Advanced Industrial SCIence
Vol. 3, No. 2, pp. 8-16, 2024

ISSN 2951-2476
https://doi.org/10.23153/AI-Science.2024.3.2.008

A Study on the Drug Classification Using Machine Learning Techniques

Anmol Kumar Singh[1], Ayush Kumar[1], Adya Singh[1], Akashika Anshum[1], Pradeep Kumar Mallick[2*]

[1]Student, School of Computer Engineering, Kalinga Institute of Industrial Technology, India

[2*]Senior Associate Professor, Kalinga Institute of Industrial Technology, India

머신러닝 기법을 이용한 약물 분류 방법 연구

Anmol Kumar Singh[1], Ayush Kumar[1], Adya Singh[1], Akashika Anshum[1], Pradeep Kumar Mallick[2*]

[1]인도 부바네스와르대학교 컴퓨터공학부 학생, [2*]인도 부바네스와르대학교 컴퓨터공학부 교수

Abstract : This paper shows the system of drug classification, the goal of this is to foretell the apt drug for the patients based on their demographic and physiological traits. The dataset consists of various attributes like Age, Sex, BP (Blood Pressure), Cholesterol Level, and Na_to_K (Sodium to Potassium ratio), with the objective to determine the kind of drug being given. The models used in this paper are K-Nearest Neighbors (KNN), Logistic Regression and Random Forest. Further to fine-tune hyper parameters using 5-fold cross-validation, GridSearchCV was used and each model was trained and tested on the dataset. To assess the performance of each model both with and without hyper parameter tuning evaluation metrics like accuracy, confusion matrices, and classification reports were used and the accuracy of the models without GridSearchCV was 0.7, 0.875, 0.975 and with GridSearchCV was 0.75, 1.0, 0.975. According to GridSearchCV Logistic Regression is the most suitable model for drug classification among the three-model used followed by the K-Nearest Neighbors. Also, Na_to_K is an essential feature in predicting the outcome.

Keywords : Drug Classification, Data Preprocessing, Label Encoding, Hyperparameter Tuning, GridSearchCV.

요 약 본 논문에서는 인구통계학적, 생리학적 특성을 기반으로 환자에게 가장 적합한 약물을 예측하는 것을 목표로 하는 약물 분류 시스템을 제시한다. 데이터 세트에는 적절한 약물을 결정하기 위한 목적으로 연령, 성별, 혈압(BP), 콜레스테롤 수치, 나트륨 대 칼륨 비율(Na_to_K)과 같은 속성들이 포함된다. 본 연구에 사용된 모델은 KNN(K-Nearest Neighbors), 로지스틱 회귀 분석 및 Random Forest이다. 하이퍼파라미터를 최적화하기 위해 5겹 교차 검증을 갖춘 GridSearchCV를 활용하였으며, 각 모델은 데이터 세트에서 훈련 및 테스트 되었다. 초매개변수 조정 유무에 관계없이 각 모델의 성능은 정확도, 혼동 행렬, 분류 보고서와 같은 지표를 사용하여 평가되었다. GridSearchCV를 적용하지 않은 모델의 정확도는 0.7, 0.875, 0.975인 반면, GridSearchCV를 적용한 모델의 정확도는 0.75, 1.0, 0.975로 나타났다. GridSearchCV는 로지스틱 회귀 분석을 세 가지 모델 중 약물 분류에 가장 효과적인 모델로 식별했으며, K-Nearest Neighbors가 그 뒤를 이었고 Na_to_K 비율은 결과를 예측하는 데 중요한 특징인 것으로 밝혀졌다.

주제어 : 약물 분류, 데이터 프로세싱, 라벨 인코딩, 하이퍼파라미터 튜닝, GridSearchCV

[*]Corresponding Author : Pradeep Kumar Mallick(pradeepmallick84@gmail.com)

Received May 23, 2024
Revised June 3, 2024
Accepted June 20, 2024
Published June 30, 2024

1. Introduction

The drug management and classification in today's healthcare are not only vital but rather they directly affect treatment outcomes. Drug management and classification in today's modern health care play an indispensable role in patient care. Accuracy for classification of drugs in health care directly affects treatment outcomes. Along with machine learning (ML) techniques, there occurs a chance to renew customs, the drug classification is based on using prescience models creation and big data analysis [1,2]. The study is focused on ML applications in drug classification with the intention of developing better and faster systems of classification that are actually used in various medical centers that are being established these days[3].

