

Leveraging Cloud Computing for Drug Review Analysis

Shiva Teja Pecheti¹*, Nithin Kodurupaka¹, Basavadeepthi H M¹, Talari Tanvi¹, and Dr. Beena B.M.¹

¹ Department of Computer Science & Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bangalore, Karnataka, India, 560035

Abstract

This research focuses on creating a Drug Information and Recommendation System using Amazon Web Services (AWS) for sentiment analysis of pharmaceutical reviews. The methodology encompasses data collection, preprocessing, sentiment analysis, and drug prediction. AWS services such as EC2, S3, and IAM are utilized to ensure compatibility, scalability, and security. The resultant platform, Sentiment AI, is deployed seamlessly on AWS, demonstrating efficient resource utilization and real-time monitoring through CloudWatch. This work signifies a significant step forward in pharmaceutical sentiment analysis, harnessing the power of cloud computing to provide reliable and effective insights into user perceptions of medications.

Keywords

Sentiment analysis, Amazon Web Services (AWS), Natural Language Processing (NLP), User-generated reviews, Cloud services, Pharmaceuticals.

1. Introduction

Pharmaceutical product awareness could be completely transformed in the ever-changing healthcare industry by combining state-of-the-art technology with insightful data analysis. The goal of this research is to create a sophisticated drug information and recommendation system by combining cloud computing's transformative capabilities with sentiment analysis and drug prediction. The primary goal is to provide consumers and healthcare professionals with an effective tool that can extract insightful information from an extensive collection of pharmaceutical reviews, eventually leading to better informed decision-making. Our approach is based on our belief that utilizing cloud computing, particularly via Amazon Web Services (AWS), can greatly improve the security, scalability, and effectiveness of sentiment analysis models used to analyze medication reviews. We propose that integrating cloud services presents a complete solution for retrieving vital pharmacological information, while also enhancing technical capabilities and streamlining deployment and management procedures.

Symposium on Computing & Intelligent Systems (SCI), May 10, 2024, New Delhi, INDIA

* Corresponding author.

† These authors contributed equally.

✉ shivatejapecheti@gmail.com (S. T. Pecheti); nithinkodurupaka@gmail.com (N. Kodurupaka); basavadeepthihm@gmail.com (B. H. M); tanvitalari2002@gmail.com (T. Tanvi); bm_beena@blr.amrita.edu (Dr. B. B.M.)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The methodology starts with careful collection of data from several web platforms, which prepares the background for a precise preprocessing stage. This essential phase consists of eliminating unnecessary words, tokenizing, changing text to lowercase, and cleaning HTML tags. primary objective is to improve the textual data's consistency and quality, which will provide an appropriate foundation for further analysis. Sentiment analysis and the drug prediction were the system's primary objectives. Sentiment analysis utilizes extensive natural language processing (NLP) techniques to identify sentiments in drug reviews and categorize them into neutral, positive, and negative categories. Additionally, the drug prediction module adds a layer of predictive insight to the system by classifying drugs according to characteristics taken from the preprocessed text data, generated by machine learning models. The outcomes of drug prediction and sentiment analysis are combined and stored in CSV files, and these are kept as essential parts of the medication Information and Recommendation System. Most importantly, we also investigate cloud computing, specifically focusing on AWS services. Restructuring the sentiment analysis approach makes it easily integrated with the AWS architecture. Since the parts are systematically arranged into modular modules, the deployment procedure can be accelerated. System efficiency is improved by optimizing data storage configurations utilizing Amazon Simple Storage Service (S3) together with customized settings for Amazon Elastic Compute Cloud (EC2) instances. In addition to these technological developments, the study highlights the significance of the user interface. This application, which is a representation of Sentiment AI, acts as an interface for users to engage with the abundance of data obtained from reviews of pharmaceuticals. It has capabilities including obtaining comprehensive information about pharmaceuticals and looking up medication suggestions depending on medical conditions. With the help of cloud-backed analytics and a user-friendly interface, there is an exceptional opportunity to bridge the knowledge gap between end users and complex data, giving up important insights to a wider audience.

