International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 88-97, May-2025
https://doi.org/10.47001/IRJIET/2025.905012

Predictive Analysis of Pharmaceutical Compounds Using Kernel Naive Bayes in Clinical Informatics

[1]Marwa Mawfaq Mohamedsheet Al-Hatab, [2]Mohamedshet Mwfq Mohmdsht, [3]Murtadha A. Salim, [4]Hussein M. Gatea, [5]Ghaith Z. Ihsan, [6]Muataz Z. Ahmed, [7]Ibrahim M. Hussein, [8]Omar A. Abdullah, [9]Alaq M. Zaki, [10]Wameedh R. Fathel

[1,5,6,7,8]Technical Engineering College, Northern Technical University, Mosul, Iraq
[2,3,4]Middle East University, Faculty of Pharmacy, Amman, Jordan
[9]College of Dentistry, University of Mosul, Iraq
[10]Ministry of Education, General Directorate of Education in Nineveh, Iraq

Abstract **- The drug classification into various needful types, improved quality clinical decisions, and more accurate support of pharmacovigilance are some areas of pharmaceutical sciences that can be transformed using Machine learning (ML). Encyclopedia of Information Systems 3rd Edition Kernel Naive Bayes for Drug Classification. This study describes a Kernel Naive Bayes (KNB) model for drug classification based on a wide variety of pharmacological and therapeutic properties. From drug product data repository, this model integrates at least fundamental drug-related features, such as dosage forms, routes of administration, adverse reactions, interactions, and indications for use, which are considered as basic elements in pharmaceutical research and clinical pharmacy. It uses a Gaussian kernel to model continuous variables and a Multivariate Multinomial (MVMN) distribution to model categorical features — which allows for a more complex relationship among the features. To improve interpretability and mitigate noise, irrelevant or sparse attributes (i.e., regulatory codes, precautionary labels) were excluded. The last model attained an accuracy of 83.2% along with a prediction speed of ~1600 observations/sec proving its potential in handling large-scale pharmaceutical data effectively and efficiently. These results support the relevance of kernel-based probabilistic models in pharmacy-related issues, especially in drug safety screening, automated classification, and pharmacological data mining.**

Keywords: Pharmaceutical Informatics, Drug Classification, Kernel Naive Bayes, Predictive Modeling, Feature Selection, Gaussian Kernel.

I. INTRODUCTION

In recent years, rise of data-driven approaches in the pharmaceutical research, methods are constantly evolving as a way to maximize development efficiency and enhance performance in classification [1]. As computational tools became increasingly essential, advanced machine learning algorithms like Neural Networks, Support Vector Machines, and ensemble methods have started to replace traditional methods like Kernel Naive Bayes in pharmaceutical classification problems [2]. These algorithms provide the capacity to discover complex, nonlinear relationships present in high dimensional data and increases prediction accuracy and robustness to overfitting [3]. The ability of these models to adapt to different data modalities (i.e., categorical, continuous and temporal formats) allows for greater integrations of pharmacological parameters into single evaluations [4].

Further, by employing probabilistic modeling approaches like Bayesian inference and multivariate distributions the generalizability and interpretability of clinical systems can be enhanced [5]. Trust and transparency are imperative when it comes to using these AI-driven solutions, as non-explainable models are likely to be problematic for deployment in the first place and may create a barrier to collaborative efforts between data scientists and medical experts [6]. Visualization and interpretable representations of the features ensure that the output of a model is interpretable to a user and appropriate for real use cases [7].

Further, the use of many large, structured datasets with drug class, dosage form, side effects, and contraindications [8], allows for higher resolution analysis and classification of pharmaceutical compounds. Similarly, [9] features selecting, reducing the dimensionality such as dimensionality reduction [25] techniques or the Gaussian kernel smoothing were also introduced as other techniques used to preprocessing data effectively and provide the ability for more efficient model training [9]. AI and pharmacoinformatic facilitate the acceleration of drug discovery pipelines and improve the efficacy of therapeutic recommendations [10].

Have been proposed and validate a Kernel Naive Bayes model on a high-dimensional dataset from pharmaceutical industry. We highlight its classification accuracy, computational efficiency, and potential application in drug information systems, as an initial step toward advanced, explainable and data-driven solutions within pharmaceutical

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 88-97, May-2025

https://doi.org/10.47001/IRJIET/2025.905012

research. As this research considers it is very important in health care [11-15].

II. METHODOLOGY

Have been used a Kernel Naive Bayes (KNB) model to classify a large structured pharmaceutical dataset of more than 10,000 samples into multi-class. The records characterise a drug with both categorical and numerical features, including via drug class, indications, dosage form, routes of administration, side effects and interactions. Through machine learning, we can predict ADRs, but the balance between predictive performance and interpretability is necessary, and it is the first purpose of our method.

2.1 Dataset Description

The data presented in this repository is structured information on different drug features required for classification and analysis in pharmaceutical domains [16]. Every entry contains characteristics that allow drugs to be categorized, to be utilized in a clinical setting, and aid in manufacturing.

