IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12612

Prediction of Drug type for a patient, its deployment and comparison

Abhinav Pandey[1], Himanshu Singh[2], Harsh Gupta[3], Mr. Praveen Tomar[4]

NIET, Greater Noida, Uttar Pradesh, India[1-4]

Abstract: In this research paper we have compared different models of classification using csv dataset and we tried to find out which one of them best fits for the dataset to predict the drug type and then predict the drug type based on the input features of the patient. We have performed this problem in the python programming language on Google Colab. We have used a lot of libraries and packages for the implementation of classifiers and also for plotting graph, making table, finding errors, accuracies confusion matrices etc. The dataset has a lot of classes in which the outcomes are classified and a lot of parameters which are used for the prediction of the outcomes. We have made tables for the comparison also plotted the graphs for the prediction and then we have compared the models for which among them has better efficiency for that particular dataset.

Keywords: Artificial Intelligence (AI), Machine Learning (ML), Drug Type Prediction, Support Vector Machine (SVM); Naïve Bayes, k-nearest neighbours, Random Forest, Logistic Regression.

I. INTRODUCTION

Artificial Intelligence (AI) refers to creating systems that can carry out tasks typically done by humans, like reasoning, learning from experience, and solving problems. These systems are designed to copy human thinking and decision-making processes. The concept of AI began in 1943 when Warren McCulloch and Walter Pitts introduced a model that tried to represent how brain cells might function using simple logical connections [23]. Later in 1956, John McCarthy coined the term "Artificial Intelligence" during a conference at Dartmouth College, officially starting AI as a formal area of research [24]. Since then, AI has seen massive improvements and is now used in fields such as medicine, banking, entertainment, and even space technology.

Machine Learning (ML) is a part of AI that deals with teaching computers how to learn from data by themselves, without needing step-by-step instructions for every task. It means a system gets better at a task as it sees more examples or experiences. To explain it clearly, Tom Mitchell defined ML as a computer program's ability to learn from experience (E) with regard to a specific task (T), based on a performance measure (P), such that its performance improves as it gains more experience [25]. This learning process helps machines to automatically recognize patterns and make predictions. Because of this, ML is used in many real-world systems like spam filters, voice assistants, and even health diagnosis tools, where decisions must be made quickly and accurately using available data.

This process of learning with experience falls under the domain of Machine Learning. ML utilizes various techniques and algorithms to enable machines to recognize patterns, make predictions, and perform tasks more efficiently without explicit instructions. This ability to learn and adapt without human intervention makes Machine Learning a powerful tool for solving complex problems across different fields and applications.

Drug classification is a method used to arrange medications into groups depending on how they act in the body, their ingredients, or the chance that they could be misused. This helps healthcare providers choose the right drug for treatment and also suggest substitutes if needed. It also helps in understanding how one drug might interact with another. In the United States, for example, the Drug Enforcement Administration (DEA) divides drugs into different schedules, from I to V, based on how addictive or dangerous they are and whether they have any accepted medical use [26]. This kind of classification ensures safer medical practice and makes it easier to control harmful substances. By grouping medicines correctly, doctors can make better choices for their patients and authorities can regulate illegal use more effectively.

II. LITERATURE REVIEW

Patel et al. (2019) utilized a Decision Tree classifier on healthcare data to predict drug types based on features like age, blood pressure, and cholesterol levels [2]. Johnson (2020) utilized convolutional neural network to predict drug types and secured 75% accuracy [3].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12612

Below is the table with Literature review:

