The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395

# Hybrid Feature Extraction and Selection Techniques for Drugs Classifier System Based on Machine Learning

**Anuja Gaikwad**
Asst. Prof.: Atharva College of Engineering
anujahodage@gmail.com
**Nidhi Bhavsar**
Asst. prof. : Atharva College of Engineering
nidhibhavsar2861988@gmail.com
**Ashwini Gaikwad**
Asst. Prof.: Atharva College of Engineering
ashupict@gmail.com

**Abstract**: Nowadays, there are thousands of approved drugs that can be used for treating people who have medical problems. Therefore, drug warnings and precautions are denoted to recognize a discrete set of adverse effects and other implied protection uncertainties that are useful for patient control. Methods/analysis/findings: In this study, the intended framework is divided into two principal stages: data retrieval and data processing. Firstly, in the data collection stage, drug reports, drug interactions, malfunctions, number of deaths, and other factors had been obtained from various references, including RxNorm and Drug Bank using web service. Secondly, in the data processing phase, different data mining algorithms used to classify drugs into suitable drugs and non-suitable drugs. Application/improvements: According to the experimental results, we found that the decision tree has more accuracy (97.9%) than other state-of-art methods.
**Keywords**: Drug Interactions, Drugs Classification, Naïve Bayes, Support Vector Machine, Decision Tree.

## Introduction

Usually, several data mining methods have been utilized in healthcare, such as classification, association, analysis, clustering, and regression, as shown in Figure 1. A short explanation of each one of them is presented next.

### 1.1. Classification Techniques

Data units are separated into new categories thanks to classification. Several data points have their target class predicted by the classification process. Patients, based on their sickness model, may be categorised as "great danger" or "normal" utilising data organisation strategies. After establishing social stratification, a supervised training technique is carried out. Combinatorial and binary.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395

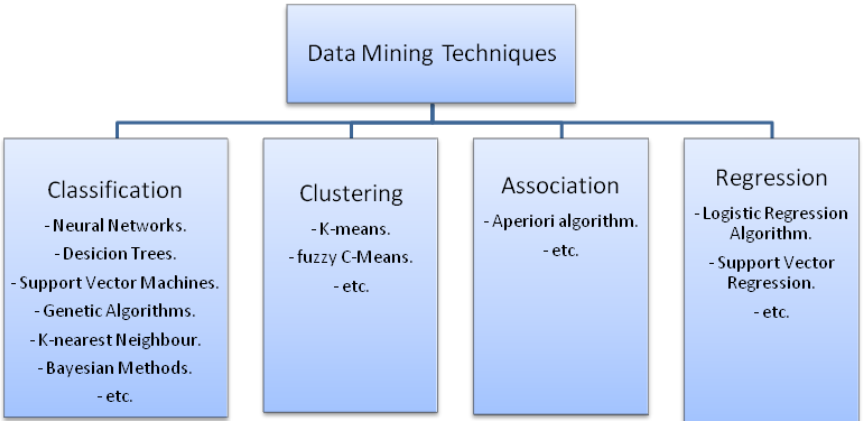


**Figure 1 : Different techniques used in the management of patient care.**

There are the two different ways that categories may be organised. The multiclass technique serves more than two goals, such as distinguishing between a "large," "moderate," and "fading" danger prisoner [1-3], whereas the binary arrangement can only evaluate two circumstances, such as "true" or "false" danger inmate. There are two phases in classification. A database's training records may be better understood after an early phase of structure design. The next stage is to devise a mechanism for classifying data using the built model. Effectiveness is measured by the percentage of test units or test datasets that are correctly categorised [4, 5]. Some of the many techniques used in healthcare management to achieve this state of unified care are as follows: Methods such as the J-48, SVM, K-nearest neighbour, neural networks, Bayesian approaches, etc.

### 1.1.1. Decision Tree Algorithm

The process of decision trees is one that is used rather often within the realm of data mining. The strategy consists of producing a set of rules that, when applied to a collection of input data, can accurately predict a specific variable related to the question being asked. Vertices and edges make up the components of a Decision Tree. The edges communicate a path or a choice that leads to the next set of vertices, maybe a pendant vertex (a pendant vertex is a set of vertices from which there are no more edges to go), which may characterise the next query or statement. The C4.5 algorithm is implemented in J48, which is a public source Java library. The C4.5 technique is an extension of the ID3 algorithm and is used to initialise a decision tree so that it may be used for classification. It is used to initialise a decision tree when the ID3 algorithm is used.

