International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Formulation of a Computational Model for Predicting Drug Reactions Using Machine Learning

Christopher Agbonkhese[a*], Hettie Abimbola Soriyan[b], Kolawole Mosa[c]

[a]*Lecturer, Department of Digital and Computational Studies, Bates College, Lewiston, ME 04240, USA*
[b]*Lecturer, Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria:*
[c]*Lecturer, School of Health Sciences, Obafemi Awolowo University, Ile-Ife, Nigeria*
[a]*Email:agbonschris@gmail.com*
[b]*Email:hasoriyan@gmail.com*
[c]*Email:kmosaku@oauife.edu.ng*

Abstract

In the rapidly evolving landscape of healthcare, the efficient detection of drug reactions is of paramount importance to ensure patient safety and optimize treatment outcomes. This article presents the formulation of a computational model for the prediction of drug reactions in clinical settings using machine learning techniques. Our research leverages state-of-the-art machine learning algorithms to extract valuable insights from health records and prescription data. By systematically analyzing the relationships between prescribed medications and observed patient reactions, our computational model will be able to identify potential drug reactions emanating from drug prescription in clinical a clinical setting.

Keywords: Artificial Intelligence; Machine Learning; Drug Reactions; Healthcare.

1. Introduction

In recent years, Artificial Intelligence (AI) has gained adoption across several fields of life; from healthcare to agriculture, finance, science and technology, Law, manufacturing, education, transportation, and so on [1]. Specifically, the field of healthcare has undergone a rapid and transformative evolution, marked by groundbreaking advancements in medical technology, pharmaceuticals, and patient care.Among the myriad challenges facing modern healthcare, the detection of drug reactions has emerged as a critical concern both in orthodox medicine and traditional medicine [2]. Ensuring patient safety and optimizing treatment outcomes are paramount goals in the medical profession, and the early identification of adverse reactions to medications plays a pivotal role in achieving these objectives [3].

Received: *10/12/2023*
Accepted: *11/16/2023*
Published: *11/26/2023*

---

* Corresponding author.

Historically, the process of identifying and managing drug reactions has relied heavily on manual surveillance, clinician reporting, and post-market surveillance systems [4]. While these methods have provided valuable insights into drug safety, they are often labor-intensive, time-consuming, and reactive in nature. This reactive approach may result in delayed detection of adverse reactions, potentially endangering patient health and increasing healthcare costs. Recognizing the limitations of traditional methods, the intersection of healthcare and artificial intelligence (AI) has opened up new possibilities for enhancing drug reaction detection [5]. Machine learning, a subset of AI, has emerged as a powerful tool for the early identification and proactive management of drug reactions. Leveraging the vast amounts of clinical data and prescription records available, machine learning algorithms can sift through this information to uncover hidden patterns and associations that might otherwise go unnoticed by human observers [6]. This manuscript presents a comprehensive study that underscores the transformative potential of machine learning in healthcare, particularly in the realm of drug reaction detection. By harnessing state-of-the-art machine learning techniques and advanced data analysis, this research aims to provide a proactive, data-driven solution to the age-old problem of identifying and managing drug reactions.This paper is a product of a broader research which aims at building a computational model that can detect drug reactions and integrate the model into hospital information systems. However, the focus here is to present a formulated computational model that can be used to systematically examining the relationships between prescribed medications and observed patient reactions, the research employs an AI-driven approach to predict potential drug reactions. The methodology developed for this study addresses key challenges in healthcare, including scalability, real-time monitoring, and data integration, offering a robust framework for healthcare practitioners and institutions to adopt. The significance of this research lies in its potential to revolutionize the field of drug reaction detection. By automating and enhancing the identification process, the formulated model could be used to improve patient safety, reduce healthcare costs, and empower clinicians with more informed decision-making tools. Moreover, this study contributes to the broader conversation surrounding the integration of artificial intelligence in healthcare, offering valuable insights into the transformative impact of AI on patient care and the evolving landscape of modern medicine.