2. Related Work

The reviewed literature encompasses a broad spectrum of advancements in computational biology, bioinformatics, and healthcare applications. Notably, studies have been conducted to enhance drug discovery and prediction of drug-related interactions. Innovative computational pipelines, such as DCMGCN, NGDTP, and DDI-IS-SL, integrate various data sources and methodologies to predict novel drug combinations and drug-target interactions, demonstrating superior performance over existing methods. Additionally, research efforts focus on personalized recommendation systems, exemplified by GCFM, which employs graph-convolved factorization machines for interpretable and effective recommendations in financial product scenarios [1, 2, 3, 4], [5]. In the context of medical emergencies, machine learning-driven drug recommendation systems, like the one proposed by Silpa et al., offer valuable and accurate suggestions based on patient symptoms and conditions [6]. Furthermore, there is a growing emphasis on the integration of artificial intelligence and cloud computing in healthcare, as explored in studies like Gupta and Sharma's investigation into cloud-based solutions for data analytics [7]. The literature also delves into disease prediction and drug recommendation prototypes, such as Nayak et al.'s multi-approach model, showcasing the potential of machine learning in providing efficient and personalized healthcare solutions [8]. In the realm of diabetes prediction, the Adaptive Weighted Decision Forest algorithm proves effective in analyzing vital condition data [9]. DeepDrug introduces an innovative computer-aided drug design software utilizing artificial

intelligence to expedite the identification of new compounds [10]. Moreover, the use of deep attention neural networks, as seen in Liu et al.'s DANN-DDI, enhances drug-drug interaction predictions [11]. DeepSide, a deep learning approach for drug side effect prediction, stands out for its multi-modal architecture, offering accurate and interpretable predictions [12]. Finally, studies such as Galeano et al.'s framework for predicting the frequencies of drug side effects and Zhang et al.'s FS-MLKNN method contribute to the understanding of drug risk-benefit assessment and multi-label learning [13, 14]. The integration of AI, particularly evident in Leora's role as an AI-powered chatbot for mental health support, showcases the potential for digital mental health services to address global public health concerns [15]. Swathi et al. proposed a methodology for predicting drug side-effects by analyzing data sourced from open health forums, utilizing UiPath for data extraction and machine learning techniques for classification [16]. In a similar vein, Palanivinayagam and Sasikumar introduced a model that identifies diseases based on symptoms and recommends drugs aiming for minimal side effects, optimizing their recommendation model for efficiency [17]. Roy et al. delved into drug-resistant ovarian cancer, employing transcriptional profiling to understand gene co-expression and uncover resistance mechanisms, particularly focusing on PARP inhibitor resistance [18]. Thomas and Zachariah conducted computational analysis to investigate the potential of 4-phenyl-4H-chromeno derivatives as anticancer and anti-inflammatory agents, identifying compounds with promising drug-like properties [19]. Furthermore, Radhika et al. explored Insilco analysis techniques for nano Polyamidoamine (PAMAM) dendrimers in cancer drug delivery, emphasizing the utilization of computational methodologies in drug development and predictive modeling for drug discovery [20]. These diverse studies [21] collectively contribute to the advancement of predictive modeling, drug recommendation systems, and understanding drug resistance mechanisms.

3. Methodology

The system methodology comprises several steps that amalgamate data collection, analysis, modeling, and prediction within a structured framework as shown in Fig.1.

3.1. Data Collection and Preparation

In the initial phase of our study, we carefully compiled diverse datasets comprising essential information about drugs. This included establishing connections between drug names and specific medical conditions, drawing from reputable medical databases and pharmaceutical records. Simultaneously, we gathered sentiment scores or reviews associated with various drugs, originating from sources. Additionally, we collected average user ratings reflecting the general feedback on the effectiveness and side effects of different medications. Once we collected this information, the subsequent step involved data cleansing. We conducted a thorough consistency check to ensure accuracy in drug-condition associations, sentiment scores, and user ratings. Handling missing values was another crucial aspect, where we either imputed missing data or removed incomplete entries based on stringent quality standards. Standardizing data formats and structures across different datasets was undertaken to facilitate seamless integration. Furthermore, normalization was applied to certain data, ensuring a uniform scale for comparative analysis.