### 1.1.2. Naive Bayesian Algorithm

The process of data mining makes use of Bayesian classification, which may predict the likelihood of a class being associated with the data. Machine learning makes extensive use of Bayesian classification, which is founded on Bayes's theorem and is routinely utilised in a variety of contexts. There are several different applications of the Bayesian classification system, the most common of which being the Naive-Bayes model.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395

### 1.1.3. Drug Interactions

Interactions between medications are known as drug–drug interactions (DDIs), and they may have a major impact on how well an individual is protected against sickness [6]. A drug-drug interaction (DDI) occurs when one drug has an effect on another. The identification of DDI is necessary for the security of inmates as well as the efficient management of fitness programmes [7]. DDIs may be broken down into three primary categories: those with no contact, those with influence, and those with guidance [8]. On the other hand, the side effects of medicine are a contributor to the necessary kinds of morbidity and destruction in the United States, taking into consideration the over 700,000 calls made to crisis hotlines and the 120,000 people who are incarcerated each year. It has been suggested that adverse drug interactions (ADIs) may be latent causes of morbidity and sickness, in addition to contributing to increased pharmaceutical expenses and instances of carelessness [9]. We are encouraged to employ Data Mining methods to infer DDIs, major adverse reactions, and clinically important responses linked with medications since these approaches identify and infer hidden patterns from massive volumes of data in a variety of sectors, including the medical profession, and this allows us to find previously hidden patterns.

Although there is a high estimate of medication datasets and semi-structured sources (e.g. Stockley [10–11]) with knowledge about DDIs, these datasets are insufficient, and the percentage of their content is constrained, making it difficult to identify novel clinical consequences to every interaction. This is because there are a large number of medication datasets and semi-structured sources (e.g. Stockley [10–11]) with knowledge about DDIs.

The primary challenge that is addressed in this article is recommending medications and determining which medications are more effective than others. Because of this, appropriate pharmaceutical recommendations have to be made for the offenders. However, doctors should be able to categorise pharmaceuticals based on their knowledge of drug specifics such as side effects, patient complaints, drug warnings, and drug precautions. This is a challenging undertaking since there are so many medications available.

The major contribution that this study has made may essentially be summed up in two different aspects. To begin, a great number of pharmaceuticals contain substantial adverse reactions, warnings, precautions, and other elements that, among other things, have the potential to endanger human lives or create severe medical issues. When it comes to the process of prescribing pharmaceuticals, it is essential for doctors to have a comprehensive understanding of the many precautions and warnings that are linked with each medication, so that they can place patients in the appropriate category. Second, despite the evident significance of drugs in both the process of prescription medications and the treatment of patients, there is not yet a single comprehensive source of information on drug risks and warnings.

The following is the order in which the remaining sections of this text are presented: In the second part of this study, a review of the outstanding efforts that have been made to analyse and coordinate the drug reports is presented. The suggested system is presented in Part 3, along with a full description of each step that went into the creation of the recommended method. The implementation practise and assessment that are described in the recommended

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395
method are outlined and illustrated in Part 4. The conclusion will be presented in the fifth and final installment.

## Review Of Literature

### 2.1. Drug to Drug Extraction and Classification Approaches

In the next subsection, we will discuss some of the research projects that have been carried out in the area of DDI extraction and categorization. According to Reference [12], the purpose of the study was to gather the dispersed drug information that may be found on the internet inside many databases that may have insufficient information on drug counselling.

As a result, the purpose of this effort is to integrate a variety of drug resources in order to develop an ontology for drug interactions that will include information regarding harmful drug reactions as well as drug precautions, side effects, and uses.

The authors introduced a novel kernel-based features technique to extract and analyse drug interactions that are published in the biological literature in the reference [13], which can be found here. Their technique, like that of many other works that came before it, consists of two phases. After finding pairings of drugs that are known to interact with one another, scientists put each extracted pair into one of the four categories of drug-drug interactions. After that, they utilised a binary classifier (the LIBSVM classifier, which is used with the RBF kernel), in order to find drug pairings that interact with one another. Their method received a score of 71.14% on the DDIExtraction 2013 challenge corpus during the evaluation that was conducted on it.