The growing numbers of patient's data along with the fact that treatment decision-making has reached a higher level of complexity have elevated the need for enhanced computational tools, for instance, drug classification. The classic classification mechanism which utilizes standard rules together are based on manual assessment found to be not scalable, adaptable and accurate enough. An inaccurate classification or adverse drug reactions, delay in treatment, and, potentially, further concerns for patient safety can occur as a result of diagnostic errors or harm. Hence, there is an argument to be made about utilizing the emerging technologies to enhance drug classification methods which have been proactive in recognition of the ever-evolving nature of drug abuse[4].

Through the deployment of ML algorithms, scientists and medical professionals can make good use of big data of patients which can build the models of classifying drugs based on the patient's individual profile. Marks e.g. age, sex, your medical history, and biochemical index are values that the analyst incorporates into the model allowing for personal medication classification according to the patient's unique needs. Furthermore, ML techniques offer the opportunity to continuously adapt and improve in accordance with the individual characteristics as well as with the rapid developments in medical knowledge, which is a way to impede the dynamic nature of healthcare data and consequent treatment protocols[5].

Through this study, ML-based drug classification, the theoretical basis, and the methodologies, as well as experimental findings will be examined. To address this, we begin by analysing the current status evidenced in the scientific literature and then focus on the emerging challenges and as well as the future scenario for drug classification. The paper attempts to counter the above arguments by providing its own perspectives based on the analysis of different sources and thus, aims at contributing to the now ongoing discussion on the usage of ML in healthcare and suggesting important implications for further research and practice[6].

2. Literature Review

Healthcare growth has generated extensive crucial information on drugs and compelled explaining the classifying system to be dynamic. In the current times, the employment of ML approaches in healthcare has become immensely essential owing to its exponential potential to shift and grade the medical cases effectively. Medicine afore mentioned context, the applications of ML technologies to drug classification as an encouraging way of improving patient health outcomes and healthcare processes enhancements are being tried as[1,4].

Beforehand, the scholarly research has indicated that the conventional drug grouping methods can have some restrictions. The doctors' evaluations (in accordance with these standards) are used in such cases and some errors can occur. Such as, existing systems may end up struggling to effectively manage the complexity of obtaining and deciphering patient data from all the digital technologies nowadays. In turn, this can lead to incorrect treatment decisions and higher risk for adverse drug reactions (ADRs). The National Institute of Health (NIH) gave frightening figures that indicated the death of many people annually due to the error in medication, which demonstrates the importance of research that aims to help in improving drug classification so that these errors can be reduced[1-5].

Consequently, researchers have been directed towards creating ML methods to develop predictive models that are reliable at classifying the drugs according to patient-matched parameters. The core attributes of age, sex, blood pressure, cholesterol levels, and sodium to potassium ratio have been found to be the most essential factors in prepared to machine learning algorithms. Through big data and complex algorithms, ML-driven approaches provide the opportunity to enhance the accuracy and effectiveness while at the same time reducing the workload of the drug classification system with the detection of existing patterns in the data.

Heterogeneous ML algorithms have been experimented while doing the drug classification part and these include: K-nearest neighbors (KNN), Random Forest, Logistic Regression. Of course, each algorithm has its own advantages and disadvantages, and the researchers seek to find the way for particular healthcare area to use the best model. Besides, one has presented methods like GridSearchCV for example, which enable model fitting to adjust to best performance and increase prediction accuracy[7,8].

Although the application of ML-based drug one-classification exhibit is astonishing, the challenges and opportunities are by no means absent[9,10]. The next research phase might involve improving algorithms, enhancing the integration between data sources, and completing studies that were not accomplished before for reasons of privacy, security and methodology. In the end, more advanced ML techniques will lead to the disruptive pattern in drug classification concepts and, as result, community healthcare will be improved.

