is an output that solely considers information from the previous many times. Long Short-Term Memory (LSTM) is used as a unit in the intermediate layer to store long-term information to solve this problem (Okubo, Yamashita, Shimada, & Ogata, 2017).

The interior state of the LSTM is kept in memory. By internal state at a prior time, the memory information collected in LSTM is maintained useful data or removed remove information. The RNN’s middle layer unit is the LSTM (Okubo, Yamashita, Shimada, & Ogata, 2017).

All above techniques are very useful for identify the Predicting student performance very accurately.

$$a(t) = b + Wh(t-1) + Ux(t) \qquad (14)$$

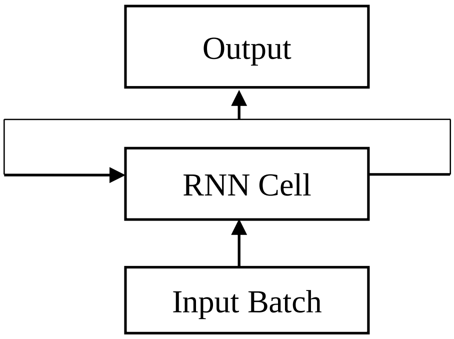

Figure 3: Rolled version of RNN (Okubo, Yamashita, Shimada, & Ogata, 2017).

Prediction and knowledge discovery are two types of challenges that data mining programmers can handle. It is suggested that you employ some related approaches for each of these issues. It also uses classification or regression for prediction, and association rules, clustering, database segmentation, sequence analysis, or visualization for knowledge discovery. Figure 3 shows the rolled version of RNN in which an input batch is connected to the RNN cell, then the RNN cell provides the Output. A classification rule tries to predict the value of a discrete dependent variable based on a set of known characteristics. The categorization based on decision tree is one of the most often used approaches. By tracing a path from the root to a leaf node, the decision tree may anticipate a new data instance. One of the benefits of decision trees is that it can be simply translated into a series of 'IF –THEN' rules that are simpler to comprehend. Clustering, also known as unsupervised learning, is a method for discovering patterns in data without the use of supervision. Unsupervised algorithms, as the name indicates, are capable of discovering structures on their own by exploiting comparisons or differences among individual data points in a data collection. Association rules mining is a popular data mining technique for uncovering interesting connections in huge datasets. Interesting connections between data points can lead to knowledge that can be utilized to make decisions. For merging rule-based reasoning (RBR) with case-based reasoning, there are three fundamental sorts of techniques (CBR). The prominence of each of the 2 component systems in the inference process is used to categories them. Following are the approaches that are used in this (Okubo, Yamashita, Shimada, & Ogata, 2017).

- **Rule-Dominant Approach**

This method concentrates on the rule-based component and only employs the case-based component when the rules are unable to handle unique scenarios in the process of data mining. The rules are used as a starting point for problem-solving, and then case-based reasoning is used to deal with exceptions to the rules (Okubo, Yamashita, Shimada, & Ogata, 2017).

- **Case-dominant approach**

The case-based reasoning (CBR) module takes precedence here, followed by the rule-based reasoning (RBR) module. The rules provide a supporting role to case-based reasoning in this paradigm, which is effective when the case library has a limited amount of cases (Okubo, Yamashita, Shimada, & Ogata, 2017).

- **Balanced approach**

Balanced techniques use a ‘mixed' paradigm, in which the integrated components' invocation order is not fixed, and one component dynamically invokes the other during inference (Okubo, Yamashita, Shimada, & Ogata, 2017).

At the start, choose a random Centre point for the cluster.

III. KNOWLEDGE DISCOVERY PROCESS

The process of discovering information in given data sets, regardless of its qualities or size attributes, is known as knowledge discovery. Many processes are involved in comprehending and extracting the pattern from the provided datasets. When choosing a database for data analysis, there are five primary factors to consider. These are the factors that the database belongs to the Knowledge Prerequisite that is necessary to comprehend the database, the Application Knowledge that is required to acquire the desired qualities, and many more. The goal that must be met following data pattern extraction, as well as the level of achievement that must be attained when the pattern is identified in the database. Also, must first choose the relevant dataset and construct the appropriate variables that are required for database matching and analysis. If the concerned variable is not properly entered into the database, the data pattern output can be skewed. After the variable in question has been corrected, the data set is cleaned and prepared for pre-processing. In the data cleansing stage, all noise and incompleteness are removed from the data (Cios, Pedrycz, & Swiniarski (2012).

