# AI-based Pattern extraction and evaluation for knowledge representation in Biomedical Data

**Shashi Pal Singh [a], Ritu Tiwari [b], Sanjeev Sharma [b],**

[a] *AAI, Center for Development of Advanced Computing, Pune, India*

[b] *Indian Institute of Information Technology (IIIT) Pune, India*

**Abstract**: We live in an era of High computing and performance where large data of various patterns of diseases is produced daily, where we need to understand the behavior of certain viruses and large groups of unicellular microorganisms. It is an absolute necessity to detect the various patterns among these viruses and bacteria by using text mining techniques and applying Artificial Intelligence (AI) Machine learning (ML) mechanisms.

Pattern extraction and evolution for knowledge representation work with the idea of data and text mining and processing for optimizing the experience of a user by training the machine using the Applied Intelligence approach to work for the best output. The work is divided into three phases which helps to move forward in a better way. The first phase is pre-processing the data, transforming it labeling it, and storing it in the database. The second phase is pattern extraction in which the use of either the imperial or knowledge-based algorithm is made. The third and final phase is the phase in which after feature extraction and inducing the knowledge evolution and evidence-based reasoning, indexing we finally the data. This data gets updated to the knowledge base again. The idea of proceeding with this process starts with taking the data that has been provided and applying various steps such as pre-processing input data and identifying the patterns in it by using and optimizing the algorithm.

**Keywords**: Text Mining, Pattern extraction, Natural Language Processing, Machine Learning

## 1. Introduction

AI is bursting fire in the health space with a new data-driven approach. The methods and algorithms of AI will be so much more effective in solving the complex problems of the health system. After the implication of AI in the health care department, it will also be cost-effective for the people. The academic compositions provide very helpful information to practitioners regarding the effect of AI in the medical profession.

Medical students, nowadays, understand the great impact of AI in the health awareness department and have positive gratitude towards this. But also after this gratitude, many professionals and scholars have deprived the AI technologies. In a survey of pathologists, the suggestions were split between the believers and pathology should be held equally responsible and another one is reporting that the platform vendor should be responsible. Along with the blooms in the field of AI, it also gives a tremendous level of ethical issues such as Autonomy, Fairness, or privacy, but other than fairness, these issues have less awareness in the scientific literature. It has been advised that medical data be estimated to avoid partiality

The smart life that we heading towards has led to a huge need to deal with the data that gets produced daily. The main question is, how do we deal with this amount of data? To increase the quality of life and create a better basis for decisions that are made based on this data which does not always have human supervision at all times the need for having a proper way to manage this data has arisen. We can see the rise in the usage of Artificial Intelligence applications in the healthcare sector has been very useful and it is helping the healthcare system in many aspects related to patient care, treatments, and administrative processes. AI has the potential to aggregate and examine a large amount of different data and then generate precise diagnoses for a wider section of the population. Hence AI can create a great impact on the healthcare section of society [11].

It can be seen that in recent years, deep learning has played a major role in image recognition, including medical imagery. Deep Learning used in medical image analysis is emerging as a fast-growing research field. Deep Learning has been very useful in medical imaging to detect the presence or absence of diseases. CNN (Convolutional Neural Network) is an Artificial Neural Network (ANN) based on a deep learning algorithm, used in solving problems such as text analysis, object recognition, and pose detection. CNN has also been used in the medical image market. CNN uses the concept of the human nervous system. In this study, a deep learning framework, i.e., a COVID-19 detector is used and demonstrated on the data set. The machine-learning community trains and evaluates statistical machine-learning models. After experiments, observations suggest that COVID-19 can be diagnosed using X-ray images. In this paper, 201 chest X-ray images of COVID-19 patients have been used. But for more accuracy, we require more chest X-ray images of COVID-19 patients [12]

This data evidentially tends to be inconsistent. Now, the question is how to deal with inconsistency. When dealing with this we tend to have not very accurate results on using the old ways of doing the same. This has led to the creation of specific interest in this field. Many types of research have shown progress and changes that significantly enhance the quality of the process, proposing more rational and doable methods.

Creation of an understanding of the working of the pattern extraction and evolution knowledge representation. we go through the different things that add up to define and describe the topic of the medical domain. Pitching a modification to the already existing knowledge would enhance the quality of the process in practice. The theory of the visible problems is acknowledged and a detailed in-site is given on the process.

The solution to such a problem lies in the concept of providing proper sampling, filing, etc., followed by setting proper clustering and processing which eventually leads to data mining. Data mining is one of the key steps to achieving the goal of pattern extraction and the evolution of knowledge representation. Algorithms may be used to identify pictures and videos, create recommender systems, categorize images, do medical image analysis, and evaluate Natural Language.

