Leveraging predictive modeling and explainable AI to understand veterinary safety profiles and health outcomes using OpenFDA data

by

Hossein Sholehrasa

B.Sc., Tehran University, Iran, 2023

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Co-Major Professor
Doina Caragea

Approved by:

Co-Major Professor
Majid Jaberi-Douraki

Copyright

Abstract

The safe use of pharmaceuticals in food-producing animals is essential for protecting animal welfare and human food safety. Adverse events (AEs) in veterinary medicine may reflect unexpected pharmacokinetic or toxicokinetic effects that cause violative residues in the food chain. This study proposes a predictive modeling framework to classify animal health outcomes as either "Death" or "Recovered", using the U.S. Food and Drug Administration's OpenFDA Center for Veterinary Medicine database, containing approximately 1.25 million reports from 1987 to 2024 Q3. The preprocessing pipeline linked relational tables and mapped AEs and drugs to standardized ontologies, including the Veterinary Dictionary for Drug Regulatory Activities (VeDDRA) and the World Health Organization's Anatomical Therapeutic Chemical (ATC) classification system for veterinary use. It further standardized units, handled missing values, and simplified high-cardinality categorical variables. Multiple supervised learning models were evaluated, including Random Forest, CatBoost, XGBoost, and the transformer-based ExcelFormer. Class imbalance was addressed using random undersampling and the Synthetic Minority Oversampling Technique (SMOTE), with a particular focus on improving recall for fatal outcomes. A semi-supervised learning approach with Average Uncertainty Margin (AUM)-based pseudo-labeling was applied to utilize reports with uncertain outcomes, incorporating high-confidence predictions into retraining. XGBoost with undersampling achieved the best overall performance (accuracy 91.60%, F1-score 92.24%, death recall 91.56%), while adding AUM-filtered pseudo-labeled data improved death recall to 93.34% with minimal loss in overall accuracy. Model explainability was achieved through Gain-based feature importance and SHapley Additive exPlanations (SHAP) analysis identified biologically plausible predictors, including specific organ system disorders (e.g., bronchial and lung disorders, heart disorders) that were strongly linked to Death outcomes, as well as animal demographic information and drug characteristics, which contributed to

distinguishing between “Death” and “Recovered” outcomes. This work demonstrates that combining rigorous data engineering, advanced machine learning, and explainable artificial intelligence enables accurate, interpretable prediction of veterinary safety profiles and health outcomes, supporting earlier detection of high-risk profiles and informing evidence-based regulatory and clinical decision-making.

Table of Contents

List of Figures vii

List of Tables viii

Acknowledgements ix

1 Introduction 1

2 Related work 4

3 Data source 6

3.1 Dataset curation 6

3.2 Terminology standards and classification systems 7

3.3 Dataset summary statistics 7

4 Methodology 13

4.1 Research design 13

4.2 Data preprocessing 14

4.2.1 Dataset integration and merging 14

4.2.2 Terminology standardization 14

4.2.3 Data cleaning and normalization 15

4.2.4 Feature engineering and representation 17

4.2.5 Dataset partitioning 17

4.2.6 Handling class imbalance 17

4.3 Supervised learning models 18

4.3.1 Machine learning . . . 18
4.3.2 Ensemble learning techniques . . . 18
4.3.3 Transformer-based DL for tabular data - ExcelFormer . . . 19
4.4 Semi-supervised learning approach . . . 19
4.4.1 Pseudo-labeling framework . . . 19
4.4.2 AUM-based confidence metric . . . 20
4.4.3 Integrated SSL pipeline . . . 20
4.5 Evaluation metrics . . . 21
4.6 Model explainability . . . 21
4.6.1 Gain-based feature importance (XGBoost) . . . 22
4.6.2 SHAP-based model interpretability . . . 22

5 Results . . . 23
5.1 Model performance . . . 23
5.1.1 Improved results using pseudo-labeling . . . 24
5.2 Feature importance analysis . . . 25
5.2.1 Gain-based feature importance (XGBoost) . . . 25
5.2.2 SHAP-based model interpretability . . . 25

6 Conclusions and future work . . . 28
6.1 Conclusions . . . 28
6.2 Future work . . . 29

Bibliography . . . 30

List of Figures

3.1 Distribution of affected animals by species. 9

3.2 Visualization of top 12 most frequently reported veterinary drugs. 10

3.3 Distribution of animal health outcomes. 11

3.4 Distributions of AEs (VeDDRA HLT level) and drugs (ATCvet therapeutic subgroups). 12

4.1 Overview of AUM-based pseudo-labeling workflow. 21

5.1 SHAP summary plot of feature impacts in predicting Death and Recovered outcomes. 27

List of Tables

4.1 Dataset features and preprocessing steps. 16

5.1 Supervised and semi-supervised classification results with undersampling. . . 24

5.2 Top 10 features ranked by Gain in XGBoost 26

Acknowledgments

I would like to express my deepest gratitude to my co-supervisors, Dr. Jaberi and Dr. Caragea, whose unwavering guidance, encouragement, and expertise have been invaluable throughout the course of my graduate studies. Their insightful feedback and thoughtful advice consistently challenged me to think critically, refine my ideas, and maintain high standards in both research and academic writing. Beyond their technical knowledge, their patience, mentorship, and commitment to my development as a researcher have been instrumental in shaping the direction of this thesis, strengthening my academic foundation, and preparing me for the next stage of my professional career.

I am also sincerely grateful to Dr. Xu, whose thoughtful perspectives and careful evaluations enriched this work. His involvement encouraged me to approach my research from new angles and contributed significantly to the depth and coherence of this thesis.

My appreciation further extends to the members of my thesis committee, Dr. Amtoft and Dr. Neilsen, for their constructive comments and valuable suggestions. Their contributions have been essential in broadening the scope of my research and enhancing the overall quality of this work.

I would also like to acknowledge the support and companionship of my colleagues and peers, whose collaborations, discussions, and encouragement greatly enriched my graduate experience. Finally, I am profoundly thankful to my family and friends for their patience, unwavering belief, and constant encouragement throughout this journey; their support has been a continuous source of strength and motivation.

This research was funded by the USDA via the FARAD program [Award No.: 2020-41480- 32497, 2021-41480-35271, 2022-41480-38135, and 2023-41480-41034] and supported through the 1DATA Consortium at Kansas State University.

Chapter 1

Introduction

The safe use of pharmaceuticals in animals is a fundamental component of public health protection, safeguarding both animal welfare and, in the case of food-producing animals, ensuring human food safety. In food-producing animals, veterinary drugs, including antimicrobials, antiparasitics, anti-inflammatories, and biologics, are integral to maintaining food-producing animals' health and productivity (Mesfin et al., 2024). Yet, their use carries the risk of leaving residues in edible tissues, milk, and other animal-derived products. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA) set maximum residue limits (MRLs) and withdrawal periods to ensure that these residues remain at or below levels considered safe for human consumption (Zad et al., 2023). Exceeding these limits can result in chronic dietary exposure, which may contribute to antimicrobial resistance (AMR) development, trigger allergic or toxic reactions in sensitive individuals, and, for certain compounds, pose potential carcinogenic or endocrine-disrupting risks (Mesfin et al., 2024). Beyond food safety, adverse events (AEs) in veterinary medicine are also a major concern for companion animals such as dogs and cats, where they can cause morbidity, mortality, and compromised welfare. In both food-producing animals and companion species, AEs may reflect unexpected pharmacokinetic or toxicokinetic behaviors that alter drug safety profiles. For example, impaired liver or kidney function can delay drug clearance, prolonging residue persistence and withdrawal times in

food animals or increasing toxicity risks in pets (Khalifa et al., 2024). Furthermore, cross-species differences in drug metabolism can complicate the extrapolation of safety data from one species to another, reinforcing the need for ongoing monitoring in real-world agricultural contexts (Margiotta-Casaluci et al., 2024).

Conventional pharmacovigilance (PV) largely depends on disproportionality analyses of spontaneous reporting system data, which are valuable for retrospective signal detection but insufficient for predicting emerging risks or managing complex, multidimensional datasets (Dijkstra et al., 2024). In contrast, machine learning (ML) and deep learning (DL) approaches can integrate diverse, high-dimensional datasets, such as demographic, clinical, pharmacological, and AEs information, and automatically learn complex, non-linear relationships among variables. Unlike conventional statistical models that typically test predefined hypotheses and rely on simplified assumptions, ML/DL can uncover patterns without prior specification and adapt to new data for forecasting and real-time risk prediction. This shift enables proactive assessment and classification of animal health outcomes, with our predictive models for veterinary drugs advancing earlier detection of safety signals and exposure risks relevant to human consumers.

