Pattern Recognition in Medical Decision Support and Estimating Redundancy in Clinical Text

Kalluri Shanmukha Sai[1], Rishi Reddy Thokala[2]

{Shanmukhasai9999.nani@gmail.com [1], rishireddythokala@gmail.com[2]}

SRM Institute of Science and Technology, Potheri, SRM Nagar, Kattankulathur, Tamil Nadu[1,2]

Abstract. Feature sets selected for multidimensional pattern classification are estimated using a novel criterion. The priority of features for each class is generally optimized to maximize relevancy and minimize redundancy between each class. While mutual information can be used to estimate relevancy information, redundancy information cannot be estimated since its dynamic ambit is determined by feature and class. In addition to assisting in adjust of changing patterns, support vector machines are helpful in classifying normal and abnormal patterns. In pattern matching, Random Forest Logic's pattern recognition algorithm lends itself easily to pattern matching algorithms. An algorithm for evaluating the classification performance of health care medical data is proposed. A variety of experimental results confirm that the proffered technique is more veracious than conventional algorithms relating to the classification of accuracy when the number of selected features is varied.

Keywords: Feature selection, Redundancy information, medical health care data, support vector machines.

1 Introduction

Healthcare analytics (HA) is an interdisciplinary field that uses statistical abilities and computerized sufferer data to assist physicians. Understanding a patient population, we require to analyse a vast mass of data. Recognizing patterns are a key tool for HA task, as it can be used to automatically identify patterns and regularities in data. Deep learning and machine learning are being utilized in healthcare to predict risk, track disease progression, and classify patients. Despite this, pattern recognition is challenged by healthcare. Heterogeneity, multidimensionality, nonlinearity, temporality, and distribution are some of the characteristics of the data, which can make it difficult to apply traditional techniques. To address these challenges, researchers in the pattern recognition domain are developing novel techniques that are specifically designed for healthcare applications. These techniques often involve the use of ML procedures can be learned from big amounts of records and identify patterns that would be difficult to detect by humans. Pattern recognition is a powerful tool that has the potential to revolutionize healthcare. By automating the identification of patterns in patient data, pattern

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343233

recognition can help physicians to make better decisions about diagnosis, treatment, and prevention.

Patients with limited medical knowledge often find it difficult to assess the appropriateness of their doctor's recommendations. This can lead to suboptimal care and waste of resources [1]. To address this problem, we need to develop an "independent doctor" that can provide objective advice on diagnostic testing. This could be achieved using an automated clinical decision-making mechanism that is trained on a large dataset of medical records. In current years, there has been a developing interest using machine learnings and artificial intelligence to develop such mechanisms. The task remains challenging, however. For example, it is difficult to define objective criteria for over-testing, and the quality of the training data can vary widely. Despite these challenges, I believe that automated clinical decision-making mechanisms have the potential to revolutionize healthcare [2]. By providing objective advice on diagnostic testing, these mechanisms can help to ensure that patients receive the best possible care while minimizing costs.

We present a new approach to identify over-testing using statistical modelling methods to simulate the heuristic reasoning process. This framework is based on information extracted from PubMed, the largest database of biomedical literature. We have developed a new way to identify over-testing that is more objective and accurate than previous methods. PubMed is a database of biomedical literature hosted by the National Institute of Health's. It contains chronological data for approximately 28 million blogs, as well as highlights and hyperlinks to the entire text. We believe that our work has the potential to make a significant contribution to the field of healthcare by helping to reduce over-testing. This could save money and improve patient outcomes. Our framework addresses this gap by providing a systematic approach to identifying over-testing based on the clinical thinking process.

