International Journal of Academic Information Systems Research (IJAISR)
ISSN: 2643-9026
Vol. 10 Issue 6 June - 2026, Pages: 64-74

Drug Risk Segmentation Based on Side Effects Severity Using K-Means Clustering

Fatima M. Salman, Samy S. Abu-Naser

Department of Information Technology, Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine

Abstract: *Pharmacovigilance is a critical pillar of modern healthcare, yet the rapid expansion of drug inventories makes manual risk classification increasingly impractical. This paper proposes a validated unsupervised machine learning framework for automated drug segmentation based on the severity of reported side effects. Using a dataset of 11,498 unique pharmaceutical products, we apply a hybrid feature extraction pipeline combining TF-IDF text vectorization, Truncated Singular Value Decomposition (SVD), and a domain-informed severity scoring model. To ensure clinical accuracy, the scoring model was validated by a panel of clinical pharmacists, achieving high inter-rater reliability (Cohen's Kappa = 0.82). K-Means clustering (k=4) partitions the drug portfolio into four distinct risk strata: Low Risk (42.4%), Moderate Risk (35.0%), High Risk (18.4%), and Critical Risk (4.2%). Cluster validity is confirmed through Silhouette, Davies-Bouldin, and Calinski-Harabasz indices, while sensitivity analysis demonstrates the model's robustness against varying weight configurations. Results reveal an empirical "Review Paradox"—where high-risk medications receive superior patient ratings—which analysis of the "Uses" field attributes to selection bias in life-saving treatments, such as oncological agents. The framework's external validity is further established through high concordance with U.S. FDA Black Box Warnings, particularly within the Critical Risk segment, and demonstrates clear alignment with global Anatomical Therapeutic Chemical (ATC) classification systems. By replacing generic token counts with a refined "Residual Side Effect Density" metric, the model maintains high interpretability and scalability. This research provides a robust, explainable methodology for drug safety stratification, offering a scalable adjunct to regulatory oversight and clinical decision support systems.*

Keywords: Drug Segmentation, K-Means Clustering, Pharmacovigilance, Side Effect Severity, TF-IDF, NLP, Machine Learning

1. Introduction

Adverse drug reactions (ADRs) are responsible for a significant proportion of hospital admissions worldwide, representing both a patient safety crisis and a substantial economic burden on healthcare systems [2]. The World Health Organization defines pharmacovigilance as the science dedicated to the detection, assessment, understanding, and prevention of adverse effects related to medicines [1]. Despite the critical importance of risk stratification, current drug classification systems rely predominantly on pharmacological mechanism rather than empirical safety profiles derived from patient experience.

The proliferation of large-scale pharmaceutical databases, combined with advances in Natural Language Processing (NLP) and unsupervised learning, opens new avenues for data-driven drug safety analysis. In this context, clustering algorithms — particularly K-Means — offer an interpretable and computationally efficient approach to grouping drugs based on shared characteristics of their adverse effect profiles.

Figure1: A representative illustration of the pharmaceutical risk stratification process

This study addresses the following research questions:

- Research question 1: Can unsupervised clustering algorithms meaningfully segment pharmaceuticals by side effect severity?
- Research question 2: What is the optimal number of drug risk clusters, and what clinical interpretation do they carry?
- Research question 3: Is there a measurable correlation between pharmacological risk level and patient-reported outcomes?

Our contributions are threefold: (1) a domain-informed severity scoring model for adverse effects, (2) a hybrid NLP-numerical feature pipeline for drug representation, and (3) a validated four-cluster risk taxonomy with clinical and regulatory implications.

1.1 Definitions of Terms

The following terms are used throughout this paper in accordance with their standard definitions in pharmacovigilance and data science literature:

Adverse Drug Reaction (ADR): Any unintended, harmful response to a pharmaceutical product occurring at doses normally used in humans for prophylaxis, diagnosis, or treatment of disease. ADRs are the primary analytical unit in pharmacovigilance and constitute the core textual signal used in this study's segmentation model.

Pharmacovigilance: The science and set of activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problems, as defined by the World Health Organization (WHO, 2002). It serves as the regulatory and clinical framework within which this research is positioned.

Drug Segmentation: The process of partitioning a pharmaceutical portfolio into homogeneous subgroups based on shared characteristics. In this study, segmentation is performed on the basis of side effect severity profiles using unsupervised machine learning, resulting in distinct risk clusters that carry clinical and regulatory meaning.

K-Means Clustering: An unsupervised machine learning algorithm that partitions a dataset into k non-overlapping clusters by iteratively minimizing the within-cluster sum of squared distances to the cluster centroid (Lloyd, 1982). In this study, K-Means is applied with k-means++ initialization and 20 independent restarts to ensure stable convergence.

TF-IDF (Term Frequency–Inverse Document Frequency): A numerical text representation technique that assigns weights to terms based on their frequency within a document relative to their frequency across all documents in the corpus. Terms that are frequent in a specific drug's profile but rare across the dataset receive higher weights, enabling the model to identify discriminative side effect descriptors.

Severity Score (S): A composite numerical index developed in this study to quantify the overall risk level of a drug's adverse effect profile. It is computed as S = 3H + 2M + L, where H, M, and L denote the count of critical, moderate, and mild effect keywords respectively. Scores range from 0 (no documented side effects) to 22 in this dataset.

Silhouette Score: A cluster validity metric that measures how similar each data point is to its own cluster compared to the nearest neighboring cluster (Rousseeuw, 1987). Values range from -1 to +1, where values approaching +1 indicate well-separated, compact clusters. It is one of three metrics used in this study to determine the optimal number of clusters k.

Unsupervised Learning: A category of machine learning in which models are trained on unlabeled data, discovering inherent structure without predefined output categories. This approach is particularly appropriate for pharmaceutical risk stratification tasks where no ground-truth risk labels exist at scale, allowing data-driven discovery of natural groupings based solely on adverse effect characteristics.

Natural Language Processing (NLP): A subfield of artificial intelligence concerned with enabling computers to understand, interpret, and process human language in textual or spoken form. In this study, NLP techniques — specifically TF-IDF vectorization and bigram extraction — are applied to the unstructured Side_effects field to transform free-text adverse event descriptions into structured numerical feature vectors suitable for clustering algorithms.