3. Methodology

3.1 Data set collection

The dataset used in this study is retrieved from Kaggle. The dataset contains 200 rows and 6 columns. Rows represent individual patients, and columns represents characteristics of each patient. Attributes of patients are in form of age, sex, blood-pressure level, cholesterol level and sodium to potassium ratio, drug. Drug is the target feature which is classified into 5 different types:

- drugA
- drugB
- drugC
- drugX
- drugY

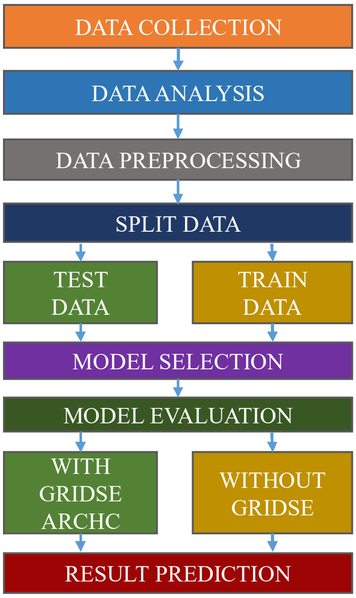

Fig. 1. Proposed Model in this research

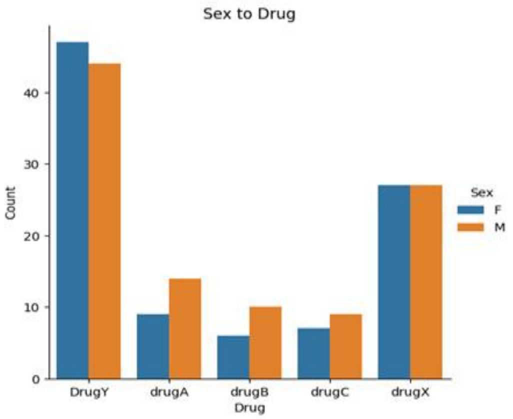

Fig. 2. Patients number per class

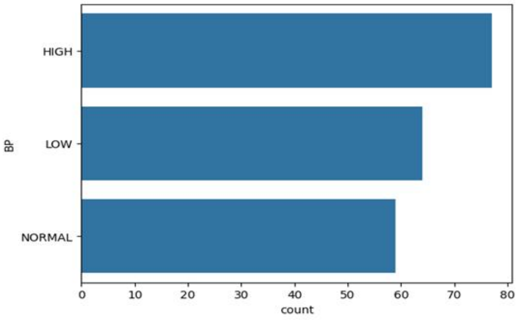

Fig. 3. BP feature Distribution

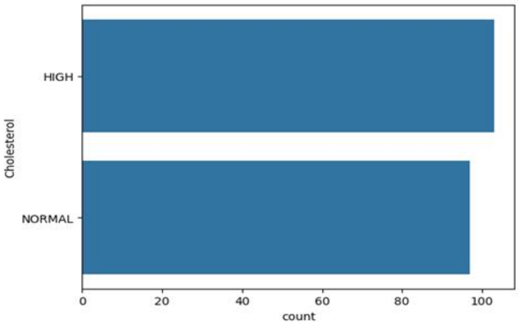

Fig. 4. Cholesterol Distribution

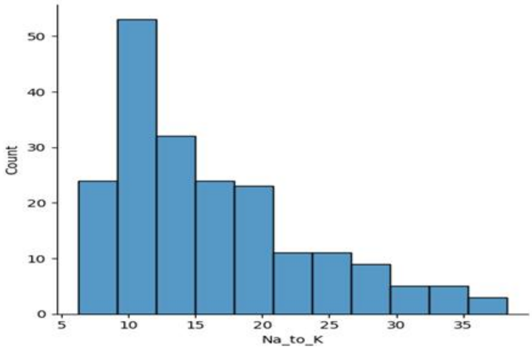

Fig. 5. Na to Potassium Ration Distribution

3.2. Feature Engineering and Label Encoding

To ensure the effective utilization of the Sodium-to-Potassium ratio (Na_to_K) feature in our analysis, we divided it into two classes: 1 and 0. Code 0 equals values less than 15 and code 1 equals values at least 15 and above. This categorization, then, serves the purpose of developing a more insightful underpinning for the subgroup effect on drugs.

