DEVELOPING AN LSTM-BASED FRAMEWORK FOR ASSESSING THE RISK OF ADVERSE DRUG REACTIONS IN POLYPHARMACY PATIENTS USING MULTIMODAL DATA

by

FOTINI DROUMALIA

B.Sc., Informatics Engineering, Hellenic Mediterranean University, 2022

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

HELLENIC MEDITERRANEAN UNIVERSITY

2025

Approved by:

Major Professor
Manolis Tsiknakis

Abstract

Adverse drug reactions (ADRs) represent a significant challenge in clinical care due to their rarity, heterogeneity, and potential severity. Accurate prediction of rare ADRs is critical to prevent harm, guide safer prescribing practices, and optimize patient outcomes, reducing both clinical complications and healthcare costs. This study aimed to develop a predictive model for tailored ADR risk assessment using the UK Biobank dataset, comprising over 2 million samples, including demographic, prescription, diagnosis, and genotype data, while special attention was given to the severe class imbalance, with only 0.39% of samples representing positive ADR cases.

We trained a Long Short-Term Memory (LSTM)-based predictive model using multi-modal input features and applied isotonic regression for probability calibration. Although overall accuracy was high, this metric was misleading due to extreme class imbalance, which led us to concentrate on evaluating the model based on metrics mainly focusing on the minority class, such as precision, recall, F1/F3 scores, and PR AUC. Calibration substantially improved precision and reduced false positives, but recall remained low, reflecting the challenge of detecting rare ADRs in the presence of severe class imbalance and incomplete prescription, genetic, and other patient-level information. Another limitation was the lack of interpretability of the deep learning model because of its complex nature, which reduces its practical usefulness in real-world clinical decision-making.

The results indicate that increasing model complexity by incorporating multiple data modalities did not substantially improve predictive performance, since many positive cases were eventually missed, despite the model being able to identify some rare ADRs. Limitations include incomplete drug exposure information, missing patient-level features, and the inability to incorporate intended genetic features due to challenges in converting imputed genotypes into usable alleles. These limitations, combined with extreme class imbalance, restricted the model's capacity to fully capture rare ADR patterns, highlighting the challenges of applying deep learning models to real-life clinical datasets.

Despite the fact that predictive performance in this study was limited, the findings provide insights into challenges for patient-specific risk assessment and highlight avenues for future improvement. Future work should focus on integrating richer clinical, pharmacological,

and genomic information, leveraging ADR-focused resources, as well as reformulating the prediction problem as an anomaly detection task supported by ensemble techniques. Addressing class imbalance and improving data completeness and quality could enhance both the performance and interpretability of ADRs, thereby increasing their clinical relevance and interpretability for real clinical settings.

Table of Contents

Abstract ii
Table of Contents iv
List of Figures vi
List of Tables vii
Acknowledgements viii
Preface ix
Chapter 1 - Introduction 1
Chapter 2 - State of the Art 3
Clinical and Biomedical Databases 3
ADR-Focused Databases 7
Adverse Drug Reactions Studied 9
Approaches to Adverse Drug Reaction Prediction 11
Interpretability 17
Data Imbalance 18
Time-to-Onset of ADRs 20
Chapter 3 - Data 21
Data Sources 21
Demographic Data 21
Prescription Data 25
Diagnoses Data 28
Genomic Data 30
Harmonization 31
Data Preprocessing 35
Diagnoses Preprocessing 36
Demographics Preprocessing 37
ADRs List 38
Windows Extraction 38
Chapter 4 - Methodology 48

Model Input Preparation 48
Model Architecture 49
Model Training 51
Chapter 5 - Results 54
Chapter 6 - Discussion 58
Chapter 7 - Limitations 61
Chapter 8 - Future Work 63
Chapter 9 - Conclusion 65
References 67