Singular Value Decomposition (SVD): A linear algebra technique that factorizes a matrix A into three component matrices (U, Σ, V^T) such that $A = U\Sigma V^T$, enabling dimensionality reduction by retaining only the top k singular values. In this study, Truncated SVD is applied to the 200-dimensional TF-IDF matrix to produce a compact 50-dimensional representation, capturing 85.2% of the total explained variance while substantially reducing noise and computational cost. Unlike PCA, Truncated SVD operates directly on sparse matrices without requiring prior mean-centering, making it particularly well-suited for text feature matrices.

K-Means Algorithm (Operational Definition): The K-Means algorithm proceeds through four steps: (1) Initialization — k centroids are seeded using the k-means++ strategy, which selects initial centers with probability proportional to their squared distance from existing centers, reducing sensitivity to poor initialization; (2) Assignment — each data point is assigned to the nearest centroid based on Euclidean distance; (3) Update — each centroid is recomputed as the mean of all points assigned to its cluster; (4) Convergence — steps 2 and 3 are repeated until cluster assignments stabilize. In this study, k is set to 4, and the algorithm is run with 20 independent restarts (n_init=20) to ensure the globally optimal solution is identified.

PCA Projection (Principal Component Analysis): A dimensionality reduction technique that projects high-dimensional data onto a lower-dimensional subspace defined by the directions of maximum variance, known as principal components. In this study, PCA is applied exclusively for two-dimensional visualization purposes — reducing the 55-dimensional feature matrix to two principal components (PC1 and PC2) to produce the scatter plot in Figure 1. It is important to note that PCA is not used in the clustering process itself; the K-Means algorithm operates on the full 55-dimensional feature space. The 2D PCA projection therefore serves as an interpretive aid rather than a modeling step.

2. Related Work

2.1 Text Mining in Pharmacovigilance

Early work in automated ADR detection relied on rule-based extraction from clinical notes and electronic health records. Subsequent efforts applied supervised classifiers to structured drug databases, achieving moderate F1 scores in binary safe/unsafe classification tasks. More recent approaches leverage transformer-based language models such as BioBERT for named entity recognition of adverse events in scientific literature and social media posts [3]. **Comment on the previous study:** While BioBERT demonstrated strong performance in biomedical named entity recognition tasks, its application requires large annotated corpora for fine-tuning and is computationally intensive, limiting accessibility in resource-constrained environments. Moreover, the model focuses on free-text literature extraction rather than structured drug databases, reducing its direct applicability to large-scale drug portfolio classification. In contrast, our framework deliberately prioritizes interpretability and scalability, offering a practical solution suitable for regulatory and clinical deployment contexts where explainability is a key requirement.

2.2 Unsupervised Drug Clustering

Unsupervised approaches to drug classification have explored chemical structure similarity, gene expression profiles, and clinical indication overlaps [4]. However, clustering based on textual side effect descriptions remains underexplored. Most existing work treats ADR information as a secondary feature rather than the primary signal for risk stratification. This paper addresses this gap by placing side effect severity at the center of the segmentation task. **Comment on the previous study:** This systematic review covers 72 studies applying statistical and machine learning methods to ADR detection, and highlights that while drug similarity-based approaches have been explored, they rely predominantly on binary presence-absence encoding of adverse events without differentiating severity levels. The absence of severity weighting produces flat representations that conflate clinically distinct risk

profiles. Furthermore, most reviewed methods target ADR signal detection rather than drug risk segmentation at scale. Our framework directly addresses this gap by introducing a tiered severity scoring model and applying unsupervised clustering to classify entire drug portfolios.

2.3 Severity Scoring Approaches

The WHO-UMC causality assessment system and the Naranjo algorithm provide frameworks for individual ADR severity evaluation [5]. These systems, however, require clinical expertise and are not readily scalable to large drug portfolios. We propose a lightweight keyword-based severity model that approximates clinical risk stratification and enables automated processing of thousands of drug profiles. **Comment on the previous study:** The Naranjo scale remains one of the most widely cited tools for adverse drug reaction causality assessment, and its clinical validity is well established. However, it was designed for individual patient-level evaluation requiring structured clinical data inputs including dechallenge and rechallenge information, prior ADR history, and laboratory confirmation. These requirements make it inherently unsuitable for automated large-scale classification based on product-level descriptions. Additionally, the scale produces a causality score rather than a severity stratification, which is the primary dimension of interest in our study. Our domain-informed severity scoring model directly addresses this limitation by introducing a tiered, keyword-driven weighting scheme that can be applied systematically across entire drug portfolios without requiring case-level clinical data.

3. Methodology

3.1 Dataset

The dataset comprises 11,825 pharmaceutical records sourced from the 1mg drug platform (India), publicly available via Kaggle [9]. After removing 327 duplicate entries by medicine name, 11,498 unique drugs were retained. Each record includes nine features: Medicine Name, Composition, Uses, Side Effects, Image URL, Manufacturer, and patient review percentages (Excellent, Average, Poor). The dataset covers 759 manufacturers, with Sun Pharmaceutical Industries Ltd being the most represented (820 drugs).

3.2 Severity Scoring Model

We define a three-tier severity hierarchy informed by clinical pharmacology literature:

- Tier 1 — Critical Effects (weight: 3): includes death, cardiac arrest, stroke, seizures, anaphylaxis, hemorrhage, organ failure, and malignancy.
- Tier 2 — Moderate Effects (weight: 2): includes hypertension, depression, hepatitis, pancreatitis, dyspnea, and cardiac dysrhythmias.
- Tier 3 — Mild Effects (weight: 1): includes nausea, vomiting, diarrhea, headache, rash, and fatigue.

A composite severity score S is computed for each drug as: $S = 3*H + 2*M + L$, where H, M, and L represent the count of critical, moderate, and mild effect keywords present in the drug's side effect description, respectively. The linear weighting scheme (3, 2, 1) was selected to represent a clear hierarchy of clinical risk while maintaining feature balance. This distribution ensures that a single 'Critical' effect carries the same analytical weight as three 'Mild' symptoms, providing sufficient signal for life-threatening reactions to influence the clustering process without being statistically overwhelmed by the higher frequency of minor side effects.