Also, after that, label encoding occurs to all attributes column templates to prepare the dataset for analysis. Encoding in label form is a vital pre-processing step that transforms the data into a numerical format which supports the machine learning algorithms in their process of interpretation and learning. We do that by transforming categorical variables into different numbers, so that our analysis can catch different machine-learning models and make thought-processing easier.

In this set of data preprocessing steps, the systematic approach has been used to adapt them into a dataset that is ready for classification tasks. The outcome of this is the building of a robust and accurate predictive model.

3.3. Model Background

In the current practice, the ML algorithms received a lot of current interest due to the fact that it is capable of converting medical diagnosis, treatment plan, and care for patients entirely. K-Nearest neighbor algorithm, logistic regression, and random forest are not only some of the major algorithms used in machine learning competency but also in the bibliometric task of drug classification as shown in Table 1.

K-Nearest Neighbor (KNN)

The KNN (K-Nearest Neighbor) algorithm is a simple yet very potent tool that performs both classification and associate tasks. Basically, KNN operates on the basic of similarity-based reasoning, where the class of an unlabelled data point is known from the class labels of the most similarly appearing of the remaining data points in the feature space. KNN works with the assumption that instances with similar attributes usually (do) belong to the same class (being) is thus useful for drug classification tasks where the outcome of drug response varies non-linearly or in complex relationships between the patient and the drugs.

KNN: **y = mode(y_i)**, where y_i represents the class label of the k nearest neighbors.

Logistic Regression

The method of logistic regression is one the best statistical methods for binary prediction which are those which have the task of predicting the probability that a particular known event may occur upon the basis of 1 or more predictor dimensions. Contrary to its name, Logistic Regression is one of the classification algorithms which symbolize interrelation between a specific portion of the outcome variable and a high class. The application of Logistic Regression to drug classification includes developing the probability that a patient will be prescribed a particular medication based on his/her demographic as well as clinical characteristics.

Logistic Regression:

$$P(Y=1 \mid X)=1/(1+e^{-\Theta_t}X)$$

where Θ are the model parameters.

Random Forest

Random Forest is a technique that involves a group of decision trees. It aims to produce more efficient classification and regression than the single decision tree. The individual decision tree in Random Forest is trained by a random subset of the training set and takes a random subset of the features at each split resulting in uncorrelated and diversified trees. Through this greater number of trees, and their unique randomized decision trees, the Random Forest model will have its classification and regression performance enhanced through comparison, and become less prone to overfitting. Drug classification is among Random Forest is that it can effectively capture complicated relationships among patient's characters and drug response thus making it extremely versatile and strong algorithm (probably market the only algorithm) for predictive medication.

Random Forest: Ensembles of decision tree with majority voting average.

For improved accuracy and generalization, optimal parameter settings are done by enhancing the models using GridSearchCV for hyperparameter tuning. This study contributes to healthcare and pharmaceutical research by providing accurate and reliable drug classification solutions through careful selection and evaluation of these algorithms.

Table 1. Model Comparison

Model	Pros	Cons
KNN	Simple to implement and understand.	Sensitive to irrelevant features.
Logistic Regression	Provides probabilistic interpretation of result.	Assumes linear relationship between feature and target.
Random Forest	Handles non-linear relationships well.	Prone to overfitting with noisy data.

4. Result and Discussion

The result of this analysis show that K-Nearest Neighbor algorithm achieved more accuracy with grid than without grid, Logistic Regression achieved 100% accuracy.