A diagnosis of a disease involves specifying tests. Both have their basis in clinical thought and are intimately tied together: over-medicalization is the mirror image of over-depathologization. Motivated by the effectiveness of predictive models in therapeutic decisions in determining illness, we designed an allied statistic-based sets for assessing the appropriateness of diagnostical tests. To automate medical procedures, clinical decisions support systems (CDSSs) have been suggested to help clinicians make correct decisions promptly [3]. They help machine learning techniques analyze electronic medical records (EMRs) via statistical methodology, and then automatically make a diagnosis. Classifications are certainly one of the most extensively researched area in medical diagnosis [4]. By improving the performance of pattern recognition algorithms and methods, clinicians can make better-informed decisions in a timelier manner, thus improving the health-care outcome. In stressful environments such as intensive care units, this is particularly important for rapid clinical decisions[5].

In order to predict a clinical measurement and monitoring system's response, predictive computational models and pattern recognition algorithms must be developed to rapid evolving clinical environment. This requires continuous updates on the latest advances in this field. Pattern recognition for healthcare analytics is the desire of this Research Topic, such as feature extraction. Health care procedures can be thought of as another sort of 'learning from data," but patients' data raise particular difficulties that may not respond to conventional pattern processing techniques. We welcome submissions of well-designed clinically interpretable papers focused on meeting these challenges. Our call is for novel approaches to prediction

models, feature engineering, time series, presentation of the data (e.g., tables, charts), Machine Learning methods, and explainability of the predictions based on patient data in tables, text and images format. We also welcome interactive tools that make it easier for clinical researchers to use pattern recognition techniques. The next section discusses the literature review, the third section discusses the proposed algorithm, the fourth section discusses the result analysis, and the fifth section provides the conclusion.

2 Literature Survey

In order to improve clinicians' decision-making, medical decision support systems (MDPs) integrate affirmation -based knowledge and patient-specific info onto a computerized platform. Biomedical data (such as signals and images) have been subjected to a variety of pattern recognition techniques over the last decade to support automated and machine-based clinical diagnosis and therapy. This review focuses on the types of diseases that can be diagnosed using pattern recognition techniques and the challenges of storing and retrieving records in a database. A study by Lucia A. Carrasco-Ribelles et al. [6] described the goals of the study, (1) to enhance a method for identifying local congruences midst PTs before malaises occur, so that morbidities can be predicted for new query individuals; (2) to validate the methodology for predicting cardiovascular diseases (CVD) occurrence among diabetic patients. Sequences of longitudinal multiscale data are used to define PTs in a novel way. Additionally, PT alignments are identified through dynamic programming to predict morbidity in the future.

A model called adversarial neural networks with sentiment-aware Attentions (ANNSA) was proposed by Zhang et al. [7] to enhance social media sentiment and improve network performance using augmented data. The sentiment-aware attention mechanism learns undertaking-related statistics by the way of optimizing a task-specific loss associated with sentiment word and extract the word-stage sentiment functions related to sentiment word's. The use of a sentiment-**conscious interest** mechanism for **adverse** networks, these model for ADR detection can effectively incorporate sentiment features from social media texts into attention scores, thereby enhancing its robustness. As a result of the sentiment-**conscious interest** mechanism, analysts suggest that the models can concentrate on ADR mentioned, and competing mastering in addition to decorate its performance.

In 2023, B M Vinjit et al. introduced "handwritten character recognitions" (HCR) as an automatic process of recognizing characters drawn on paper/painted or printed material and converting them into digital content. Research on HCR has matured significantly to-date; however, further improvements in accuracy and efficiency may enhance the utility of these methods. In today's world Digitisation of hand written documents is very useful as the data available at anywhere and any-point of time. Text digits can be made available to use for commerce and it's eco-friendly compared with manual text .Currently, all handwritten character recognition systems require human intervention because the process is semi-automatic. It has not been possible to develop an automated system for HCR up to this point. Researchers Shadnaz Asgari and colleagues [9] created novel pattern recognition algorithms with high accuracy and/or time complexity that improve clinical outcomes by facilitating clinicians' efficient and more informed decisions. Particularly in stressful environments like intensive care units, where rapid clinical decisions are necessary, this is of vital importance. Recent

developments in pattern recognition methodologies in clinical decision systems are outlined in this special issue.