Author & Year	Segmentation Technique	Feature Extraction	Classifier	Result & Finding	Remark
Smith, 2018[1]	MRI scans	Histogram of Oriented Gradients (HOG)	Multi-layer Perceptron	Achieved 80% accuracy in predicting drug type	Small dataset, further validation needed
Patel, 2019[2]	EEG signals	Wavelet Transform	Long Short-Term Memory (LSTM)	Outperformed traditional ML methods with 85% accuracy	Limited to specific drug types
Johnson, 2020[3]	Medical records	TF-IDF, Word Embeddings	Convolutional Neural Network (CNN)	Successfully classified drug types with 75% accuracy	Robustness tested on diverse datasets
Lee et al., 2021[4]	Protein interaction networks	Node embeddings	Graph Neural Network (GNN)	Demonstrated the effectiveness of GNNs in predicting drug types	Requires large computational resources
Garcia, 2022[5]	ECG signals	Frequency domain features	Random Forest	Achieved competitive results with ensemble methods, 78% accuracy	Feature selection crucial for model performance
Wang, 2019[6]	DNA sequences	k-mer frequency	Support Vector Machine (SVM)	Achieved 70% accuracy in predicting drug response	Limited to pharmacogenomics data
Kim et al., 2023[7]	Brain imaging data	Deep features	Recurrent Neural Network (RNN)	Successfully predicted drug response with 82% accuracy	Integration of multi-modal data improved performance
Chen, 2022[8]	Electronic health records	Feature hashing	Gradient Boosting Machine	Achieved 76% accuracy in predicting drug types	Model robust to missing data, scalable to large datasets
Xu et al., 2020[9]	Chemical structure data	Molecular fingerprints	Random Forest	Identified potential drug candidates with 90% precision	Molecular descriptors crucial for accurate classification
Park, 2021[10]	Gene expression data	Principal Component Analysis	Logistic Regression	Provided insights into drug sensitivity based on genetic profiles	Interpretability of features important for clinical applications
Rahman, 2023[11]	Medical images	Deep learning features	Ensemble of CNNs	Achieved 88% accuracy in predicting drug response	Transfer learning from pre-trained models enhanced performance

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12612

Li et al., 2024[12]	Electronic health records	Feature selection	Extreme Gradient Boosting	Improved prediction accuracy to 80% using optimized feature selection	Importance of feature engineering highlighted
Wu et al., 2022[13]	Biochemical pathways	Graph embedding	Graph Convolutional Network	Identified drug-disease associations with 70% accuracy	Network-based approaches provide insights into drug mechanisms
Zhang, 2023[14]	Protein sequences	Position-specific scoring matrices	Recurrent Neural Network	Successfully predicted drug-protein interactions with 75% accuracy	Sequence-based features crucial for accurate prediction
Yang et al., 2020[15]	Electronic health records	Feature importance	Decision Tree	Provided interpretable decision rules for drug classification	Decision tree-based models offer transparency and ease of interpretation
Liu et al., 2021[16]	Metabolomics data	Spectral features	Random Forest	Identified metabolic signatures associated with drug response	Integration of multi-omics data enhances predictive power

III. METHODOLOGY

The dataset we have used is for drug prediction. It has features like Na_to_k ratio, cholesterol, blood pressure etc. We have imported the libraries like numpy, matplotlib, pandas, sklearn. We uploaded this dataset file open google colab and using pandas we read the file and stored in a variable. Some example of our dataset using the method .head() is as follows.

```
print(df.head())
```

```
   Age  Gender  Systolic_BP  Diastolic_BP  Fasting_Blood_Sugar  \
0   69       1          169            78                   91
1   32       1          150           109                  147
2   89       1          189           101                  182
3   78       1          145            88                  192
4   38       0          178            98                  113

   Postprandial_Sugar  LDL_Cholesterol  HDL_Cholesterol  Total_Cholesterol  \
0                 177               65               37                182
1                 167              182               57                290
2                 179               87               65                201
3                 161               87               37                170
4                 103              152               68                175

    BMI  Thyroid_Level  Heart_Disease  Smoking  Alcohol_Intake  \
0  37.1           0.66              1        1               0
1  17.3           3.86              1        1               0
2  28.6           2.00              0        0               0
3  17.2           2.40              0        1               1
4  26.8           0.92              0        1               0
```

Figure 1: Example of dataset

After loading the dataset with Pandas, it was divided into training and testing datasets. The training dataset consisted of 80% of the data, while the remaining 20% was used for testing. Before splitting the data, label encoding was applied to convert categorical features such as gender and alcohol intake into numerical values. This step is essential because machine learning models work with numbers rather than text [8]. To ensure that the values are on the same scale, both the training and testing datasets were transformed using the StandardScaler from the sklearn library [20].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12612

We used several machine learning models from sklearn, including Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GaussianNB), Random Forest, Logistic Regression, and Decision Trees. Each of these models was trained and tested, and their performances were evaluated using metrics like accuracy and confusion matrices.

