Privacy-Preserving Framework for Smart Prediction of Drug Side Effects

Eka Asibong Ibeh
Department of Computer Science, Federal Polytechnic of Oil and Gas, Bonny, Rivers State.Nigeria

Alalibo Ralph Fiberesima
Department of Computer Science, Federal Polytechnic of Oil and Gas, Bonny, Rivers State. Nigeria

ABSTRACT
Side effects and adverse reactions are major public health concerns and major causes of drug failure. Predicting side effects is crucial for controlling development costs, time, and launching effective drugs for patient health recovery and preventing immediate withdrawal from the market. The new system is smart drug side effect prediction using deep neural network with the chemical structure of drugs. As it is a common experience that a single drug can cause multiple side effects, that's why we have proposed a deep neural network model that can predict multiple side effects for a single drug. This study has examined two side effects: Dizziness and Headache. The dataset was collected for the drug side effects information from the SIDER database. We have achieved an accuracy of '0.9494' with our multi-label classification based proposed model. The proposed model can be used in different stages of the drug development process. The methodology adopted the Agile software design methodology and unified modeling language as design tool. Python programming language was used for its implementation.

Keywords
Drug, Side Effects, Healthcare, Classification Algorithm, Deep Neural Network

1. INTRODUCTION
Drugs, natural, semi-synthetic, or synthetic chemicals, can treat diseases but often come with side effects. Recognizing potential side effects helps reduce costs and avoid risks in drug discovery, as many approved drugs have been withdrawn. Extensive testing is conducted on medicines before they enter the market, but side effects may not become apparent until a large number of users experience them. The United States Food and Drug Administration (FDA) runs a program called MedWatch, a voluntary post-marketing surveillance program that allows consumers to report drug side effects. However, wet experiments are costly and time-consuming. Since researchers collected drug data and compile them in the public databases, computational methods were developed for the side effect prediction. Many drugs approved by National Agency for Food and Drug Administration and Control (NAFDAC) were recalled each year after some unexpected side effects were discovered; for example, in 2010, Aldrin, Binapacryl Fungicide and Captafol Fungicide were withdrawn. According to the Drug recall twenty thousand people had taken the drugs before it was later withdrawn from the market. Developing and producing drugs that were later found having serious side effects would be a disaster to a pharmaceutical company. For instance, the withdrawal of the aforementioned antiobesity drug has cost Wyeth more than $21 billion in America alone [1]. Therefore, it will not only avoid causing harm to patients but also avoid wasting lots of money if we can discover the side effects of a drug compound in the early phase of drug discovery. At the same time, drug side-effects, or adverse drug reactions, have become a major healthcare concern. As an illustration to the extent of this problem, serious drug side-effects are estimated to be the fourth leading cause of death in US, resulting in 100,000 deaths per year. The identification of potential severe adverse side-effects is a challenging issue at many stages of the drug development process [2]. A useful experimental approach for predicting side-effects is preclinical in-vitro safety profiling which tests compounds with biomedical and cellular assays, but experimental detection of drug side-effects remains very challenging in terms of cost and efficiency. To address the problem, it is highly demanded by pharmaceutical industries to develop computational methods for predicting the side effects of drugs.

2. RELATED WORK
In recent years, machine learning methods were applied to the drug side effect prediction, because of their capability of dealing with complicated data.

[3] designed a new enhanced Singular Value Decomposition (SVD) scheme to reduce the dimensionality of drug-disease data. First, an integrated model for Hepatocellular Carcinoma (HCC) subordinate was developed using the Multi-source Bat Algorithm-based Random Walk (MBARW) to differentiate novel drugs and diseases. Then, a drug-drug similarity grid and disease-drug similarity chain were created based on the multi-source random stroll in gene-gene weighted correlation order. Moreover, all drugs in the drug-drug similarity chain and disease-drug bipartite graph chain were ranked by considering the known drugs for HCC. But, it was not suitable for the database containing many genes.