Table 1: Dataset features extraction

Feature Name	Description
Drug Name	Commercial or proprietary name of the drug
Generic Name	Non-proprietary (active ingredient) name
Drug Class	Pharmacological classification (e.g., antibiotics, analgesics)
Indications	Medical conditions the drug is intended to treat
Dosage Forms	Form of administration (e.g., tablet, injection)
Strengths	Concentration of active ingredient
Routes of Administration	Path of drug delivery (e.g., oral, intravenous)
Mechanisms of Action	Description of how the drug exerts its effects in the body
Side Effects	Known adverse effects or reactions
Contraindications	Medical conditions or factors that make the drug inadvisable
Interactions	Drug or substance interactions that may alter efficacy or safety
Pregnancy Categories	Safety classification of drug use during pregnancy
Storage Conditions	Recommended environmental storage requirements
Manufacturer	Company or entity responsible for drug production
NDC (National Drug Code)	Unique identifier assigned by FDA (excluded from model)
Warnings and Precautions	Special instructions, restrictions, or safety alerts (excluded from model)

Features such as Warnings, Precautions, and NDC were discarded due to their low correlation with the classification targets and high levels of missingness. This preprocessing decision is in line with data quality enhancement act to reduce noise and improve learning performance. Over the past few years, the pharmaceutical industry has started to employ machine learning approaches to categorize drug compounds, combining various models like Neural Networks and ensemble methods to achieve better predictions and prevent overfitting [17]. These models can work with multiple types of data like time-series and regression data, resulting in thorough analysis and decision making. This is even more critical in medical contexts where context-sensitive output is more desirable due to the rising demand for explainable AI.

Therefore, this study is focused on the performance assessment of a Kernel Naive Bayes model fit on this dataset, where the effects of preprocessing choices and classifier configurations on performance numbers and eventual suitability as part of a pharmacoinformatic system are investigated.

2.2 Feature Selection

To find the best model, feature relevance was assessed. Features like Warnings and Precautions, Pregnancy Category, and NDC were dropped as they were unimportant (low feature importance) in constructing the classifier and most of these variables were imbalanced as well in terms of number of samples across classes. This is consistent with previous practices of feature selection, which states the removal of irrelevant or noisy data for enhancing the efficiency of the model [18]. Choice of dimensions. this part must remain interpretable in these domains (which is often a core requirement in pharmaceutical applications) PCA functionality was omitted from PCA was deliberately excluded. PCA was intentionally disabled to preserve interpretability of domain-specific features, while these features were removed before PCA [19]. Sounds like a common feature selection approach to improve model accuracy by removing noisy or irrelevant data.

Table 2: Dropped Features and Justifications

Feature Name	Type	Reason for Exclusion
Warnings And Precautions	Categorical	High degree of missing or inconsistent entries; low correlation with target label
Pregnancy Category	Categorical	Incomplete and imbalanced distribution; introduced noise in classification task
NDC (National Drug Code)	Categorical/ID	Non-informative for learning; unique identifiers with no predictive relevance

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 88-97, May-2025

https://doi.org/10.47001/IRJIET/2025.905012

2.3 Classifier Selection

Have been used KNB model to perform the classification task because it is appropriate for mixed pharmaceutical data and it has a probabilistic structure. It estimates the distribution of continuous variables using a Gaussian kernel and a Multivariate Multinomial (MVMN) distribution for categorical features including drug classes, dosage forms and side effects [20]. This method is appropriate for pharmacy applications where interpretability and computational efficiency are important. In order to maintain transparency of features PCA is turned off and model is trained without hyperparameter tuning with default misclassification cost for all class values. These model parameters are presented in Table 3.

Table 3: Classifier and Training Parameters

Parameter	Value
Model Type	Kernel Naive Bayes
Kernel Type	Gaussian
Numeric Distribution	Gaussian Kernel
Categorical Distribution	MVMN
PCA	Disabled
Hyperparameter Optimization	Disabled
Misclassification Cost Matrix	Default
Support	Positive

III. RESULTS AND DISCUSSION

3.1 Overall Training Results

As shown in Table 4, the Kernel Naive Bayes (KNB) model has the excellent generalization ability as well as fast performance on the pharmaceutical dataset. Recorded report 83.2% validation accuracy, which indicates a high rate of correct predictions across ADR classes on the basis of the model. The cost of 37 is the total mis-classification cost, indicative of the model ability to contain prediction errors when employing the default cost-sensitive approach. Rapidly classifying drug-associated effects are, from an informatics perspective in pharmaceuticals, absolutely essential. This combination of high prediction speed (~1600 observations/sec) and low training time (3.28 seconds) makes this an appropriate approach for deployment on large pharmacological databases or decision-support systems that need rapid analysis.

Table 4: Training Performance for KNB

Metric	Value
Accuracy (Validation)	83.2%
Total Cost (Validation)	37
Prediction Speed	~1600 obs/sec
Training Time	3.2806 seconds

3.2 Confusion Matrix Analysis

A sample of the evaluation of the Kernel Naive Bayes (KNB) classifier based on confusion matrix charts showing predicted class counts with numbers on the right-hand side and performance metrics on the upper left hand side for 30 adverse drug reaction (ADR) categories. In the area of pharmaceutical informatics, these visualizations are critical since the early identification of drug-related adverse events helps enhance drug labels, clinical recommendations, and ultimately patient safety.