List of Figures

Figure 1. Basic approach on windows extraction 40
Figure 2. Windows extraction scenario where no diagnoses are reported at all 40
Figure 3. Windows extraction scenario where an ADE diagnosis is reported within the candidate non-ADE window period 41
Figure 4. Multimodal Attention LSTM Pipeline for ADE Risk Prediction 51

List of Tables

Table 1. Clinical and Biomedical Databases Used in ADR Research 6
Table 2. ADR-Focused Databases 8
Table 3. Approaches to ADR Prediction 16
Table 4. Demoraphic data obtained from UKB Biobank 23
Table 5. UK Biobank Encoded Values for Smoking and Alcohol Status 24
Table 6. UK Biobank Encoded Values for Alcohol Style 24
Table 7. UK Biobank Encoded Values for Ethnic Background 24
Table 8. UK Biobank Encoded Values for Qualifications 25
Table 9. Prescription Data Extracted from UK Biobank 26
Table 10. Prescription data with ATC codes 28
Table 11. ICD-10 Diagnoses Extracted from UK Biobank 29
Table 12. Dates of First In-Patient Diagnoses Extracted from UK Biobank 29
Table 13. Combined ICD-10 Diagnoses with First Diagnosis Dates from UK Biobank 29
Table 14. Demographic Data Features Description 32
Table 15. Genotypes Data Features Description 34
Table 16. Diagnosis Data Features Description 35
Table 17. Prescription Data Features Description 35
Table 18. Patient Window Dataset 43
Table 19. Class Distribution of Windows in the Full Dataset 46
Table 20. Gender Distribution of Patients in the Dataset 46
Table 21. ADE Occurrence by Gender (Patients With at Least One ADE vs. None) 46
Table 22. Descriptive Statistics of Participant Ages 47
Table 23. Distribution of Negative and Positive Windows Across Train, Validation, and Test Sets 49
Table 24. Model performance before and after calibration 56

Acknowledgements

I would like to sincerely thank my supervisor, Manolis Tsiknakis, for giving me the opportunity to work in this exciting research field and for their guidance throughout the project. I would also like to thank my advisor, Vassilis Kilintzis, for working with me on resolving issues, for helping me understand the problem in depth, and for providing valuable clinical perspectives that complemented the technical aspects of this work.

I would also like to acknowledge the members of the broader SafePolyMed team, including clinicians and technical experts, whose insights, advice, and support were vital in shaping the research and deepening my understanding of the clinical and technical challenges.

Preface

This thesis is the result from my participation in a collaborative project, SafePolyMed, focused on Adverse Drug Reaction (ADR) risk assessment, in which we worked closely with clinicians, pharmacological experts, and federated learning specialists. The project aimed to improve patient safety by identifying potential ADRs, giving me the opportunity to have hands-on experience with real-life healthcare applications.

During this work, several real-world challenges came to light, among which was the severe class imbalance of the dataset, which complicated the analysis and required careful handling. Additionally, the data often lacked important information, like, for example, we had access to prescription dates but not the complete medication plans for patients, therefore we did not have any details regarding the date, or even dates, the medication was actually consumed. These limitations highlighted the practical difficulties of translating ADR research into valuable insights for clinical practice.

Looking back on the experience, I gained knowledge into how ADR risk assessment is conducted in practice. I learned to communicate effectively with healthcare and pharmacology experts and to tackle real-world constraints such as incomplete data and interpretability concerns. SafePolyMed not only shaped the direction of this thesis but also deepened my understanding of patient-centered approaches in pharmacovigilance and the practical challenges involved in improving medication safety.