In this study, we develop an integrated predictive modeling framework for animal health outcome classification. Leveraging OpenFDA reports, we apply supervised learning, semi-supervised Average Uncertainty Margin (AUM)-based pseudo-labeling, and explainable artificial intelligence (XAI) techniques to predict fatal versus non-fatal outcomes and identify the most influential predictors. Our work advances veterinary PV and human toxicology by demonstrating how predictive modeling can improve early detection of safety signals and provide insights relevant to protecting both animal and human health.

From a computational perspective, this research situates veterinary PV within the broader field of applied ML for complex biomedical datasets. The OpenFDA Center for Veterinary Medicine (CVM) dataset presents challenges common to large-scale, real-world health data, including heterogeneous formats, high-cardinality categorical variables, class imbalance, and incomplete or noisy entries. To address these issues, we developed a robust preprocessing pipeline that integrates relational datasets, maps data to standardized on-

tologies, incorporates the Veterinary Dictionary for Drug Regulatory Activities (VeDDRA), and applies hierarchical drug classification using the WHO Anatomical Therapeutic Chemical (ATC) system for veterinary use (ATCvet codes). We then employed a multi-model strategy spanning classical algorithms, ensemble methods (Random Forest, CatBoost, XGBoost), and transformer-based DL (ExcelFormer) to model non-linear interactions between demographic, pharmacological, and clinical features. Semi-supervised AUM-based pseudo-labeling enables the expansion of the training set with confidently predicted unlabeled cases, improving minority-class representation. At the same time, SHAP (SHapley Additive exPlanations) and Gain-based feature importance analyses ensure interpretability and clinical relevance. This integration of rigorous data engineering, advanced predictive modeling, and explainable AI provides a scalable and transparent computational framework for proactively identifying high-risk drug–event profiles in veterinary medicine.

Chapter 2

Related work

Computational PV research increasingly leverages large-scale AE databases to detect safety signals through advanced statistical analysis and ML. In the veterinary domain, Xu et al. (2019) applied ML to the FDA's veterinary AE dataset for dogs and cats, aiming to identify hidden drug–event patterns and improve PV signal detection. Their study compared five similarity measures, Pearson, Spearman, cosine, Yule, and Euclidean, to quantify associations between AEs, then used Graphical Least Absolute Shrinkage and Selection Operator (GLASSO) and hierarchical clustering to construct sparse AE networks and visualize organ-system groupings via Circos plots and heatmaps. Cosine and correlation-based metrics consistently outperformed others in recovering known label-listed events, while Euclidean distance performed poorly due to instability and noise. Focusing on two drugs (selamectin and spinosad), they demonstrated that similarity and network-based analysis can reveal clusters of related events, distinguish severe from non-severe reactions, and validate labeling information, providing a robust framework for post-market veterinary drug safety assessment. Most human-focused PV studies using the U.S. Food and Drug Administration's Event Reporting System (FAERS) and the OpenFDA platform rely on disproportionality-based signal detection. Examples include (Pan et al., 2024) for ticagrelor, (Yang et al., 2024) for memantine + donepezil, and (Zhong et al., 2025) for finasteride, all of which employed rigorous FAERS preprocessing and statistical algorithms such as the Proportional Report-

ing Ratio, Reporting Odds Ratio, and Bayesian Confidence Propagation Neural Network to characterize safety profiles. In contrast, predictive modeling has been less explored; Farnoush et al. (2024) integrated FAERS data with DrugBank and PubChem to train Random Forest and DL models for AEs prediction using demographic and drug-chemical features. Building on these methodological advances, our work applies similar predictive modeling principles to the OpenFDA CVM AE dataset, extending PV research from retrospective human AE characterization to forward-looking veterinary safety profile and health outcome prediction. More recently, Xu et al. (2024) reviewed *in-silico* strategies to assess AE of high-level drug-drug and drug-disease interactions, integrating frequentist and Bayesian disproportionality methods, AI-driven clustering, and pharmacogenomics to enhance signal detection and enable personalized medicine. Their approach, systematic curation of spontaneous reporting data, standardized terminology mapping, comparative statistical modeling, and advanced visualization, is applicable to this work. In this study, we adopt a similarly rigorous preprocessing pipeline for CVM OpenFDA data, map drugs and reactions to standardized ontologies (ATCvet, VeDDRA), and apply ML models with SHAP-based interpretability to uncover clinically meaningful patterns.

Chapter 3

Data source

3.1 Dataset curation

The primary dataset utilized in this research was obtained from the Animal and Veterinary Adverse Event Reporting System (AERS) maintained by the FDA, accessible through its public OpenFDA platform (U.S. Food and Drug Administration). This open-access repository provides one of the most comprehensive collections of AE records in veterinary medicine. It serves as the foundation for data-driven analysis of clinical outcomes in animals. The dataset contains 1,246,745 reports for 93 species submitted between 1987 and 2024 Q3, sourced from veterinarians, animal owners, pharmaceutical companies, and academic institutions. Updated quarterly, the data includes structured information such as animal demographics (species, breed, sex, age, weight), clinical outcomes (recovery, death, ongoing or unknown), drug details (active ingredients, brand name, dosage, route of administration, formulation), AEs (coded using standard medical ontologies for medical symptoms), and reporter type. Each report is uniquely identified by a `unique_aer_id_number` to ensure consistent referencing across related tables. The raw data, distributed in quarterly compressed JSON files following the OpenFDA API schema, was programmatically downloaded, parsed, and converted into four structured CSV files: the Main Table (metadata and demographics), Reactions Table (AEs), Outcome Table (clinical outcomes and affected animals counts), and

Drug Table (drug-specific information). These CSVs were imported into a PostgreSQL relational database to support efficient querying and integration during data preprocessing and modeling.

3.2 Terminology standards and classification systems

The dataset adopts internationally recognized veterinary coding systems to ensure consistency, comparability, and regulatory compliance in representing AEs and drug information. Particularly, AEs are categorized through the VeDDRA hierarchy to ensure consistency and regulatory compliance. This hierarchical structure begins with Lowest Level Terms (LLT), which include specific descriptors and synonyms—these map to Preferred Terms (PT), representing standardized clinical concepts. PTs are then grouped under High-Level Terms (HLT) based on anatomical or functional relationships. At the top of the hierarchy are System Organ Class (SOC) categories, which encompass major physiological systems (Bousquet et al., 2005).

Drug information in the dataset is classified using ATCvet codes, which were developed by the WHO Collaborating Centre for Drug Statistics Methodology. This system organizes veterinary drugs into five hierarchical levels: the Anatomical Main Group, followed by the Therapeutic Subgroup, Pharmacological Subgroup, Chemical Subgroup, and finally the Chemical Substance (Dahlin et al., 2001). For example, the drug carprofen is classified under the ATCvet code QM01AE91, which reflects its full classification path through these five levels.

3.3 Dataset summary statistics

The analyzed dataset comprises 1,086,612 reports for selected animal species, including chicken, turkey, cattle, dog, cat, pig, horse, sheep, and goat. Each report provides structured information encompassing animal demographics such as species, age, and sex; drug characteristics including product name, active ingredients, and administration details; the

reported AEs; the recorded health outcomes following the event; and the number of affected animals.

Figure 3.1 shows the total number of affected animals, as reported within each report. Chickens make up the vast majority of affected animals (111,562,852, representing $\approx 89\%$ of total animals), followed by turkeys (5,990,558), cattle (3,809,008), and dogs (3,347,777). This distribution reflects a dataset heavily weighted toward poultry-related reports. Drug usage patterns (Figure 3.2) reveal that Narasin is the most frequently reported drug (58,972,798 affected animals), followed by Diclazuril (16,052,334).

Outcome distribution analysis (Figure 3.3a) indicates that Outcome Unknown is the most common category (49.4%), followed by Recovered/Normal (23.8%), Ongoing (17.1%), Died (5.4%), Recovered with Sequela (2.9%), and Euthanized (1.5%). When restricting the analysis to only Died vs. Recovered (Recovered/Normal + Recovered with Sequela) (Figure 3.3b), survival outcomes dominate, with recovery representing 83.3% of cases in this subset.