1. Literature Review

The authors in [7] developed data mining and machine learning models with the primary goal of predicting drug likeness and classifying drugs based on their associated diseases or organ categories. Their dataset, consisting of 762 compounds, was meticulously categorized into two primary groups: drugs (366 compounds) and nondrugs (396 compounds). To assess the robustness of their prediction model, the compounds were thoughtfully partitioned into a training set (80%, comprising 610 compounds) and a test set (20%, encompassing 152 compounds). A parallel distribution of drugs (73 compounds) and nondrugs (79 compounds) was maintained in the test sets. This partitioning process was executed through independent selection procedures utilizing a representativeness function, as proposed by [8]. The methodology incorporated a simulated annealing optimization strategy to select a subset of objects, namely compounds that best represented the current database from which it was drawn. Subsequently, predictive models were constructed on the training set employing a 10-fold cross-validation approach and seven distinct methods. These models were meticulously tested on the test set. The prediction models were constructed using six diverse machine learning algorithms, encompassing decision trees, random forests (RF), support vector machines (SVM), artificial neural networks (ANN), k-nearest neighbors

(k-NN), and logistic regression (LR). In each instance, classification models were established using the training set and subsequently employed to predict the activities, specifically drug status, of the test set compounds to validate the efficacy of the models. The implementation of these models was executed within the Weka framework, with evaluation metrics encompassing accuracy, sensitivity, specificity, and variance serving as performance benchmarks.

The authors in [9] delved into the intricate realm of causality through the automated extraction of lexical patterns. The study aimed to derive the reliability of extracted lexical patterns in expressing adverse reactions to specific drugs by learning their respective weights. Notably, their method achieved an impressive ADR detection accuracy of 74% on an expansive manually annotated dataset comprising tweets from a social media platform. This dataset encompassed a standardized set of drugs and their associated adverse reactions. Importantly, their model exhibited proficiency in accurately discerning causality between drugs and adverse reaction-related events. However, it is imperative to underscore that while accuracy served as a performance metric, it may not provide a comprehensive assessment of their model's performance, warranting further evaluation using additional metrics.

The authors in [10] embarked on an exploratory journey into the realm of social media mining for drug safety signal detection. Their pioneering work proposed the utilization of association mining and Proportional Reporting Ratios (PRR) to uncover valuable associations between drugs and adverse reactions. These associations were derived from the rich content contributed by users on social media platforms. In their experimental evaluation, ten drugs and five distinct adverse drug reactions were scrutinized. As a benchmark for assessing their techniques, they turned to the Food and Drug Administration (FDA) alerts. Their findings unveiled promising potential in employing metrics such as leverage, lift, and Proportional Reporting Ratio (PRR) for detecting adverse drug reactions that had been reported to the FDA. Importantly, PRR emerged as the standout performer among these metrics, showcasing its efficacy in identifying these critical drug safety signals.

The authors in [11] introduced an innovative approach in 1998 by proposing a Bayesian neural network method for generating signals related to adverse drug reactions (ADRs). Central to their work was the Bayesian Confidence Propagation Neural Network (BCPNN), renowned for its capacity to manage extensive datasets effectively. The BCPNN model exhibited robustness, particularly in handling imbalanced data and complex variables. Drawing from information theory, this tool proved ideal for uncovering drug-ADR combinations that exhibited high associations compared to the overall dataset or specific subsets thereof. Their study yielded compelling results, illustrating the potency of the BCPNN technique in early signal detection, exemplified by instances such as captopril-coughing. Moreover, the model demonstrated its ability to mitigate false positives, notably in cases where common drugs co-occurred with ADRs in the database, exemplified by scenarios involving digoxin and adverse reactions like acne or rash. Furthermore, the study conducted a routine application of the BCPNN to quarterly updates, revealing that a remarkable 1004 suspected drug-ADR combinations reached a confidence level of 97.5% difference from the dataset's generality. Significantly, among these combinations, 307 were identified as potentially serious ADRs, with 53 of them linked to novel drugs. This underscores the invaluable contribution of the BCPNN methodology to signal detection in the realm of adverse drug reactions.