3.2.1 Expert Validation and Reliability

While the severity scoring model is domain-informed, an additional validation layer was integrated to ensure clinical accuracy. A panel of clinical pharmacists reviewed the keyword-to-severity mapping (H, M, L), and their classifications were compared against the automated model. The resulting Inter-rater Reliability (IRR), measured via Cohen's Kappa, demonstrated a high level of agreement ($\kappa = 0.82$), confirming that the keyword-based heuristic effectively mirrors professional clinical judgment in risk stratification.

3.3 Feature Engineering

The feature extraction process follows a hybrid architecture, integrating unstructured text signals with domain-specific numerical indices to create a robust multi-dimensional representation for each pharmaceutical product.

Textual Vectorization and Dimensionality Reduction: Adverse effect descriptions from the Side_effects field are first transformed into numerical vectors using TF-IDF vectorization with bigram support *ngram_range=(1, 2)* This captures both individual symptoms and complex descriptors (e.g., "cardiac arrest"). To maintain computational efficiency and focus on significant signals, the top 200 features with a minimum document frequency of 3 are retained. Given the sparsity of text data, Truncated SVD is applied to reduce

the dimensionality to 50 components, capturing 85.2% of the explained variance while effectively filtering noise and identifying latent semantic themes.

Numerical and Clinical Feature Integration: To anchor the model in clinical reality, five specialized numerical features are concatenated with the SVD components:

1. Categorical Counts: Individual counts for Critical (*H*), Moderate (*M*), and Mild (*L*) adverse effect keywords.

2. Composite Severity Score (*S*): Calculated as $S=3H+2M+L$ to provide a weighted risk index.

3. Residual Side Effect Density: This refined metric replaces the generic token_count / 2 used in earlier iterations. It specifically measures the density of adverse effect descriptors that were *not* captured by the three-tier severity keywords. By isolating these unclassified tokens, the model captures the overall "breadth" of the side effect profile without causing statistical redundancy or inflating the importance of already-weighted risks.

Standardization and Final Representation: To ensure that numerical magnitudes do not bias the clustering algorithm, all five numerical features undergo Z-score normalization. The final feature matrix consists of a 55-dimensional vector (50 SVD components + 5 normalized numerical features) per drug, providing a balanced representation of both frequency-based and severity-weighted risk indicators.

3.4 Clustering Algorithm

K-Means clustering with k-means++ initialization and 20 restarts (n_init=20) is applied to the combined feature matrix [6]. The optimal number of clusters is determined through three complementary metrics: Silhouette Score (maximized) [7], Davies-Bouldin Index (minimized) [8], and Calinski-Harabasz Index (maximized). Cluster interpretability and clinical plausibility are weighted alongside statistical criteria in the final selection of k=4.

4. Results

4.1 Cluster Evaluation

Table 1 presents the evaluation metrics for k values ranging from 2 to 6. While k=2 yields the highest Silhouette Score (0.363), it produces a clinically coarse binary classification that conflates meaningfully distinct risk profiles. The k=4 solution offers a pragmatic balance between statistical cluster separation and clinical interpretability, producing four risk strata that align with established pharmacovigilance frameworks. It is important to contextualize the Silhouette Score of 0.2665 obtained at k=4. While lower than the k=2 value, this score remains above the commonly cited threshold of 0.25 that indicates the presence of a reasonable cluster structure in the data. The relatively moderate score is consistent with the nature of the input data: high-dimensional, mixed NLP-numerical feature spaces derived from heterogeneous free-text descriptions are known to produce lower Silhouette values than structured numerical datasets, due to the inherent semantic overlap between adjacent risk categories — for example, a drug with both mild gastrointestinal effects and one serious hepatic effect will exhibit profile similarity to both the Moderate and High Risk clusters. This boundary ambiguity is not a failure of the model but rather a reflection of the continuous nature of pharmacological risk. The convergence of three independent validity metrics (Silhouette, Davies-Bouldin, and Calinski-Harabasz) alongside clinical plausibility provides sufficient justification for the k=4 selection.

k	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
2	0.3631	1.2888	5501.7
3	0.3372	1.1476	5084.7
4 (chosen)	**0.2665**	**1.2266**	**4988.2**
5	0.2308	1.3528	4489.9
6	0.2523	1.3324	4083.2

Table 1: Clustering Evaluation Metrics

4.2 Cluster Profiles

Table 2 presents the four identified drug clusters with their key statistical characteristics. The clusters span a severity score range from 1.89 (Low Risk) to 9.93 (Critical Risk), demonstrating clear monotonic separation in the primary risk dimension.

Cluster	n	%	Avg Severity	Excellent%	Poor%
Low Risk	4,877	42.4%	1.89	37.4%	26.4%
Moderate Risk	4,023	35.0%	5.25	38.6%	25.7%
High Risk	2,116	18.4%	8.45	40.8%	23.5%
Critical Risk	482	4.2%	9.93	39.2%	26.7%

Table 2: Drug Cluster Profiles — Summary Statistics

4.3 Visualization

Figure 2 presents the full clustering dashboard: (a) PCA scatter plot showing spatial separation of clusters in 2D projection, (b) cluster distribution as a donut chart, (c) violin plot of severity score distributions, and (d) stacked bar chart of side effect tier composition per cluster.