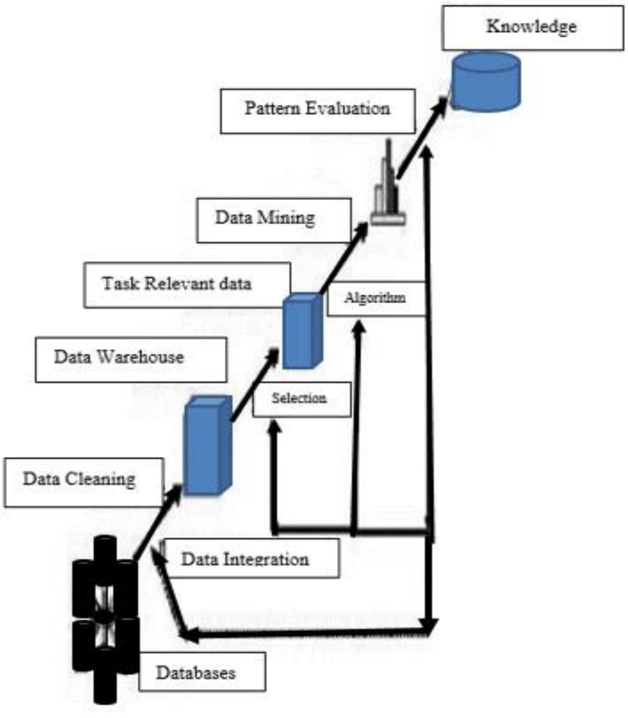

Figure 4: KDD ((Hemanth, Vastrad & Nagaraju, (2011))

IV. LITERATURE REVIEW

This section shows the several related work of many authors:

Shankar, Naresh & Agrawal, (2021) suggested a concept in which a revolutionary hybrid data mining approach combines Clustering with a Modified Apriori Algorithm to increase software defect prediction efficiency and reliability. This method reduces the amount of association rules that are created. The results are obtained through the use of a measure of interest known as spread. The findings of the innovative approach are compared to the results of the current hybrid Clustering and Apriori technique in the publication.

Nega, & Kumlachew, (2017) offers a hybrid intelligent system that use data mining as a method for acquiring knowledge. By delivering extracted information to a rule-based reasoning system, data mining explains the challenge of rule-based reasoning knowledge acquisition. WEKA is used to build and evaluate models, Java NetBeans is used to integrate data mining findings with rule-based reasoning, and Prolog is used to describe knowledge. J48, BFTree, JRIP, and PART are used in four studies to find the optimum model for illness diagnosis. The PART classification method is chosen as the best classification algorithm, and the rules provided by the Planning of Activities, Resources, and Technology (PART) classifier are utilized to construct the hybrid intelligent system's knowledge base. The suggested system has an accuracy of 87.5 percent and a usability of 89.2 percent in this investigation.

Korovin, Khisamutdinov, Schaefer & Kalyaev, (2016) explained, look at how a qualitative enhanced oil recovery (EOR) application might help increase the effectiveness of heavy oil production. A fresh technique is offered, which is based on the study of data from effective events. A synthesized model is created in especially for automated well search for EOR applications. In this authors describes a unique data processing method related to a hybrid implementation of NN analysis and evolutionary algorithms [20]. The suggested method allows for the selection of EOR in oilfield settings that are ambiguous and difficult to codify, reducing the reliance on human variables.

Chamatkar & Butey, (2014) suggested to evaluate, manage, and make decisions with such a large volume of data, data mining techniques are required, which can alter numerous fields. Data sets can be evaluated in Data Mining to uncover hidden and undiscovered estimates which can be utilized in the future to make more informed decisions. Data mining is a process for searching data warehouses that combines pattern recognition, mathematical, and statistical tools to assist analysts in discovering key patterns, factual correlations, and anomalies.