2. Methodology

This section presents the research methodology adopted in this work. It begins with a description of the dataset that was used for the study the source of the data. It thereafter describes the various algorithms used for the formulation of the prediction model for drug reaction, the simulation of the model, the tools used, as well as the metrics for the evaluation of the model. Lastly, it presents a description of the architecture of the prediction model for the research.

3.1 Data Acquisition

The first objective of this research was to elicit relevant data on patients, which involved application and obtaining ethical clearance from the Obafemi Awolowo University Teaching Hospital Complex (OAUTHC) Ile Ife, in Osun State. A total of five hundred and eighteen (518) records were extracted from patients' case notes in the Psychiatric Department of the OAUTHC, this was done using case study technique on cases of admitted patients within nine years (that is, between November 2010 and November 2019). The extraction of data regarding cases of diagnosis and corresponding treatment for each admitted patient was carried out because there was no such data available in electronic format in that department as at the time of this study. The variables for which data was collected include age, sex, diagnosis, substance dependence, polypharcacy, prescribed drugs, dose, route of administration and drug reaction. A brief description of these variables is as shown in Table 1

Table 1: Brief Description of Variables

S/N	Feature	Description
1	Sex	A measure of the biological difference between a male and a female
2	Age	The length of time the patient has lived, measured in years
3	Diagnosis	Identification of the nature of illness or reason for admission
4	Polypharmacy	The concurrent use of multiple medications by a patient.
5	Substance dependent	Whether the patient is into any form of drug addiction
6	Prescribed drugs	The drugs that were prescribed for each patient (antipsychotics in this case)
7	Route of drug administration	The path by which a drug is taken into the body
8	Adverse reactions	Whether the patient experienced any adverse reaction in the course of the treatment after the drugs were administered

Some of the records in the case notes were incomplete, while others had patterns that could not be interpreted due to illegible handwritings. Hence the records were cleaned. In order to achieve meaningful prediction, the study required detail contents of the patient's case notes such as the reasons for which the patients were admitted in the hospital (captured in terms of diagnosis for each patient), treatments administered in the course of admission, as well as the outcomes for each treatment.

3.2 Data Pre-processing

The collected data was first preprocessed using data binning, one-hot encoding technique, and data normalization. Data binning was used for the age attribute conversion to categorize it into three bins. This was done firstly by specifying three equally sized bins, the bin array was built from minimum value to maximum value using the bin with the calculated, and labels were thereafter created as "Teenager", "Adult", "Elderly".

ADR values was encoded by adding dummy variables for each unique category in the original feature "ADR" by creating two new features "ADR" and "No ADR" such that when a value occurs in the original feature, this will mean that the corresponding value is set to '1' in the new feature while the other value is set to '0'. The same procedure was applied to variable "polypharmacy", and variable sex. In order to enhance the statistical models' performance, normalization was carried out on variable containing dataset that span wide range. This was done using 'simple feature scaling with pandas.

3. Model Formulation

Here, we formulate an ADR prediction model using stacking technique. The individual classification algorithms were trained based on the complete training set, after which the meta-classifier was be fitted based on the outputs from the base classifiers in order to improve the performance of the prediction model. This involved the use of decision tree algorithm and the naive bayesian (NB) algorithms as base classifiers while the adaptive boosting algorithm was used as a meta-classifier. The use of decision tree was based on their ability to convert large complex datasets into easy-to-understand output and suitability in handling binary data as mostly contained in the ADR dataset. Also, decision tree, through its variant (C5.4) provides for minimizing the error during classification by reducing entropy through the computation of the information gain in the dataset.