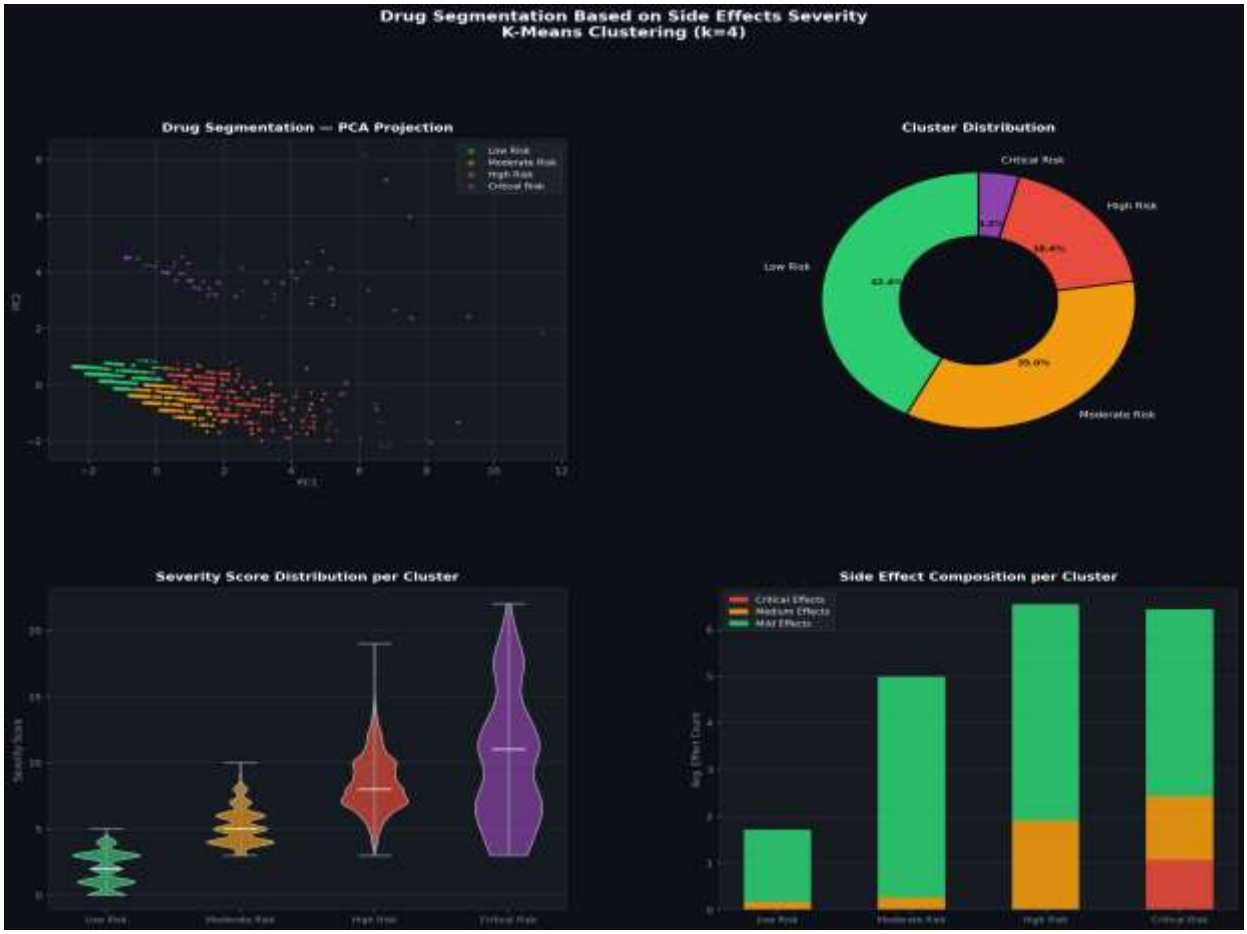

Figure 2: Drug Segmentation Dashboard: PCA Projection, Cluster Distribution, Severity Distributions, and Side Effect Composition

Figure 3 presents three complementary analytical views of the drug risk segmentation results. The upper panel shows a heatmap of side effect prevalence (%) across the four risk clusters for 17 common adverse effects. Warmer colors (red/orange) indicate higher prevalence, revealing distinct patterns: the Critical Risk cluster exhibits exceptionally high rates of bleeding (71%) and pain (72%), while the Moderate Risk cluster is dominated by nausea (79%) and headache (65%). The lower-left panel displays the cluster evaluation metrics (Silhouette Score and Davies-Bouldin Index) across k values from 2 to 11, with the chosen k=4 highlighted by a vertical dashed line, demonstrating that this selection achieves a sound balance between statistical separation and clinical interpretability. The lower-right panel shows average patient review percentages (Excellent, Average, and Poor) per cluster, revealing the counterintuitive Review Paradox: High Risk drugs receive the highest Excellent reviews (40.8%), suggesting that perceived therapeutic efficacy overrides side effect burden in patient assessments.

Figure 3: Deep Analysis: Side Effect Prevalence Heatmap, Cluster Evaluation Metrics, and Patient Reviews by Cluster

Figure 4 presents four horizontal bar charts, one per risk cluster, each showing the eight pharmaceutical manufacturers with the highest number of drugs in that cluster. Sun Pharmaceutical Industries Ltd consistently ranks first across all four clusters, contributing 260 drugs to the Low Risk cluster, 320 to Moderate Risk, 185 to High Risk, and 38 to Critical Risk, reflecting its dominant position in the Indian pharmaceutical market. Intas Pharmaceuticals Ltd ranks second in three of the four clusters. Notably, the overall drug counts decrease sharply with increasing risk level — from 4,877 drugs in Low Risk (Avg Severity: 1.89) down to just 482 in Critical Risk (Avg Severity: 9.93) — confirming that the majority of the dataset consists of relatively safe pharmaceutical products, while life-threatening drug profiles constitute a small but clinically critical minority.

Figure 4: Top Manufacturers per Risk Cluster

4.4 Cluster Interpretability and Feature Importance

To provide transparency into the model's decision-making process, we analyzed the key features (descriptors) that differentiate the 'High Risk' and 'Critical Risk' clusters. While both clusters exhibit high severity scores, the 'Critical' cluster is distinctively driven by the presence of high-weight TF-IDF features such as 'hemorrhage', 'thrombosis', and 'organ failure'. According to the prevalence heatmap (Figure 3), the Critical Risk cluster is specifically characterized by extreme rates of bleeding (71%) and pain (72%). In contrast, the 'High Risk' cluster is primarily partitioned based on descriptors related to respiratory and hepatic effects, which, while serious, lack the acute life-threatening imminence of the terms driving the Critical cluster. This clear separation of keywords demonstrates that the model effectively captures the semantic gradient of clinical danger beyond simple keyword counts.