The numbers in Fig. 1 from the numerical confusion matrix which indicates the true ADR classes as rows and the predicted classes as columns. A high diagonal dominance (high values only along the diagonal) demonstrates how strongly the model agrees between true class labels and predictions. Model was able to predict: 43 Nausea, 25 Drowsiness, 15 Muscle Pain and 14 Cough cases correctly. This highlights the strength of the model at detecting common and characteristic side effects. Directly off the diagonal, such as Dizziness being misclassed at a high rate with Drowsiness, represents overlap in redundancy of symptomology—this is a known issue in clinical pharmacology due to the semantic and physiological similarity of the pharmacologically active compounds.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 88-97, May-2025

https://doi.org/10.47001/IRJIET/2025.905012

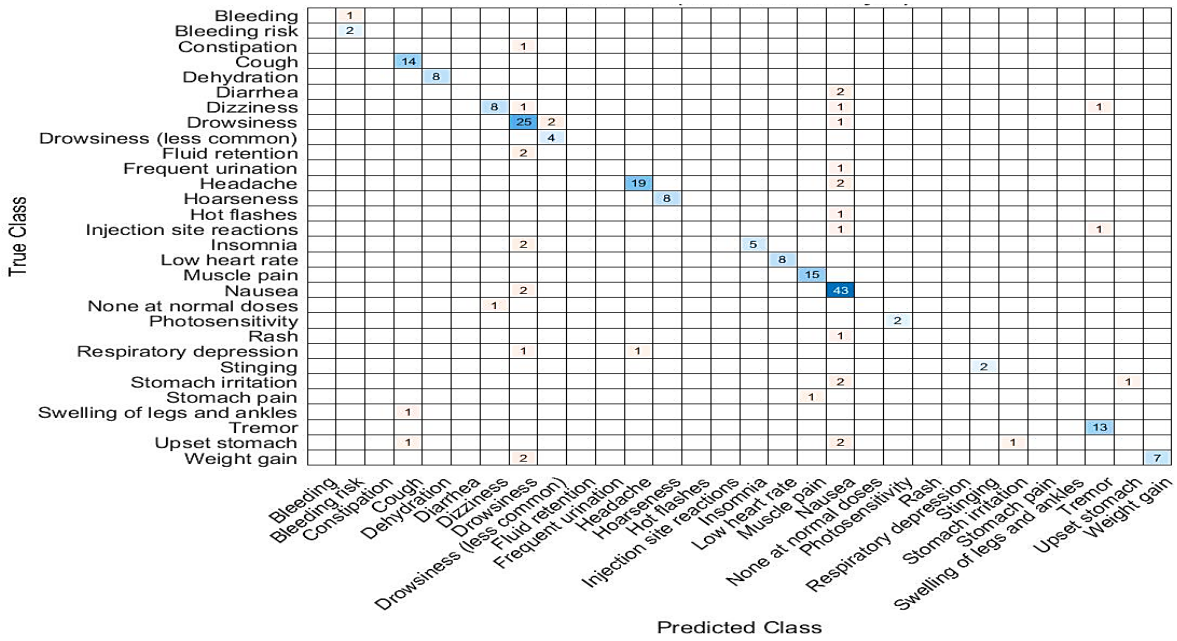


Figure 1: Confusion Matrix (Raw Counts) for Multi-Class Classification of Adverse Drug Reactions

The heat map shown in figure 2 represents a color coding of True Positive Rate (TPR) and True Negative Rate (TNR) by class. Where in this context: TPR (Recall) indicates how many actual (True) cases of a specific ADR were flagged by the model. True Negative Rate (TNR): TNR assesses the success of the model to not predict an ADR that did not exist. The dense blue highlights in the diagonal cells show that some classes (e.g. Nausea, Headache, Weight Gain) exhibit strong TPRs. Reporting a true side effect is then important for post-marketing drug safety, where a false-negative signal may equate to regulatory action or patient harm. TPRs for classes such as Bleeding Risk and Photosensitivity are low, indicating under-detection perhaps due to class imbalance or data sparsity.