The authors of Reference [14] studied the potential of coupling multiple machine-learning approaches to generate DDI. This included (i) a feature-based approach adopting an SVM with a collection of characteristics extracted from texts, and (ii) a kernel-based approach mixing three distinct kernels. Both of these methods were used to create DDI. Our technique was used to the DDIExtraction2011 challenge corpus, and the results indicate that it is useful for choosing DDIs with F1 scores of 0.6398 or above.

The authors of Ref. [15] incorporated and supported a method to pick DDI for drug specific combinations that were seen in biological papers. This method is described. This strategy places a significant emphasis on extensive syntactic parsing as the primary method for representing the connections between pharmacological remarks.

They evaluated the compatibility of text-based and database-obtained features for DDI discovery as part of the explanation of the DDI extraction procedure. When it came to machine learning, they investigated both the SVM and RLS techniques, focusing their studies on determining the components and training strategy that were most effective. The implementation of their plan resulted in a score of 62.99% F for the DDI Extraction 2011 assignment.

In the study referred to as Reference [16], the researchers produced a corpus of data that were approved by the Food and Drug Administration for inclusion on medication containers. These records were then manually analysed by a pharmacologist and a medicine information expert in order to determine pharmacokinetic DDIs.

Then, they estimated three different machine learning algorithms (SVM, and J48) for their

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395
experience to 1) recognise pharmacokinetic DDIs in the package insert corpus and 2) analyse pharmacokinetic DDI records by their modality (that is, whether they report a DDI or no interaction between medication pairs). They did this in order to 1) recognise pharmacokinetic DDIs in the package insert corpus and 2) analyse pharmacokinetic DDI records by their modality. [17].

### 2.2. Web Services Concepts

When developing the solution that was suggested, we relied primarily on online services to get the necessary domain information. A concise overview of the fundamental ideas behind web services may be found in the next subsection. There are two primary classifications that may be used to classify web services; these are the SOAP API and the REST API web services. The architectural style of this organisation that was employed in the configuration phase of the implementation.

SOAP is a method that is based on OOP and it sets a standard rule that is used when transferring information that is based on XML. In the process of establishing Web Services with machine interfaces, it is depicted as a protocol designation for transmitting structured data. The protocol offers a collection of controls for translating platform-specific data models into XML descriptions, whereas the designation describes a scenario for transporting information that is based on XML.

The phrase "Representational State Transfer" (REST) refers to a source-oriented method, and it is characterised by fielding in as a structural form that contains of a collection of scheme guidelines that establish the suitable behaviour for applying web patterns such as HTTP. REST also implies a fielding in as a structural form that includes of a collection of scheme guidelines that determine the appropriate behaviour. Despite the fact that REST is primarily presented within the context of the web, it is quickly becoming a common way of implementation for the generation of web services.

RESTful applications are built using Web models (URI, HTTP, and XML) in conjunction with REST sources. Connectivity, addressability, and statelessness are all components of REST policies. RESTful was implemented in order to figure out precise operations that were performed on URL sources.

### 2.3. Drug Databases

As a typical representation of contradicting findings, extensive examination purposes have been accomplished in order to realise DDI information. This is because of the important influence that DDIs have on the expense of prisoner health care and the protection of inmates. In this section, we will demonstrate that we have mastered the amazing aspects of the various fields. In recent years, there has been an explosion in the number of easily available datasets and semi-structured references that include drug reports and knowledge about DDIs. Two examples of these are Drug-Bank and RxNorm [18].

DrugBank is a comprehensive online database that includes broad pharmacological and biochemical data on medications, which justifies their treatment techniques and their goals. The database is organised in a way that makes it easy to search. It is generated, controlled,

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395

and developed via the expression of thorough research investigations by field-specialist trained curators.

RxNorm is a vocabulary that includes the normalised titles of therapeutic drugs. Its original purpose was to be used for the treatment of all recognised drugs in the United States. It is concerned with the actual component, dose, interactions, and strengths of a particular medication.

**Proposed System Design**

Figure 2 displays, for the purpose of this section, a block schematic of the system-based interactive tool that has been suggested. Following the recommended structure, we progressed through two steps that were dubbed "data retrieval" and "data processing." Two stages of the suggested system will be discussed in the following subsections.