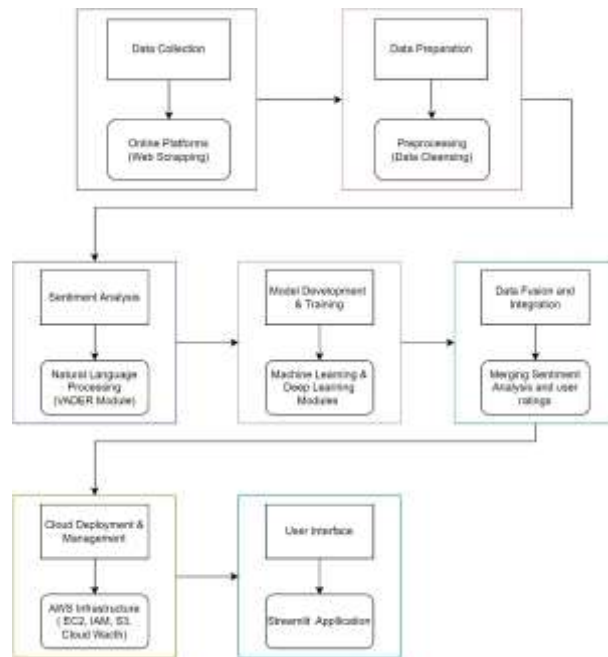


Figure 1: Block diagram of proposed research methodology

The process also involved eliminating any duplicate entries or redundant information that could potentially skew analytical results. The final step included preparing the datasets in a structured format ready for analysis, employing tools such as Pandas in Python for effective management. This comprehensive compilation and cleansing of datasets establish a robust foundation for the subsequent phases of in-depth analysis and modeling.

3.2. Sentiment Analysis

In the phase of Sentiment Analysis, we delve into Natural Language Processing (NLP) techniques, leveraging the VADER (Valence Aware Dictionary and sentiment Reasoner) module available through NLTK (Natural Language Toolkit). This powerful approach enables us to decipher the sentiment embedded within user-generated drug reviews. By employing NLP methodologies, we can systematically break down these reviews, identifying nuanced sentiments such as positive, negative, or neutral tones expressed by users. The VADER module specifically aids in this process by assigning polarity scores to individual words or phrases within the reviews, allowing us to gauge the overall sentiment conveyed. The goal here is not just to identify sentiments but also to extract profound insights from this analysis. Through sentiment insights, we aim to unravel the varied user perceptions and experiences associated with different drugs. This involves a deeper understanding of how users express their satisfaction, concerns, or experiences regarding the efficacy, side effects, or overall impact of specific medications. By comprehensively analyzing these sentiments, we can glean valuable information about user sentiment

trends, common issues reported, or standout positive experiences. These insights become instrumental in understanding the user landscape and can guide future decision-making processes in healthcare, pharmaceuticals, and patient care.

3.3. Web Scraping for Additional Information

In the stage of Web Scraping for Additional Information, we employ web scraping tools, notably Selenium, to access and extract crucial data from prominent online platforms such as 1mg.com. This process involves navigating through the website's pages, specifically targeting sections that contain valuable information about various drugs. Our primary focus lies in collecting comprehensive data encompassing drug names, their corresponding generic equivalents, and alternative brands available on the platform. The aim is to gather a robust dataset that includes detailed information about the drugs listed on the website. This includes not only their specific names but also their generic counterparts, which are vital in understanding the broader spectrum of medications available. Moreover, we gather details about alternative brands, taking into consideration user ratings, as these serve as viable alternatives for the same medication. Through this web scraping endeavor, we streamline the process of data collection, ensuring we acquire extensive and relevant information from reputable online sources. This consolidated dataset becomes a foundational asset for our research, facilitating a comprehensive analysis and providing a richer context for understanding drug-related information and user preferences.