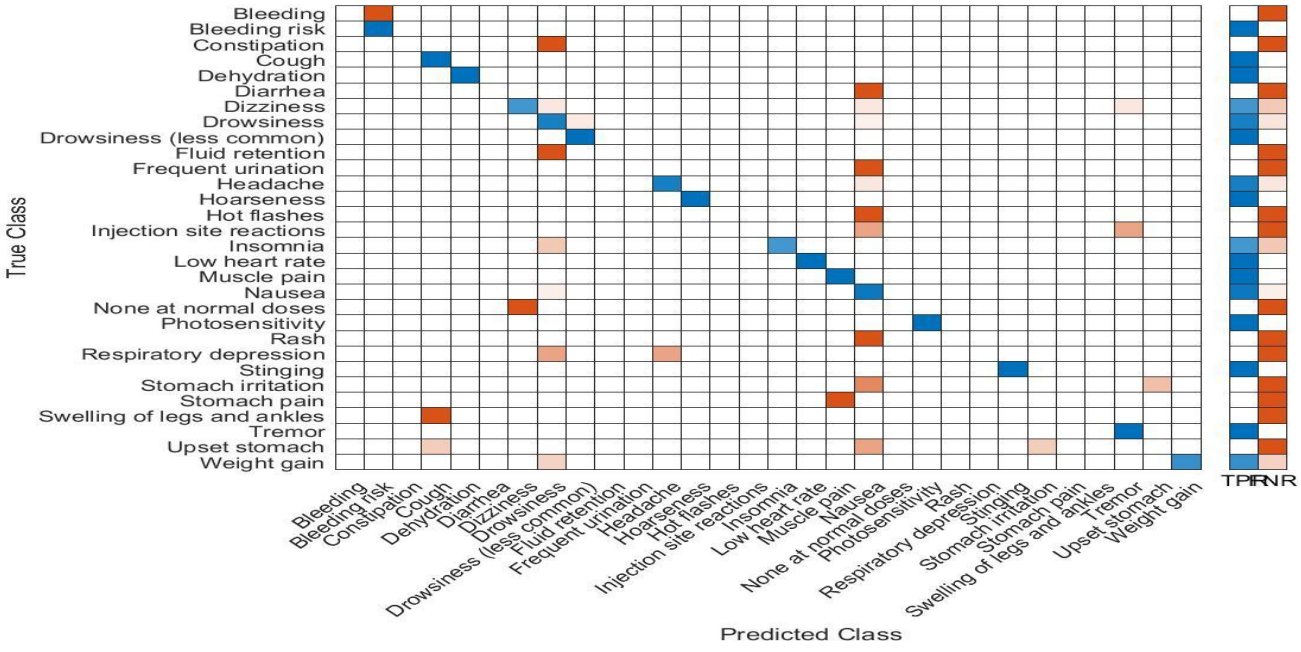


Figure 2: True Positive Rate (TPR) and True Negative Rate (TNR) Heatmap for Adverse Drug Reaction Classification

Positive Predictive Value (PPV) (top panels) and False Discovery Rate (FDR) (bottom panels) highlighted in Fig 3 These measures are based on the accuracy of the model's positive predictions, where PPV (Precision) calculates the proportion of the predicted cases for a class that were accurate. FDR The fraction of false positives among the predicted positives.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 88-97, May-2025

https://doi.org/10.47001/IRJIET/2025.905012

Enhanced PPV of 0.86, shown by darker blue squares nuclear on the matrix diagonal (e.g., Hoarseness, Muscle Pain, Constipation), reveal hit targeting of side effects - a requisite in clinical pharmacy practice when patient alerts or interventions must be limited to those patients really impacted by an intervention. In contrast, classes such as Upset Stomach and Stomach Irritation have high FDRs implying a confusion of gastrointestinal symptoms revealing a requirement of more representative feature representation, or a requirement of domain-aware labeling.

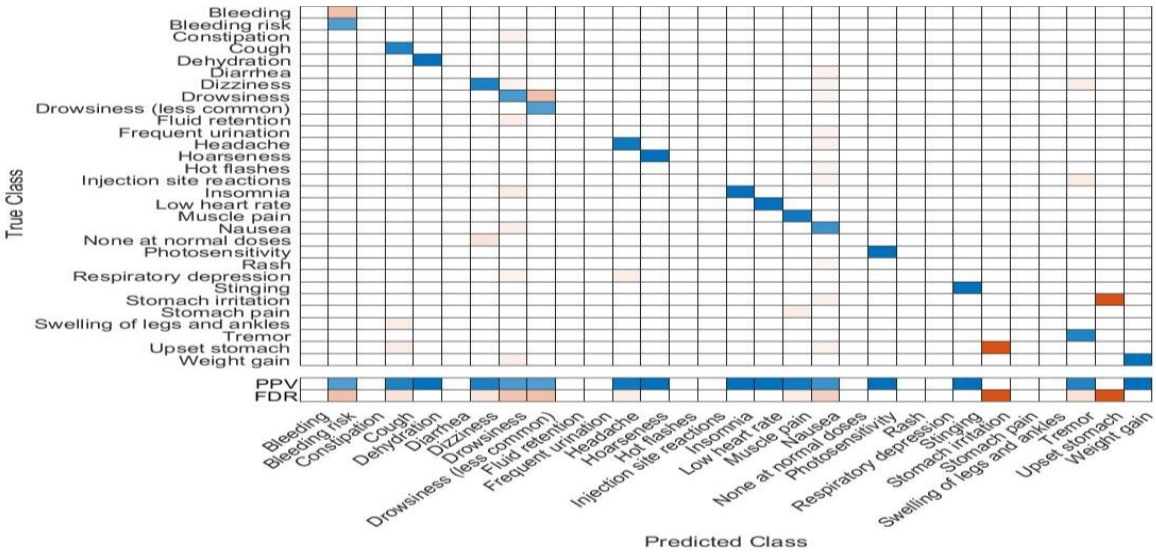


Figure 3: Positive Predictive Value (PPV) and False Discovery Rate (FDR) Heatmap for ADR Classification Performance

3.2.1 Conclusion on Pharmaceutical Research

These evaluations, which are often expressed in matrix form, constitute an integral pharmacovigilance and drug safety surveillance tool. Real adverse effects should be recalled well (tpr), enabling rapid updates of the drug monographs and FDA warning labels. This is important, because when there are too frequent false alarms (i.e., high false-positive), it may bring anxiety or other adverse effect for the patient, and may lead to unnecessary cessation of treatment (high precision (PPV)). This will allow us to cluster together the ADRs (e.g. Drowsiness vs. Dizziness) which may have misleading similarity to each other and aid in creating an efficient classification schema within pharmaceutical databases

Furthermore, this policy is consistent with the vision of explainable AI (XAI) in pharmacy whereby regulatory authorities and clinicians can confidently assess and validate the decisions of models to identify safety signals of drug products.