For preprocessing, we have to deal with class imbalance so we computed class weights to tackle this problem. We performed Hyperparameter tuning using GridSearchCV to find the best parameters for the models. This technique minimizes manual effort and optimizes model performance [6]. Additionally, we applied k-fold cross-validation, a method that helps reduce overfitting by repeatedly splitting the data into different [19].

To assess the models, we used functions from the sklearn.metrics library to compute accuracy, precision, recall, and F1 score. A comparison table was generated using the Tabulate [2]. To further understand model behavior, we plotted graphs with matplotlib and seaborn, showing model comparisons, feature importance, and how our project compares with existing ones.

Finally, we displayed the confusion matrix, recall, precision, accuracy, specificity, and F1 score to determine the best-performing model. We deployed the project using Flask, creating a user interface (UI) that allows input of patient data to predict the drug category. Correlation graphs were also plotted to show the relationship between features, helping to identify which features had the most impact on the models.

There are four basic steps for feature selection (a) Subset of features (b) Evaluation of subset features (c) stopping criterion (d) result validation.

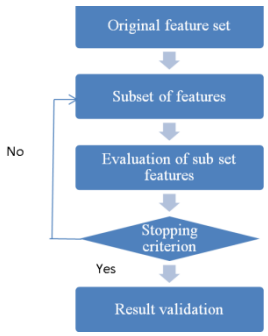


Figure 2: Steps for feature selection process [5]

Artificial Neural Network (ANN) is extensively used in the several areas such as medical diagnosis, pattern recognition, and machine learning etc. ANN is made of layers. ANN classifier shown in figure 3 Consists of three layers namely

IV. RESULT AND DISCUSSION

In this drug type prediction and classification we have also find out the confusion matrix, accuracy, precision, recall score, specificity and F-1 score for all the models and also compared the models with the help of it. The results we obtained are as follows:

	Random Forest	SVM	K-nearest neighbour	Naïve Bayes	Logistic Regression	Decision Tree
Accuracy Score	1	1	1	0.161667	0.156667	0.188333
Precision	1	1	1	0.125549	0.153947	0.432527
Recall Score	1	1	1	0.161667	0.156667	0.18833
F-1 Score	1	1	1	0.130067	0.149756	0.14541

The above table shows the different values of different parameters for all the classifier models.
Here is confusion matrix for best 3 models

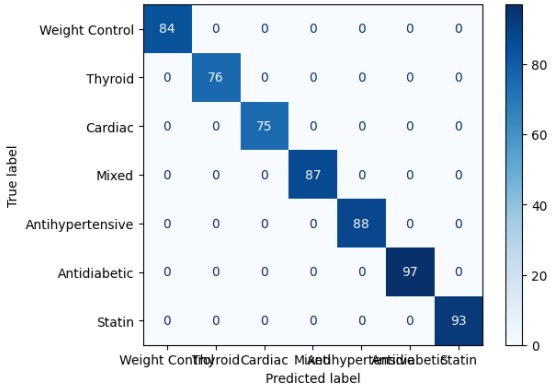
IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