Chapter 1 - Introduction

Adverse drug reactions (ADRs) are defined by the World Health Organization (WHO) as harmful, unexpected reactions to medications that occur at doses usually administered in treatment [1]. These reactions can be dangerous and may occur either due to the pharmacological qualities of the medicine, drug-drug interactions, or underlying medical conditions [2]. While ADRs are a subset of adverse drug events (ADEs), a more general term that also incorporates harm from overdoses and prescription errors, this study does not make a distinction between the two terms. According to [3], ADEs occur after the administration of a medication at any dosage level and may or may not result in patient harm, like for example when over-dosage causes increased monitoring but no harm, whereas ADRs occur at normal therapeutic doses and lead to harmful and unintended effects. It is important to note that 'normal dose ranges' are defined according to standard dosing guidelines specific to each drug, which consider typical therapeutic ranges. For this study, ADRs and ADEs are treated as the same, since the focus of our research is on predicting clinically significant medication-related harm regardless of classification nuances.

ADRs are a complicated problem with many clinical and biological contributors and the risk of an ADR occurring is significantly increased in cases of polypharmacy, which is typically defined as the use of five or more drugs at the same time, according to a review by de Souza et al. [4]. Beyond polymedication and comorbidities, an individual's likelihood to experience an ADR can be greatly influenced by patient age, improper medication usage, as well as variations in genetics, such as differences in enzymes that metabolize drugs and drug targets [5]. With that said, ADRs are especially common among elderly people, because the older a person is, the more the comorbidities they suffer from, and, therefore, the more the different drugs they need for treatment (i.e., polypharmacy) [6].

Assessing and predicting the likelihood of ADRs has become a crucial area of study in clinical practice, due to the fact that these events can lead to serious health complications and increased healthcare costs, even though they can be typically prevented [7]. Precise ADR prediction improves patient safety and quality of life while also advances the pharmaceutical research pipeline by promoting safer medication development and monitoring methods [8].

Recent advances in electronic health records, data integration, and health informatics have increased the availability of real-world clinical datasets, providing new opportunities for research in the healthcare sector. These datasets, which highlight the complexity of clinical data, often include both static and time-dependent patient information, that need to be properly examined using sophisticated analytical techniques, like multimodal approaches. These approaches handle several types of data and can be very effective because they may extract complementary knowledge from each data modality. Deep learning models can be used for this cause due to their ability to effectively uncover hidden patterns and model complex connections within and across these many data sources, making them particularly well-suited for clinical risk prediction tasks.

In order to ascertain whether a patient is at risk for ADRs, this study incorporates a number of factors, including genetic information, drug prescriptions, diagnoses, and demographic data (such as age, sex, and lifestyle). To achieve this, our multimodal design uses two attention-based LSTM networks to extract temporal relationships from longitudinal data, such as prescription and diagnosis data. Static data is processed by two Multilayer Perceptrons (MLPs): one for genomic data and one for demographic data. With this differentiation, every modality may be evaluated according to its distinct characteristics before being integrated into a unified representation for prediction.

Chapter 2 - State of the Art

In medical settings, ADRs remain a major concern, especially for patients with polypharmacy, where the risk is increased by variation among individuals and drug-drug or drug-gene interactions. Given that the same medication can have significantly differing impacts in different people and because interactions between drugs or between drugs and genes may alter outcomes in unpredictable ways, detecting ADRs is a particularly challenging task. According to [9], about 3 out of every 1000 hospital admissions lead to death as a result of an ADR, making early detection of ADRs crucial as they can significantly increase patient negative effects. In addition to death, ADRs extend hospital stays, and raise healthcare expenses, which is why prompt diagnosis and prevention are critical clinical and scientific priorities. As a result, there is increasing interest in using machine learning (ML) and deep learning (DL) techniques, which may combine multimodal patient data to find patterns and more accurately predict the probability of ADRs.

Clinical and Biomedical Databases

In ADR research, a number of extensive clinical and biological databases are frequently employed. With around 500,000 individuals, the UK Biobank [10] is a unique population-based cohort that offers a wealth of biomedical data. Biomedical data in the UK Biobank includes, among other things, imaging, biomarkers, genetics, medical records, physical measurements, demographic and lifestyle information, questionnaire responses, and environmental exposures. To preserve participant privacy, all identifying information is anonymised and researchers access the data via a secure portal. Comprehensive patient-level datasets like these enable multidimensional analyses that relate clinical, environmental, and genetic factors to the risk of ADRs.