In [10] Najarian et al. introduced a technique for pattern identification and analysis from the complexity of multiple clinical data sources such as physiologic signals and images. It's crucial because doctors very frequently have to make fast judgments that are dependent on their surrounding environment and the doctor can't see everything going on inside the more complicated information." These reasons emphasize the desire for advances in the ability to analyze large amounts of patient data — such as new signal and image processing algorithms with the ability to provide recommendation/prediction systems for clinicians. For more than 10 years now, scientists have utilized numerous pattern recognition strategies on bio-signals/images (from biomedical information) to deliver unattended and computer-aided medical detection to therapy. The advancement of new pattern detection strategies and algorithms which exhibit superior performance (in terms of precision or time efficiency) has positively impacted the health care outcome with enabling physicians to take better informed decisions in the shorter duration. A continuing research area in developing predictive computable designs and identification of patterns techniques with performance and capability equal to or surpassing the sophistication of healthcare evaluation and tracking technologies are developing is the need for periodic updates on the state-of-the-art advance. Pattern recognition is the process of automatically identifying patterns in data using a machine or computer. It is a field of computer science that studies how to give computers the ability to "see" and "understand" the world around them. Pattern recognition algorithms are used to classify objects, recognize faces, read handwriting, and make predictions about future events.

The purpose is to

- Identify familiar patterns quickly and accurately with a pattern recognition system
- Unfamiliar objects can be recognized and classified
- Recognizing objects and shapes from different perspectives accurately
- Even partially hidden objects and patterns can be identified
- Pattern recognition is quick, easy, and automatic.

The proposed algorithm uses feature vectors and a greedy algorithm uses to generate key features for pattern recognition and pattern matching. The algorithm is flexible and can be applied to any given data set. The output of the pattern generation process is classified using a ML algorithm, such as support-vector machine. The classifier classifies both normal and pathological patterns by analysing data redundancy present in the data. The proper patterns are detected at the output of the proposed method, which can predict data duplication using data redundancy techniques.

3 Proposed Methodology

The proposed method can detect data redundancy in medical data stored in a database. The objective of this work is to predict feature extraction after the pre-processing, filtering of input

data and the database. Patterns are generated using the key features with the help of a pattern recognition algorithm. In the feature extraction process, the variation in features is arranged using the greedy algorithm, and the variation in features is taken as key attributes. The flexible pattern matching algorithm can accept the variation in feature vectors. The pattern recognition algorithm uses the random forest logic. In which M is the wide variety of feature vectors, N is the bigness of feature vector, and m is the key attributes based on the greedy algorithm.

Random Forest Logic in pattern matching

Following is a description of the random forest algorithm:

- N observations have been collected. A random sample of N observations with replacement will be taken.
- Feature or input variables M are present. At each node, m will be selected randomly from the total number of features, M. In the forest, nodes are split according to the best splits on these m variables.
- It is the forest's goal to grow each decision tree to its full potential.
- Using the combined predictions of all the trees in the forest, the forest will produce a prediction. (Most votes or the average)