The Table 2 shows the accuracy of all the models used without and with GridSearchCV:

Table 2. Model's Performance

Model	Accuracy (without GridSearchCV)	Accuracy (with GridSearchCV)
K-Nearest Neighbor	70%	75%
Logistic Regression	87.5%	100%
Random Forest	97.5%	97.5%

The algorithm Random Forest condemn accuracy in testing as a plus side and it scores well during testing with an accuracy of 0.975. Nevertheless, the precision of neural network is fluctuant during training session. It was initially 0.98375, then moved to 0.99375. This higher level of accuracy asserts the power and effectiveness of the algorithm, shows that it can learn from the training data and in case of new data created a generalization. By tying in stable accuracy in test data to higher performance in training, we summarize what the Random Forest model does well via accurate classification of drugs based in patient characteristics. The analysis from table 2 shows that grid can be used to achieve high accuracy in drug classification and the Logistic Regression and Random Forest could be a good choice for drug classification, as it showed the best performance in the study.

Table 3. Confusion Matrix (without GridSearchCV)

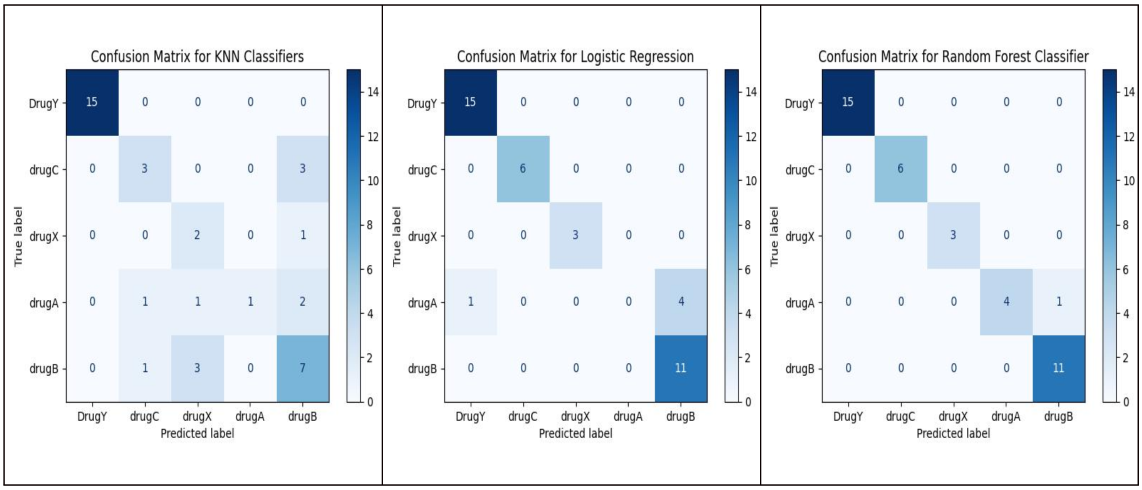

Table 4. Confusion Matrix (with GridSearchCV)

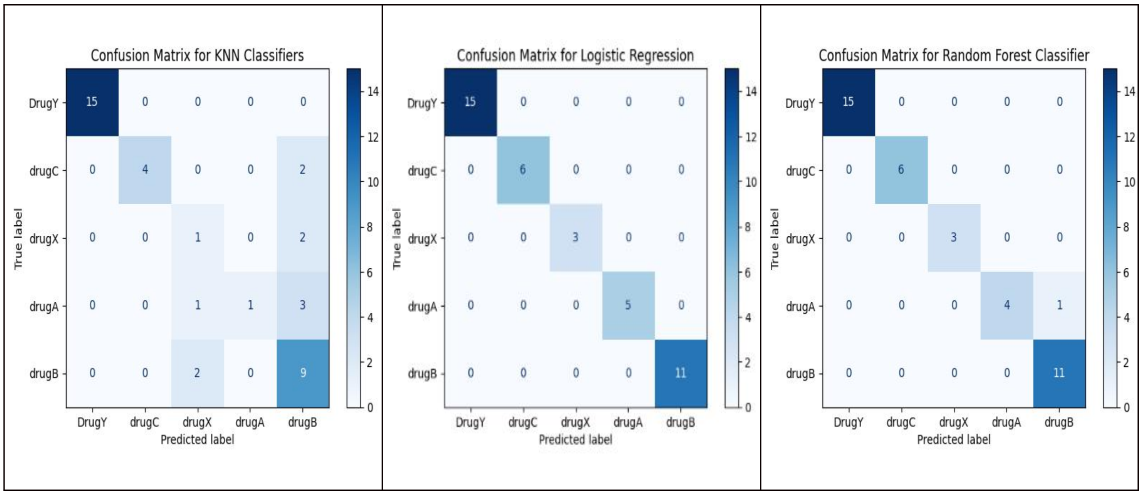

Table 5. Precision, recall and f1-score (without GridSearchCV)