The AEs term distribution in HLT level (Figure 3.4a) highlights "lack of efficacy" (18.9%), "general signs or symptoms" (14.8%), and "stomach disorders" (14.2%) as the most frequently reported clinical signs, with a long tail of lower-frequency conditions. The ATCvet code distribution at the high-level therapeutic group (Figure 3.4b) shows a strong predominance of Endectocides (53.8%), followed by Ectoparasiticides, Insecticides, and Repellents (15.9%) and Antiinflammatory and Antirheumatic Products (7.0%).

Together, these statistics underscore the skewed nature of the dataset toward poultry-related reports, specific high-use drugs, and certain AE categories, highlighting the need for balanced sampling and appropriate feature handling in downstream modeling.

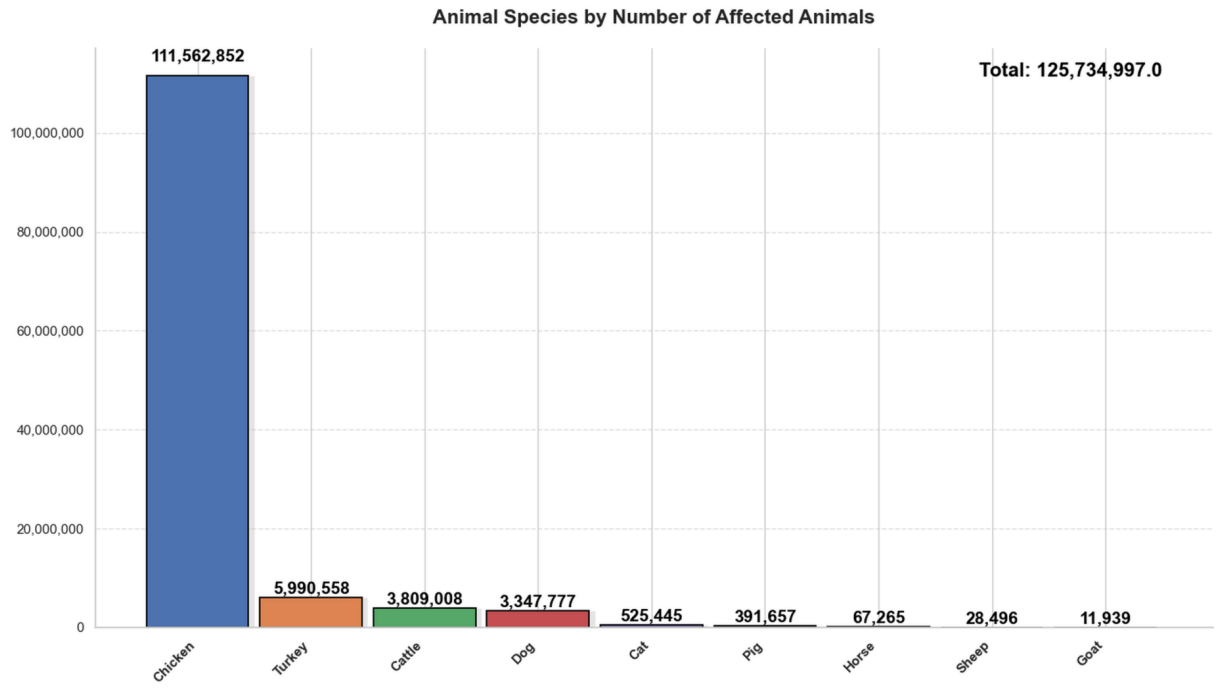


Figure 3.1: Distribution of affected animals by species, based on counts reported in the OpenFDA veterinary dataset, showing a strong predominance of poultry cases, particularly chickens (∼89%), followed by turkeys, cattle, and dogs.

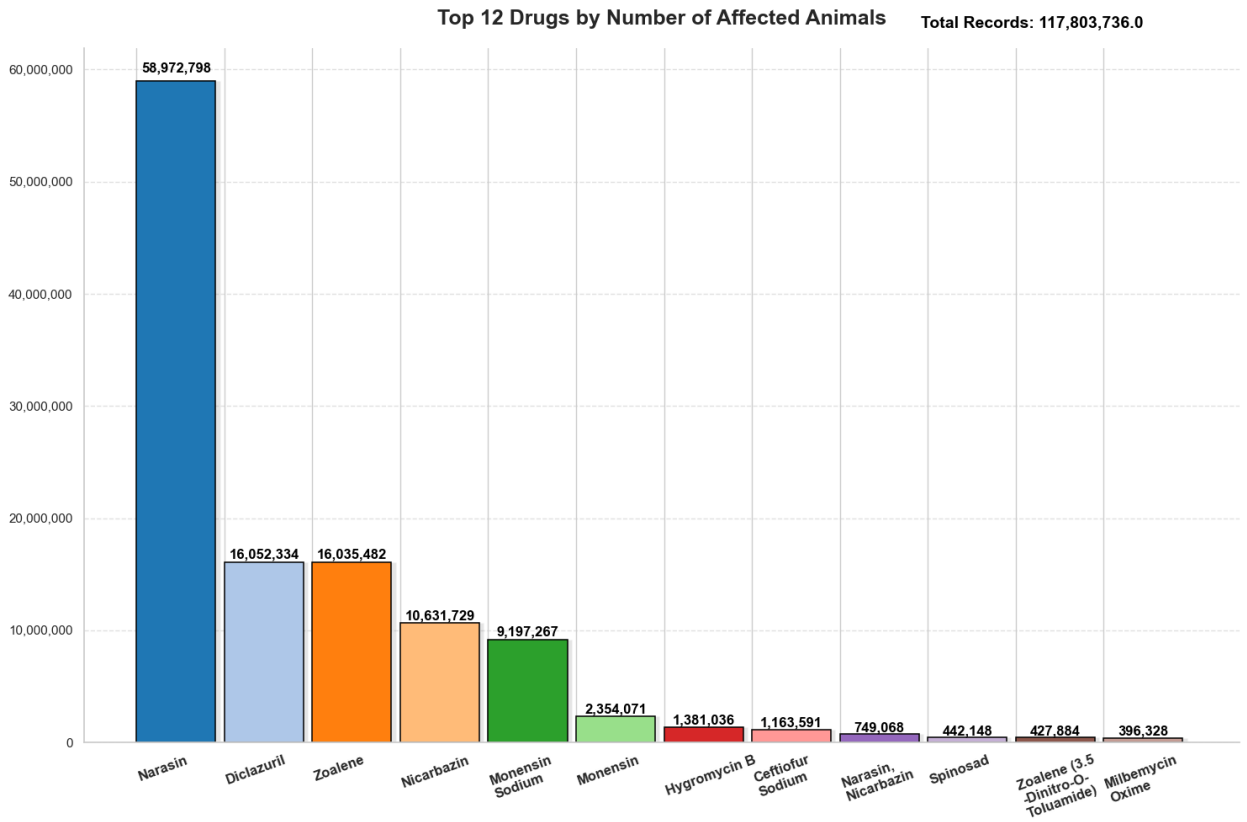
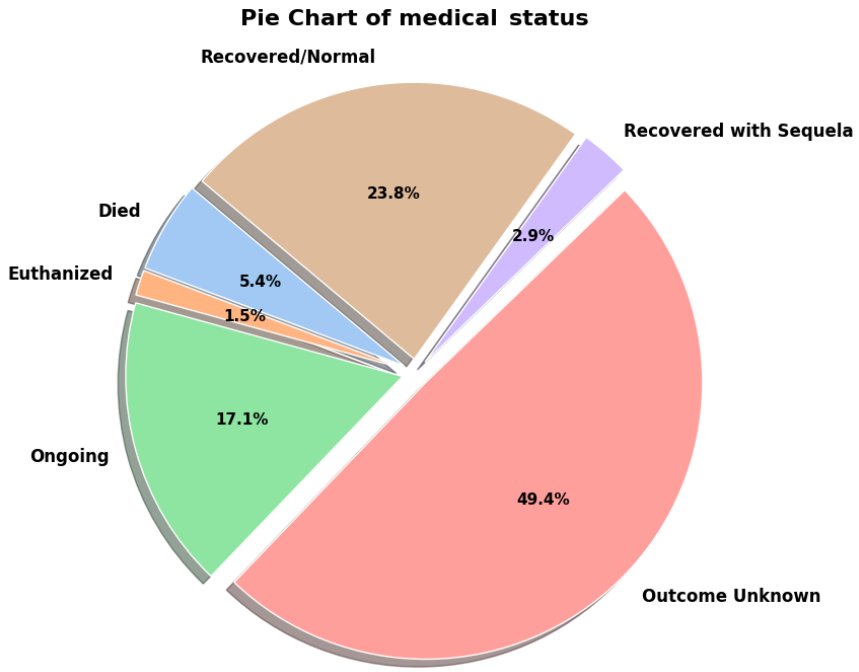
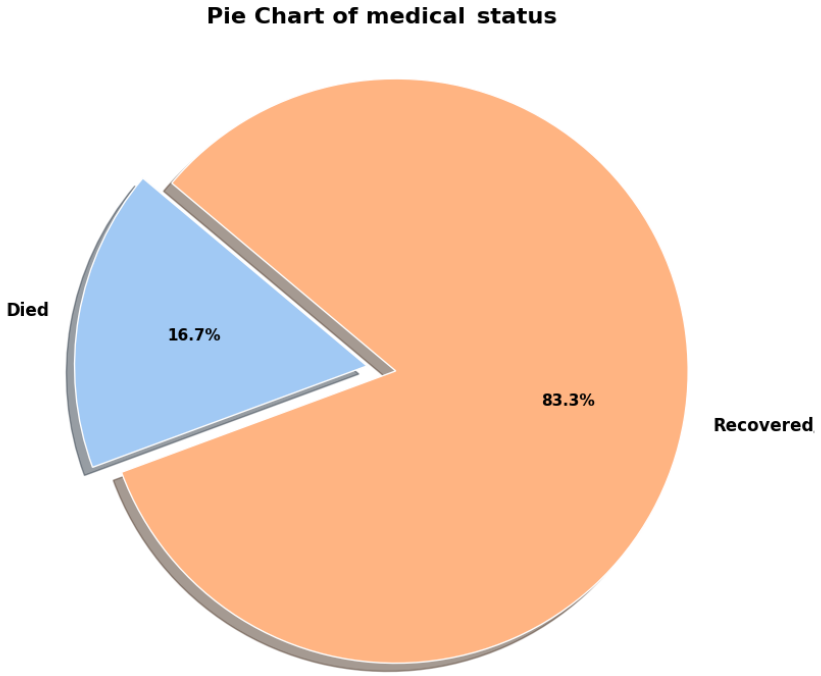