[4] developed a Sigmoid Kernel and CNN (SKCNN) to train new attributes efficiently, defining drug disease correlations using its hidden layer. Initially, the similarity metric of drugs was created by sigmoid drug similarity, drug structural similarity and that of disease utilizing disease sigmoid similarity and disease semantic similarity. According to the fused similarities of drugs and diseases, the SKCNN was utilized for training hidden interpretations for all drug-disease pairs whose tags were predicted by the random forest categorizer. But, the training data was comparatively inadequate, influencing the prediction accuracy.

[5] developed a learning-based technique depending on feature interpretation training and deep learning called DTI-CNN to predict drug-target correlations. Initially, the Jaccard similarity coefficient extracted the related attributes of drugs and proteins from heterogeneous networks and restarted the random walk scheme. After that, a denoising AE was used to minimize the dimension and detect the important attributes. According to these attributes, the CNN was created to predict the correlation between drugs and proteins. But, it needs more related data and

network design to contain more sophisticated input networks.

[6] developed a Similarity Network Fusion and Collective Variational AE (SNF-CVAE) model to predict new drug-disease correlations using drug-associated similarity data and known drug-disease correlations. This model integrated similarity measures, similarity choice, SNF and CVAE to perform a non-linear analysis and enhance the drug-disease correlation prediction. But, it has a high training period due to the huge amount of mixtures of learnable hyperparameters.

3. METHODOLOGY

Agile software design methodology is a combination of iterative and incremental process models with focus on process adaptability and customer satisfaction by rapid delivery of working software product. Agile Methods break the product into small incremental builds. These builds are provided in iterations. Each iteration typically lasts from about one to three weeks. Every iteration involves cross functional teams working simultaneously on various areas like

i. Planning
ii. Requirements Analysis
iii. Design
iv. Coding
v. Unit Testing and
vi. Acceptance Testing.

At the end of the iteration, a working product is displayed to the customer and important stakeholders.

4. PROPOSED SYSTEM

The proposed model is deep neutral network (DNN), several data regarding patients and drugs are gathered from NAFDAC website, which involves many attributes like drug name, ingredients, size, shape, capsule shell, weight, surface region, dissolution period, predisposition of medema, side effect, drug mode, drug price, drug category, dosage, brand name, drug feedback, patient age, sex, weight and height. Then, those attributes are fed to the DNN as input to predict drugs and their side effects for certain diseases. We have optimized features for a DNN model, utilizing a 1D multiple layers scheme, to predict future drug side effects, using the activation function as the transfer function.

Figure 1 show the architecture of the system. predicting the medicinal uses of natural compounds based on the trained DNN learning model. When the input features are complex and heterogeneous, deep learning can improve the performance of the predictor by learning high-level representation from low-level features.

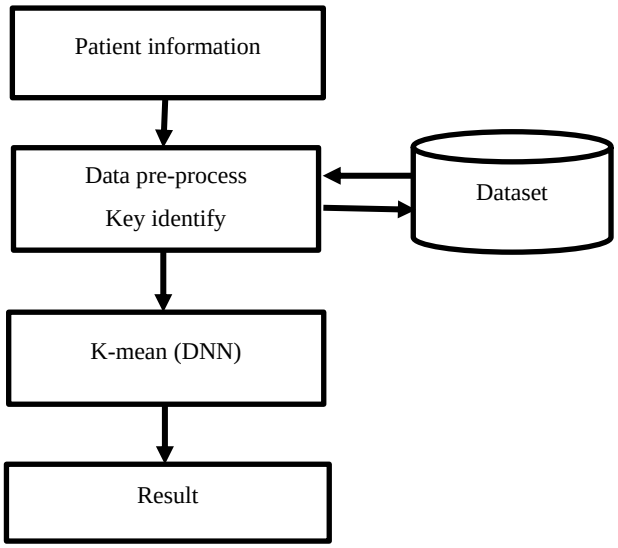

Figure 1 Architecture Design of the Proposed System

The DNN proposed model consists of four sequential layers Figure 2:

1) input layer,
2) partially connected hidden layers,
3) fully connected hidden layers, and
4) output layer.