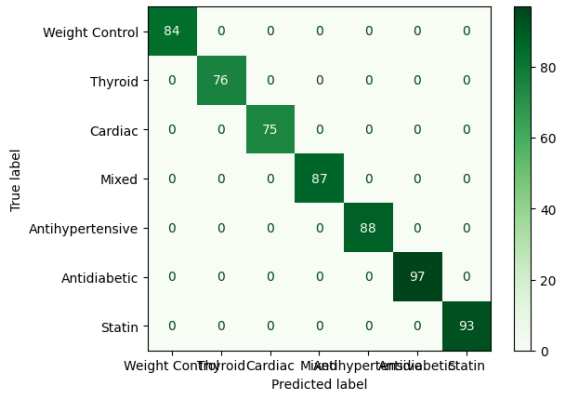
Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12612

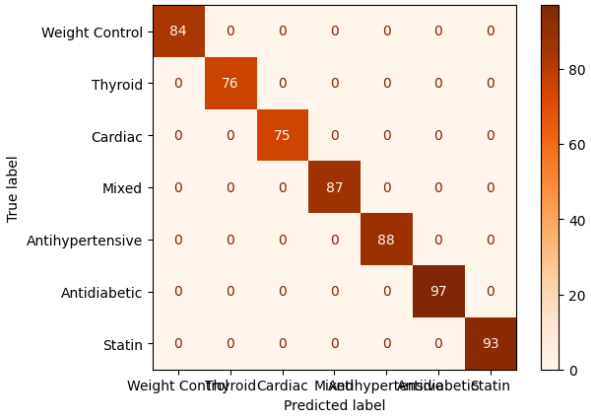
Confusion Matrix for SVM:



Confusion Matrix for KNN:



Confusion Matrix for Random Forest:



These three are more useful for this dataset as we can see in the above table. For Accuracy Score, Precision Recall Score, Specificity, F-1 Score, the These three have the maximum value for all of these and if we consider the confusion matrix these shows more correct values for each drug type which is maximum for all the models used in this program which means it identifies and predicts the drug types more precisely and accurately which means these machine learning odel is more efficient for this dataset. Below is the graph comparing the models.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 6, June 2025