Figure 3.2: Visualization of the top 12 most frequently reported drugs linked to affected animals in the OpenFDA veterinary dataset.

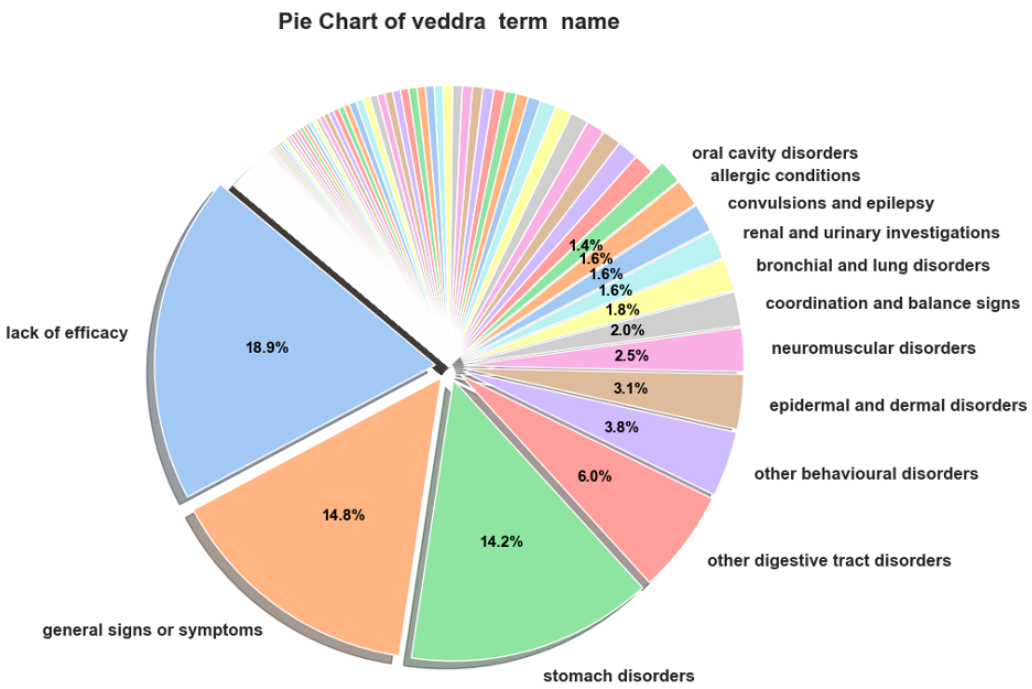


(a) Overall distribution of animal health outcomes for all classes.

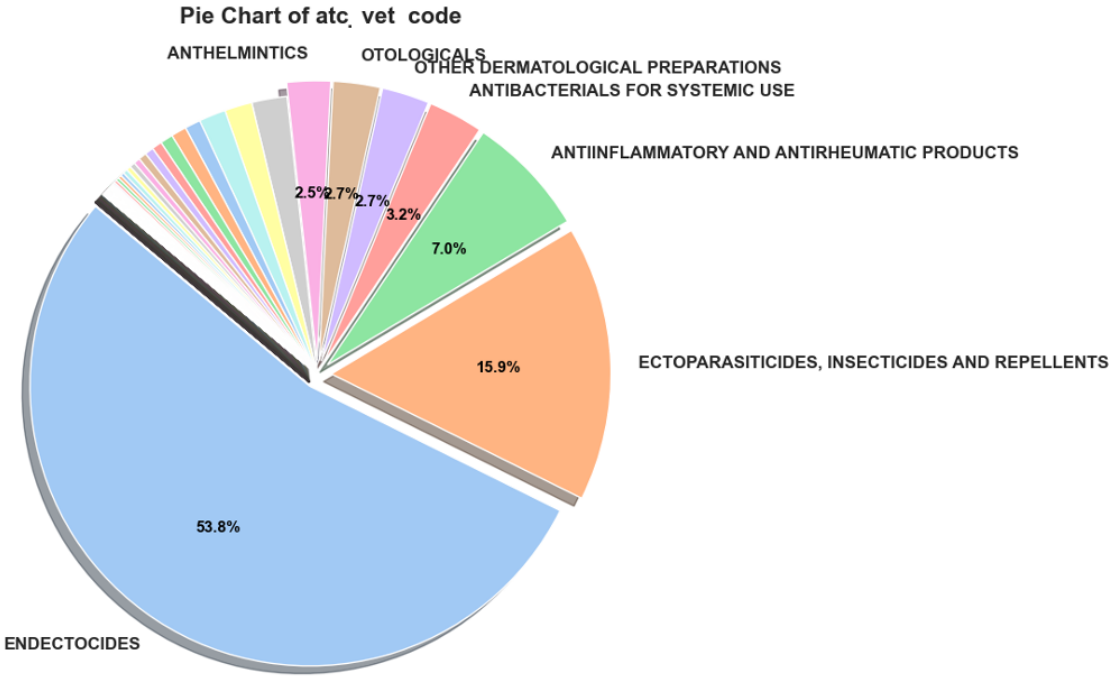


(b) Distribution of Died vs. Recovered (Recovered/Normal + Recovered with Sequela) animal health outcomes.

Figure 3.3: Pie charts illustrating the distribution of animals' medical health outcomes in the OpenFDA veterinary dataset.



(a) Distribution of AEs categorized at the VeDDRA HLT level.



(b) Distribution of ATCvet at the therapeutic subgroup level.

Figure 3.4: Summary distributions from the OpenFDA veterinary dataset, illustrating the proportional breakdown of reports by AEs in HLT level categories and therapeutic group classifications.

Chapter 4

Methodology

This chapter describes the methodological framework employed to predict clinical outcomes in animals exposed to veterinary drugs using data from the FDA's openFDA platform. The proposed methodology involves data preparation, feature engineering, supervised and semi-supervised learning (SSL), ensemble modeling, and explainability. A comprehensive pipeline was developed to clean and standardize heterogeneous AE reports, build predictive models, and interpret their decisions using modern ML and DL models and explainability techniques.