The models were generated for 15 diseases, respectively, to predict the potential effects list from input features. For each drug, and chemical property features and used them as the inputs of the model. Hidden layers generalized their outputs by providing a high-level representation that was more abstract than the previous layer by discovering nonlinear relationships between the low- and high-level data. Let Xl is the output of the lth hidden layer. The forward propagation of the neural network with lth hidden layer can be represented as follow.

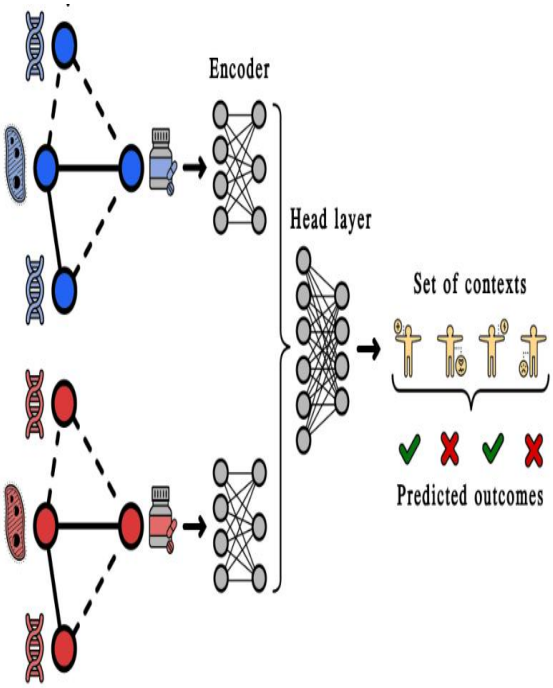

Figure 2 DNN Model Design

5. ALGORITHM

In the prediction of drug side effects, all the classifier model design was achieved using python programming. Because of the different pattern of how the SIDER dataset was presented from the Connectivity Map, scripts to extract the desired information. Here are the pseudo codes:

read the SIDER data table

Algorithm 1: DNN learning

Function DNN LEARNING

Pre-learn DNN;

Initialize coarse attribute layer;

Pre-learn fine attribute layers;

Adjust the entire DNN

End Function

make the initial matrix (0, no of unique drugs × no of unique side effects)

for each unique drug in the matrix

for each unique side effect matrix

if the side effects are related to that drugs exists in the SIDER data table, the assign 1 to the corresponding elements in the matrix

output the matrix

get the label for classification

6. TRAINING THE NETWORK

To validate the network at regular intervals during training, validation data should be specified by choosing the validation frequency value so that the network is validated about once per epoch.

Table 1: Key data structures of DNN Predictive System:

Data	Description	Type
Training Samples	Number of Training samples	integer
Epoch's	iterations	Integer
Input Unit	Data input	String
Learning Rate	Used for improving Prediction	double

outputEpochs 100

Leaning rate 0.01

The predictive system is centred on the provision of convolutional neural network which shows number of hidden neurons (hidden unit size), number of training epochs, letter size, maximum number of trials.

Table 2. Key predictive System Network Default Parameters

System Parameter	Value
Hidden unit size	40 by 40
Epochs	100
Learning rate	0.01
Soft-max sample temperature	0.10
Letter	5
Maximum trials	3

7. SEQUENCE DIAGRAM

A sequence diagram depicts in figure 3. Sequence diagram show how object interact with each other via messages in the execution of an operation. The patient information and symptom is capture to map with the Drug chemical Structure for the system to make it prediction.