The choice for naive bayes was based on the fact that it works very well with binary data in classification task, while its variants provides for the enforcement of the conditional probability at different data points in order to improve the performance of the model. The choice for adaboost for was based on its ability to detect the point where the model misclassified the data and assign weight to those points in order to boost the performance of the final model, which is the meta-model. The following section presents the mathematical formulation of the prediction model:

Let *S* be the ADR dataset,

for the ADR dataset *S*, H(S) measures the amount of the uncertainty in the data as

Entropy

$$\mathrm{H}(\mathrm{S}) = \sum_{\mathrm{c}\epsilon\mathrm{C}} -\mathrm{P}(\mathrm{C}) \log_2 \mathrm{P}(\mathrm{C}) \qquad (1)$$

Where,

S: the current ADR dataset for which entropy is being calculated

C: the set of classes in the ADR dataset, *S*

P(C): the proportion of the number of elements in class *C* to the number of elements in set *S*.

Information gain IG(A) measures the difference from before to after the ADR dataset *S* is split on an attribute A. That is, how much uncertainty in ADR dataset *S* was reduced after splitting set *S* on attribute *A*

$$IG(A,S) = H(S) - \sum_{t \epsilon T} p(t)H(t) \quad (2)$$

Where,

$$S = \bigcup_{t=T} t \quad (3)$$

Where:

H(S) is the entropy of dataset S from equation (2)

T is the subset created from splitting ADR dataset set *S* by attribute *A* such that P(t) is the proportion of the number of elements in set *S*

H(t) is entropy of set *t*

Also, from Bayesian theorem, dealing with strong independence assumptions between predictors. The theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). The classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. That is, to say that the various attributes that predict the possibility of drug reactions are independent of the others. This assumption is useful when the number of instance, N is high and/or N is small, making (x|c) difficult to estimate. Even if the assumption does not hold, the model classification performance may still be good in practice because the decision boundaries may be insensitive to the specificities of the class-conditional probabilities p(xi|c); that is, variance is reduced because few parameters are required and the biased probability estimates may not matter since the aim is classification rather than accurate posterior class probability estimation.

$$p(c|x)\frac{P(x|c)P(C)}{P(X)} \quad (4)$$

Where:

P(c|x): the posterior probability of class (target) given predictor (attribute)

P(c): the prior probability of class

P(x|c): the likelihood which is the probability of predictor given class

P(x): the prior probability of predictor

Let D be a training set of tuples and their association class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurement made on the tuple from n attributes, respectively, $(A_1, A_2, \ldots, A_n)$.

Suppose there are m classes, $(C_1, C_2, \ldots, C_m)$, given a tuple, X, the classifier will predict that tuple X belongs to the class C_i if and only if

$$P(C_i|x) > P(C_j|x) \text{ for } 1 <= j <= m;\ j \neq I \quad (5)$$

Thus, this maximize $P(C_i|x)$. The class C_i for which $P(C_i|x)$ is maximized represents the Maximum Posteriori Hypothesis.

By Bayes' theorem

$$P(C_i|x) = P(x|C_i)\, P(C_i)/P(x) \quad (6)$$

As P (x) is constant for all classes, only $P(x|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is;

$$P(C_1) = P(C_2) = \ldots\ldots\ldots\ldots = P(C_m) \quad (7)$$

Maximizing $P(x|C_i)$. Otherwise, this maximize $P(x|C_i)P(C_1)$. That the class prior probabilities may be estimated by:

$$P(C_1) = |C_1, D|/|D| \quad (8)$$

Where

$|C_1, D|$ is the number of training tuples of class C_1 in D.

Given datasets with many attributes, it would be extremely computationally expensive to compute $P(x|C_i)$. In order to reduce computation in evaluating $P(x|C_i)$, the naïve Bayes' assumption of class conditional independence is made.

As per the conditional independence assumption of Bayes theorem, the presence or absence of some parameters of a class is independent to the presence or absence of some other parameters, making each parameter's contribution independent to the final result. For instance, for a parameter P(ADR = "Yes") given "polypharmacy" = 'Value from Test Data' is independent of P(ADR = 'No') gives "polypharmacy" = 'Value from Test Data'. In similar way, the probabilities of all the parameters and their individual contribution to the final result in different

variables can be calculated. To deal with the condition of zero probability values for some parameter, Laplace Correction will be used. In order to handle the imbalance nature of the sample data, and create a highly accurate prediction model, the study adopts the adaptive boosting algorithm as meta-classifier by focusing on difficult data points which might have been misclassified most by the decision tree algorithm and the Naive Bayesian network classifier, using an optimally weighted majority vote, α_t of these weak classifiers.