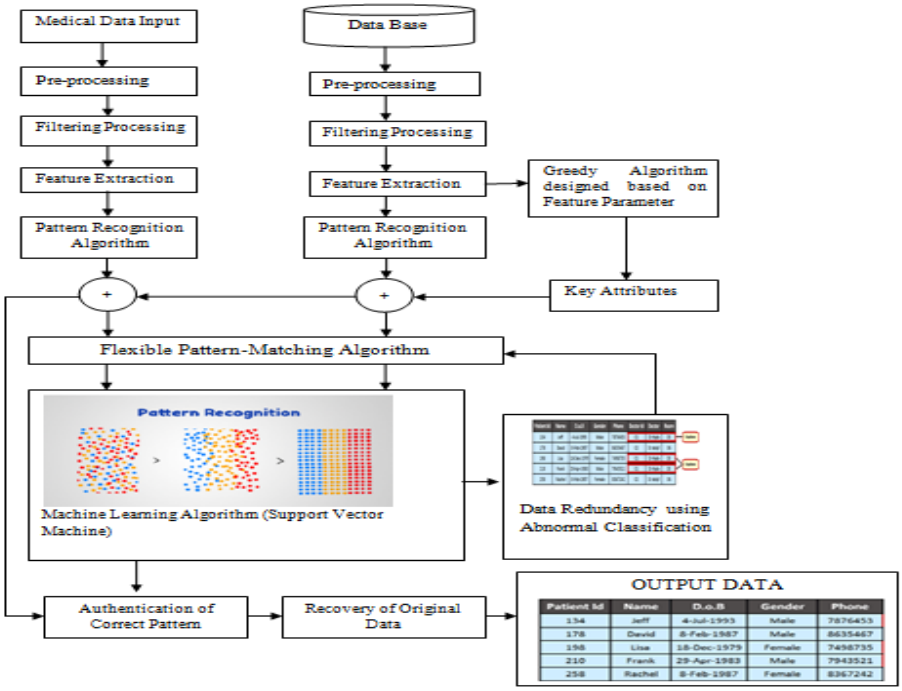


Fig. 1. Block Diagram for pattern recognition algorithm in medical data analysis

Utilizing recognition of patterns techniques, duplicate records are shown in figure 1 to be identified.

Medical Database: Database technology, which studies, oversees, and uses databases, is used in the medical field.

Pre-processing: data pre-processing is a phases inside the record mining's and statistics assessment process that converts raw records right into a shape that computer and ML algorithm can apprehend and scrutinise.

Filtering: Filtering is the process of changing an image's look. Applying special effects, including blurring, sharpening, or colour correction, can achieve this. An image's appearance can be changed with filters to make it seem more intriguing or realistic.

Feature Extraction: by wringing new feature from the contemporary ones (and then eliminating the authentic functions), feature extractions provoke to lessen the number of statistics in the datasets.

Flexible Pattern Matching Algorithm Dynamic Programming

Simple Matching program:

1) Assume that city 1 serves as both the beginning and the conclusion.

2) Produce every permutation of end (n-1).

3) Determine the estimate of all permutation and record the changeover with the least values.

4) Reinstate the vicissitude with the lowest viable price.

Time Complexities is indicated as Θ(n!).

Dynamic Programming

Assume the list of vertexes consist of 1, 2, 3, 4 and so on. Using 01 as input & the output starting and finishing points. We locate a low-price paths with 01 as the start point, i as final points, and all vertex occurring exactly once for every other vertex I (except from 0). If the cost of this path were to be (i), the cost of the matching cycle would be (i) + dist (i, 1), where (i, 1) is distance between (I, 1) and (i, 1) is the cost of the cycle. The lowest value of all [cost(i) + dist(i, 1)] values is then returned. At first glance, this seems easy.

How can I obtain cost(i) at this point? Dynamic programming requires some sort of recursion link in terms of smaller issues in order to determine the cost(i). Let's define the word C(S, i) as the cost of the cheapest path, starting at 1 and finishing at i, perusing each vertex in set S precisely once. Starting escorted by subsets of size 2, we compute C(S, i) for all subsets in which S is the subset. From there, we compute C(S, i) for all subsets in which S is the subset of size 3, and so on. Take note that each subset must include the value 1.

If size of S is 2, then S must be {1, i},

C(S, i) = dist(1, i)

Else if size of S is greater than 2.

C(S, i) = min { C(S-{i}, j) + dis(j, i)} where j belongs to S, j != i and j != 1.

Support vector Machine (SVM)

SVM uses linear decision boundary to classify the info into multiples classes. These methods are best suited for problems where data can be linearly separable, i.e., if there exists a hyperplane which divides all data points into two separate groups (in two dimensions) or more (in higher dimensions). The classifier chooses the decision boundary to the maximum margin atwix the classes, meaning class points are taken most apart from decision boundaries [11].

SVMs can be used to predict if cancer is benign or malignant. This can be done by training an SVM model on historical data about patients diagnosed with cancer. The model will learn to distinguish between malignant and benign cases based on the independent attributes of the data, such as the patient's age, tumor size, and tumor grade.