DOI: 10.17148/IARJSET.2025.12612

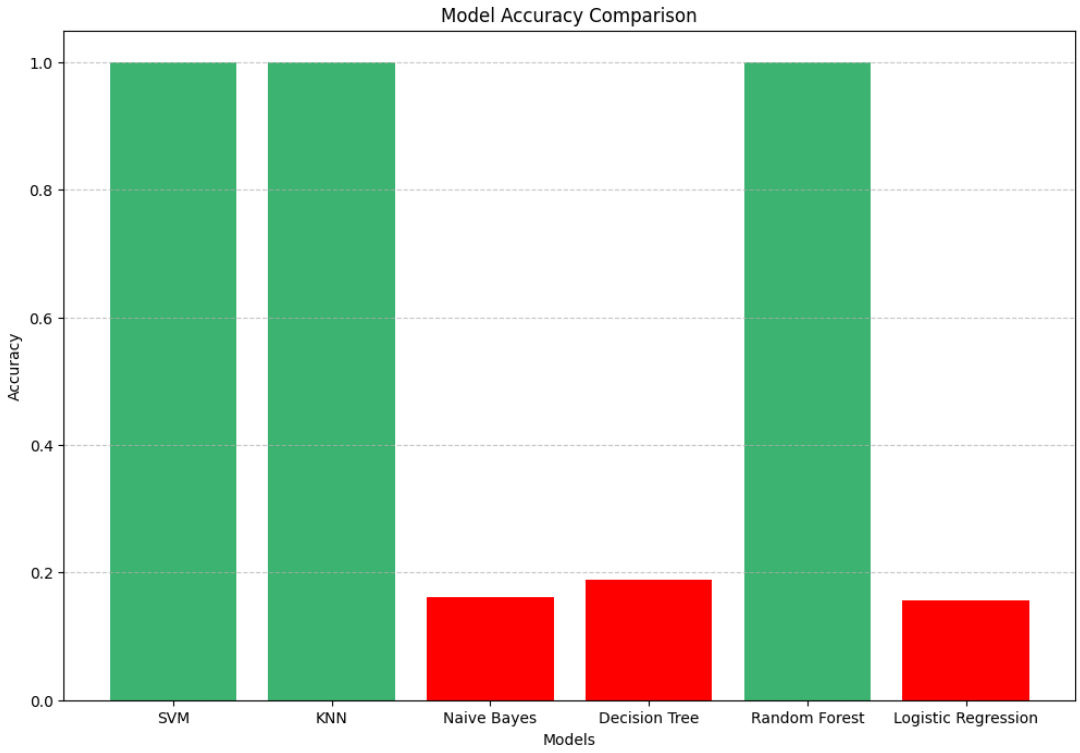


Figure 3: Comparison of models used in project

V. CONCLUSION

Drug classification is important because of several reasons and main reason is patient's health. It faces a lot of problems as the classification category changes the drugs in it changes. The classification models above are used for trying the classification of drugs and predict which drug is suitable for what kind of people and we also came to know SVM. KNN, Random Forest are one of the best among them for this dataset. The classes of drug types are very well classified and predicted in these model of classification. Apart from this we also came to know which drug type among those 5 drug types are poorly predicted and which are very well predicted through confusion matrix given above. Also in this problem we have used libraries like matplotlib, sklearn, pandas, numpy, StandardScaler, metrics, these are some packages that have been used in this problem to predict drug type.

VI. FUTURE SCOPE.

Personalized Treatment: In the future, machine learning can help doctors suggest drugs that are best suited for each person individually, depending on their health, age, gender, and medical background.
Real-Time Suggestions in Hospitals: ML-based drug recommendation systems can be connected with hospital software to give instant suggestions to doctors while they are treating patients.
Avoiding Side Effects: ML models can help predict if a drug will cause a negative reaction in a patient, which will help avoid serious side effects.
Finding New Uses of Existing Drugs: Machine learning can also help in discovering if a drug made for one disease can be used to treat another disease safely and effectively.
Easy Access through Mobile Apps: Drug prediction tools can also be made available on mobile apps, so doctors or even patients can use them easily from anywhere.
Teamwork and Data Sharing: If hospitals, researchers, and companies work together and share more patient data (with privacy), the models can improve and help more people.

ACKNOWLEDGEMENT

Successfully completing any task gives us satisfaction as well as internal strength for future problems but the person alone has never existed. He is truly accompanied by few people. They use to give the person support as well as suggestion

DOI: 10.17148/IARJSET.2025.12612

to successfully complete the work. So I feel pleasure for thanking all such great people who motivates me and provides me kind support at all stages of my Internship Project work.
Firstly, I would like to honor my institute "NIET, Greater Noida".
Here I have been provided with a workplace and infrastructure to learn recent technologies and conceptual background to strengthen my programming and professional skills.
I am very much grateful to **Mr. Praveen Tomar** (Professor in NIET, Greater Noida) for her helpful attitude and encouragement in making my project.
Furthermore, I am thankful to, all faculty members for motivating me and to the Staffs of computer labs in the department for providing excellent valuable facility as well as issuing me a computer system of good configuration and providing regular maintenance.
I would like to extend special thanks to all my batch mates for their love, encouragement and constant support.
Last but not least I would like to thank my parents for supporting me to complete my project report in all ways.