Hemanth, Vastrad & Nagaraju, (2011) suggested a knowledge discovery system for the choice of engineering materials that match design parameters is created using a predictive data mining approach and a ML algorithm. Materials categorization and selection are done using a predictive approach (Nave Bayesian classifier) and a ML algorithm (Pearson correlation coefficient method). The information acquired from engineering materials data sets is offered for use in advanced engineering materials design purposes for effective decision making. It also cover the relevance of data mining, as well as several challenge regions and application areas in data mining, in this suggested work.

The goal of this research is to create a hybrid methodology that combines data mining methods like association rules and classification trees. The methodology is tested by comparing it to other methodologies using real-world emergency data obtained from a hospital. Physicians should be able to classify chest pain disorders more quickly and accurately using this system (Ha & Joo,(2010). Table 1 shows the summarize table of Literature Review.

Table 1. Summarize Table of Literature Review

References	Technique	Outcomes
Shankar, Naresh & Agrawal, (2021)	Modified Apriori Algorithm	findings of the innovative approach are compared to the results of the current hybrid Clustering and Apriori technique in the publication
Nega, & Kumlachew, (2017)	hybrid intelligent system	The suggested system has an accuracy of 87.5 percent and a usability of 89.2 percent in this investigation.
Korovin, Khisamutdinov, Schaefer & Kalyaev, (2016)	qualitative enhanced oil recovery (EOR) application	The suggested method allows for the selection of EOR in oilfield settings that are ambiguous and difficult to codify, reducing the reliance on human variables
Chamatkar & Butey, (2014)]	Data mining	uncover hidden and undiscovered predictions which can be utilized in the future to make more informed decisions
Hemanth, Vastrad, & Nagaraju . (2011)	knowledge discovery system	cover the relevance of data mining
(Ha & Joo,(2010)	create a hybrid methodology	combines data mining

V. COMPARATIVE STUDY

Following table 2 show the comparative study. It demonstrates the Publication, Technique, Major Contribution and performance metrics evaluated.

Table 2. Comparative Study

References	Publication	Techniques	Major Contribution	Performance metric Evaluated
Cheng, Wu, Yuan & Wan, (2016).	IEEE, 2016	Semi-Supervised Learning	Deals with the class-imbalance problem separately and then performs classification of defects	F-value
Özturk & Zengin (2016)	IEEE, 2016	Sampling	Hybrid approach for dealing with class imbalance problem for very huge project with large datasets.	G-Mean
Liu, Li, Shao & Liu (2015)	IEEE 2015	Fuzzy Logic	Consider the interaction among the attributes to improve the defect prediction quality.	Accuracy, Recall, Precision and F value.
Rana, Mian & Shamail (2015).	Elsevier 2015	Association Rule Mining	This hybrid method shows a huge improvement of forty percent against the execution of NB alone.	Recall
Arar, & Ayan (2015)	Elsevier 2015	Artificial Neural Network	The hybrid method of ANN and Artificial Bee Colony is a revolutionary strategy that has been tested and shown to increase performance over previous classifiers	Balance, NECM, Accuracy, AUC, Probability of Detection, Probability of False Alarm
Laradji, Alshayeb & Ghouti, (2015)	Elsevier 2015	Classification	For the goal of fault classification, a hybrid method combining feature selection and ensemble learning is used.	G-Mean, AUC

VI. RESULT ANALYSIS

This section suggested the accuracy of the different algorithm. Table 3 shows the accuracy of several algorithm.

Table 3. Accuracy

Reference	**Algorithm**	**Accuracy**
Singh & Jindal (2018)	LR	77%
Ba-Alwi & Hintaya (2013)	Naïve Bayes	96.52
Singh & Jindal (2018)	NN	89.01%
Bansal,Shrama & Kaur (2020	RF	92.06%

Table 4 shows the precision value of several algorithms.

Table 4. Precision

Reference	Algorithm	Precision
Bansal,Shrama & Kaur (2020)	KNN	80.98
Bansal,Shrama & Kaur (2020)	LR	78.57
Bansal,Shrama & Kaur (2020)	SVM	71.12
Bansal,Shrama & Kaur (2020)	Random Forest (RF)	95.18
Bansal,Shrama & Kaur (2020)	Hybrid approach	96.56

Figure 5 shows the accuracy of several algorithms in which the highest accuracy is achieved by Naïve Bayes.