Steps

• Open sklearn.datasets and load the breast cancer dataset.

• Keep goal variables and input characteristics apart.

• Create and refine the RBF kernel-based SVM classifiers.

Gaussian RBFK(w,x)=exp(-γ‖x_i-x_j ‖^n) (1)

• Identify the input feature scatter plot.

• Identify a decision boundary on a map.

• Identify a decision border.

Authentication of correct pattern

I&A (identification and authentication) establishes identity using a protocol. I&A serves as the foundation for both authorization and accessing since it offers responsibility. The system may offer a proof of authentication after identification has been confirmed in order to prevent repeated authentications.

Data redundancy over duplicate data

Duplicate data can cause errors, inconsistencies, and inaccurate reporting of care to spread. In order to properly assess innovations based on clinical narratives, techniques to quantify information redundancy are crucial. This study examines redundant data in EHR notes quantitatively [12].

Performance metrices:

The choice of mattress depends on the specific task that are attested such as disease diagnosis patient risk assessment or text analysis. Always the clinical significance and the consequences of falls positive and false negative or considered when selecting the most appropriate matrices for the application.

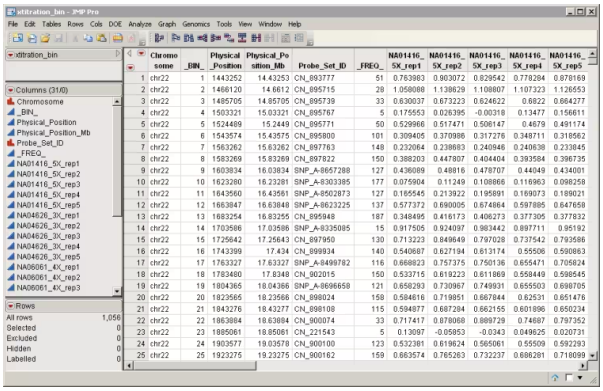
For disease diagnosis: accuracy measures the proportion of correctly classified instance and is suitable when the classes are balanced. However, in imbalanced data set it may not be a reliable metric. Precision is defined as the proportion of two correct predictions irrespective of all correct predictions. It is essential when minimising falls positive is critical e.g, in disease diagnosis. Sensitivity proceeding the fraction of two +ve forecasts out of all real +ve intenses. It is critical as reducing false negatives is an objective, such as identifying critical medical conditions.

For regression tasks (e.g, patient risk assessment): mean-absolute-error calculates the mean absolute variance among predicted and observed values. It's interpretable and gives equal weight to all errors. Mean square error computes the squared variance among Predator and the true values. It enhances the impact of greater blunders which may be important in some medical decision support scenarios.R^2 measures the fraction of variance explained by the variable of interest. by the model. It is a good indicator of how well the model fits the data.

Estimating redundancy in clinical text: Precision, recall, score: these mattresses can be adapted for text classification task, such as identifying redundant clinical text. Precision is the portion of right positives; recall is the portion of true positives; and F1 scores combine between accuracy and recollection

4 Result Analysis

As indicated in figure 2, the input dataset was examined using the Python 3.11.2 programme.