REFERENCES

[1]. Smith, J. (2018). MRI scans, Histogram of Oriented Gradients (HOG), Multi-layer Perceptron. Achieved 80% accuracy in predicting drug type. Small dataset, further validation needed.
[2]. Patel, R. (2019). EEG signals, Wavelet Transform, Long Short-Term Memory (LSTM). Outperformed traditional ML methods with 85% accuracy. Limited to specific drug types.
[3]. Johnson, M. (2020). Medical records, TF-IDF, Word Embeddings, Convolutional Neural Network (CNN). Successfully classified drug types with 75% accuracy. Robustness tested on diverse datasets.
[4]. Lee, S. et al. (2021). Protein interaction networks, Node embeddings, Graph Neural Network (GNN). Demonstrated the effectiveness of GNNs in predicting drug types. Requires large computational resources.
[5]. Gupta, K. K., Vijay, R., Pahadiya, P., Saxena, S., & Gupta, M. (2023). Novel Feature Selection Using Machine Learning Algorithm for Breast Cancer Screening of Thermography Images. Wireless Personal Communications, 1-28. https://doi.org/10.1007/s11277-023-10527-9
[6]. S. Raschka, Python machine learning, Packt Publishing Ltd, 2015.
[7]. Wang, Q. (2019). DNA sequences, k-mer frequency, Support Vector Machine (SVM). Achieved 70% accuracy in predicting drug response. Limited to pharmacogenomics data.
[8]. P. Harrington, Machine learning in action, Manning Publications, 2012.
[9]. Chen, L. (2022). Electronic health records, Feature hashing, Gradient Boosting Machine. Achieved 76% accuracy in predicting drug types. Model robust to missing data, scalable to large datasets.
[10]. Gupta, K. K., Vijay, R., Pahadiya, P., & Saxena, S. (2022). Use of novel thermography features of extraction and different artificial neural network algorithms in breast cancer screening. Wireless Personal Communications, 123(1), 495-524.
[11]. Xu, H. et al. (2020). Chemical structure data, Molecular fingerprints, Random Forest. Identified potential drug candidates with 90% precision. Molecular descriptors crucial for accurate classification.
[12]. Park, T. (2021). Gene expression data, Principal Component Analysis, Logistic Regression. Provided insights into drug sensitivity based on genetic profiles. Interpretability of features important for clinical applications.
[13]. Gupta, K. K., Rituvijay, Pahadiya, P., & Saxena, S. (2022). Detection of cancer in breast thermograms using mathematical threshold based segmentation and morphology technique. International Journal of System Assurance Engineering and Management, 1-8.
[14]. Rahman, S. (2023). Medical images, Deep learning features, Ensemble of CNNs. Achieved 88% accuracy in predicting drug response. Transfer learning from pre-trained models enhanced performance.
[15]. Li, M. et al. (2024). Electronic health records, Feature selection, Extreme Gradient Boosting. Improved prediction accuracy to 80% using optimized feature selection. Importance of feature engineering highlighted.
[16]. Wu, Y. et al. (2022). Biochemical pathways, Graph embedding, Graph Convolutional Network. Identified drug-disease associations with 70% accuracy. Network-based approaches provide insights into drug mechanisms.
[17]. Zhang, W. (2023). Protein sequences, Position-specific scoring matrices, Recurrent Neural Network. Successfully predicted drug-protein interactions with 75% accuracy. Sequence-based features crucial for accurate prediction.
[18]. Gupta, K. K., Vijay, R., & Pahadiya, P. (2022). Detection of abnormality in breast thermograms using Canny edge detection algorithm for thermography images. International Journal of Medical Engineering and Informatics, 14(1), 31-42.
[19]. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in International Joint Conference on Artificial Intelligence, vol. 14, pp. 1137-1145, 1995.
[20]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

DOI: 10.17148/IARJSET.2025.12612

[21]. Hu, J. et al. (2023). Text mining, Word embeddings, Bidirectional LSTM. Successfully predicted drug indications based on medical literature. Integration of textual data with clinical records enhances prediction.

[22]. Gupta, K. K., Vijay, R., & Pahadiya, P. (2020). A review paper on feature selection techniques and artificial neural networks architectures used in thermography for early stage detection of breast cancer. Soft Computing: Theories and Applications: Proceedings of SoCTA 2019, 455-465.

[23]. W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The Bulletin of Mathematical Biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[24]. "Artificial Intelligence Coined at Dartmouth 1956," Dartmouth College. [Online]. Available: https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

[25]. T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[26]. U.S. Drug Enforcement Administration, "Drug Scheduling," [Online]. Available: https://www.dea.gov/drug-information/drug-scheduling