4.1 Research design

This study aims to apply AI to predict veterinary clinical outcomes from AE reports using a two-phase approach. In the first phase, supervised learning models are trained on records with known outcomes, specifically the Death and Recovered classes. In the second phase, the trained model is applied to cases originally labeled as Ongoing or Outcome Unknown to assign them to one of the two definitive outcome categories (Death or Recovered). The overall methodology comprises three core components: preprocessing and standardization of structured AE data for selected species, development of supervised and semi-supervised ML models, and interpretation of predictions using feature importance metrics and explainability methods such as SHAP.

4.2 Data preprocessing

Effective preprocessing is crucial for producing a clean, consistent, and analysis-ready dataset for downstream ML applications. In this study, the pipeline integrated multiple CVM OpenFDA reports, standardized heterogeneous formats, and addressed missing values to ensure high-quality reports for analysis

4.2.1 Dataset integration and merging

The four CSV files generated during data curation were merged using the unique adverse event report identifier (`unique_aer_id_number`), resulting in a unified and relationally consistent dataset. This integration allowed each row to comprehensively reflect all aspects of a given report, including the species demographic, drug characteristics, reactions, and final clinical outcomes.

4.2.2 Terminology standardization

Building on the established VeDDRA and ATCvet frameworks, we applied a standardized terminology mapping process to harmonize AEs and drug classes across reports, ensuring consistency, regulatory compliance, and comparability in downstream analyses. Specifically, reported AEs were standardized by mapping them to HLTs. Matching PTs were linked directly, while unmatched terms were compared against LLTs and converted accordingly. This ensured consistent categorization for structured modeling, and reports with incomplete or malformed AE data were excluded to maintain data integrity.

For drug classification, each ATCvet code was mapped to its corresponding high-level therapeutic group. We mapped these codes to the Therapeutic Subgroup using web scraping of the World Health Organization Collaborating Centre (WHOCC) website (WHO Collaborating Centre for Drug Statistics Methodology, 2025). By aligning all codes to a single classification level, the analysis could focus on broader drug categories rather than overly specific compounds, enabling more generalizable pattern discovery.

4.2.3 Data cleaning and normalization

The dataset comprises reports, each documenting one or more affected animals, with every animal described by 14 structured features (Table 4.1). The features vary in cardinality, from low-dimensional attributes such as Animal Gender (4 categories) to high-cardinality fields such as Animal Breeds (> 9,500 categories) and AEs (> 2,900 categories). To ensure consistency across records, several columns experienced systematic normalization and cleaning during preprocessing. For example, animal age values were converted to a consistent unit of years by transforming entries reported initially in days, weeks, or months. Records with missing units, such as C17998, were excluded. Animal weight was standardized to kilograms. When both minimum and maximum weights were provided, their average was used. For missing weights, the mean weight served as the imputation. Dosage information was also normalized, with all values converted to milligrams. Beyond numerical data cleaning, several categorical features in the dataset contained missing or noisy values, which were addressed through targeted cleaning strategies. Missing entries in columns like gender, dosage form, and route of administration were filled using the mode, or the most frequently occurring value, within each respective column. To reduce feature sparsity, categories with extremely low frequency, defined as representing less than 1% of the total, were excluded from the dataset. Additionally, entries with ambiguous or uninformative labels, such as "unknown" for gender, were filtered out to improve overall data quality and ensure meaningful analysis. In addition, the "Euthanized" category was removed due to insufficient information for accurate classification. The "Recovered with Sequela" and "Recovered/Normal" categories were combined to avoid overlap between recovery types.

In sum, numerical values like dose, weight, and age were standardized, while missing data in numerical and categorical fields were imputed using mean or mode values. Reaction terms were mapped to standardized VeDDRA codes, rare categories were removed or consolidated, and categorical features were encoded appropriately, resulting in a clean, consistent dataset ready for robust modeling and analysis.

Table 4.1: Summary of dataset features, including feature types, number of unique values, percentage of missing data, and preprocessing steps applied after filtering out selected animals.

Feature Name	**Type**	**# Unique Values**	**Missing (%)**	**Preprocessing step**
unique_aer_id _number	Identifier	1,086,612	0	-
Medical Status (outcome)	Categorical	6	0	Target variable
Number of Affected Animals	Numeric	515	0	-
AEs	Categorical	2900+	0	One-hot; map to HLT codes
ATC VET Code	Categorical	724	2	Convert to Therapeutic Subgroup; drop missing
Administrative Route	Categorical	58	4	Impute mode; drop $< 1\%$ categories
Dosage Form	Categorical	55	2	Impute mode; drop $< 1\%$ categories
Year	Categorical	38	0	Remove
Animal Species	Categorical	9	0	-
Animal Gender	Categorical	4	4	Impute mode; drop "mixed" ($< 1\%$)
Animal Breeds	Categorical	9500+	$< 1\%$	Drop missing
Animal Age	Numeric	931	5	Convert to years; impute mean
Animal Weight	Numeric	9100+	7	Convert to kg; impute mean
Dose	Numeric	229	0	Convert to mL

4.2.4 Feature engineering and representation

To improve the interpretability of the model and minimize noise, a manual feature filtering technique was applied. The Year feature was excluded to avoid introducing temporal bias, and reports with AE of "Lack of Efficacy" were removed since they reflect treatment failure rather than a physiological response, which could potentially leak outcome labels. The AEs feature was transformed using one-hot encoding to preserve the multi-label nature and categorical specificity of reported AEs. All other categorical features were encoded numerically (label encoding) to represent their categories as integer values while maintaining compatibility with the ML models.

4.2.5 Dataset partitioning

After cleaning and feature engineering, the dataset contained approximately 660,000 reports, and it was then divided into two main groups based on outcome labels. The first group contained reports with definitive outcomes, specifically Death and Recovered cases, while the second group consisted of reports with uncertain outcomes categorized as Ongoing or Unknown. Following a common approach, 80% of the labeled data was allocated for training, and the remaining 20% was reserved for testing (Géron, 2022). The training set was used to develop the initial model, and the test set served to evaluate the final model's performance.

4.2.6 Handling class imbalance

The dataset showed a class imbalance, with Recovered cases outnumbering Death cases. To address this, two sampling strategies were evaluated independently during model training. The Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) was applied to generate synthetic samples for the minority class, while Random UnderSampling (He and Garcia, 2009) was used to reduce the number of majority class examples. Each method was applied separately to assess its impact on model performance, enabling the selection of the most effective approach for improving generalization and enhancing recall for the minority class.

4.3 Supervised learning models

Several classical supervised ML and ensemble learning approaches, as well as transformer-based DL supervised learning models, were trained and evaluated to classify reports into either Death or Recovered.

4.3.1 Machine learning

The ML component of this study incorporated a diverse set of classification algorithms. These included traditional linear models such as Logistic Regression, and tree-based methods including Decision Tree, Random Forest, AdaBoost, CatBoost, and XGBoost. Additionally, the K-Nearest Neighbors (KNN) algorithm was employed to leverage instance-based learning for outcome prediction.

4.3.2 Ensemble learning techniques

To strengthen the robustness and generalizability of the predictive models, two ensemble learning strategies were incorporated. These approaches combine the outputs of multiple base learners to leverage their complementary strengths and mitigate individual model biases.

Voting classifier

This method combines predictions from multiple base models through ensemble voting. The final prediction is determined by majority rule based on predicted class labels in hard voting. In soft voting, the predicted probabilities from each model are averaged, and the class with the highest average probability is selected as the final output (Géron, 2022).

Stacking classifier

Stacked generalization works by combining multiple base generalizers (logistic regression, random forest, gradient boosting) using a meta-level generalizer (logistic regression) that

operates on their outputs, often yielding superior performance compared to winner-takes-all or simple averaging methods (Wolpert, 1992).

4.3.3 Transformer-based DL for tabular data - ExcelFormer

ExcelFormer (Chen et al., 2023), as one of the state-of-the-art transformer-based architectures optimized for tabular data, was evaluated as part of this study to benchmark performance against classical ML and ensemble learning models. Designed to address the limitations of traditional neural networks on structured datasets, ExcelFormer effectively handles mixed categorical and numerical inputs. Its core innovation, Semi-Permeable Attention (SPA), enables the model to filter irrelevant features selectively, enhancing focus on informative signals (Chen et al., 2023). ExcelFormer incorporates data augmentation strategies such as Swap-Mix and Hidden-Mix, which promote robustness in scenarios with limited training data to improve generalization (Chen et al., 2023). Furthermore, its attention-based feedforward layers can capture complex, non-linear interactions between features, making it well-suited for high-dimensional classification tasks in healthcare and veterinary PV.