	Chromosome	_BIN_	Physical_Position	Physical_Position_Mb	Probe_Set_ID	_FREQ_	NA01416_5X_rep1	NA01416_5X_rep2	NA01416_5X_rep3	NA01416_5X_rep4	NA01416_5X_rep5
1	chr22	1	1443252	14.43253	CN_893777	51	0.763983	0.993072	0.829542	0.778284	0.878169
2	chr22	2	1466120	14.6612	CN_895715	28	1.058088	1.138629	1.108807	1.107323	1.126553
3	chr22	3	1485705	14.85705	CN_895739	33	0.630037	0.673223	0.624622	0.6822	0.664277
4	chr22	4	1503321	15.03321	CN_895767	5	0.175553	0.026395	-0.00318	0.13477	0.156611
5	chr22	5	1524489	15.2449	CN_895771	50	0.529966	0.517471	0.506147	0.4679	0.491174
6	chr22	6	1543574	15.43575	CN_895800	101	0.309405	0.370986	0.317276	0.348711	0.318562
7	chr22	7	1563262	15.63262	CN_897763	149	0.232064	0.238693	0.240946	0.240639	0.233845
8	chr22	8	1583269	15.83269	CN_897822	150	0.388203	0.447807	0.404404	0.393584	0.396735
9	chr22	9	1603834	16.03834	SNP_A-8657288	127	0.436089	0.48816	0.478707	0.44049	0.434001
10	chr22	10	1623280	16.23281	SNP_A-8303385	177	0.075904	0.11249	0.108866	0.116963	0.098258
11	chr22	11	1643560	16.43561	SNP_A-8502873	127	0.165545	0.213922	0.195891	0.169073	0.189021
12	chr22	12	1663847	16.63848	SNP_A-8623225	137	0.577372	0.690005	0.674864	0.597885	0.647658
13	chr22	13	1683254	16.83255	CN_895948	187	0.348495	0.416173	0.406273	0.377305	0.377832
14	chr22	14	1703586	17.03586	SNP_A-8335085	15	0.917505	0.924097	0.983442	0.897711	0.95192
15	chr22	15	1725642	17.25643	CN_897950	130	0.713223	0.849649	0.797028	0.737542	0.793586
16	chr22	16	1743399	17.434	CN_899934	140	0.540687	0.627194	0.613174	0.55508	0.590863
17	chr22	17	1763327	17.63327	SNP_A-8499782	116	0.668023	0.757375	0.750136	0.655471	0.705824
18	chr22	18	1783480	17.8348	CN_902015	150	0.533715	0.619223	0.611869	0.559449	0.596545
19	chr22	19	1804365	18.04366	SNP_A-8696658	121	0.656293	0.730987	0.749931	0.655503	0.698705
20	chr22	20	1823565	18.23566	CN_898024	159	0.584616	0.719851	0.667844	0.62531	0.651478
21	chr22	21	1843276	18.43277	CN_898108	115	0.594877	0.687284	0.662155	0.601896	0.650234
22	chr22	22	1863894	18.63894	CN_900074	33	0.717417	0.878068	0.889729	0.74687	0.797352
23	chr22	23	1885061	18.85061	CN_221543	5	0.13097	-0.05853	-0.0343	0.049625	0.020731
24	chr22	24	1903577	19.03578	CN_900100	123	0.532381	0.619624	0.565061	0.55509	0.592393
25	chr22	25	1923275	19.23275	CN_900162	159	0.663574	0.765263	0.732237	0.686281	0.718099

Fig. 2. Input Dataset or patient details

Applying the pattern recognition technique shown in figure 3 below to the initial processing dataset.