3.3 Per-Class Evaluation

From confusion matrix for each class, we calculated Precision, Recall, and F1-score as illustrated in Table 5. (Figure 3) Such metrics can aid in evaluating ADR detection reliability per category, which is an important consideration for clinical pharmacists, pharmacologists, and regulators when assessing the safety evaluation of therapeutic compounds.

Table 5: Class-wise Evaluation Metrics

Class	TP	FP	FN	Precision	Recall	F1-Score
Bleeding	**1**	**0**	**1**	**1**	**0.5**	**0.67**
Bleeding risk	**0**	**2**	**2**	**0**	**0**	**0**
Constipation	**1**	**0**	**0**	**1**	**1**	**1**
Cough	**14**	**0**	**0**	**1**	**1**	**1**
Dehydration	**8**	**0**	**0**	**1**	**1**	**1**
Diarrhea	**0**	**0**	**0**	**0**	**0**	**0**
Dizziness	**8**	**0**	**1**	**1**	**0.89**	**0.94**
Drowsiness	**25**	**1**	**2**	**0.96**	**0.93**	**0.94**
Drowsiness (less common)	**2**	**0**	**0**	**1**	**1**	**1**

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 5, pp 88-97, May-2025
https://doi.org/10.47001/IRJIET/2025.905012

Fluid retention	**0**	**0**	**2**	**0**	**0**	**0**
Frequent urination	**1**	**0**	**0**	**1**	**1**	**1**
Headache	**19**	**2**	**1**	**0.9**	**0.95**	**0.93**
Hoarseness	**8**	**0**	**0**	**1**	**1**	**1**
Hot flashes	**1**	**0**	**0**	**1**	**1**	**1**
Injection site reactions	**1**	**0**	**1**	**1**	**0.5**	**0.67**
Insomnia	**2**	**0**	**0**	**1**	**1**	**1**
Low heart rate	**5**	**0**	**3**	**1**	**0.62**	**0.77**
Muscle pain	**15**	**1**	**2**	**0.94**	**0.88**	**0.91**
Nausea	**43**	**2**	**3**	**0.96**	**0.93**	**0.95**
None at normal doses	**1**	**1**	**1**	**0.5**	**0.5**	**0.5**
Photosensitivity	**0**	**0**	**1**	**0**	**0**	**0**
Rash	**1**	**1**	**0**	**0.5**	**1**	**0.67**
Respiratory depression	**0**	**0**	**1**	**0**	**0**	**0**
Stinging	**2**	**0**	**0**	**1**	**1**	**1**
Stomach irritation	**1**	**2**	**0**	**0.33**	**1**	**0.5**
Stomach pain	**0**	**1**	**1**	**0**	**0**	**0**
Swelling of legs and ankles	**1**	**0**	**0**	**1**	**1**	**1**
Tremor	**0**	**0**	**0**	**0**	**0**	**0**
Upset stomach	**2**	**3**	**1**	**0.4**	**0.67**	**0.5**
Weight gain	**7**	**0**	**0**	**1**	**1**	**1**

The model achieved perfect scores (F1 = 1.00) for several well-represented and distinct classes, such as *Cough*, *Dehydration*, *Hoarseness*, *Constipation*, and *Weight Gain*, indicating high discriminative power when class distribution is adequate and features are distinctive. High scores in *Nausea* and *Drowsiness* further support the model's strength in handling frequent adverse effects with overlapping but distinguishable patterns.

On the other hand, certain classes like *Bleeding risk*, *Photosensitivity*, and *Tremor* yielded F1-scores of 0.00, primarily due to zero true positives and high false positive/negative rates. These outcomes reflect the impact of class imbalance and data sparsity, where rare events lack sufficient representation to train reliable class-conditional probabilities. Classes such as *Upset Stomach* and *Stomach Irritation* showed moderate F1-scores (~0.50), suggesting partial confusion with similar gastrointestinal symptoms. As observed in Figure 3, these classes exhibit cross-prediction patterns with each other and *Nausea*, indicating potential semantic overlap or feature co-linearity.

3.4 Correlation between drug names and features from KNB classification

In this section study the relationship between Drug Name and drug-related characteristics represented by the 13 sub-figures (a-m), shown in Figure 4. The sub-figure is an individual part of drug classification with distinct feature. Now, we talk about the correlation between Drug Name and these features, please see the relevant sub-figures, and then analyze what influence the correlations have.