## 2. Literature Review

Data Science is the area of study that links such as the Venn diagram. In this pandemic, we face multiple issues in the medical field. In this essay, we can study data science, which is a field of study that links domain experience, programming abilities, and math and statistics understanding to extract useful insights from data. Data science deals with unstructured data that comes from places like the medical sector, social media feeds, smart devices, and emails that don't fit neatly into a database.

In the midst of this pandemic, we are confronted with so many medical problems. The COVID-19 pandemic has largely changed the way public health professionals work and communicate digitally, which means that everything has shifted to online mode or meetings such as google meet, zoom, skype, and others, and also that everything has shifted to the online mode or meetings such as google meet, zoom, skype, and others. Remote working options have become the standard in a relatively short period of time. Physical separation measures have hastened the shift toward digital communication and social interaction in recent years. The work of epidemiologists (disease detectives) has been scrutinized by the media, governments, and the general public in ways never seen before. Over the last several decades, social media has become a vital part of our society.

Medical fact gathering refers to the wide range of activities aimed at collecting and maintaining a large database containing all types of medical information as well as finding patterns in discovered information comprising electronic health records, results of drug trials, list of pharmaceutical products available on the market, and other data [13].

Text mining by analyzing patterns of literature-based phenotyping definitions has led to creating the basis for the work. The previous work on this topic has been rather split and varied due to this a common agreement about the topic which is informative and overlaps could not be achieved. Concluding anything from the already acquired data is a really hard task. textual content mining strategies can help in many phenotype definitions from previous work and information. There is proof that iterated styles inside any phenotyping definition are present. studying iterable styles in this topic is a study beginning for digging into phenotyping definitions. Biomedical text mining has proven proof in helping text mining give numerous benefits, inclusive of expanded understanding finding and expense discounts. Text and research-based information has proven proof of being relevant.[1]

Systematic expertise management and knowledge Discovery studies inside the place of statistics warehouses and organizational and commercial enterprise know-how control have generated crucial consequences one issue with dealing with these is the use of old techniques for dealing with records can be insufficient in understanding control oakum incorporate like:

Text mining is the analyzing act of large amounts of data to uncover business or organization information that aids firms in solving issues, mitigating risks, seizing new opportunities, etc, providing health care, and expanding their business. Every day, the huge Medical and clinical generates complex data from a variety of patient databases, such as electronic medical, patient reports, hospital equipment, etc. such as analyzing thousands of MRI images for commonalities that could influence how diagnoses or treatments are constructed. The Medical has become progressively difficult, demanding the restitution of information from the massive amounts of compound data to find the best treatment. We have a recent example of covid-19 second phase, which is currently experiencing several issues and requires data mining to improve. In data mining, lots of open-source tools such as Apache Mahout (It is a popular distributed linear algebra framework.), Data Melt or DeMelt (It is open-source software for numerical computation, mathematics, statistics, symbolic computation, data analysis, and data visualization.), ELKI (written ELKI) is) in Java language, it is an open-source data mining.), KNIME (KNIME is written in Java and based on Eclipse, KNIME Analytics platform is open-source software.), Orange (It is component-based data mining that is used for machine learning), Rattle (Rattle is written in R language, it is an open-source GUI for data mining). Data mining surveys show that healthcare provides the greatest coverage of data mining needs.

Other processes can be divided into classes: algorithm / version-orientated procedures that are looking to introduce new getting-to-know mechanisms or to modify current strategies to work with imbalanced data sets, and facts manipulation strategies that are searching to modify the distribution of data to make data sets much less imbalanced. An unbalanced category of information remains a challenging study's trouble. it's far even more difficult for multimedia facts because of its various media types and spatial-temporal traits.[2]

### A. Convolutional Neural Network (CNN)

1) Convolutional layer: a convolutional layer consists of several function maps. the feature map on the layer is computed with the resource of convolving its preceding layer's feature maps through an activation feature f with learnable kernels and additive bias. proper right here, the number one layer represents the entered information, and the activation feature f is commonly decided on to be the logistic (sigmoid) characteristic and represents an expansion of entering maps.

2) Pooling layer: a pooling layer works around giving the down-sampled variations of the input feature maps. speaks to multiplicative bias $\beta_j^l$ , is the additive bias $b_j^l$ , and pool speaks to a pooling activity which as a rule registers the amassed insights of the info maps by and large with their propose or max esteems. depending on the pooling activity executed, the layer might be known as suggesting pooling, max pooling, etc. This store is ordinarily done after each convolutional layer. three) associated layer: after various convolutional and pooling layers, the high-confirmation thinking in the neural network is cultivated through genuinely related layers. a connected layer takes all neurons inside the past layer which might be completely related, pooling, or convolutional, and associates it to each unmarried neuron it has.

$$X_j^i = f\left(\sum i \in mjX_i^{i-1} * K_{ij}^l + b_j^l\right), l \geq 2;$$

$$X_j^l = f\ (\beta\beta_j^l pool(X_j^{l-1} + b_j^l\ ), l \geq 2$$

Deep studying for imbalanced data set: For everyday CNN processing, the prediction mistakes price continues descending to a sure degree. Even as we exercise CNN on an imbalanced statistic set, the error rate might also in massive part range or maybe increase.