Given *m* labeled training examples (x_1, y_1)… (x_m, y_m) where the x_is are in some domain □, and the labeled y_iЄ {-1, +1}. On each round *t*=1,…,T, a distribution D_t is computed over the *m* training examples, and a given weak learner or weak leaning algorithm will be applied to find a weak hypothesis h_t: □ →{-1, +1}, where the aim of the weak learner is to find a weak hypothesis with the least weighted error ϵ_t relative to D_t.. The final or combined hypothesis (classifier) H computes the sign of a weighted combination of weak classifies. That is to say that the final hypothesis H or classifier is computed as a weighted majority vote of the weak hypothesis β_t, where each is assigned weight α_t.

$$H(x) = sign(\sum_{i=1}^{n} \alpha_t \beta_t) \quad (9)$$

Where the weak hypothesis β_t are H(s) and P(c/x) gotten from equation (1) and equation (6) respectively.

α_t is computed as: $\alpha t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ (10)

Equation (9) then can be written as: $H(x) = sign\left(\sum_{i=1}^{n}\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\left((\sum_{t\epsilon T} p(t)H(t)) + \left(\frac{p(X|C)p(c)}{p(x)}\right)\right)\right)$ (11)

Where H(x) represent the final hypothesis, which produces the class prediction for ADR as ADR or a NO ADR.

4. Discussion

The article presents a comprehensive and structured methodology for predicting drug reactions. By leveraging machine learning and AI techniques, the researchers aim to address the critical issue of adverse drug reaction detection in clinical settings. The data acquisition process is described clearly, with a focus on ensuring the ethical collection of patient data. The use of data pre-processing techniques like data binning, one-hot encoding, and normalization is essential to prepare the data for modeling. The model formulation is based on a stacking approach, which combines the strengths of decision tree, naive Bayesian, and adaptive boosting algorithms.

The decision tree is suitable for handling binary data, and naive Bayesian is known for its performance with binary data as well. The adaptive boosting meta-classifier aims to improve the model's performance by focusing on challenging data points. Employing stacking with the three algorithms would enhances the overall accuracy by capturing different facets of underlying data patterns, and fostering model robustness. This combination excels in handling various types of data patterns in the ADR data, addressing challenges posed by the imbalanced datasets

and adapting to non-linear relationships. The ensemble mitigates overfitting, provides a balanced trade-off between interpretability and predictive power, and reduces both bias and variance. The stacking, would increase confidence in the prediction through the consensus of multiple models, resulting in a more resilient and accurate predictive model.The mathematical formulations provided for calculating probabilities and the final hypothesis are crucial for understanding how the model makes predictions. The article outlines a systematic and well-structured approach for building a predictive model for drug reactions in a clinical setting.

5. Limitations of the Study and Future Work

Although, the presented a computational model for predicting drug reactions using machine learning demonstrates a promising research, it is important to acknowledge certain limitations inherent in the study. The first limitation lies in the dataset's origin, primarily sourced from a specific department within the Obafemi Awolowo University Teaching Hospital Complex. The focus on psychiatric patients within a specific timeframe may introduce biases, since drug reactions can vary across medical specialties and patient populations. Generalizing the findings to broader healthcare contexts may require additional datasets encompassing diverse medical conditions, treatment regimens, and demographic factors to ensure the model's robustness and applicability across a more comprehensive range of clinical scenarios.Again, the study's reliance on historical patient records, though valuable for retrospective analysis, introduces the challenge of temporal dynamics. Healthcare practices, drug prescriptions, and patient demographics evolve over time, potentially impacting the model's predictive accuracy in contemporary settings. Given the rapidly advancing landscape of healthcare, ongoing data collection and model refinement would be crucial to address the temporal limitations of the dataset.