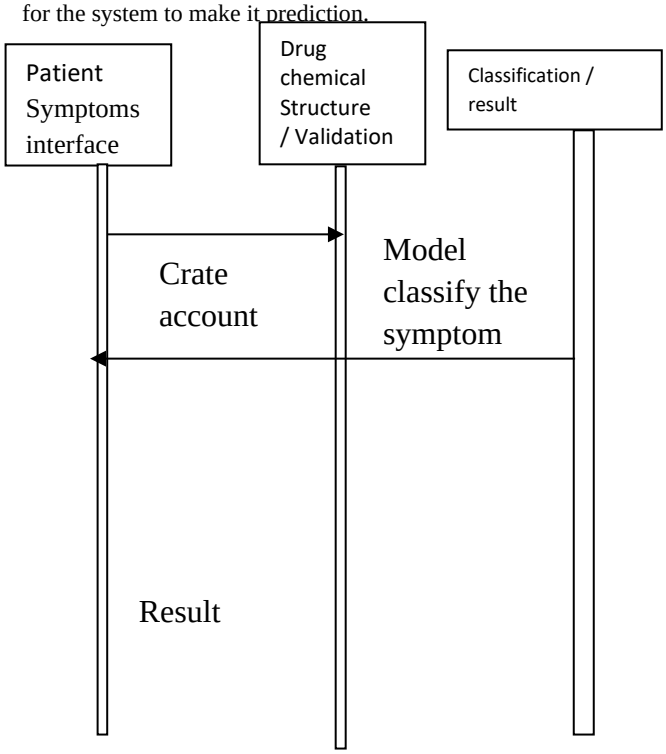

Figure 3. Sequence Diagram of the proposed system

8. RESULT

In this study, DNN computational method was developed to predict the side effects of drug compounds like the chemical structure. The parameter are quinoline-related compounds and antifolates for the treatment of malaria the side effect like dizziness and headache. A training dataset and test datasets were constructed from the benchmark dataset that contains 835 drug compounds to evaluate the method. Then, such data are fed to the DNN for prediction. In the DNN model, DNN is embedded into the attribute hierarchy. It segregates simple classes using a coarse classifier, whereas fine classifiers

differentiate complex classes. In the learning phase, an element-wise pre-learning is supported by global fine-tuning with a multinomial logistic loss normalized by a coarse coherence factor. By a jackknife test on the training dataset, the 1st order prediction accuracy was 86.30%, while it was 89.16% on the test dataset. It is expected that the new method may become a useful tool for drug design, and that the findings obtained by hybridizing various interactions in a network system may provide useful insights for conducting in-depth pharmacological research as well, particularly at the level of systems biomedicine. The system was implemented using python library . For this experiment, different details of patients and drugs are gathered from NAFDAC. Figure 4 show the Result chart of the proposed system