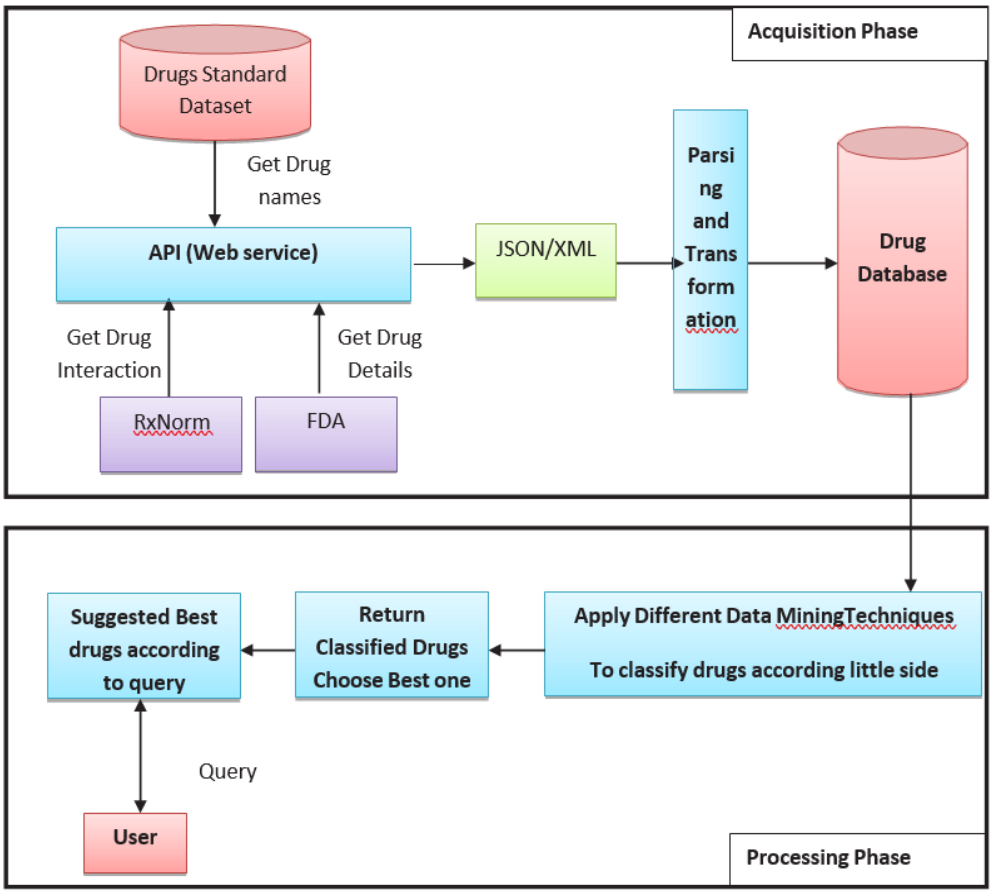


**Figure 2: The block structure of the proposed framework.**

are broken down into their component parts. The following statement, which has been formatted for better readability, is the definition of the primary two blocks in the proposed system. The first part of the proposed system is the acquisition of pharmaceuticals, and the second phase is the analysis of data.

In the first phase, known as the Acquisition Phase, information on the medicine is gathered from a variety of sources and then recorded in a relational database.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395

Processing Phase: During this phase, we want to create a classification module utilising data mining methods so that we may categorise pharmaceuticals based on information about drugs that was acquired from a variety of sites.

### 3.1. Drug information Collection

The reports on the drugs had been acquired from a variety of sources, such as the FDA, RxNorm, and the central drug dataset. The first step in the process of retrieving information about drugs is to take the titles of the drugs from the core standard dataset for drugs. After that, for a number of different drug titles, an HTTP request is sent to the RxNorm dataset, and the dataset is investigated by means of a web service in order to comprehend the various drug titles and drug interactions.

In addition, the FDA provides data on drug anticipations, warnings, conflicting effects, proof, and use. The FDA also includes material that is proposed by drug generators and suppliers regarding their stocks. It is vital that the labelling include a discussion of the important scientific information that is required for the effective and dependable use of the medicine. The open FDA pharmaceuticals stock labelling API gives data from this obedience for both prescribed and over-the-counter products. The information is also divided down into divisions, such as recommendations for use (prescription meds) or purpose, contradictory effects, and so on. HTTP Requests that are expressly addressed to the medication labelling endpoint utilising the URL.