3.4. Data Fusion and Integration

The phase of Data Fusion and Integration involves merging and unifying diverse datasets, specifically integrating the outcomes derived from sentiment analysis with the user ratings. This process aims to create a cohesive and consolidated dataset that amalgamates varied aspects of drug performance and user sentiments. Initially, sentiment analysis outcomes obtained through Natural Language Processing techniques, particularly VADER from NLTK, are harmonized with the existing dataset. These sentiment insights, derived from analyzing user-generated drug reviews, capture the emotional tone and sentiment expressed towards different medications. Subsequently, this sentiment-inclusive dataset is combined with the user ratings dataset, aligning the sentiments expressed in reviews with the quantitative ratings provided by users. This integration is pivotal as it brings together qualitative sentiment analysis and quantitative user ratings, providing a comprehensive and nuanced understanding of drug performance from multiple perspectives. The objective of this fusion and integration process is to create a singular, enriched dataset that encapsulates both qualitative and quantitative aspects of user feedback. This integrated repository becomes a valuable resource for conducting a holistic assessment of drug performance, offering insights into user sentiments alongside numerical ratings, ultimately contributing to a more comprehensive evaluation of various medications.

3.5. Model Development

Sentiment analysis and drug prediction are the two primary models of the Drug Information and Recommendation System. Evaluating the sentiment expressed in reviews of drugs is the primary objective of the sentiment analysis model. Data is initially collected from many online resources, and then the text is thoroughly cleaned up, with tokenization,

lowercasing, and the removal of HTML tags and special characters. The model uses natural language processing (NLP) techniques to classify reviews as good, negative, or neutral by assigning sentiment scores. The drug prediction model simultaneously aims to classify medicines according to properties taken from the preprocessed text. Collectively, these models provide an in-depth understanding of the emotional tone of reviews and drugs categorizations by sharing the same dataset and preprocessing processes. The findings are combined, stored as CSV files, and used as the basis for an extensive Drug Information and Recommendation System. To improve analytical power and usability, cloud computing services are integrated into the system. The model evaluation includes metrics such as accuracy, mean squared error (MSE), and a confusion matrix, providing a comprehensive assessment of classification and regression performance. These metrics offer a comprehensive evaluation of the models' performance in predicting drug-related categories. Ultimately, they contribute to enhancing the understanding of how user sentiments align with numerical ratings, offering predictive capabilities for drug performance assessment.

4. Experimental Results

The drug prediction task involves predicting the drug based on the review. The neural network model's training over 30 epochs shows significant improvement and potential as shown in Fig.2. Validation accuracy falls at 50.55%, indicating a possibility of overfitting despite a training accuracy of 61.99%. But this common challenge provides a chance for optimization via regularization strategies or model adjustments. Computational efficiency is demonstrated by the effective training time of 16 seconds for every batch.

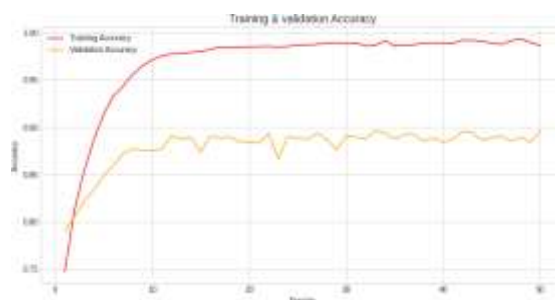


Figure 2: Drug Prediction task training and validation accuracy

The sentiment analysis task involves predicting whether a review has a positive, neutral, or negative sentiment. The neural network performed well overall, with an accuracy of about 88.20% on the validation set as shown in Fig.3. Neural networks are an excellent choice for tasks involving complicated interactions because they can effectively capture complex patterns in data. With its several deep layers, the selected architecture offers flexibility in terms of learning from the input features.

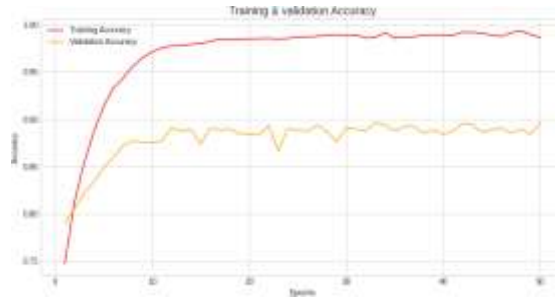


Figure 3: Sentiment analysis task training and validation accuracy

4.1. Cloud Deployment:

Model and Application Structuring: The model architecture and application design have been refactored to ensure compatibility with AWS infrastructure. The components have been organized into modular and scalable units, allowing for easy deployment and management on AWS. This modular approach enables efficient utilization of cloud resources and simplifies maintenance and updates.