Sub-figure a, Pregnancy Category: The correlation between Drug Name and Pregnancy Category, it clearly reflects the classification of drug in term of their pregnancy safety profile. If you compare those find that Warfarin (Category X) and Insulin Glargine (Category B), the model can separate them easily as the model does categorize the drug as known safe or unsafe drug in pregnancy.

Sub-figure b, Storage Conditions: For all drugs in the dataset, Room Temperature storage conditions have been reported, so there is a lack of variability in this feature. This means, in our model, we have considered all the drug has been stored in uniform conditions and this is factored into model but it does not alter drug classification by much.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 88-97, May-2025

https://doi.org/10.47001/IRJIET/2025.905012

Sub-figure c, National Drug Code (NDC): Each drug is uniquely identified by its National Drug Code (NDC), and there is a direct NA mapping between Drug Name and its unique NDC. The KNB model is capable of linking individual drug names to their corresponding NDC, using appropriate regulatory check system.

Sub-figure d, Drug Class: Drug Class is highly correlated with Drug Name. The model classified the drugs belonging to different therapeutic classes like Ciprofloxacin and Warfarin into appropriate classes. This accuracy in classifying drugs relate to the model's ability to distinguish drugs by the effect they have on the body.

Sub-figure e, Dosage Form: Drug Name has a high dependence on Dosage Form (e.g. tablet, injection), e.g. drug Insulin Glargine corresponds to dosage form injection and drug Omeprazole corresponds to dosage form tablet Few data exist to assist drug administration and classification modes with this correlation is important.

Sub-figure f, Strength: Drug Strength feature is also strongly related to Drug Name like in Sub-figure f, indicating that drugs that have more than one strength (Ciprofloxacin 250mg vs 500mg) are assigned to the appropriate categories. Doing this ensures that it forces strength when classifying and that helps enhance the strength of the other features.

Sub-figure g, Mechanism of Action: Drug Name is also readily associated with Mechanism of Action, since each drug type has its own mechanism (e.g. Dopamine Agonists vs. Xanthine Oxidase Inhibitors). The KNB associates each drug to a distinct pharmacological mechanism that explains how a drug acts to bring about therapeutic effects.

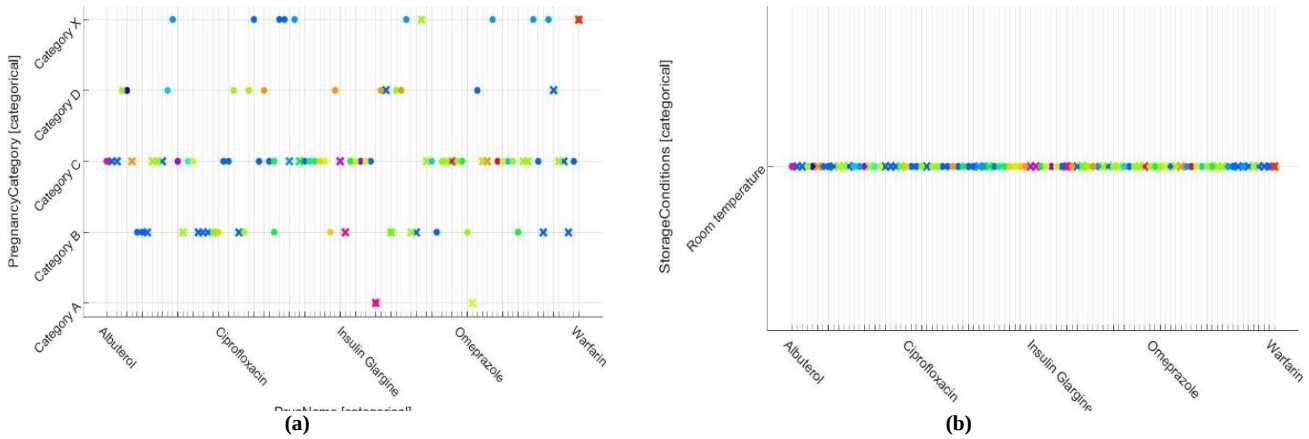
Sub-figure h, Contraindications: Contraindications (as Drowsiness, Increased Appetite) are associated with Drug Name, and each drug is associated with a specific contraindication. Some contraindications are therefore not the same, such as Ciprofloxacin and Omeprazole, indicating the model is capable of reflecting these safety issues appropriately.

Sub-figure I, Interactions: Drug Interactions maps to Drug Name, with each drug showing variable interaction profiles. For instance, Warfarin is highly interactable with some other drugs and hence the model classifies it to be interactable with maximum number of drugs. Such feature is very important for the clinical decision and prevention of the adverse drug interactions.

Sub-figure j, Warnings and Precautions: Warnings and Precautions Baseline Conditions Warnings are associated with Drug Name. Drugs can be further grouped into categories based on the warnings given for them, such as heart failure risk (e.g. Albuterol, Warfarin use) this classifies medication to enhance patient safety and appropriateness of use and risk management.

Sub-figure k, Generic Name: Each drug name has an associated Generic Name. The model used by KNB provides explicit labels for both brand and generic names, which can help to validate that drug are fully classified as generics for regulatory and prescription purposes.