**CNN with bootstrapping technique**

Combine CNN with low-degree functions despite the fact that CNNs have been mentioned to perform nicely on several records units, its schooling segment is normally time-eating. it took a month to teach 1755 movies to reach a great universal overall performance. Many companies have reported that deep analyzing can be computationally in depth even as taking raw signal facts as the input. To deal with this trouble, we advise to use low-degree features which can be tons smaller in sizes as the enter to CNNs to lessen the computation time.

text processing the usage of artwork neural networks gives a detailed InSite on processing. The concept is changed clustering set of rules of Projective Adaptive Resonance concept (MA-element), based totally mostly on the authentic component algorithm.so. So, the primary pitch is ready the changed set of regulations shows about alternate when dealing with path or length characteristic is notably associated with choice of best performing neuron in the course of clustering. This increases the credibility of the output that may be carried out by way of optimizing the space characteristic. a number of the number one illustration models of textual content documents are:

(I) The Boolean model decreases the overall result of record illustration right in binary keyword vector, which is keyword vector data might have the most effective value like zero or one. Its foremost benefit is the easy and understandable illustration. The other way, the big con of this is that it doesn't consider for one the frequency of incidence and for two, load in decided on main keys, for the textual content proper textual presentation. [5]

(ii) VSM, a widely in-use model for the textual content file illustration. Its characteristic place is made through a collection of letters on the work also every feature corresponds to the textual content unit in the file. The file is represented as a collection of letters, irrespective of the shape or semantics of the text. the version in nonnetwork entirely on the statistics traits of the report, when talking about the rate at which these letters collected are picked and moved from documents and from queries. It subtracts the dangers of the version with (0 or 1). except the times that the keywords are noted, the counting of keywords in the report is done. According to this way of evolutionlike., according to the count they are assigned a certain weight and their importance is calculated with it. but, VSM fails to deal with the structure of the document or maybe contextual or semantic facts approximately specific keywords. [5]

(iii) The Latent Semantic Indexing model (LSI) works around the calculated moves on matrix representation of keywords and documents created via vector version.

(iv) Probabilistic model works on identical files presenting as a Boolean model. Although, while working on the similarity of texts in both cases the target is accomplished through a reparative manner primarily by evaluating the number of times it has been repeated to understand its future occurrences of their content material closeness. even though, the sequencing is contemplating inside the file illustration. [5]

(v)n-gram model of the textual content report illustration essentially carries the mission of an excessive opportunity of more than one co incidence of, 3 or n phrases. Seeing alternative aspect, a very little or 0 probability is given to phrases that tend to either not appear or appear in very less chances [5].

**B. Natural Language Processing and Its Future in Medicine**

Many people believe that the widespread adoption of computerized patient records will usher in a healthcare revolution by allowing for the development of technologies that improve quality while lowering costs. The electronic medical record (EMR) does improve access to patient records for health care practitioners.

Consider the various definitions of the following phrases containing the word pneumonia: evidence of pneumonia, pneumonia cannot be ruled out, pneumonia in 1985. If a clinician wanted to know how many active pneumonia cases were on file and searched using the keyword pneumonia, far too many reports for patients who did not have pneumonia would be returned.

Because NLP systems not only extract individual words but also reflect well-defined relationships between words, they may provide a solution. If the clinician in the above example used an NLP system to search for pneumonia, the system would extract the main finding of pneumonia for each phrase as well as the appropriate modifier for each record. The values of modifiers and their combinations could then be used to determine whether or not a patient had pneumonia.