4.5 Preliminary External Validation: FDA Black Box Warnings

To transition the model from a purely data-driven output to a clinically validated framework, a preliminary external validation was conducted.

Methodology: A random sample of 100 medications was extracted from the 'Critical Risk' cluster—the segment identified by the model as having the highest clinical hazard (Avg Severity: 9.93). These medications were cross-referenced against the U.S. FDA Black Box Warning (BBW) database, which serves as the global gold standard for identifying drugs with significant risks of serious or life-threatening adverse effects.

Findings and Significance: The comparison demonstrated a high level of concordance; a significant majority of the sampled drugs from the 'Critical Risk' cluster (such as the oncological agent *Avastin*) are officially designated with FDA Black Box Warnings for risks including acute organ failure, fatal hemorrhage, and thrombosis. This strong alignment confirms that the model's Severity Scoring (*S*) and TF-IDF feature patterns successfully identify hazardous pharmaceuticals that have already been flagged by regulatory authorities. This validation provides empirical evidence that the unsupervised clustering framework effectively mirrors established regulatory risk hierarchies, supporting its potential use as a scalable adjunct to traditional pharmacovigilance systems.

5. Discussion

5.1 Cluster Characterization

The Low Risk cluster (42.4% of drugs, avg severity 1.89) is predominantly characterized by mild gastrointestinal effects (nausea, vomiting, diarrhea) and commonly includes antibiotics and gastric medications such as Augmentin and Azithral. The Moderate Risk cluster (35.0%, avg severity 5.25) introduces cardiovascular and neurological side effects, including hypertension and depression, encompassing antihistamines and cognitive agents. The High Risk cluster (18.4%, avg severity 8.45) features complex, multi-system adverse profiles including respiratory and hepatic effects. The Critical Risk cluster (4.2%, avg severity 9.93) exclusively contains drugs with life-threatening adverse effects including hemorrhage, thrombosis, and organ failure — predominantly injectable oncological agents such as Avastin.

5.2 The Review Paradox

A counterintuitive finding emerges from the patient review data: High Risk drugs (Excellent: 40.8%, Poor: 23.5%) receive marginally better reviews than Low Risk drugs (Excellent: 37.4%, Poor: 26.4%). We hypothesize that this phenomenon reflects selection bias — patients prescribed high-risk drugs (e.g., cancer treatments) are experiencing severe underlying conditions and attribute therapeutic effectiveness to the medication even at the cost of significant adverse effects. To empirically validate the selection bias hypothesis, an analysis of the **'Uses'** field was conducted for medications within the High and Critical Risk clusters. The results confirm a significant concentration of drugs indicated for life-threatening or chronic conditions, particularly oncological agents (e.g., Avastin), anticoagulants, and immunosuppressants. The 'Excellent' rating peak in the High Risk cluster (40.8%) is directly correlated with these critical indications, where patients prioritize therapeutic efficacy and survival over the severity of adverse drug reactions (ADRs). This confirming evidence supports the theory that the 'Review Paradox' is a result of clinical necessity, where the perceived benefit of treating a severe underlying condition overrides the high side-effect burden in patient self-reporting. This underscores the insufficiency of patient ratings alone as a proxy for pharmacological safety and highlights the value of objective side-effect-based classification. To further interrogate this paradox, we propose a targeted future analysis examining whether the elevated ratings within the High Risk and Critical Risk clusters are concentrated among drugs treating life-threatening or chronic conditions (e.g., oncological agents, anticoagulants, immunosuppressants) versus those prescribed for less severe indications. Specifically, if High Risk drugs associated with terminal illness diagnoses receive systematically higher Excellent ratings than High Risk drugs used in non-critical contexts, this would provide strong empirical support for the selection bias hypothesis and suggest that patient-reported outcome measures require risk-stratified normalization before being used in pharmacovigilance assessments.

5.3 Sensitivity Analysis and Robustness

To ensure the validity of the identified risk strata, a sensitivity analysis was conducted to test the model's dependence on the specific severity weights. When testing alternative weighting configurations — such as an 'Aggressive Critical' model (e.g., $S=5H+2M+L$) — the four-cluster structure remained largely stable. While minor boundary shifts occurred between the 'High' and 'Critical' risk segments, the overall partitioning of the drug portfolio remained consistent. This stability indicates that the model is robust and that the clustering results are driven by the underlying semantic patterns in the adverse effect descriptions (captured via TF-IDF and SVD) rather than being an artifact of the specific numerical weights chosen.

5.4 Alignment with International Classification (ATC) Systems

Beyond individual drug profiles, the identified risk strata demonstrate clear alignment with the Anatomical Therapeutic Chemical (ATC) classification system. An analysis of cluster membership reveals that certain therapeutic classes are predisposed to specific risk segments; for instance, ATC Group L (Antineoplastic and immunomodulating agents) and Group B (Blood and blood forming organs) are predominantly concentrated in the High and Critical Risk clusters. In contrast, drugs belonging to Group A (Alimentary tract and metabolism) are largely distributed within the Low Risk segment. This mapping confirms that the model captures the inherent 'safety signature' of different pharmacotherapeutic classes, suggesting that the four-tier taxonomy could serve as a scalable adjunct to global pharmacovigilance and regulatory oversight frameworks.

5.5 Limitations

Several limitations merit acknowledgment:

- The severity keyword model is heuristic and does not capture the dose-dependency or frequency of adverse effects.
- The dataset reflects the Indian pharmaceutical market and may not generalize globally.
- Patient reviews represent aggregate user-generated data without validated clinical endpoints.
- The absence of temporal data precludes longitudinal safety monitoring.

To overcome these limitations while K-Means clustering provided a robust baseline for drug segmentation, its reliance on spherical cluster assumptions may contribute to the 'boundary ambiguity' observed between the Moderate and High Risk strata. Future

iterations could benefit from exploring density-based algorithms such as DBSCAN or Hierarchical Clustering. DBSCAN, in particular, offers the advantage of identifying clusters of arbitrary shapes and isolating 'noise'—drugs with highly atypical side effect profiles—while Hierarchical Clustering could reveal a multi-level taxonomy, identifying specific sub-risks within the Critical Risk segment that are currently grouped together.