The extensive, publicly accessible MIMIC-III database contains deidentified health information linked to 53,423 unique hospital admissions of adult patients (age 16 and up) to the Beth Israel Deaconess Medical Center's critical care units between 2001 and 2012 [11]. The database contains data on demographics, laboratory test results, procedures, drugs, radiology reports, caregiver feedback, mortality (including post-hospital discharge), and vital sign metrics (approximately one data point per hour) [12]. MIMIC-III facilitates a wide variety of

scientific studies in the areas of electronic tool design, clinical decision-rule optimization, and epidemiology. It is noteworthy due to the fact that it is openly accessible to researchers anywhere, incorporates a wide range of ICU patients, and includes extremely detailed information such as vital signs, test results and prescriptions. [13] provided an example of how this dataset was used to extract ADEs from clinical notes.

Another large publicly accessible database, MIMIC-IV, contains information on every patient admitted to the hospital or critical care units at Beth Israel Deaconess Medical Center between 2008 and 2019 [14]. In comparison with MIMIC-III, which solely incorporates ICU patients from 2001 to 2012, MIMIC-IV encompasses both ICU and general hospital admissions and structures data into three modules (ICU, hospital, and core), which makes it possible to conduct more thorough hospital-wide studies than when using MIMIC-III. In addition to ICU-related data, MIMIC-IV provides comprehensive information on patient demographics, laboratory values, drugs, provider orders, prescriptions, diagnoses, and microbiology measurements. Study [15] presents an example of how one can utilize this dataset by applying ML models for predicting inpatient ADEs based on ICD-10-CM codes using the MIMIC-IV database, combining sociodemographics, baseline hospital admission data, and previous prescriptions as predictive factors.

The Salzburg Intensive Care Database (SICdb) offers high-resolution temporal data that are ideal for studies that require information related to the progression of patients' health conditions in comparison to MIMIC-IV [16]. Temporal features enable more accurate and complex modeling of healthcare outcomes, such as predicting ICU re-admission of a patient or severity of a patient's disease, using deep learning approaches like LSTM networks [38,39]. Detailed time series data can also offer valuable insights into patient pathways in the ICU and help predict adverse events. Additionally, SICdb includes data on ICU patients during the COVID-19 pandemic, allowing researchers to study adverse effects on patient care that are related to a pandemic, with a preliminary comparison study demonstrating that COVID-19 patients in SICdb had significantly longer lengths of stay than patients in MIMIC-IV. However, unlike MIMIC-IV, SICdb records only primary ICD-10 codes, which makes it challenging to capture secondary diagnoses and to calculate combined scores, such as comorbidity indices, and, in addition, the current version of SICdb does not contain clinical notes or imaging data. Despite these limitations, the publication of SICdb represents a significant advancement in the

availability of diverse, high-resolution ICU data, with strong potential to support the development of artificial intelligence (AI) applications in healthcare [16]. However, during our study we could not find any studies that employed SICdb for ADR prediction tasks, which underscores its potential as an underutilized resource in this domain.

Collectivelly, these clinical and biomedical databases provide to researchers access to large-scale, diverse patient information that is essential for studying ADRs. While UK Biobank enables population-level analyses linking lifestyle, genetics, and environmental factors to ADR risk, the MIMIC databases provide detailed, time-stamped hospital and ICU records that support fine-grained predictive modeling of adverse events. The more recent SICdb expands these opportunities by offering high-resolution temporal ICU data and coverage of the COVID-19 pandemic period, although it has not yet been widely adopted in ADR prediction research.