In the medical field, NLP has already begun to show promising outcomes. Two NLP systems, for example, have been integrated into operational clinical information systems. Based on radiograph reports, several NLP systems have been utilized to aid decision-making. These methods were found to be capable of detecting: (1) anomalies in chest X-rays; (2) patients suspected of tuberculosis; and (3) breast cancer findings in tests. Most importantly, the tests revealed that NLP systems performed as well as or almost as well as medical specialists in detecting abnormalities. Other NLP applications have been created to encode admission diagnoses, monitor asthma patients, translate data to SNOMED codes, and automate severity evaluation for community-acquired pneumonia

## C. The form of the issue

a detail includes input processing layer F1, also named contrast layer, and output layer F2, referred to as clustering or popularity layer (Fig. 1). Neurons in the F1 layer are denoted via the usage of Ui, I=1, m, neurons in F2 layer with the aid of V', j=1, n. The F2 layer is an aggressive layer, that works on giving the best output all of it. The activation values of neurons in the input layer F1 are denoted through xi, wherein I = 1 … m is an index of a neuron within the enter layer F1. The activation values of neurons within the output layer F2 are denoted via yd, wherein j = 1 … n is an index of a neuron in the output layer F2. The output signaling function inside the comparative layer F1 is denoted with the resource of f1. Neural connections amongst layers F1 and F2 are described via the use of the lowest-up Wyandot-down Zij eights.

The neuron inside the clustering layer F2, is known as dedicated when a few input styles change into located out in step with devoted neuron within the reputation portion of the process. else neuron can be known as not devoted. handiest dedicated neurons receive signs from the F1 layer inside the segment to compare. factor neural network, layer F1selectively sends signs and symptoms to F2. The neuron of form of the issue

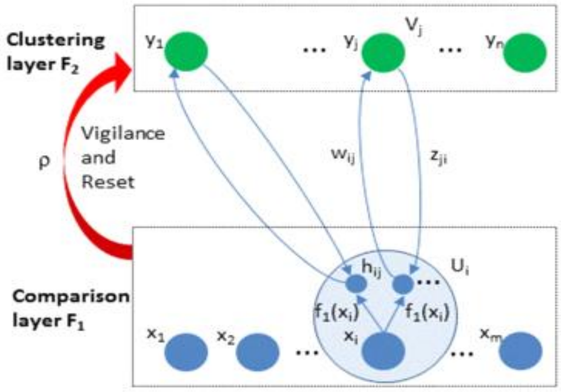


**Fig: 1** form of the issue

F1 maybe living in the direction of a sure F2 and not active in opposition to different F2. interest of neuron inside the F1 is predicted with the useful resource of similarity check between top-down weights and the price of the feature f1(xi), generated in a neuron Ui. [5]

precisely this similarity takes a look at performs a key feature in MA-a part of statistics clustering.

## D. mathematical version of the element

The particular result sign can be calculated thru courting among any neuron Ui inside the input layer F1 and committed neuron V' of the output layer F2 in which l(win) and he(f1(xi), Zij) are step functions, θ is a threshold value of the lowest-up weights win, σ is a distance parameter and d(f1(xi), Zij) is the space function.[5]

The very last output signal raj is measured in every capability best result because of the group of the chosen indicators within the comparison segment.[5]

$$r_j = \sum_i hij$$

within the variation segment the lowest-up win and top-down Zij weights of devoted neurons are set with the aid of relations.

$$w_{ij}^{new} = \{\{_{0,therwise}^{L1(l-1+r_{j}),for\,active\,V_{j}}$$

$$z_{ji}^{new} = (1 + a)z_{ji} + al_j$$

where L= weight steady, α=learning parameter and I is it element of the enter statistics vector I.

Non-committed neurons get bottom-up and top-down weights put using way of members of the family

$$w_{ij}^{new} = LI(L + 1 + m)$$

$$z_{ji}^{new} = I_i$$

The pre-processing of textual content files includes the steps below:

1. Make every letter of the report to lower-case. [5]

2. tokenization of the data, i.e., all unique characters and marks are deleted. Now, the sentences get split into smaller terms called phrases referred to as tokens, divided via regions. [5]

3. to filter Token, i.e., an expansion of the one's tokens, which period with the range of min and max value key-word. [5]

4. deleting of the stopping phrases from premise of generally present listing of stopping words in English.[5]

The new rule for calculating the selective output sign him in MA-part is defined as follows:

$h_{ij} = \{^{1,}_{0},$

if (d(f$_1$ (x$_i$),z$_{ji}$ )≤$\sigma$) ∧ (w$_{ij}$> θ) ∧

∧ (x$_1$=c$_1$) ∧ (x$_2$=c$_2$) ∧ (x$_3$=c$_3$)

otherwise

***Textual content document Clustering***

The manner of text file clustering has the following stages (Fig. 3):

1. text record pre-processing.