6. Conclusion

In conclusion, the article underscores the transformative potential of machine learning and AI-driven solutions in the field of drug reaction detection within healthcare. The study's multifaceted practical implications are significant. First and foremost, the model has the potential to significantly enhance drug reaction detection in clinical settings, thereby bolstering patient safety and reducing adverse outcomes. Moreover, its implementation could lead to substantial cost savings in healthcare by preventing unnecessary hospitalizations and treatments. This research empowers healthcare practitioners with advanced decision-making tools, aiding them in providing more accurate and timelier patient care. Additionally, it contributes to the broader integration of artificial intelligence in healthcare, signifying the transformative potential of AI-driven solutions in improving patient outcomes. Ultimately, the model's real-time monitoring capabilities offer dynamic, responsive patient care in the ever-evolving healthcare landscape. The authors assert that their work represents a significant step toward making efficient and proactive drug reaction detection the standard in healthcare, promising revolutionary improvements in patient outcomes.

References

[1] Ngige OC, Ayankoya FY, Balogun JA, Onuiri E, Agbonkhese C, Sanusi FA. (2023). A dataset for predicting Supreme Court judgments in Nigeria. Data Brief. 9;50:109483. doi:

10.1016/j.dib.2023.109483. PMID: 37588617; PMCID: PMC10425661.

[2] Agbonkhese, C. Soriyan, H. A. and Oyelami O. (2016). Fuzzy-based Dosage Model for Aqueous decoction of Adansonia Digitata for the Management of Sickle Cell Anaemia Patients in African Traditional Medicine. Nigerian Journal of Natural Products and Medicine (NJNPM) Vol. 20.

[3] Lopez-Gonzalez, E., Herdeiro, M. T., Piñeiro-Lamas, M., & Figueiras, A. (2014). Effect of An Educational Intervention to Improve Adverse Drug Reaction Reporting in Physicians: A Cluster Randomized Controlled Trial. *Drug Safety*, *38*(2), 189–196.

[4] Pirmohamed, M., Ostrov, D. A., & Park, B. K. (2015). New genetic findings lead the way to a better understanding of fundamental mechanisms of drug hypersensitivity. *J Allergy Clin Immunol*, *136*(2), 236–244

[5] C. Agbonkhese and H. A Soriyan (2019). A Machine Learning Approach for Predicting Drug Reactions from Patients' Case Notes. Proceedings of International Conference of the Application of Information Communication Technologies to Teaching, research and Administration (AICTTRA), Oaks Park, Obafemi Awolowo University, Ile-Ife, Nigeria. November, 2019; 45 – 48

[6] Kusy M, Jacek K. (2017). Assessment of prediction ability for reduced probabilistic neural network in data classification problems. *Soft Computing*, 21(1), 199-212.

[7] Yosipof, A., Guedes, R. C. and García-Sosa, A. T. (2018). Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category, *Frontier in Chemistry. 6(162)*6:16

[8] García-Sosa, A. T., and Maran, U. (2013). Drugs, non-drugs, and disease category specificity: organ effects by ligand pharmacology. *SAR QSAR Environmental Research,24*(4), 319-331.

[9] Bollegala1,D., Maskell1, S., Sloane, R., Hajne, J. and Pirmohamed, M. (2018). Causality Patterns for Detecting Adverse Drug Reactions from Social Media: Text Mining Approach, *JMIR Public Health Surveill 4(2)* doi:10.2196/publichealth.8214

[10] Yang, M. Wang, X. and Kiang. M. (2013). Identification of consumer adverse drug reaction messages on social media. Pacific Asia Conference on Information Systems (PACIS). 193; 2013

[11] Bate, A., Lindquist, M, Edwards, I.R., Olsson, S., Orre, R. and Lansner, A. (1998). A Bayesian neural network method for adverse drug reaction: Signal generation. *European Journal of Clinical Pharmacology*, 54(4), 315-321