Figure 4: Creating EC2 Instance

EC2: Customized settings and configurations specific to Amazon Elastic Compute Cloud (EC2) have been implemented to ensure optimal performance and scalability. The EC2 instances have been provisioned with appropriate specifications and resources to support the sentiment analysis model and application requirements. This includes defining instance types, storage options, and network configurations.



Figure 5: AWS S3 Bucket creation

Data Storage Setup: Amazon Simple Storage Service (S3) has been utilized to establish storage buckets for securely storing model artifacts and datasets. The S3 buckets have been configured with appropriate access policies and encryption settings to protect the data at rest and in transit. Versioning has been enabled to maintain data consistency and integrity, allowing for easy retrieval and management of different versions of datasets and model artifacts.

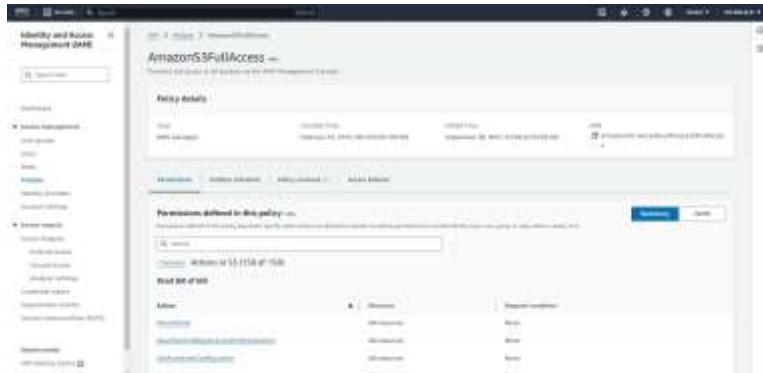


Figure 6: Creating IAM Role

IAM Policies and Access Control: IAM policies have been configured to manage user permissions and restrict access, ensuring the security of sensitive data and resources. Access controls have been defined to limit privileges and enforce data confidentiality and system integrity. This includes assigning roles and permissions to different user groups and implementing multi-factor authentication for enhanced security.

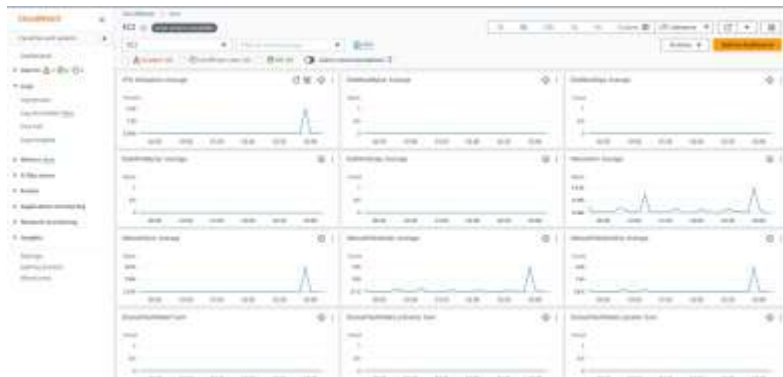


Figure 7: Cloud Watch monitoring

Monitoring and Management Configuration: AWS CloudWatch has been set up to monitor system metrics, log data, and performance indicators. This enables proactive management by tracking resource utilization, identifying bottlenecks, and detecting anomalies. Alerts and notifications have been configured to promptly notify administrators of critical events or performance deviations, facilitating timely responses and issue resolution.