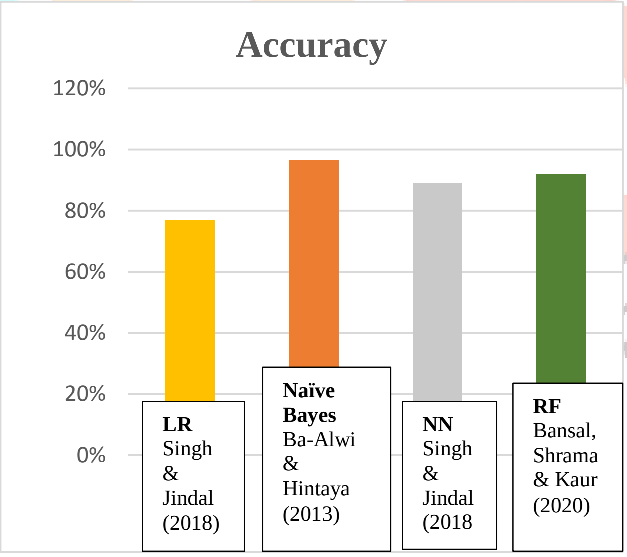

Figure 5: Accuracy of different algorithms

Figure 6 Shows the bar graph of Table 4. In which, the highest precision value is achieved by Hybrid approach and RF and lowest accuracy is achieved by LR.

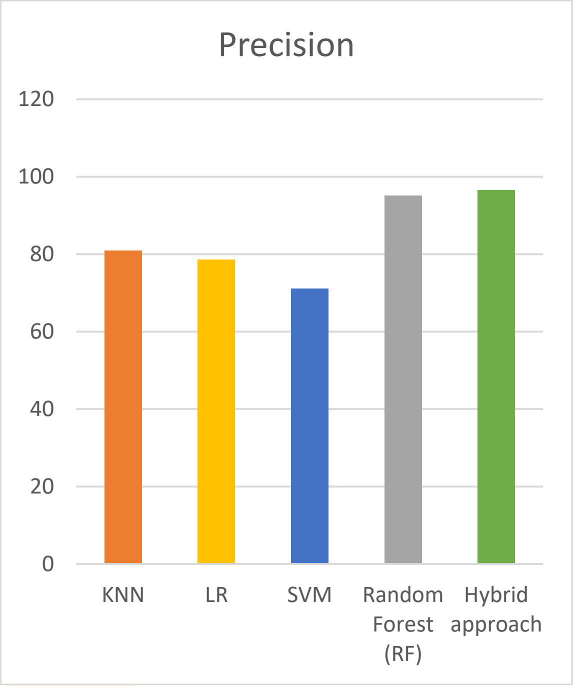

Figure 6: Precision of different algorithms

VII. CONCLUSION

A hybrid predictive data mining approach for KDD is discussed in this work. Materials informatics uses a predictive NB classifier and many algorithms for materials categorization and variety. The suggested method for extracting knowledge from a materials collection is reviewed. The naïve Bayesian classifier's method is applied in steps to explain classification issues and the result is the suggested materials class. The results are comparable and positive, indicating that the suggested techniques can be useful in data mining for high accuracy and Precision. With more data instances, the accuracy can be improved even further.

REFRENCES

[1] Hemanth, K. S., Vastrad, C. M., & Nagaraju, S. (2011, January). Data mining technique for knowledge discovery from engineering materials data sets. In International Conference on Computer Science and Information Technology (pp. 512-522). Springer, Berlin, Heidelberg.

[2] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. Ieee Access, 5, 15991-16005.

[3] Nagy, H. M., Aly, W. M., & Hegazy, O. F. (2013). An educational data mining system for advising higher education students. World Acad. Sci. Eng. Technol. Int. J. Inf. Sci. Eng, 7(10), 175-179.

[4] Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.

[5] Madni, H. A., Anwar, Z., & Shah, M. A. (2017, September). Data mining techniques and applications—a decade review. In 2017 23rd international conference on automation and computing (ICAC) (pp. 1-7). IEEE.

[6] Shukla, R. K., Sharma, P., Samaiya, N., & Kherajani, M. (2020, February). Web usage mining-a study of Web data pattern detecting methodologies and its applications in data mining. In 2nd International Conference on Data, Engineering and Applications (IDEA) (pp. 1-6). IEEE.