4.4 Semi-supervised learning approach

4.4.1 Pseudo-labeling framework

To incorporate reports with Unknown or Ongoing outcomes into the modeling process, we adopted a pseudo-labeling strategy (Lee et al., 2013). In this approach, a baseline model was first trained on labeled data (Death and Recovered) and then used to assign labels to the unlabeled cases. The resulting pseudo-labels were combined with the original labeled set for retraining, allowing the model to iteratively improve its performance by leveraging a larger training set.

4.4.2 AUM-based confidence metric

To enhance the reliability of pseudo-label selection, the high-confidence filtering step was implemented using the AUM metric (Pleiss et al., 2020). Unlike fixed probability thresholds, AUM assesses prediction stability across training epochs, making it less sensitive to noisy samples. Specifically, for each pseudo-labeled sample, the AUM score was computed as

$$\text{AUM} = \frac{1}{T} \sum_{t=1}^{T} \left(P_{\text{top}}^{t} - P_{\text{second}}^{t} \right),$$

where P_{top}^{t} and P_{second}^{t} are the predicted probabilities of the most likely and second most likely classes at epoch t, and T is the total number of training epochs. All pseudo-labeled samples were then ranked according to their AUM scores, and the top 80% were retained. This threshold was determined through validation experiments, which demonstrated that retaining 80% of the highest-AUM samples yielded superior performance compared with 30% and 60% thresholds. By prioritizing samples with consistently large probability margins between the top two predicted classes across epochs, the AUM metric ensures that pseudo-labels are both stable and confident, thereby reducing the propagation of label noise.

4.4.3 Integrated SSL pipeline

The final SSL pipeline integrated the pseudo-labeling framework (Section 4.4.1) with AUM-based high-confidence filtering (Section 4.4.2). As illustrated in Figure 4.1, the model was first trained on labeled data, then used to generate pseudo-labels for unlabeled cases. High-confidence pseudo-labels, defined as the top 80% ranked by the AUM metric, were merged with the labeled dataset under the undersampling setting for retraining. This integration leverages pseudo-labeling to expand the training set while safeguarding against unstable predictions, resulting in a more robust SSL pipeline.

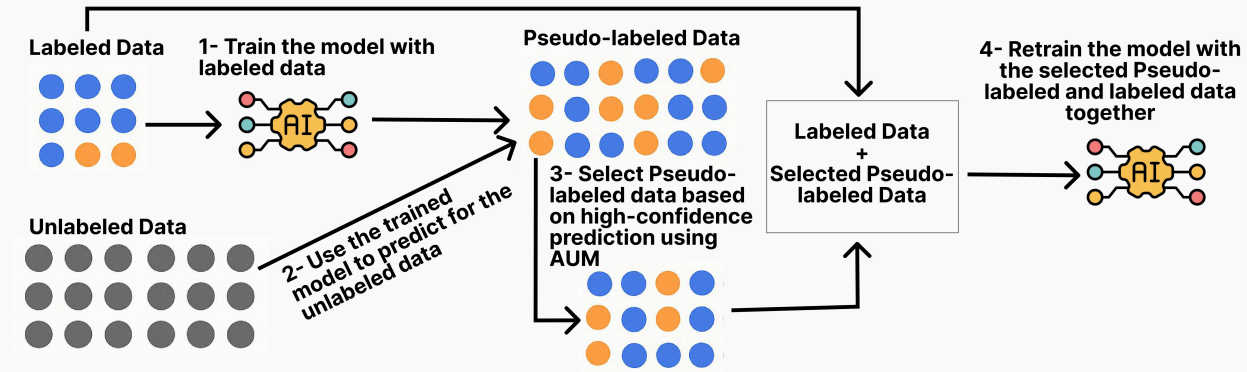


Figure 4.1: Overview of AUM-based pseudo-labeling workflow. The model is first trained on labeled data, then used to predict outcomes for unlabeled cases. High-confidence predictions are merged with the original labeled set for retraining.

4.5 Evaluation metrics

All models were evaluated using four weighted metrics: accuracy, F1-score, precision, and recall for both the Death and Recovered classes. These metrics provided a balanced assessment of overall performance, class-specific sensitivity, and the ability to handle class imbalance effectively.

4.6 Model explainability

Understanding the factors influencing model predictions is essential to ensure transparency, trust, and actionable insights in veterinary PV. This section presents two complementary approaches for model interpretation. First, XGBoost's internal Gain metric is used to rank features by their overall contribution to classification performance. Second, SHAP values provide local, instance-level explanations, highlighting the specific features that drive individual predictions (Lundberg et al., 2018). Together, these methods enable global and local interpretability, supporting model validation and offering domain-relevant insights into outcomes.

4.6.1 Gain-based feature importance (XGBoost)

The internal Gain metric from XGBoost was employed to assess each feature's relative contribution to the classification task (Chang et al., 2021). In the context of gradient-boosted decision trees, the Gain represents the average improvement in the model's objective function brought by splits using a given feature, averaged over all trees in the ensemble (Chang et al., 2021). Higher Gain values indicate features that contribute more to reducing classification error, making this metric suitable for ranking the importance of predictors in structured datasets. This analysis provides an interpretable ranking of features, guiding domain understanding and subsequent feature selection steps.

4.6.2 SHAP-based model interpretability

SHAP values were also computed to interpret individual predictions and enhance the transparency of the classification model. SHAP is grounded in cooperative game theory and provides a principled way to attribute the contribution of each feature to the model's output (Lundberg et al., 2018). For a given instance, the SHAP value of a feature represents the marginal contribution of that feature when added to all possible feature subsets. In this study, SHAP was applied to the test dataset to identify factors driving the model's decisions and to diagnose sources of error. The analysis was conducted locally, explaining individual predictions through visualizations that illustrate how features increase or decrease the likelihood of Death versus Recovered outcomes. Key observations from SHAP analysis are as follows: SHAP effectively attributed each prediction to specific feature contributions, enabling a clear decomposition of the model's decision process. Visual inspection of the SHAP plots revealed recurring misclassification patterns, such as Death predictions being assigned to cases without lethal indicators. Overall, SHAP-based interpretability improved model transparency and provided valuable clinical insights into drug safety in animals, revealing the key factors associated with Death and Recovered outcomes.

Chapter 5

Results

This chapter presents the experimental findings of this study, including the performance of various ML models, improvements through sampling and semi-supervised learning, and the explainability insights gained using the XGBoost Gain metric and SHAP values.

5.1 Model performance

Using the cleaned dataset, undersampling produced the best performance among imbalance-handling methods and was applied to equalize class sizes. Table 5.1 in the supervised models section reports all evaluated models' accuracy, F1-score, and class-wise recall. Among the evaluated models, XGBoost achieved the highest overall performance, with an accuracy of 91.60%, an F1-score of 92.24%, and balanced recall across both outcome classes. CatBoost and the stacking classifier also delivered competitive results, with accuracies above 90% and strong recall for both classes. The transformer-based ExcelFormer model demonstrated robust generalization, achieving 90.23% accuracy and a recall for the Recovered class (92.07%), confirming its suitability for tabular veterinary health data. Overall, ensemble tree-based methods (XGBoost, CatBoost, Random Forest) and the ExcelFormer DL model consistently outperformed baseline classifiers such as Logistic Regression, indicating the effectiveness of models that can capture complex, non-linear feature interactions.

Table 5.1: Supervised and semi-supervised classification results with undersampling, reported in percentages. XGBoost provided the top supervised performance and was subsequently extended with AUM-based pseudo-labeling.

Supervised Models				
Model	**Accuracy**	**F1-score**	**Death Recall**	**Recovered Recall**
Logistic Regression	87.16	88.40	83.57	87.70
Decision Tree	86.41	87.85	85.97	86.47
K-Nearest Neighbors	87.35	88.45	79.87	88.48
Random Forest	90.37	91.22	**92.09**	90.10
AdaBoost	89.26	90.22	88.11	89.43
CatBoost	91.27	91.96	91.10	91.29
XGBoost	**91.60**	**92.24**	91.56	91.60
Voting Classifier	90.16	91.02	90.70	90.07
Stacking Classifier	90.55	91.37	90.29	**92.24**
ExcelFormer	90.23	90.38	88.70	92.07
Semi-supervised (AUM-based pseudo-labeling)				
XGBoost (pseudo-label)	89.31	**93.51**	**93.34**	88.69

The undersampling configuration in XGBoost yielded the best overall performance, achieving a Death recall of 91.56% compared to 80.00% with oversampling and 79.00% without sampling. Although its Recovered recall (91.60%) was lower than the 99.00% achieved by both oversampling and without sampling approaches, the substantial improvement in fatal outcome detection underscores undersampling as the most effective strategy for the primary objective of minimizing false negatives in critical cases.