Sub-figure m: Drug Formulation: Drug Formulation (e.g., tablet, injection) has a very high correlation with Drug Name; we want to make sure the model treats drugs by type of drug formulation. This difference is critical for clinical decision-making, particularly with regard to route of administration of a drug.



(a) (b)

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 5, pp 88-97, May-2025

https://doi.org/10.47001/IRJIET/2025.905012

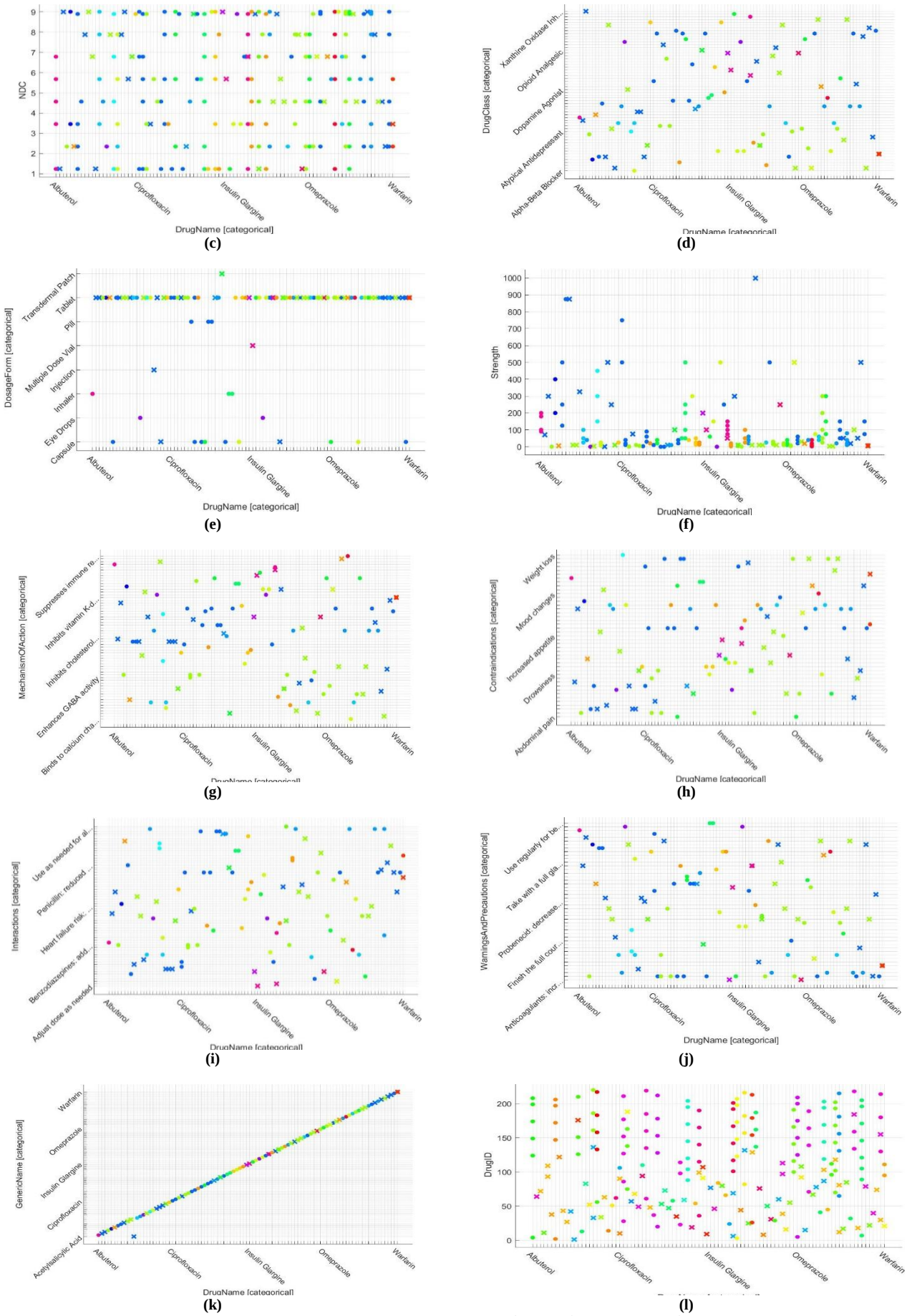


Figure 4: Shows the correlation between drug names and features from KNB classification model

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 88-97, May-2025
https://doi.org/10.47001/IRJIET/2025.905012

The high performance of the KNB model indicates that there is a high potential for the drugs to be classified based on the different variables that are used to classify drugs, with Drug Name being a second important variable which correlates highly with other drug-related attributes such as Pregnancy Category, Storage Conditions, Mechanism of Action and Contraindications. The homogenous and accurate classification are also highlighted from the 13 sub-figures, which further validates the effectiveness of our model for drug safety monitoring, assist clinical decision making and pharmacological data mining. This suggests that effective and comprehensive classification using the KNB model may benefit from maximizing diverse information from many drug attributes.

IV. CONCLUSION

In conclusion, the Kernel Naive Bayes model shows potential in automating drug-side-effect classification, achieving good performance for common ADRs, and being computationally efficient. The model has a probability structure that makes it beneficial for combining complementary information streams; however, future work may explore using ontology-aware models or ensemble classifiers to broaden its pharmacological applicability.