MODEL	PRECISION		RECALL		F1-SCORE	
	Macro Average	Weighted Average	Macro Average	Weighted Average	Macro Average	Weighted Average
K-Nearest Neighbor	0.69	0.76	0.60	0.70	0.58	0.69
Logistic Regression	0.73	0.78	0.80	0.88	0.76	0.82
Random Forest	1.00	1.00	1.00	1.00	1.00	1.00

Table 6. Precision, recall and f1-score (with GridSearchCV)

MODEL	PRECISION		RECALL		F1-SCORE	
	Macro Average	Weighted Average	Macro Average	Weighted Average	Macro Average	Weighted Average
K-Nearest Neighbor	0.76	0.82	0.60	0.75	0.62	0.74
Logistic Regression	0.73	0.78	0.80	0.88	0.76	0.82
Random Forest	1.00	1.00	1.00	1.00	1.00	1.00

The researches on the model performance between without using GridSearchCV and with the help of GridSearchCV prove to us the optimizing of hyper parameter in classification accuracy is very effective as shown in Table 3 and 4. Whereas, Logistic Regression keeps consistency in delivering standard performance indicators, and this showcases how it is certainly consistent regarding predicting drug groups.

The fact that KNN has precise recall, better classifier F1-scores and over property have demonstrated that GridSearchCV does indeed perform the optimizing process of parameters required for better classification accuracy.

On the other hand, the unaltered efficiency of Random Forest with the default option points at the fact that they used the best starting point without any pinch for this dataset.

Empirically, the present study has established that the K-Nearest Neighbor method shows real signs of improvement when the appropriate tuning is done as shown in Table 5 and 6. We have come to the point when we need to choose the algorithm that matches the study requirements, for instance computational efficiency, interpretation and specificity of the study.

5. Limitation and Future Scope

5.1 Limitation

• A limitation of this study lies in the scope of the dataset, which primarily focuses on a limited number of drug types. Expanding the dataset to include a broader range of drug types could enhance the accuracy of the results by providing a more comprehensive representation of real-world drug classification scenarios.

• Additionally, while the algorithms were trained on a single dataset, evaluating them on multiple datasets would improve their generalizability across different patient populations and healthcare settings.

• Furthermore, the absence of evaluation on a holdout set presents a limitation, as it may affect the realistic estimate of the algorithm's accuracy in practical deployment scenarios.

5.2 Future Scope

• Integration of additional features:
The model's predictive accuracy could be improved by additional patient attributes or biomarkers. For enhancing drug classification performance, the new data sources or advanced feature engineering techniques may discover important insights.

• Deployment as a WEB application:
The usability of model and accessibility would be enhanced by a user-friendly web application in which healthcare professionals can input patient data and receive drug classification predictions in real-time. Further this process could be streamlined by an electronic health record systems integration.

• Continuous model monitoring and updates:
In evolving healthcare trends and practices, we can ensure the classification data remains updated and accurate by integrating a model performance monitoring system in production environments and periodically retraining the new data in the models.

· EXPANSION TO MULTI-CLASS CLASSIFICATION:
For multiple medications requirement in a patient, the utility in clinical settings can be enhanced by including a wider range of drug categories which leads the single class drug classification study to handle multi class drug classification study.