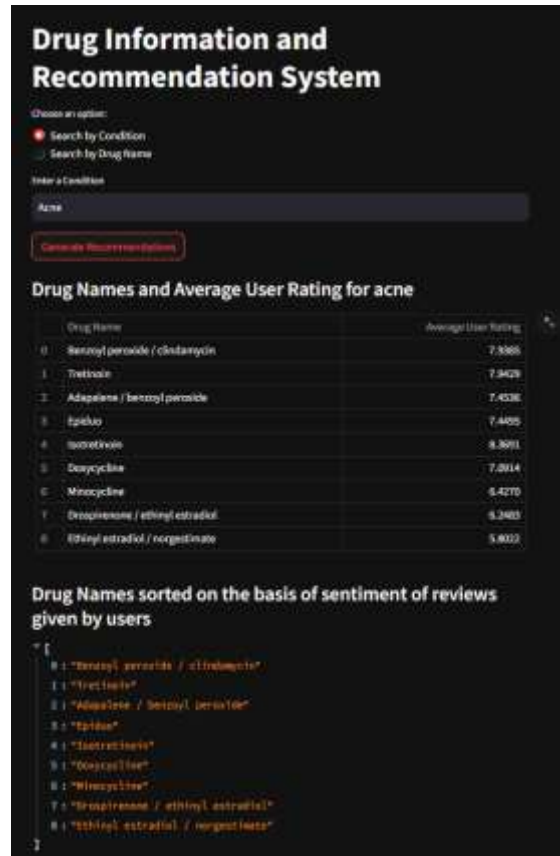


Figure 8: User Insights

By structuring the model and application to align with AWS infrastructure, configuring IAM policies for access control, setting up secure data storage with S3, and implementing monitoring and management through CloudWatch, the sentiment analysis model can be seamlessly deployed and managed on AWS as shown in Fig. 8. This ensures compatibility, scalability, security, and efficient utilization of cloud resources, enabling reliable and effective sentiment analysis for user reviews of pharmaceutical products.

5. Challenges

One significant challenge was handling a wide range of user-generated information with different languages, expressions, and emotions. This was overcome by using strong Natural Language Processing (NLP) techniques. Effective data processing strategies and data cleaning procedures were necessary to handle massive volumes of data while maintaining quality and integrity. Continuous optimization, including algorithm refining, parameter adjustment, and iterative improvement techniques, was necessary to ensure model correctness and relevance.

6. Conclusion

In summary, a significant advancement in pharmaceutical sentiment analysis has been made with the development of Sentiment AI, a Drug Information and Recommendation System that is hosted on Amazon Web Services (AWS). Sentiment AI provides compatibility, scalability, and security by utilizing AWS services including EC2, S3, and IAM. This allows for effective resource use and real-time monitoring via AWS CloudWatch. Furthermore, sentiment analysis is improved by the integration of complex machine learning (ML) and deep learning (DL) models, offering insightful information on how users view pharmaceuticals. This platform is a solid pharmaceutical sentiment analysis tool that has the potential to revolutionize healthcare decision-making.

7. Future Scope

Future scope involves addressing challenges related to data variability, volume handling, and model accuracy improvement. Continuous optimization strategies, including refining algorithms and incorporating feedback loops, are essential for enhancing predictive accuracy. By harnessing sentiment insights, businesses can refine customer experiences, healthcare benefits from patient feedback, and policymakers can comprehend public sentiments for responsive governance.

References

- [1] H. Chen, Y. Lu, Y. Yang, Y. Rao, A drug combination prediction framework based on graph convolutional network and heterogeneous information, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2022).
- [2] P. Xuan, B. Chen, T. Zhang, Y. Yang, Prediction of drug–target interactions based on network representation learning and ensemble learning, *IEEE/ACM transactions on computational biology and bioinformatics* 18 (2020) 2671–2681.
- [3] C. Yan, G. Duan, Y. Zhang, F.-X. Wu, Y. Pan, J. Wang, Predicting drug-drug interactions based on integrated similarity and semi-supervised learning, *IEEE/ACM transactions on computational biology and bioinformatics* 19 (2020) 168–179.
- [4] Y. Zheng, P. Wei, Z. Chen, Y. Cao, L. Lin, Graph-convolved factorization machines for personalized recommendation, *IEEE Transactions on Knowledge and Data Engineering* (2021).
- [5] D. Parasar, A. Ali, N. M. Pillai, A. Shahi, B. S. Alfurhood, K. Pant, Detailed review on integrated healthcare prediction system using artificial intelligence and machine learning, in: *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, 2023, pp. 682–685.
- [6] C. Silpa, B. Sravani, D. Vinay, C. Mounika, K. Poorvitha, Drug recommendation system in medical emergencies using machine learning, in: *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, 2023, pp. 107–112.
- [7] U. Gupta, R. Sharma, A study of cloud based solution for data analytics in healthcare,