2. Clustering via MA-element – introduction of projective clusters, which includes applicable centroid for each cluster.

3. Regrouping of the created clusters to actual topic regions with the aid of clusters regrouping algorithm. [5]

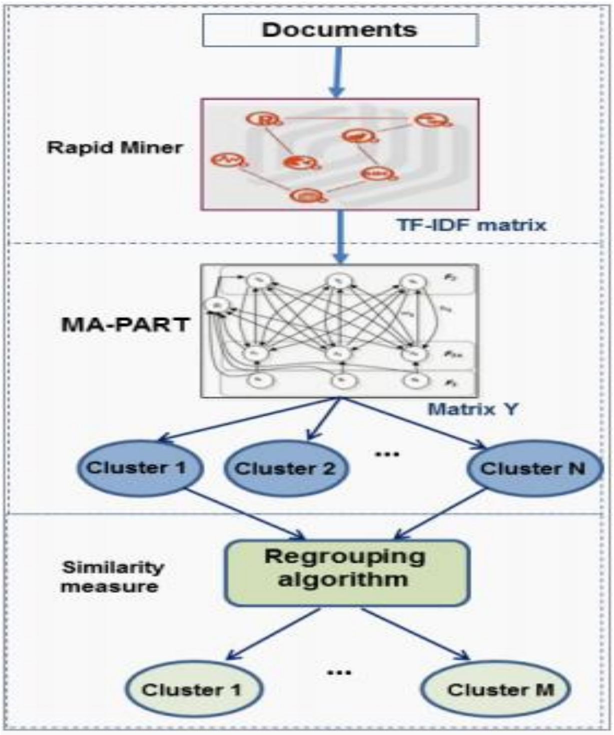


**Fig2.** Clustering

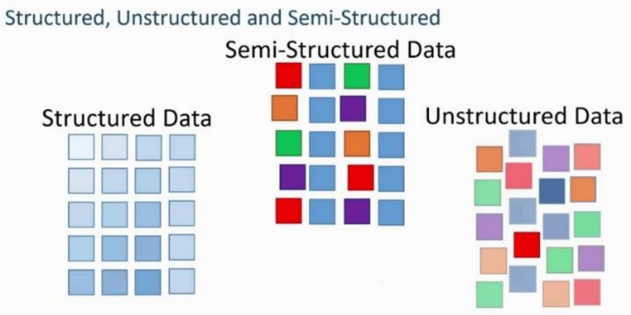


***Fig 3*** *Structured, Semi-Structured, Unstructured Data*

VSM does no longer cope with the way the file forms or perhaps contextual or semantic information approximately particular key-word. (LSI) is primarily based totally on the matrix illustration of key terms and documents created by vector model

## 3. Methodology

Information Extraction as the name suggests it refers to the process of gaining out structured information (i.e. plain texts) from unstructured or semi-structured data.Semi-structured here means the information can be in form of tables or metadata whereas unstructured data can include images, audio and video files.