[7] Omar, T., Alzahrani, A., & Zohdy, M. (2020). Clustering approach for analyzing the student's efficiency and performance based on data. Journal of Data Analysis and Information Processing, 8(03), 171.

[8] Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., ... & Durda, K. (2012). Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nature genetics, 44(3), 312-318.

[9] Agrawal, G. L., & Gupta, H. (2013). Optimization of C4. 5 decision tree algorithms for data mining application. International Journal of Emerging Technology and Advanced Engineering, 3(3), 341-345.

[10] Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., & Kohli, P. (2011, November). Decision tree fields. In 2011 International Conference on Computer Vision (pp. 1668-1675). IEEE.

[11] Wakefield, J. (2013). Bayesian and frequentist regression methods (Vol. 23). New York:: Springer.

[12] Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017, March). A neural network approach for students' performance prediction. In Proceedings of the seventh international learning analytics & knowledge conference (pp. 598-599).

[13] Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (2012). Data mining methods for knowledge discovery (Vol. 458). Springer Science & Business Media.

[14] Kumar, N., Jain, S., & Chauhan, K. (2019). Knowledge Discovery from Data Mining Techniques. International Journal of Engineering Research & Technology (IJERT), 7(12), 1-3.

[15] Shankar, S. P., Naresh, E., & Agrawal, H. (2021). Optimization of association rules using hybrid data mining technique. Innovations in Systems and Software Engineering, 1-11.

[16] Nega, A., & Kumlachew, A. (2017). Data mining-based hybrid intelligent system for medical application. International Journal of Information Engineering and Electronic Business, 9(4), 38.

[17] Korovin, I., Khisamutdinov, M., Schaefer, G., & Kalyaev, A. (2016, May). Application of hybrid data mining methods to increase profitability of heavy oil production. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 1149-1152). IEEE.

[18] Chamatkar, A. J., & Butey, P. (2014). Importance of data mining with different types of data applications and challenging areas. Journal of Engineering Research and Applications, 4(5), 38-41.

[19] Hemanth, K. S., Vastrad, C. M., & Nagaraju, S. (2011, January). Data mining technique for knowledge discovery from engineering materials data sets. In International Conference on Computer Science and Information Technology (pp. 512-522). Springer, Berlin, Heidelberg.

[20] Ha, S. H., & Joo, S. H. (2010). A hybrid data mining method for the medical classification of chest pain. International Journal of Computer and Information Engineering, 4(1), 33-38.

[21] Cheng, M., Wu, G., Yuan, M., & Wan, H. (2016). Semi-supervised software defect prediction using task-driven dictionary learning. Chinese Journal of Electronics, 25(6), 1089-1096.

[22] Özturk, M. M., & Zengin, A. (2016, September). HSDD: a hybrid sampling strategy for class imbalance in defect prediction data sets. In 2016 Eleventh International Conference on Digital Information Management (ICDIM) (pp. 225-234). IEEE.

[23] Liu, L., Li, K., Shao, M., & Liu, W. (2015, November). Fuzzy integral based on mutual information for software defect prediction. In 2015 International Conference on Cloud Computing and Big Data (CCBD) (pp. 93-96). IEEE.

[24] Rana, Z. A., Mian, M. A., & Shamail, S. (2015). Improving Recall of software defect prediction models using association mining. Knowledge-Based Systems, 90, 1-13.

[25] Arar, Ö. F., & Ayan, K. (2015). Software defect prediction using cost-sensitive neural network. Applied Soft Computing, 33, 263-277.

[26] Laradji, I. H., Alshayeb, M., & Ghouti, L. (2015). Software defect prediction using ensemble learning on selected features. Information and Software Technology, 58, 388-402.

[27] https://www.ijariit.com/manuscripts/v4i2/V4I2-1487.pdf

[28] Ba-Alwi, F. M., & Hintaya, H. M. (2013). Comparative study for analysis the prognostic in hepatitis data: data mining approach. International Journal of Scientific & Engineering Research, 4(8), 680-685.

[29] https://www.ijaiem.org/Volume9Issue7/IJAIEM-2020-07-31-35.pdf