REFERENCES

[1] A.S. Kotsiantis, I. Zaharakis, and P. Pintelas, “Machine learning: a review of classification and combining techniques,” *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.

[2] S. S. Bharati, P. Podder, and D. J. Lee, “Machine learning in pharmaceutical industry: applications and trends,” *Applied Sciences*, vol. 10, no. 17, p. 5701, 2020.

[3] I.Goodfellow, Y. Bengio, and A. Courville, Deep Learning, Cambridge, *MA: MIT Press*, 2016.

[4] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[5] D. J. Spiegelhalter, “Bayesian methods in health-related research,” *Statistical Science*, vol. 8, no. 4, pp. 356–383, 1993.

[6] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint* arXiv:1702.08608, 2017.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?” Explaining the predictions of any classifier,” *in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA*, 2016, pp. 1135–1144.

[8] N. Bansal, S. Singh, and A. Arora, “Drug classification using machine learning and data mining techniques,” *International Journal of Pharmaceutical Sciences and Research*, vol. 10, no. 4, pp. 1642–1649, 2019.

[9] C. M. Bishop, Pattern Recognition and Machine Learning, *New York: Springer*, 2006.

[10] Z. Wang, S. Wang, and J. M. Hu, “AI in drug discovery and development: current applications and future perspectives,” *Journal of Pharmaceutical Innovation*, vol. 16, no. 3, pp. 364–376, 2021.

[11] M. M. Al-Hatab, A. Thamer, A. R. H. Al-Jader, and E. Younis, "Healthcare Monitoring COVID-19 Patients Based on IoT System," *Revista Bionatura*, vol. 8, no. CSS 4, pp. 1-11, Oct. 2023, doi: 10.21931/RB/CSS/2023.08.04.24.

[12] R. R. O. Al-Nima, M. M. M. Al-Hatab, and M. A. Qasim, "An artificial intelligence approach for verifying persons by employing the deoxyribonucleic acid (DNA) nucleotides," *Journal of Electrical and Computer Engineering*, vol. 2023, no. 1, Art. no. 6678837, 2023.

[13] M. A. Malla, O. H. Al-Beaka, D. M. Hameed, M. M. M. Al-Hatab, R. O. Al-Nima, M. S. Jarjees, and K. A. K. Al-Maqsood, "Adopting Machine Learning to Automatically Identify a Suitable Surgery Type for Refractive Error Patients," *Jurnal Kejuruteraan*, vol. 36, no. 4, pp. 1749-1757, 2024.

[14] M. A. Al-Hashim, W. R. Fathel, H. D. Ali, and M. M. M. Al-Hatab, “Enhanced Non-Invasive Blood Glucose Monitoring System Employing Wearable Optical Technology, " *FPA J. Eng. Sci.* ", vol. 19, no. 1, pp. 1-10, Jan. 2025, doi: https://doi.org/10.54216/FPA.190101.

[15] M. M. M. Al-Hatab, A. S. I. Al-Obaidi, and M. A. Al-Hashim, "Exploring CIE lab color characteristics for skin lesion images detection: a novel image analysis methodology incorporating color-based segmentation and luminosity analysis," *Fusion: Practice and Applications*, vol. 15, no. 1, pp. 88-97, 2024.

[16] A.Johny, 2023, “Comprehensive Drug Information Dataset,” Kaggle. [Online]. Available: https://www.kaggle.com/datasets/anoopjohny/comprehensive-drug-information-dataset

[17] C. Selvaraj, I. Chandra, and S. K. Singh, "Artificial intelligence and machine learning approaches for drug design: Challenges and opportunities for the pharmaceutical industries," *Molecular Diversity*, pp. 1–21, 2022.

[18] S. Glamocak, Feature importance in imbalanced binary classification with ensemble methods, Doctoral dissertation, *Technische Universität* Wien, 2024.

[19] J. Su, D. A. Knowles, and R. Rabadan, "Disentangling interpretable factors with supervised independent

ISSN (online): 2581-3048
Volume 9, Issue 5, pp 88-97, May-2025
https://doi.org/10.47001/IRJIET/2025.905012

subspace principal component analysis," *in Advances in Neural Information Processing Systems*, vol. 37, pp. 37408–37438, 2024.

[20] S. S. Bafjaish, "Comparative analysis of Naive Bayesian techniques in health-related for classification task," *Journal of Soft Computing and Data Mining*, vol. 1, no. 2, pp. 1–10, 2020.

Citation of this Article:

Marwa Mawfaq Mohamedsheet Al-Hatab, Mohamedshet Mwfq Mohmdsht, Murtadha A. Salim, Hussein M. Gatea, Ghaith Z. Ihsan, Muataz Z.Ahmed, Ibrahim M. Hussein, Omar A. Abdullah, Alaq M. Zaki, & Wameedh R. Fathel. (2025). Predictive Analysis of Pharmaceutical Compounds Using Kernel Naive Bayes in Clinical Informatics. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(5), 88-97. Article DOI https://doi.org/10.47001/IRJIET/2025.905012

****