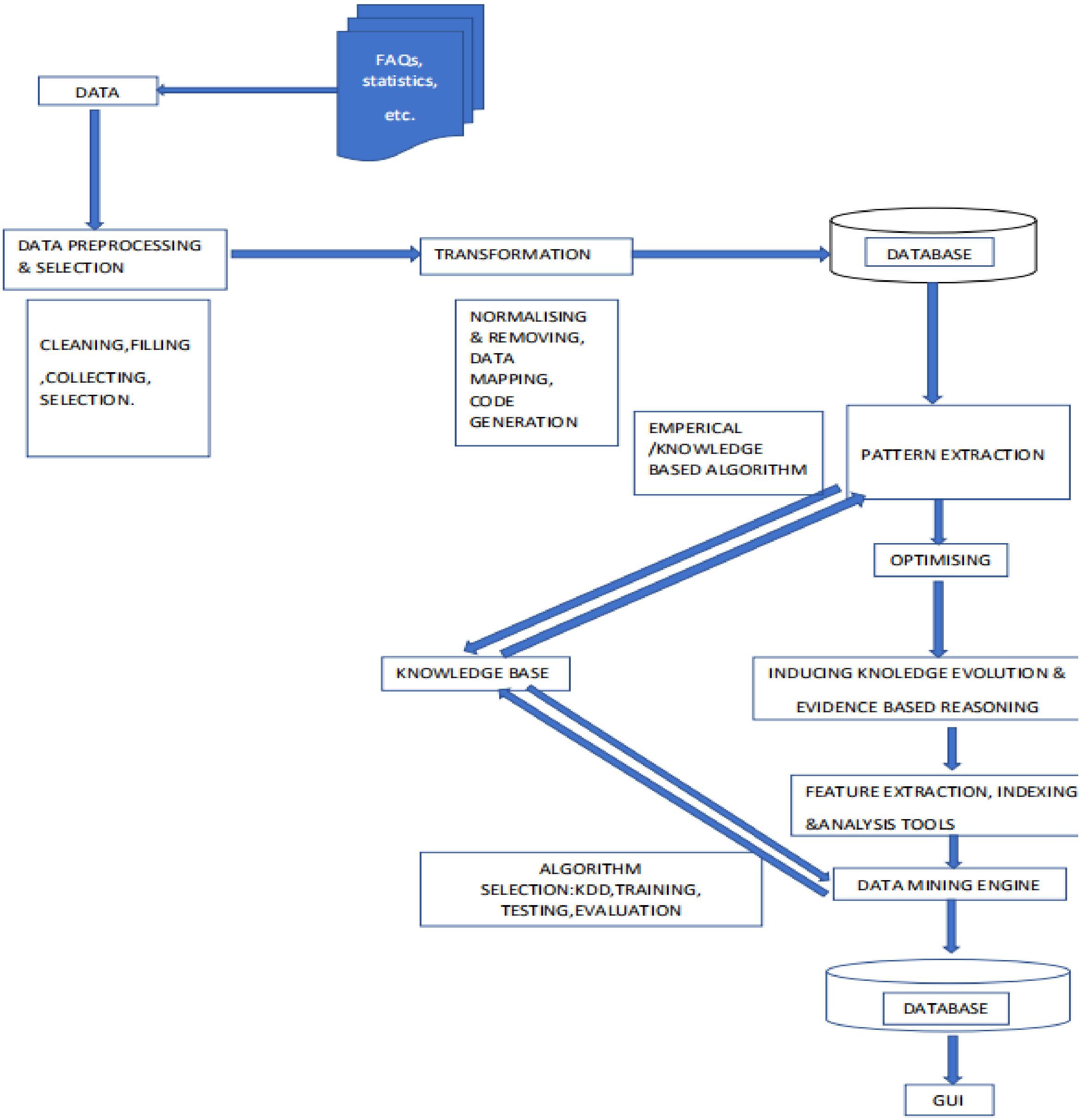

FAQs, statistics, etc.
DATA
DATA PREPROCESSING & SELECTION
CLEANING,FILLING ,COLLECTING, SELECTION.
TRANSFORMATION
NORMALISING & REMOVING, DATA MAPPING, CODE GENERATION
DATABASE
EMPERICAL /KNOWLEDGE BASED ALGORITHM
PATTERN EXTRACTION
OPTIMISING
KNOWLEDGE BASE
INDUCING KNOLEDGE EVOLUTION & EVIDENCE BASED REASONING
FEATURE EXTRACTION, INDEXING &ANALYSIS TOOLS
ALGORITHM SELECTION:KDD,TRAINING, TESTING,EVALUATION
DATA MINING ENGINE
DATABASE
GUI

**Fig3.** Methodology and WorkFlow

1) The first step is to collect our FAQs, and statistics from the preceding resources. This process performs an important part in figuring out lots of further steps. The information is extracted in this step and then similarly placed for evaluation.

2) We gather applicable information for processing it into meaningful and usable shapes. This received statistics is positioned right into a form fit for usage for pre-processing.

3) pre-processing is a process of lowering the opportunity of having facts that aren't always

4) matched for utilization. This includes steps like cleansing, submitting, gathering, and

5) choosing the stairs to smooth the statistics consist of:

a) doing away with unwanted observations out of your dataset. Replica observations most often rise up for the duration of statistics collection, together with:

- integrate datasets from a couple of locations
- Scrape statistics
- get maintain of statistics from clients/exclusive department Besides the point, observations are people who don’t certainly healthy the precise trouble that we’re trying to remedy.

b) restore Structural mistakes

records cleaning involves fixing structural errors. Structural mistakes arise during dimension, facts switch, or different types of "terrible housework. “For example, typos or inconsistent capitalization.

c) clean out undesirable Outliers

Outliers can motivate troubles with a sure variety of models. for instance, linear regression models are an awful lot much less strong to outliers than selection tree fashions. In fashionable, in case you've been given a legitimate purpose to put off an outlier, it's going to help your version’s universal overall performance. however, outliers are harmless till confirmed accountable. You must by no means dispose of an outlier in reality as it’s a "big amount." That large variety could be very informative to your model.

d) control lacking facts

lacking records is deceptively complex trouble inside the applied machine getting to know.

First, we cannot actually forget about lacking values for your dataset. we ought to manipulate them in some manner for the very practical cause that most algorithms no longer take transport of lacking values. [11]

- The algorithms used to fill consist of:
- A flood fill set of rules
- test-line polygon filling, and many others.[10]
- The statistics series consists of the subsequent steps:
- Create a statistics collection plan.
- arrange a team to collect information.
- prepare gear to gather statistics.
- overview of the collection method.
- prepare a toolkit for records pre-processing. [9]
- The algorithms used to select consist of:
- Curse of dimensionality — Overfitting
- Occam’s Razor
- garbage in rubbish out [8]

6) After considerably making the statistics extra usable we transform the information. This step involves normalizing and doing away with the available information and to make it greater studies friendly the mapping of statistics takes vicinity.