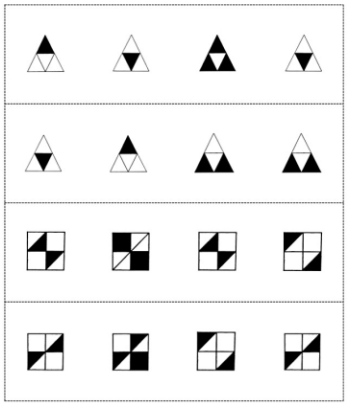


Fig. 3. Pattern images

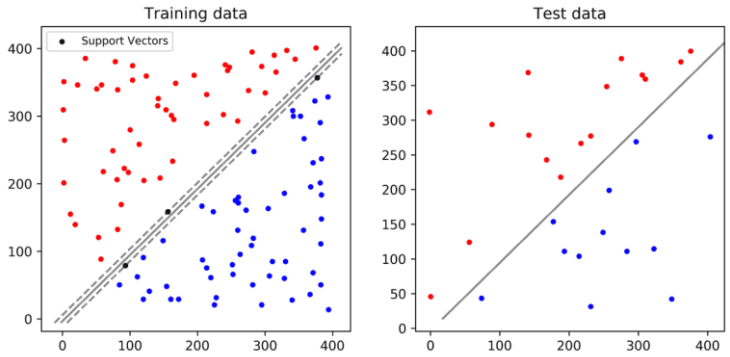


Fig. 4. Support vector machine classification of Input Dataset

5 Conclusion

In this work, we propose an inventive discrimination metric to choose the best possible features for pattern classifiers. An important matter to deal with when choosing numerous features in pattern classification is the redundancy among them. This is due to various feature types containing correlated or duplicate information methods accurately estimates the mutual information's medially chosen and candidate contours together with the class variables. This is achieved by approximating the conditional mutual information of the features without requiring the arduous regularization steps taken by conventional methods. We compared the results with classical algorithms and found experimental evidence to validate the success of our proposed approach.

References

[1] Hongxing Huo , Xuemei Sun , Yang , Xiang Wan , Yi Guan , Jingchi Jiang , Xitong Guo , "Clinical decision-making framework against over-testing based on modeling implicit evaluation criteria", Journal in Biomedical Informatics 119,103823, doi.org: /10.1016/, 2021

[2] M. T. Bahadori, A. Schuetz, W. F. Stewart, J. Sun, E. Choi, "Doctor ai: Predicting clinical events via recurrent neural networks", Machine learning for healthcare conference PMLR , (301-318) ,2016

[3] D. Spiegelhalter ,J. Wyatt, "Field trials of medical decision-aids: potential problems and solutions, in: Proceedings of the annual symposium on computer application in medical care", American Medical Informatics Association, [3], 1991

[4] Zina Ibrahim , James Teo b, Thomas Searle, Richard Dobson ,"Estimating redundancy in clinical text" , Journal of Biomedical Informatics (124), (103938), 2021, doi.org/10.1016

[5] Juan M. García-G´omez ,Lucía A. Carrasco-Ribelles, Salvador Tortajada c, Jose Ram´on Pardo-Mas a, Carlos S´aez a, Bernardo Valdivieso b, "Predicting morbidity by local similarities in multi-scale patient trajectories", Journal of Biomedical Informatics (120), (2021), doi.org:10.1016/j.jbi.

[6] Liang Yang , Bo Xu , Hongfei Lin , Tongxuan Zhang, Jian Wang , Xiaodong Duan. ,"Adversarial neural network with sentiment-aware attention for detecting adverse drug reactions" ,Journal of Biomedical Informatics (123), (103896) (2021).
[7] B M Vinjit, Sujit Kumar, Gitanjali Chalak, Mohit Kumar Bhojak,"A Review on Handwritten Character Recognition Methods and Techniques" , International Conference on Communication and Signal Processing, July (28 – 30), 2020 IEEE
[8] Fabien Scalzo, Shadnaz Asgari , Magdalena Kasprowicz,"Pattern Recognition in Medical Decision Support. Hindawi, BioMed Research International" , Volume 2019, Article ID 6048748, doi.org:10.1155/2019/6048748
[9] K. Najarian, S. Shirani and K. R. Ward," Biomedical signal and image processing for clinical decision support systems", Computational and Mathematical Methods in Medicine, Article ID (974592), 2013
[10] Kevin R.Ward ,Kayvan Najaria , Shahram Shirani " Biomedical Signal and Image Processing for Clinical Decision Support Systems", Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine , Volume [2015], Article ID(974592), .doi.org: 10.1155/2015/974592
[11] S. Ladhake, A. Khodaskar, "Pattern Recognition: Advanced Development, Techniques and Application for Image Retrieval", IEEE, International Conference on Communication and Network Technologies (ICCNT) ,2014
[12] TaeChoi,Vitaly Kober, Pablo Aguilar-González ,Victor Diaz-Ram-reznd," Pattern Recognition: Recent Advances and Applications" ,Hindawi Mathematical Problems in Engineering Volume 2018, Article ID [8510319], doi.org: /10.1155/2018/8510319.