### 3.2. Information Processing

As illustrated in Figure 2, the subsequent step in the process of producing the suggested interactive tool makes use of a variety of strategies for the categorization of appropriate medications. These strategies include Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM), in addition to algorithms.

### Results And Discussions

During the experimental study, we generate the matrices so that the system performance evaluation would be as accurate as possible. The computer has a 2.8 GHz Intel i3 processor and 4 gigabytes of random-access memory (RAM). The system was developed using an open-source Python architectural framework. Following the completion of the system's installation, a number of operating systems that are now in use were contrasted with the system that was presented. The specifics of both trials are described in further depth below;

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 1386 - 1395

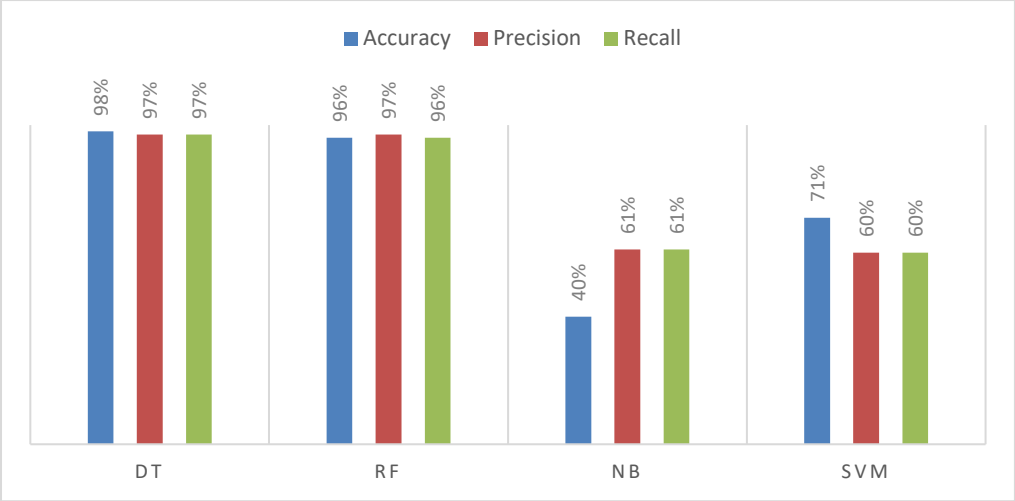


**Figure 3: detection accuracy of proposed model**

We used the aforementioned tools and the decision tree approach on our dataset to perform this experiment. Our 97.9% accuracy in patient classification is based on 455 out of 468 data. The random forest test (RF test) is an experiment where we use the aforementioned tools to apply the random forest approach to our dataset. Our 96.2% accuracy in patient classification is based on 453 out of 468 data. In SVM, we experiment with the aforementioned tools and apply the SVM technique to our dataset. With 61.5% accuracy, 288 out of 468 patient data have been accurately categorized.

## Conclusion And Future Work

In this study, we present an interactive framework that promotes the fitting to find a suitable and safe medication to the inmate before entering the patient clinical information and his or her history medication according to some circumstances such as drug interaction, the number of side effects, the number of deaths, and other similar factors. The application programming interface (API) for the online service was used to get drug databases from the Drug bank, the FDA, and RxNorm. In addition to that, we performed tests making use of a variety of data mining techniques. The decision tree has a recall percentage of 97%, an accuracy rate of 97%, and a precision rate of 98%. According to the findings of our research, this strategy is superior to the random forest, SVM, and Naive Bayes approaches.

## References

[1]. Pu L. e ToxPred: a machine learning-based approach to estimate the toxicity of drug candidates. BMC Pharmacology and Toxicology. 2019, 20(1), 2. https://bmcpharmacoltoxicol.biomedcentral. com/articles/10.1186/s40360-018-0282-6

[2]. Rezaee R. An evaluation of classification algorithms for prediction of drug interactions: Identification of the best algorithm. International Journal of Pharmaceutical Investigation. 2018, 8(2), 92–99. https://www.jpionline.org/index.php/ijpi/article/view/255

[3]. Vamathevan J. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery. 2019, 1. https://www.nature.com/articles/s41573-019-

0024-5

[4]. Tomar D, Agarwal S. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Biotechnology. 2013, 5(5), 241–266. http://dx.doi.org/10.14257/ ijbsbt.2013.5.5.25

[5]. Patel S, Patel H. Survey of data mining techniques used in the healthcare domain. International Journal of Information. 2016, 6(1/2), 53–60. http://aircconline.com/ijist/V6N2/6216ijist06.pdf

[6]. Ontology for drug-drug interactions. https://www.researchgate.net/publication/286834043_ An_ontology_for_drug-drug_interactions. Date accessed: 01/2013.