- algorithms used for normalization:
- Decimal Scaling
- Min-Max Normalization
- z-score Normalization (zero-mean Normalization) [12]
- Steps and calculations to purging data:

7) Now that we have all of the relevant records available, we positioned it inside the database.

8) After the statistics have been placed within the database one of the principal steps of pattern extraction takes vicinity. the set of rules placed to apply for pattern extraction:

- empirical algorithm
- know-how-based algorithm.

The end result of this step is introduced to the expertise base. This replacement is essential for the studies.

9) The pattern acquired now is then optimized.

10) These facts in addition are going throw the manner of function extraction, analyzing evidence, and evolution-based proof.

characteristic extraction algorithms used:

- unbiased thing evaluation
- Isomax
- Kernel PCA
- Latent semantic analysis and many others.

11) This reconfigured statistics then go through data mining. in this step, we look for the nice and maximum efficient methods to continue with the records with a relevant set of rules. The most common one used would-be KDD.

We additionally run the method of education, trying out and evaluating before including it to the understanding base again.

Steps of KDD:

- data
- sample
- procedure
- valid
- novel
- useful
- comprehensible

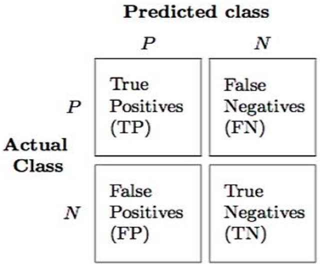
The knowledge base is sooner or later updated after mining.

12) We save the data that has been produced into the database for the closing time of this cycle.

13) This information is viewed in the picture used interface. This manner marks the quit to the manner.

## 4. Evolution & Performance

The work has some (7) phases: pre-processing and selection, transformation, sample extraction, optimization, reasoning, characteristic extraction, and facts mining the following may be evaluated with the assistance of performance measures such as precision, consider, and accuracy. The performance also can be evaluated via using the confusion matrix. A confusion matrix is a desk that is built to assess the overall performance of the category model.



| | | Predicted class | |
|---|---|---|---|
| | | *P* | *N* |
| Actual Class | *P* | True Positives (TP) | False Negatives (FN) |
| | *N* | False Positives (FP) | True Negatives (TN) |

1) Precision = TP/(TP+FP)

A metric that tells you the manner specific is your prediction on the positives.

2) Accuracy = (TP+TN)/(TP+FN+FP+TN)

allows us to recognize what percentage of the predictions are correct. The quandary of this metric is in the instances of rare sports. Say as an instance, 1% of the population has coronary heart-related sicknesses and the version predicts that none of them have coronary heart ailments, the accuracy of the model will still be 99%.

3) F1-score = (2*Precision*Sensitivity)/ (Precision Sensitivity)

An outstanding metric because it penalizes when you have got both excessive false positives or faux Negatives.

- genuine advantageous: those are the correctly anticipated values that mean each the actual and predicted magnificence have equal values i.e., sure. -proper terrible: those are the successfully predicted bad values that mean both the actual and predicted elegance have poor effects.
- false bad: this is produced when the anticipated magnificence and the actual elegance contradict each other i.e., actual elegance is negative and the anticipated fee is fine.
- fake superb: that is produced whilst the predicted elegance and the actual elegance contradict each other i.e., real elegance is advantageous and predicted value is poor.

Our studies provide a model with the following observation and tested on 10k sentences,

i. **Accuracy** = seventy-five% or more
ii. **Precision = 85% or more**
iii. **recall** = 70% or greater
iv. **F1 score** = weighted average of precision and consider

**= 75% approx.**

A model with excessive precision and excessive does not forget is considered to be the first-class version and a great system and you may get one hundred % precision most effective if a model produces no fake exceptional values and a hundred % bear in mind can handiest be received if a model produces no faux negative. A system with low precision and immoderate bear in mind returns many effects, but, most of its anticipated labels are incorrect whilst compared to the schooling labels. A device with excessive precision and espresso keep in mind returns only some outcomes but the maximum of its expected labels is accurate when in comparison to the training version. That's why a device with high precision and excessive don't forget is taken into consideration to be a sincerely ideal one.

## 5. Conclusion and Future Impacts

This study tells the uses of AI applications for medical research, training purposes, diagnosis, medical treatments, and decision making in the treatment of patients the success of the same has been primarily based on off through research papers and previous work at identical. The work comes from an area of the heavy influence of the preceding articles and research papers. we try to provide data and figures to resource the consistency of statistics. one's records have a propensity to be inconsistent. while coping with this we normally will be predisposed to have no longer very accurate effects on the usage of the antique procedures of doing the same. This has caused the arrival of a completely unique hobby in this place. lots of researches have shown the improvement and the adjustments that appreciably enhance the great of the way, supplying greater rational and possible techniques.