5.1.1 Improved results using pseudo-labeling

Using the AUM-based pseudo-labeling procedure described in Section 4.4.3, XGBoost, as the best-performing model in the initial result, was retrained on the original labeled set augmented with the top 80% of high-confidence pseudo-labeled samples from Ongoing/Unknown cases. This augmentation enlarged and diversified the dataset, notably strengthening the representation of the minority Death class. As shown in Table 5.1, Death recall increased to

93.34%, with a high overall F1-score (93.51%), while Recovered recall decreased only slightly to 88.69%. The general accuracy remained robust at 89.31%, indicating that improved sensitivity to fatal outcomes, of greater clinical importance, was achieved without substantial compromise in performance in recovered cases.

5.2 Feature importance analysis

5.2.1 Gain-based feature importance (XGBoost)

Table 5.2 lists the top 10 features ranked by Gain, highlighting variables that consistently provided the most predictive value. Notably, the highest-ranking predictors are biologically meaningful and closely associated with systemic and organ-level toxicities observed in veterinary AEs. Features such as application site reactions and injection site reactions capture local administration effects, while bronchial and lung disorders, cardiac/heart disorders aggravated, and renal and hepatic disorders represent critical organ system impacts linked to severe outcomes. Additionally, abnormal pathology reports and abnormal necropsy findings indicate post-mortem or laboratory findings consistent with severe systemic reactions. The prominence of these features aligns with known toxicological mechanisms in veterinary medicine, reinforcing the biological plausibility of the model's learned patterns.

5.2.2 SHAP-based model interpretability

Figure 5.1 presents the SHAP summary plot for the XGBoost model, illustrating the top predictive features' relative importance and directional influence. Each point represents an instance from the dataset, positioned according to its SHAP value, with negative values (left side) increasing the probability of a Death prediction and positive values (right side) increasing the probability of a Recovered prediction. Color encodes the feature value, ranging from low (blue) to high (red). The most influential feature was atc vet code, which exhibited a broad range of SHAP values on both sides of the axis, indicating its strong discriminative power for both classes. Dose, animal age, and general signs or symptoms

Table 5.2: Top 10 features ranked by Gain in the XGBoost model. Gain indicates the relative contribution of each feature to model predictions, with higher values reflecting greater influence in distinguishing between Death and Recovered outcomes.

Rank	Feature	Gain Score
1	Application site reactions	0.0517
2	Drug identity	0.0482
3	Bronchial and lung disorders	0.0449
4	Cardiac/heart disorders aggravated	0.0446
5	Abnormal pathology reports	0.0403
6	Convulsions and epilepsy	0.0314
7	Abnormal necropsy findings	0.0276
8	Injection site reactions	0.0270
9	Renal disorders	0.0261
10	Hepatic disorders	0.0257

also ranked highly, with older ages tending to shift predictions toward Death, while lower values were more associated with recovery. Physiological measurements, such as animal weight, and treatment-related attributes, like dosage form, further contributed to classification. Several AE features, including bronchial and lung disorders, convulsions and epilepsy, abnormal pathology reports, and cardiac/heart disorders aggravated, were strongly associated with Death outcomes when present at high severity (red points on the negative SHAP axis). Conversely, features like breed and animal species showed more balanced contributions, influencing predictions for both outcomes depending on their specific category values. Notably, other clinical AEs, such as epidermal and dermal disorders, allergic conditions, application site reactions, and administration errors, were displayed in concentrated clusters on the positive axis, reinforcing their link to recovery.

Overall, the SHAP analysis confirms that the model's decision-making aligns with clinically plausible patterns, where severe pathological signs and certain demographic or dosage factors are primary drivers of fatal outcome predictions. In contrast, milder conditions and lower physiological risk factors correspond to recovery predictions.

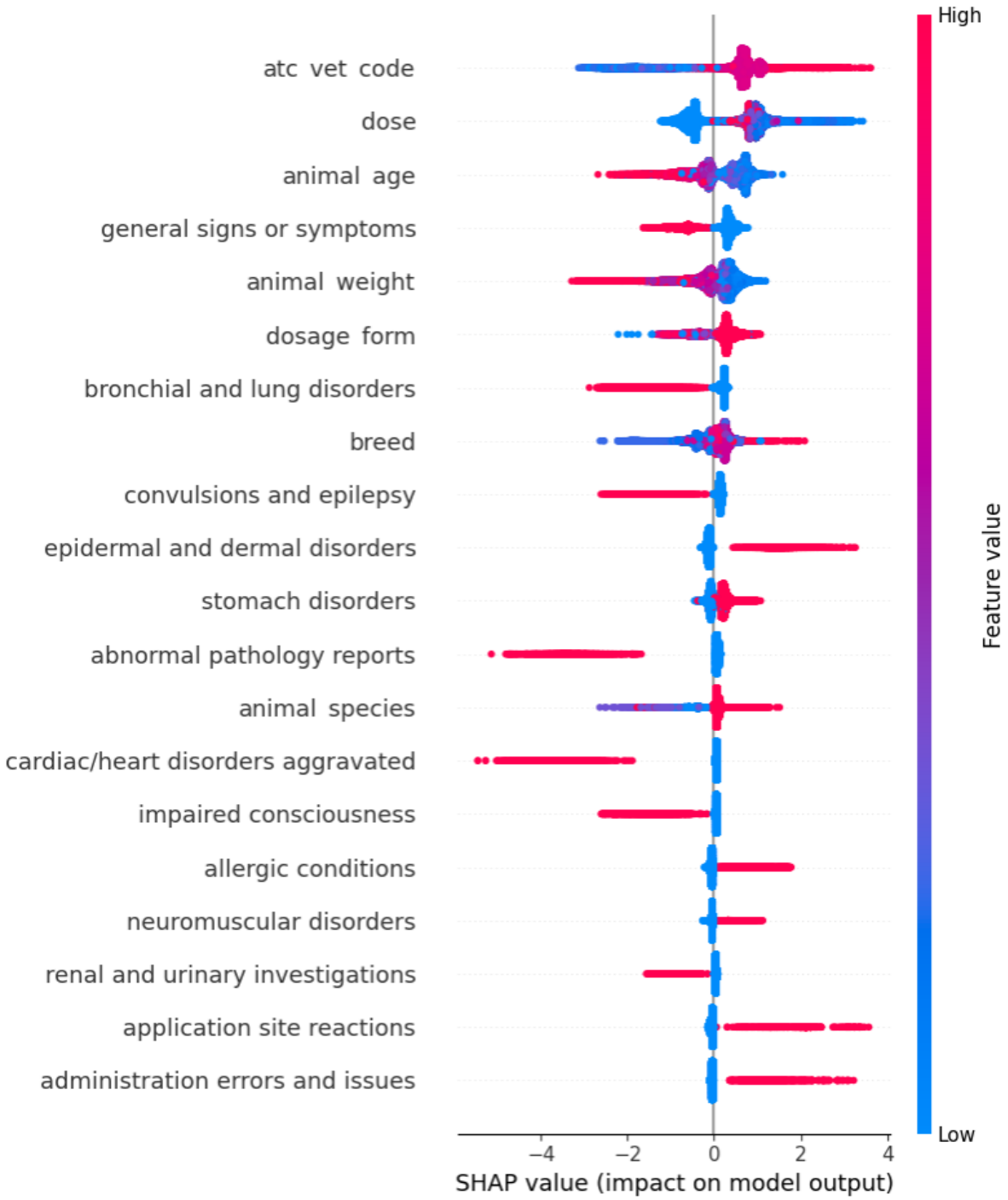


Figure 5.1: SHAP summary plot illustrating the impact of features in predicting Death and Recovered outcomes. Negative SHAP values (left) indicate a higher likelihood of predicting Death, while positive SHAP values (right) indicate a higher likelihood of predicting Recovered. Color represents the feature value, from low (blue) to high (red).