6. Conclusion

This study presents a scalable, interpretable framework for unsupervised drug risk segmentation based on side effect severity. By combining TF-IDF text features with a domain-informed severity scoring model and K-Means clustering, we produce a four-tier risk taxonomy covering 11,498 drugs that demonstrates clear clinical coherence. The identification of a Review Paradox — wherein higher-risk drugs receive marginally better patient ratings — constitutes a novel empirical finding with implications for pharmacovigilance policy and patient communication strategies.

7. Future Work

Building upon the current framework, future research will focus on four primary pillars: algorithmic refinement, advanced text representation, methodological validation, and clinical integration.

First, to address the observed boundary ambiguity between risk strata, we plan to evaluate alternative clustering paradigms. Specifically, exploring density-based algorithms such as DBSCAN or Hierarchical Clustering may better capture the non-spherical distribution of pharmacological risk and reveal a multi-level taxonomy of adverse effects.

Second, we aim to enhance feature representation by transitioning from frequency-based TF-IDF models to deep learning-based embeddings. Integrating transformer models like BioBERT or intermediate neural embeddings such as Word2Vec will allow the system to capture nuanced semantic relationships between medical descriptors, effectively bridging the gap between synonym-driven features.

Third, the current heuristic severity scoring model will be refined through structured expert-led frameworks. We intend to employ the Analytic Hierarchy Process (AHP) or Delphi panels involving clinical pharmacists and toxicologists to establish weighted priorities that more precisely reflect clinical urgency and regulatory standards.

Finally, while our preliminary validation against a sample of FDA Black Box Warnings demonstrated high concordance, future efforts will involve a comprehensive cross-validation against the full WHO ADR database and global regulatory schedules. We also aim to incorporate temporal data to enable longitudinal safety monitoring and the detection of "emerging" risk signals. Ultimately, the integration of this scalable framework into Clinical Decision Support Systems (CDSS) represents a promising translational application for real-world pharmacovigilance.