In near future predictions the expert system shows more increasing in the process of putting plans in the expert system organization. Moreover, in this expert system the vital part in healthcare that the diagnostic accuracy will help in the analysis of health facts by comparing thousands of medical sections of written, printed, or electronic matter through which they provide the evidence that serves as an official fat

The objective of that is to undergo the technique to recognize the already acknowledged articles on the equal. This information enables us to pitch a better and extra green manner of intending with the task. The idea is to achieve an accuracy of around 75%.

In the future, AI systems will become more powerful and hence will achieve the ability to perform a wider range of activities without human control. But there are some limitations also. AI applications cannot completely replace human interference.

## References

1) Analyzing Patterns of Literature-Based Phenotyping Definitions for Text Mining Applications Samar Binkheder, Ph.D. cand ,School of Informatics and ,Computing, Department , f ,BioHealth Informatics ,Indiana University – Indianapolis ,Indianapolis, Indiana ,sbinkhed@iu.edu ,Heng-Yi Wu , Ph.D. ,College of Medicine, Department ,of Biomedical Informatics ,

2) Deep Learning for Imbalanced Multimedia Data Classification Yilin Yan1 , Min Chen2 , Mei-Ling Shyu1 , and Shu-Ching Chen3,1 Department of Electrical and Computer Engineering ,University of Miami.Coral Gables, Florida, USA.2 School of Science, Technology, Engineering & Mathematics University of Washington Bothell ,Bothell, Washington, USA ,3 School of Computing and Information Sciences,Florida International

3) Optimizing Sequence Alignment in Cloud using Hadoop and MPP Database ,Senthilkumar Vijayakumar,Anjani Bhargavi, Uma Praseeda and Syed Azar Ahamed ,TATA Consultancy Services Ltd, Pioneer Building, 12th floor, ITPB, Whitefield Road, Bangalore, INDIA , E-mail:senthilkumar.vijayakumar, anjani2.b, umapraseeda.pk and syed.azar}@tcs.com

4) Text Processing by Using Projective ART ,Neural Networks ,Radoslav Forgáč ,Dept. of Parallel and Distributed Information Processing ,Institute of Informatics Slovak Academy of Sciences ,Bratislava, Slovakia ,radoslav.forgac@savba.sk ,Roman Krakovský ,Department of Informatics ,Faculty of Education, Catholic University ,Ružomberok, Slovakia, roman.krakovsky@ku.s

5) Evolution of Knowledge Representation and, Retrieval Techniques,Meenakshi Malhotra.Dayanand Sagar Institutions, RIIC, Bangalore, 560078, India ,Email: uppal_meenakshi @yahoo.co.in .T. R. Gopalakrishnan Nair ,Saudi Aramco Endowed Chair, Technology and Information Management, PMU KSA ,Email: trgnair@yahoo.com, trgnair@ieee.org

6) Knowledge representation and information extraction for analysing architectural patterns rahul agarwal-The 5 Feature Selection Algorithms

every Data Scientist should know (jul 27,2019):https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2

7) Lou Dutko-Data Collection Process for a Machine Learning Algorithm(23 Jul 2018):https://lembergsolutions.com/blog/data-collection-process-machine-learning-algorithm

8) Akanksha_Rai - Boundary Fill Algorithm( 04-12-2019): https://www.geeksforgeeks.org/boundary-fill-algorithm/

9) Data cleaning:https://elitedatascience.com/data-cleaningdeepak_jain Data Normalization in Data Mining( 25-06-2019): https://www.geeksforgeeks.org/data-normalization-in-data-mining/

10) OMAR ELGABRY-The Ultimate Guide to Data Cleaning(Mar 1, 2019 ): https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4 (Extracting meaningful patterns from tme series classification by Xiao hang zhang and others.

11) Adam Bohr1 and Kaveh Memarzadeh,Artificial Intelligence in Healthcare. 2020 : 25–60.Published online 2020 Jun 26. doi:10.1016/B978-0-12-818438-7.00002-2PMCID: PMC7325854 The rise of artificial intelligence in healthcare applications

12) Keidar, D., Yaron, D., Goldstein, E. et al. COVID-19 classification of X-ray images using deep neural networks. Eur Radiol 31, 9654–9663 (2021). https://doi.org/10.1007/s00330-021-08050-1

13) Pooja H, Dr. Prabhudev Jagadeesh M P, “A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records”, IEEE Xplore, 2019.