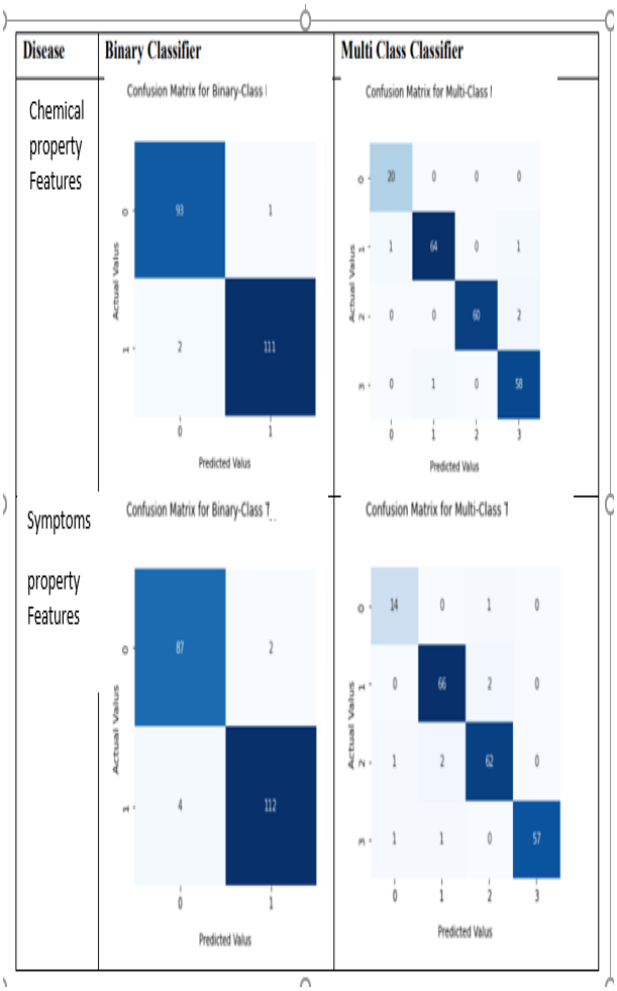

Figure 4 show the Result chart of the proposed system

Table 3: Performance metrics for malaria binary classifier and typhoid binary classifier

Metric	Chemical structure binary classifier	Patient Symptoms binary classifiers
Precision	0.990	0.980
Recall	0.995	0.990
f-score	0992	0.985
Accuracy	0.986	0.061

The performance metrics of both Chemical structure and Patient Symptoms classifier are shown in table 2. Chemical structure binary classifier has F1-Score of 0.992(99.2%) and accuracy of 0.986(98.6%), while Patient Symptoms binary classifier has F1-Score of 0.985(98.5%) and accuracy of 0.961(96.1%).

9. CONCLUSION

This study developed the DNN model for drug prediction and recommendation for specified illnesses. Primarily, the database was prepared by gathering information regarding patients, diseases, drugs and their side effects. Afterward, the created database was provided to the DNN for the prediction process. This DNN model was designed by embedding the DNN into the attribute hierarchy, which splits uncomplicated classes by the coarse categorizer and complicated classes by the fine categorizers. During the DNN learning, an element-wise pre-learning was adopted based on the global adjustment with a multinomial logistic error normalized through the coarse coherence factor. Additionally, the DNN was enhanced by the conditional executions of fine categorizers and the reduction of layer variables for huge databases. By learning the DNN, the appropriate drugs for certain diseases were predicted according to the patient's characteristics. At last, the test outcomes proved that the DNN model has 95.3%, 97.1% and 98.5% accuracy for predicting the drugs for CKD, diabetes and heart diseases, respectively, compared to the other existing models.

10. REFERENCES

[1] Fukuzaki, M. Seki, H. Kashima, and J. Sese,; 2017 "Side effect prediction using cooperative pathways," in Proceedings of the IEEE InternationalConference on Bioinformatics and Biomedicine (BIBM '09), pp. 142–147,

[2] Blower, C. Yang, M. A. Fligner 2018., "Pharmacogenomics analysis: correlating molecular substructure classes with microarray gene expression data," Pharmacogenomics Journal, vol. 2, no. 4, pp. 259–271

[3] Ibrahim, S. J. A., & Thangamani M, 2017 "Enhanced Singular Value Decomposition for Prediction of Drugs and Diseases with Hepatocellular Carcinoma Based on Multi-Source Bat Algorithm Based Random Walk," Measurement, vol. 141, 176-183

[4] Jiang, H. J., You, Z. H., & Huang Y. A, 2018 "Predicting Drug−Disease Associations Via Sigmoid Kernel-Based Convolutional Neural Networks, Journal of Translational Medicine, vol. 17, no. 1, pp. 1-11

[5] Peng, J., Li, J., & Shang X, 2019 "A Learning-Based Method for Drug-Target Interaction Prediction Based on Feature Representation Learning and Deep Neural Network," BMC Bioinformatics, vol. 21, no. 13, pp. 1-13.

[6] Jarada, T. N., Rokne, J. G., & Alhajj R, 2019 "SNF–CVAE: Computational Method to Predict Drug–Disease Interactions using Similarity Network Fusion and Collective Variational Autoencoder," Knowledge-Based Systems, vol. 212, pp. 1-23.