Chapter 6

Conclusions and future work

6.1 Conclusions

This research demonstrates that integrating robust data preprocessing, advanced ML models, and explainable AI techniques can yield high-performing, clinically interpretable predictions of veterinary safety profiles and health outcomes. Using the FDA's OpenFDA veterinary PV data, we addressed real-world challenges including heterogeneous data formats, high-cardinality categorical variables, class imbalance, and incomplete records. Our results show that the XGBoost model, enhanced with undersampling and AUM-based pseudo-labeling, achieved superior performance in detecting fatal outcomes, a critical priority for regulatory decision-making and food safety risk assessment. The inclusion of semi-supervised learning allowed the model to leverage unlabeled cases, increasing the diversity and representativeness of the training dataset without substantially degrading performance on recovered outcomes. Explainability tools such as SHAP and gain-based feature importance revealed important clinical and toxicological patterns. They identified key predictors, including specific organ system disorders (e.g., bronchial and lung disorders, heart disorders) linked to fatal outcomes, animal demographic information, and notable drug characteristics. These findings not only confirm the accuracy of the model's predictions but also highlight clinically relevant patterns that can support veterinarians, regulators, and researchers in understanding patterns across AEs, animal demographics, and drug-related attributes that are associated with reported

outcomes. The proposed framework offers a scalable and transparent approach for integrating AI into veterinary PV workflows, enabling earlier detection of high-risk drug–event profiles. This capability can inform evidence-based regulatory actions, guide veterinary prescribing practices, and ultimately protect both animal welfare and human food safety.

6.2 Future work

While this study focused primarily on structured clinical and demographic features available in the OpenFDA CVM dataset, future research could be significantly enhanced by integrating molecular-level and pharmacokinetic (PK) information for each veterinary drug. This additional layer of mechanistic detail would enrich the feature space with descriptors that capture how drugs behave in the body and why certain AEs occur. At the molecular level, descriptors such as chemical fingerprints, topological indices, molecular weight, lipophilicity (LogP), hydrogen bond donors/acceptors, and polar surface area can be obtained from public repositories like DrugBank, PubChem, and ChEMBL. These descriptors would allow the model to infer potential toxicological mechanisms based on chemical structure similarity and known toxicity profiles of related compounds (López-Pérez et al., 2024). From a pharmacokinetic perspective, parameters such as absorption rate, bioavailability, distribution volume, plasma protein binding, half-life, clearance rate, and metabolic pathways could provide crucial insight into exposure–toxicity relationships (Riviere et al., 2017). For example, drugs with long half-lives may accumulate in tissues, increasing the risk of delayed or chronic AEs. Integrating these molecular and PK features with existing clinical AE data would transform the model from a purely statistical classifier into a mechanistically informed predictive framework. This could enable the detection of high-risk drug–event profiles even for drugs with limited historical AE reports by leveraging structural and mechanistic analogies. Moreover, such integration would support explainable AI approaches, where feature importance could be tied back to well-understood toxicokinetic principles, improving interpretability for veterinarians, regulators, and toxicologists.

Bibliography

C. Bousquet, G. Lagier, A. L.-L. Louët, C. Le Beller, A. Venot, and M.-C. Jaulent. Appraisal of the meddra conceptual structure for describing and grouping adverse drug reactions. *Drug Safety*, 28(1):19–34, 2005.

W. Chang, X. Ji, Y. Xiao, Y. Zhang, B. Chen, H. Liu, and S. Zhou. Prediction of hypertension outcomes based on gain sequence forward tabu search feature selection and xgboost. *Diagnostics*, 11(5):792, 2021.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

J. Chen, J. Yan, Q. Chen, D. Z. Chen, J. Wu, and J. Sun. Excelformer: A neural network surpassing gbdts on tabular data. *arXiv preprint arXiv:2301.02819*, 2023.

A. Dahlin, A. Eriksson, T. Kjartansdottir, A. Markestad, and K. Odensvik. The atcvet classification system for veterinary medicinal products. *Journal of Veterinary Pharmacology & Therapeutics*, 24(2), 2001.

L. Dijkstra, T. Schink, R. Linder, M. Schwaninger, I. Pigeot, M. N. Wright, and R. Foraita. A discovery and verification approach to pharmacovigilance using electronic healthcare data. *Frontiers in Pharmacology*, 15:1426323, 2024.

A. Farnoush, Z. Sedighi-Maman, B. Rasoolian, J. J. Heath, and B. Fallah. Prediction of adverse drug reactions using demographic and non-clinical drug characteristics in faers data. *Scientific Reports*, 14(1):23636, 2024.

A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.", 2022.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

H. O. Khalifa, L. Shikoray, M.-Y. I. Mohamed, I. Habib, and T. Matsumoto. Veterinary drug residues in the food chain as an emerging public health threat: Sources, analytical methods, health impacts, and preventive measures. *Foods*, 13(11):1629, 2024.

D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

K. López-Pérez, J. F. Avellaneda-Tamayo, L. Chen, E. López-López, K. E. Juárez-Mercado, J. L. Medina-Franco, and R. A. Miranda-Quintana. Molecular similarity: Theory, applications, and perspectives. *Artificial intelligence chemistry*, 2(2):100077, 2024.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

L. Margiotta-Casaluci, S. F. Owen, and M. J. Winter. Cross-species extrapolation of biological data to guide the environmental safety assessment of pharmaceuticals—the state of the art and future priorities. *Environmental Toxicology and Chemistry*, 43(3):513–525, 2024.

Y. M. Mesfin, B. A. Mitiku, and H. Tamrat Admasu. Veterinary drug residues in food products of animal origin and their public health consequences: A review. *Veterinary Medicine and Science*, 10(6):e70049, 2024.

Y. Pan, Y. Wang, Y. Zheng, J. Chen, and J. Li. A disproportionality analysis of fda adverse event reporting system (faers) events for ticagrelor. *Frontiers in Pharmacology*, 15:1251961, 2024.

G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020.

J. E. Riviere, L. A. Tell, R. E. Baynes, T. W. Vickroy, and R. Gehring. Guide to farad resources: Historical and future perspectives. *Journal of the American Veterinary Medical Association*, 250(10):1131–1139, 2017.

U.S. Food and Drug Administration. OpenFDA: Data Downloads. https://open.fda.gov/data/downloads/. Accessed: 2025-07-25.

WHO Collaborating Centre for Drug Statistics Methodology. ATCvet — Anatomical Therapeutic Chemical Classification System for veterinary medicinal products. https://www.whocc.no/atcvet/, 2025. Accessed: 2025-08-12.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

X. Xu, R. Mazloom, A. Goligerdian, J. Staley, M. Amini, G. J. Wyckoff, J. Riviere, and M. Jaberi-Douraki. Making sense of pharmacovigilance and drug adverse event reporting: comparative similarity association analysis using ai machine learning algorithms in dogs and cats. *Topics in companion animal medicine*, 37:100366, 2019.

X. Xu, J. E. Riviere, S. Raza, N. I. Millagaha Gedara, R. Ampadi Ramachandran, L. A. Tell, G. J. Wyckoff, and M. Jaberi-Douraki. In-silico approaches to assessing multiple high-level drug-drug and drug-disease adverse drug effects. *Expert Opinion on Drug Metabolism & Toxicology*, 20(7):579–592, 2024.

Y. Yang, S. Wei, H. Tian, J. Cheng, Y. Zhong, X. Zhong, D. Huang, C. Jiang, and X. Ke. Adverse event profile of memantine and donepezil combination therapy: a real-world pharmacovigilance analysis based on fda adverse event reporting system (faers) data from 2004 to 2023. *Frontiers in Pharmacology*, 15:1439115, 2024.

N. Zad, L. A. Tell, R. A. Ramachandran, X. Xu, J. E. Riviere, R. Baynes, Z. Lin, F. Maunsell, J. Davis, and M. Jaberi-Douraki. Development of machine learning algorithms to estimate maximum residue limits for veterinary medicines. *Food and Chemical Toxicology*, 179: 113920, 2023.

X. Zhong, Y. Yang, S. Wei, and Y. Liu. Multidimensional assessment of adverse events of finasteride: a real-world pharmacovigilance analysis based on fda adverse event reporting system (faers) from 2004 to april 2024. *PloS one*, 20(3):e0309849, 2025.