6. Conclusion

The objective of our study has been completed by a drug classification predictive model which measure many attributes of a patient like age, sex, BP, cholesterol levels and Na-to-K ratio. The efficiency and feasibility of various machine learning techniques for the drug classification chores is shown by rigorous data analysis, and it's preprocessing for building a model. The various algorithms like K-Nearest Neighbors (KNN), Logistic Regression and Random Forest, also hyperparameter tuning with GridSearchCV has been used to optimize the performance of the model. These models have achieved high accuracy levels with and without hyperparameter tuning is concluded by the estimation of the results and it also shows the efficiency of prediction of correct drug classes.

REFERENCES

[1] Gala, D. V., Gandhi, V. B., Gandhi, V. A., & Sawant, V. (2021, October). Drug classification using machine learning and interpretability. *In 2021 Smart Technologies, Communication and Robotics (STCR)* (pp. 1-8). IEEE.

[2] Mridha, K., Bappon, S. D., Sabuj, S. M., Sarker, T., & Ghosh, A. (2023, August). Explainable Machine Learning for Drug Classification. *In*

International Conference on Electrical and Electronics Engineering (pp. 673-683). Singapore: Springer Nature Singapore.
DOI : 10.1007/978-981-99-8661-3_48.

[3] Chen, C. (2024). Research on Drug Classification Using Machine Learning Model. *Highlights in Science, Engineering and Technology, 81,* 350-355.

[4] Gururaj, H. L. et al. (2021). Classification of drugs based on mechanism of action using machine learning techniques. *Discover Artificial Intelligence, 1(1),* 13.
DOI : 10.1007/s44163-021-00012-2

[5] Saad, A. I., Omar, Y. M., & Maghraby, F. A.(2019). Predicting drug interaction with adenosine receptors using machine learning and SMOTE techniques. *IEEE Access, 7,* 146953-146963.
DOI : 10.1109/ACCESS.2019.2946314

[6] Shobana, G., & Bushra, S. N. (2020, December). Drug administration route classification using machine learning models. *In 2020 3rd International Conference on Intelligent Sustainable Systems* (ICISS) (pp. 654-659). IEEE.
DOI : 10.1109/ICISS49785.2020.9315975

[7] Lee, S., Kim, S., Lee, J., Kim, J. Y., Song, M. H., & Lee, S. (2023). *Explainable Artificial Intelligence for Patient Safety: A Review of Application in Pharmacovigilance.* IEEE Access.
DOI : 10.1109/ACCESS.2023.3271635

[8] Ponzoni, I., Páez Prosper, J. A., & Campillo, N. E. (2023). Explainable artificial intelligence: A taxonomy and guidelines for its application to drug discovery. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 13(6),* e1681.
DOI : 10.1002/wcms.1681

[9] Puneeth, G. R. et al. (2021). Analysis of drug classification using mechanism of action. *J Phys Conf Ser., 1914(1),* 01204.
10.1088/1742-6596/1914/1/012034.

[10] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition, 40(7),* 2038-2048.
DOI : 10.1016/j.patcog.2006.12.019.

Ayush Kumar [Student Member]

- Aug. 2021 ~ Current : B.Tech students of School of Computer Engineering, Kalinga Institute of Industrial Technology Deemed to be University.
- Research Interests : IoT system, CNN, Machine Learning

Adya Singh [Student Member]

- Aug. 2021 ~ Current : B.Tech students of School of Computer Engineering, Kalinga Institute of Industrial Technology Deemed to be University.
- Research Interests : IoT system, CNN, Machine Learning

Akashika Anshum [Student Member]

- Aug. 2021 ~ Current : B.Tech students of School of Computer Engineering, Kalinga Institute of Industrial Technology Deemed to be University.
- Research Interests : IoT system, CNN, Machine Learning

Pradeep Kumar Mallick [Regular member]

- Feb. 2019 ~ Current : Senior Associate Professor, School of Computer Engineering, Kalinga Institute of Industrial technology (KIIT) Deemed to be University
- Jul. 2019 ~ May. 2020 : Post Doctoral Fellow, Kongju Nat'l University
- Dec. 2016 : Siksha Ó'Anusandhan University, Computer Science & Engineering(PhD)
- Research Interests : Machine Learning, Image Recognition, IoT, Big Data
- E-Mail : pradeepmallick84@gmail.com