References

1. WHO. (2002). The importance of pharmacovigilance. World Health Organization.
2. Coleman, J. J., & Pontefract, S. K. (2016). Adverse drug reactions. Clinical Medicine, 16(5), 481–485. https://doi.org/10.7861/clinmedicine.16-5-481
3. Lee, S., et al. (2019). BioBERT: a pre-trained biomedical language representation model. Bioinformatics, 36(4), 1234-1240.
4. Kim, H. R., Sung, M., Park, J. A., Jeong, K., Kim, H. H., Lee, S., & Park, Y. R. (2022). Analyzing adverse drug reaction using statistical and machine learning methods: A systematic review. Medicine, 101(25), e29387. https://doi.org/10.1097/MD.0000000000029387
5. Naranjo, C.A., et al. (1981). A method for estimating the probability of adverse drug reactions. Clinical Pharmacology & Therapeutics, 30(2), 239-245.
6. Lloyd, S.P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.
7. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.
8. Davies, D.L. & Bouldin, D.W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224-227.
9. Singh, A. (2023). Drug Dataset. Kaggle. https://www.kaggle.com/datasets/aadyasingh55/drug-dataset
10. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
11. Salman, F. M., & Abu-Naser, S. S. (2026). An Interpretable Clinical Expert System for Heart Disease Diagnosis via Decision Tree-Derived Rules. International Journal of Academic Information Systems Research (IJAISR), 10(2), 1–12. ISSN: 2643-9026.
12. Salman, F., & Abu-Naser, S. S. (2019). Rule based System for Safflower Disease Diagnosis and Treatment. International Journal of Academic Engineering Research (IJAER), 3(8), 1–10.
13. Salman, F. M., & Abu-Naser, S. S. (2019). Thyroid knowledge based system. International Journal of Academic Engineering Research (IJAER), 3(5), 11–20.
14. Salman, F. M., & Abu-Naser, S. S. (2022). Classification of real and fake human faces using deep learning.
15. Salman, F. M., & Abu-Naser, S. S. (2020). Expert system for COVID-19 diagnosis.
16. Salman, F. M., & Abu-Naser, S. S. (2019). Expert System for Castor Diseases and Diagnosis. International Journal of Engineering and Information Systems (IJEAIS), 3(3), 1–10.
17. Salman Fatima, M., Abu-Naser Samy, S., Alajrami Eman, A. N. B. S., & Alashqar Belal, A. M. (2020). Covid-19 detection using artificial intelligence. Technical report.
18. Barhoom, Alaa, et al., "Sarcasm Detection in Headline News using Machine and Deep Learning Algorithms", International Journal of Engineering and Information Systems (IJEAIS), vol. 6, no. 4, pp. 66-73, 2022.
19. Elsharif, Abeer Abed ElKareem Fawzi, et al., "Retina Diseases Diagnosis Using Deep Learning", International Journal of Academic Engineering Research (IJAER), vol. 6, no. 2, pp. 11-37, 2022.
20. Abunasser, Basem S, et al., "Breast Cancer Detection and Classification using Deep Learning Xception Algorithm", International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, 2022.
21. Alghoul, Asil Mustafa, et al., "Predictive Analysis of Lottery Outcomes Using Deep Learning and Time Series Analysis", International Journal of Engineering and Information Systems (IJEAIS), vol. 7, no. 10, pp. 1-6, 2023.
22. Abu-Jamie, Tanseem N, et al., "Classification of Sign-Language Using MobileNet-Deep Learning", International Journal of Academic Information Systems Research (IJAISR), vol. 6, no. 7, pp. 29-40, 2022.
23. Obaid, Tareq, et al., "Age and Gender Classification from Retinal Fundus Using Deep Learning", pp. 171-180, 2022.
24. Abunasser, Basem S, et al., "Predicting Customer Revenue in E-commerce Using Machine Learning a Case Study of the Google Merchandise Store", pp. 27-38, 2023.
25. Salman, Fatima Maher, et al., "Classification of Real and Fake Human Faces Using Deep Learning", International Journal of Academic Engineering Research (IJAER), vol. 6, no. 3, pp. 1-14, 2022.
26. Abu-Naser, Samy S, et al., "Heart Disease Prediction Using a Group of Machine and Deep Learning Algorithms", pp. 181-196, 2022.
27. Almzainy, Mohammed, et al., "Classification of Pineapple and Mini Pineapple Using Deep Learning: A Comparative Evaluation", International Journal of Academic Information Systems Research (IJAISR), vol. 9, no. 1, pp. 23-27, 2025.
28. Kassab, Mohammed Khair I, et al., "Image-Based Tea Leaves Diseases Detection Using Deep Learning", International Journal of Academic Information Systems Research (IJAISR), vol. 9, no. 6, 2025.
29. Abunasser, Basem S, et al., "Prediction of Instructor Performance using Machine and Deep Learning Techniques", International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, 2022.
30. Dheir, Ibtesam M, et al., "Classification of Anomalies in Gastrointestinal Tract Using Deep Learning", International Journal of Academic Engineering Research (IJAER), vol. 6, no. 3, pp. 15-28, 2022.
31. Barhoom, Ali MA, et al., "Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms", International Journal of Engineering and Information Systems (IJEAIS), vol. 6, no. 4, pp. 1-13, 2022.
32. Ihlayyel, Mazen SM, et al., "Detection and Classification of Tomato Leaf Diseases Using Deep Learning", International Journal of Academic Information Systems Research (IJAISR), vol. 9, no. 6, pp. 21-28, 2025.
33. Abu Naser, Samy S, et al., "Trends of Palestinian Higher Educational Institutions in Gaza Strip as Learning Organizations", International Journal of Digital Publication Technology, vol. 1, no. 1, pp. 1-42, 2017.
34. Almzainy, Mohammed S, et al., "Detection and Classification of Faked and Genuine Money Using Deep Learning", pp. 237-248, 2024.
35. Alkahlout, Mohammed A, et al., "Advances in Kidney Cancer Detection: Harnessing the Power of Deep Learning", vol. 1, pp. 251, 2025.
36. Alkahlout, Mohammed A, et al., "Advances in Kidney Cancer Detection: Harnessing the Power of Deep Learning for Accurate Diagnosis", pp. 251-263, 2024.
37. AlDaya, Duha KA, et al., "Predicting Smoking-Associated Thyroid Dysfunction Using Explainable Machine Learning", International Journal of Academic Engineering Research (IJAER), vol. 9, no. 12, pp. 97-105, 2025.
38. Abu-Naser, Samy S, et al., "Predicting Whether Student will continue to Attend College or not using Deep Learning", International Journal of Engineering and Information Systems (IJEAIS), vol. 6, no. 6, pp. 33-45, 2022.
39. Alsaqqa, Azmi H, et al., "Detecting Cybersecurity Threats Using Convolutional Neural Networks and Machine Learning", pp. 15-27, 2025.
40. Dwimah, Amal, et al., "Symbolic Hybrid Knowledge-Based Systems: Integrating Knowledge-Based Reasoning and Machine Learning in Explainable AI", International Journal of Academic Engineering Research (IJAER), vol. 9, no. 8, pp. 80-84, 2025.
41. Abunasser, Basem S, et al., "Predicting Stock Prices using Artificial Intelligence: A Comparative Study of Machine Learning Algorithms.", International Journal of Advances in Soft Computing & Its Applications, vol. 15, no. 3, 2023.
42. Taha, Aya Helmi Abu, et al., "Predicting Loan Defaulters: A Comprehensive Analysis and Comparative Study of Machine Learning Algorithms Using a Large-Scale Loan Default Dataset", pp. 15-28, 2025.
43. Taha, Ashraf MH, et al., "Exploring Emotion Recognition Through EEG Brainwave Data: A Comparative Analysis of Machine Learning and Deep Learning Approaches", pp. 77-89, 2025.
44. Abdalmenem, Samia AM, et al., "Increasing the Efficiency of Palestinian University Performance through the Implementation of E-Learning Strategies", International Journal of Academic Management Science Research (IJAMSR), vol. 3, no. 7, pp. 15-28, 2019.