[7]. Chowdhury FM, Abacha AB, Lavelli A, Zweigenbaum P. Two different machine learning techniques for drug-drug interaction extraction. Challenge Task on Drug-drug Interaction Extraction. 2011, 19–26. https://www.semanticscholar.org/paper/Two-Different-Machine- Learning-Techniques-for-Chowdhury-Abacha/9998ab164023c2400c725d68d5971579 bbb19008

[8]. Automated extraction and classification of drug-drug interactions from text. https://www.researchgate.net/publication/278030155_Automated_Extraction_and_Classification_of_Drug-Drug_Interactions_from_Text. Date accessed: 01/2013.

[9]. Goldberg RM, Mabee J, Chan L, Wong S. Drug-drug and drug-disease interactions in the ED: analysis of a high-risk population. The American Journal of Emergency Medicine. 1996, 14(5), 447–450. DOI: 10.1016/S0735-6757(96)90147-3.

[10]. Stockley's drug interactions. https://about.medicinescomplete.com/publication/stockleys-drug-interactions/. Date accessed: 2010.

[11]. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A. Drug Bank 4.0: shedding new light on drug metabolism. Nucleic Acids Research. 2013, 42, 1091–1097. DOI: 10.1093/nar/gkt1068

[12]. Amer S, Mahmoud H, El-Shishtawy T. A methodology for building integrated drug interaction ontology. International Journal of Advancements in Computing Technology. 2018, 10(2), 1–9. http://www.globalcis.org/ijact/ppl/IJACT3620PPL.pdf

[13]. Raihani A, Laachfoubi N. Extracting drug-drug interaction from biomedical text using a feature-based kernel. Journal of Theoretical & Applied Information Technology. 2016, 109–120. http://www.jatit.org/volumes/Vol92No1/14Vol92No1.pdf

[14]. Chowdhury FM, Abacha AB, Lavelli A, Zweigenbaum P. Two different machine learning techniques for drug-drug interaction extraction. Conference: proceedings of DDI extraction. 2011.

[15]. Rohokale, M. S., Dhabliya, D., Sathish, T., Vijayan, V., & Senthilkumar, N. (2021). A novel two-step co-precipitation approach of CuS/NiMn2O4 heterostructured nanocatalyst for enhanced visible light driven photocatalytic activity via efficient photo-induced charge separation properties. Physica B: Condensed Matter, 610 doi:10.1016/j.physb.2021.412902

[16]. Björne J, Airola A, Pahikkala T, Salakoski T. Drug-drug interaction extraction from biomedical texts with SVM and RLS classifiers. DDI Extraction. 2016, 35–42.

https://pdfs.semanticscholar. org/d641/472d03c05dd1ae3b98acfc86e8e768711622.pdf

[17]. Boyce RD, Gardner G, Harkema H. Using SVM and decision tree to identify pharmacokinetic drug-drug interactions. Conference: proceedings of the workshop on biomedical natural language processing. 2018.

[18]. Sherje, N. P., Agrawal, S. A., Umbarkar, A. M., Dharme, A. M., & Dhabliya, D. (2021). Experimental evaluation of mechatronics based cushioning performance in hydraulic cylinder. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1021

[19]. Mumbaikar S, Padiya P. Web services based on SOAP and REST principles. International Journal of Scientific and Research Publications. 2013, 3(5), 1–4. http://www.ijsrp.org/research- paper-0513/ijsrp-p17115.pdf

[20]. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: Rx Norm at 6 years. Journal of the American Medical Informatics Association. 2011, 18(4), 441–448. DOI: 10.1136/amiajnl-2011-000116.

[21]. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. In: Pacific symposium on biocomputing. 2014, 410–421. https://www.ncbi.nlm.nih. gov/pubmed/22174296

[22]. Tatonetti NP, Denny JC, Murphy SN. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clinical Pharmacology & Therapeutics. 2011, 90(1), 133–142. DOI: 10.1038/clpt.2011.83.