in: 2023 6th International Conference on Information Systems and Computer Networks (ISCON), IEEE, 2023, pp. 1–6.

- [8] S. K. Nayak, M. Garanayak, S. K. Swain, S. K. Panda, D. Godavarthi, An intelligent disease prediction and drug recommendation prototype by using multiple approaches of machine learning algorithms, *IEEE Access* (2023).
- [9] R. Han, A study of diabetes prediction based on adaptive weighted decision forest, in: 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), IEEE, 2023, pp. 203–207.
- [10] S. Mukhopadhyay, M. Brylinski, A. Bess, F. Berglind, C. Galliano, P. F. McGrew, Deepdrug: Applying ai for the advancement of drug discovery, in: 2022 14th International Conference on COMMunication Systems & NETWORKS (COMSNETS), IEEE, 2022, pp. 667–674.
- [11] S. Liu, Y. Zhang, Y. Cui, Y. Qiu, Y. Deng, Z. Zhang, W. Zhang, Enhancing drug-drug interaction prediction using deep attention neural networks, *IEEE/ACM transactions on computational biology and Bioinformatics* 20 (2022) 976–985.
- [12] O. C. Uner, H. I. Kuru, R. G. Cinbis, O. Tastan, A. E. Cicek, Deepside: a deep learning approach for drug side effect prediction, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20 (2022) 330–339.
- [13] D. Galeano, S. Li, M. Gerstein, A. Paccanaro, Predicting the frequencies of drug side effects, *Nature communications* 11 (2020) 4575.
- [14] W. Zhang, F. Liu, L. Luo, J. Zhang, Predicting drug side effects by multi-label learning and ensemble learning, *BMC bioinformatics* 16 (2015) 1–11.
- [15] E. L. van der Schyff, B. Ridout, K. L. Amon, R. Forsyth, A. J. Campbell, Providing self-led mental health support through an artificial intelligence-powered chat bot (leora) to meet the demand of mental health care, *Journal of Medical Internet Research* 25 (2023) e46448.
- [16] D. N. Swathi, et al., Predicting drug side-effects from open source health forums using supervised classifier approach, in: 2020 5th International conference on communication and electronics systems (ICCES), IEEE, 2020, pp. 796–800.
- [17] A. Palanivinayagam, D. Sasikumar, Drug recommendation with minimal side effects based on direct and temporal symptoms, *Neural Computing and Applications* 32 (2020) 10971–10978.
- [18] S. Roy, J. Jeyalakshmi, S. Gochhait, S. Poonkuzhali, M. M. Gromiha, Metadata analysis to get insight into drug resistant ovarian cancer, *Ingénierie des Systèmes d’Information* 28 (2023).
- [19] N. Thomas, S. M. Zachariah, In silico drug design and analysis of 4-phenyl-4h-chromene derivatives as anticancer and antiinflammatory agents, *International Journal of Pharma- ceutical Sciences Review and Research* 22 (2013) 50–54.
- [20] R. Radhika, V. Rohith, N. Anil Kumar, K. Varun Gopal, P. Krishnan Namboori, O. Deepak, Insilico analysis of nano polyamidoamine (pamam) dendrimers for cancer drug delivery, *Int. J. Recent Trends Eng. Technol* 4 (2010) 142–144.
- [21] N. Sukumar, M. P. Krein, M. J. Embrechts, Predictive cheminformatics in drug discovery: statistical modeling for analysis of micro-array and gene expression data, *Bioinformatics and Drug Discovery* (2012) 165–194.