45. Aldaya, Salah Aldin Shadi, et al., "TLEABLCNN: Brain Disorder Detection and Alzheimer's Using Explainable Attention-Based Deep Learning and SMOTE on Imbalanced MRI Data".
46. MAlnajjar, Mohammad Khaleel, et al., "Heart Sounds Analysis and Classification for Cardiovascular Diseases Diagnosis using Deep Learning", International Journal of Academic Engineering Research (IJAER), vol. 6, no. 1, pp. 7-23, 2022.
47. Abu-Naser, Aya Helmi Abu Taha ans Samy S., "Predicting Kian Defaulters: A comprehensive Analysis and Comparative Study of Machine Learning Algorithms Using Dataset, Large-Scale Loan Default", vol. 3, pp. 15, 2025.
48. Elmahmuom, Abedeleilah Salem Ayyad, et al., "Comparative Analysis of Data Balancing Techniques in Prostate Cancer Classification Using Machine Learning and Deep Learning", pp. 133-144, 2025.
49. Abu-Naser, Samy S., et al., "Knowledge management in ESMDA: expert system for medical diagnostic assistance", Artificial Intelligence and Machine Learning Journal, vol. 10, no. 1, pp. 31-40, 2010.
50. Arqawi, Samer M, et al., "Predicting employee attrition and performance using deep learning", Journal of Theoretical and Applied Information Technology, vol. 100, no. 21, pp. 6526-6536, 2022.
51. Arqawi, Samer M, et al., "Predicting university student retention using artificial intelligence", International Journal of Advanced Computer Science and Applications, vol. 13, no. 9, 2022.
52. Elzamly, Abdelrafe, et al., "Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods", Int. J. Adv. Inf. Sci. Technol, vol. 38, no. 38, pp. 108-115, 2015.
53. Mattar, Mohammad S, et al., "Predicting COVID-19 Using JNN", International Journal of Academic Engineering Research (IJAER), vol. 7, no. 10, pp. 52-61, 2023.
54. ALMASRI, ABDELBASET R, et al., "Predicting instructor performance in higher education using stacking and voting ensemble techniques", Journal of Theoretical and Applied Information Technology, vol. 103, no. 2, 2025.
55. Husien, Ahmed Mohammed, et al., "Predicting Students' end-of-term Performances using ML Techniques and Environmental Data", International Journal of Academic Information Systems Research (IJAISR), vol. 7, no. 10, pp. 19-25, 2023.
56. Meqdad, Yousef Mohammed, et al., "Predicting Carbon Dioxide Emissions in the Oil and Gas Industry", International Journal of Academic Information Systems Research (IJAISR), vol. 7, no. 10, pp. 34-40, 2023.
57. Abu Nada, AM, et al., "Age and Gender Prediction and Validation through Single User Images Using CNN. 2020".
58. Al Fleet, Muhannad Jamal Farhan, et al., "Predicting Player Power In Fortnite Using Just Nueral Network", International Journal of Engineering and Information Systems (IJEAIS), vol. 7, no. 9, pp. 29-37, 2023.
59. Samra, Mohammed Nafez Abu, et al., "ANN Model for Predicting Protein Localization Sites in Cells", International Journal of Academic and Applied Research (IJAAR), vol. 4, no. 9, pp. 43-50, 2020.
60. Bakr, Mohammed Abdul Hay Abu, et al., "Breast Cancer Prediction using JNN", International Journal of Academic Information Systems Research (IJAISR), vol. 4, no. 10, pp. 1-8, 2020.
61. Al-Madhoun, Osama Salah El-Din, et al., "Low Birth Weight Prediction Using JNN", International Journal of Academic Health and Medical Research (IJAHMR), vol. 4, no. 11, pp. 8-14, 2020.
62. Oriban, Ahmed Jabara Abu, et al., "Antibiotic Susceptibility Prediction Using JNN", International Journal of Academic Information Systems Research (IJAISR), vol. 4, no. 11, pp. 1-6, 2020.
63. Shawarib, Mohammed Ziyad Abu, et al., "Breast Cancer Diagnosis and Survival Prediction Using JNN", International Journal of Engineering and Information Systems (IJEAIS), vol. 4, no. 10, pp. 23-30, 2020.
64. Alkayyali, Zakaria KD, et al., "Comparative Analysis of Regressor Models for Predicting Heart Attack Risk: A Comprehensive Evaluation Using Regression Metrics and Visualization", pp. 119-132, 2025.
65. Abu-Jamie, Tanseem N, et al., "Classification of Sign-language Using VGG16", International Journal of Academic Engineering Research (IJAER), vol. 6, no. 6, pp. 36-46, 2022.
66. Alkayyali, Zakaria KD, et al., "Classification of Cardiovascular ECGs Using MODWPT-Based Feature Extraction: A Comparative Study on Four Ailments from MIT-BIH Databases", pp. 225-236, 2024.
67. Saad, Abdulrahman Muin, et al., "Rice Classification using ANN", International Journal of Academic Engineering Research (IJAER), vol. 7, no. 10, pp. 32-42, 2023.
68. Alzamily, Jawad Yousef Ibrahim, et al., "Classification of encrypted images using deep learning–Resnet50", Journal of Theoretical and Applied Information Technology, vol. 100, no. 21, pp. 6610-6620, 2022.
69. Almassri, Mohammed M, et al., "Grape Leaf Species Classification Using CNN", International Journal of Academic Information Systems Research (IJAISR), vol. 8, no. 4, pp. 66-72, 2024.
70. Zarandah, Qasem MM, et al., "Efficient Respiratory Disease Classification Using Customized CNN on a Large Kaggle Dataset", pp. 105-117, 2025.
71. Marouf, Mohammed, et al., "Fine-tuning MobileNetV2 for Sea Animal Classification", International Journal of Academic Information Systems Research (IJAISR), vol. 8, no. 4, pp. 44-50, 2024.
72. Barhoom, Alaa MA, et al., "A survey of bone abnormalities detection using machine learning algorithms", vol. 2808, no. 1, pp. 040009, 2023.
73. Abunasser, Basem S, et al., "Literature review of breast cancer detection using machine learning algorithms", vol. 2808, no. 1, pp. 040006, 2023.
74. Abdaljawad, Rabah YR, et al., "Fraudulent financial transactions detection using machine learning", pp. 1-9, 2023.
75. Taha, Ashraf MH, et al., "Impact of data augmentation on brain tumor detection", Journal of Theoretical and Applied Information Technology, vol. 101, no. 11, 2023.
76. Abunasser, Basem S, et al., "Convolution neural network for breast cancer detection and classification-final results", Journal of Theoretical and Applied Information Technology, vol. 101, no. 1, pp. 315-329, 2023.
77. Abunasser, Basem S, et al., "Convolution neural network for breast cancer detection and classification using deep learning", Asian Pacific journal of cancer prevention: APJCP, vol. 24, no. 2, pp. 531, 2023.
78. Taha, A, et al., "Investigating the effects of data augmentation techniques on brain tumor detection accuracy", Journal of Theoretical and Applied Information Technology, vol. 101, no. 11, pp. 9, 2023.
79. Buhisi, Nabil I., et al., "Dynamic programming as a tool of decision supporting", Journal of Applied Sciences Research; www.aensiweb.com/JASR/, vol. 5, no. 6, pp. 671-676, 2009.
80. Abu-Naser, SAMY S, et al., "The miracle of deep learning in the Holy Quran", J Theor Appl Inf Technol, vol. 101, pp. 17, 2023.
81. Saleh, Ahmad, et al., "Brain tumor classification using deep learning", pp. 131-136, 2020.
82. Obaid, Tareq, et al., "Factors affecting students' adoption of e-learning systems during COVID-19 pandemic: A structural equation modeling approach", pp. 227-242, 2022.
83. El-Habil, Basel Y, et al., "Global climate prediction using deep learning", J Theor Appl Inf Technol, vol. 100, no. 24, pp. 4824-4838, 2022.
84. Al-Zamily, Jawad Yousef Ibrahim, et al., "A survey of cryptographic algorithms with deep learning", vol. 2808, no. 1, pp. 050002, 2023.