Table 1. Clinical and Biomedical Databases Used in ADR Research

Database	Size \ Coverage	Data Types	Strengths	Limitations	Example Use in ADR Research
UK Biobank	~500,000 individuals (population-based)	Imaging, biomarkers, genetics, medical records, demographics, lifestyle, environment	Multimodal population-level data; strong genetics & lifestyle info	No detailed hospital/ICU data; not ADR-specific	Linking genetics + environment + clinical risk factors for ADRs
MIMIC-III	53,423 ICU admissions (2001–2012)	Demographics, labs, procedures, drugs, vital signs, mortality, notes	Open access; rich ICU data; time-stamped vitals	ICU-only; older data	Extracting ADEs from clinical notes
MIMIC-IV	All hospital & ICU admissions (2008–2019)	Demographics, labs, prescriptions, diagnoses, microbiology, ICU & hospital records	Broader scope than MIMIC-III; structured into ICU, hospital, core modules	Still hospital-only; preprocessing complexity	ML models for inpatient ADEs based on ICD-10-CM
SICdb	ICU patients (includes COVID-19 era)	High-resolution time series, ICU data	High temporal resolution; includes COVID-19; suitable for LSTM modeling	Only primary ICD-10 codes; no notes or imaging	Used for ICU outcomes; no ADR-focused studies yet

ADR-Focused Databases

In addition to patient-centric resources, several databases focus specifically on drug–adverse-event relationships, providing essential data for pharmacovigilance and predictive modeling of ADRs.

SIDER [17] is a widely used resource that provides curated information on marketed drugs and their associated side effects. Beyond simply listing ADRs, SIDER includes frequency data that indicates how often each adverse event occurs for a given drug, enabling researchers to prioritize clinically significant reactions or study rare events. SIDER has been used in numerous studies aiming to identify ADR mechanisms, predict potential off-target drug effects, and model drug repurposing and benchmark computational methods for ADR detection. Its structured format and mappings to standard vocabularies, such as MedDRA preferred terms, are extremely useful, since they facilitate harmonization across studies and computational pipelines.

DrugBank [18] complements SIDER by providing in-detail chemical, pharmacological, and pharmaceutical data, as well as comprehensive information on drug targets such as their sequences, structures, and pathways, which are necessary for understanding how drug mechanisms work and how they influence ADRs. For example, drug enzymes can influence drug metabolism and increase the risk of ADRs, while drug transporters and carriers influence absorption, distribution, metabolism, excretion, and drug toxicity [19]. Therefore, by integrating SIDER and DrugBank features that often use standard identifiers such as ATC codes to connect datasets, ADR prediction models are enabled to take into consideration both the phenotypic effects along with the molecular characteristics of drugs that are obtained from multiple separate sources.

FDA Adverse Event Reporting System (FAERS) [20] offers another helpful source of ADR information, by collecting voluntary reports of post-marketing adverse event and medication error reports submitted by healthcare professionals or even patients. The reports are coded according to the MedDRA terminology and follow international safety reporting standards, which makes FAERS a key resource for pharmacovigilance studies. In addition to adverse event information, FAERS provides basic patient demographics, such as age and gender, as well as drug treatment details, patient outcomes, and report source data [19] and is organized into multiple linked tables, that allows researchers to examine links between drug exposures, patient characteristics, and reported adverse events.

Table 2. ADR-Focused Databases

Database	Data Content	Strengths	Limitations	Example Use in ADR Research
SIDER	Curated marketed drugs + side effects; frequency data; MedDRA mappings	Structured; frequency info; harmonized with vocabularies	Literature/document-based; not patient-level	Predicting off-target effects, benchmarking ADR detection models
DrugBank	Drug chemical & biological properties (targets, enzymes, transporters, carriers)	Rich molecular-level info; useful for mechanistic studies	Not ADR-specific; needs integration with ADR sources	Combining with SIDER for multimodal ADR prediction
FAERS	Spontaneous post-marketing ADR reports; demographics, drug therapy, outcomes	Large-scale, real-world post-market data; MedDRA coding	Voluntary reporting → bias & inconsistencies; preprocessing heavy	ML models using demographics + drug features for ADR prediction

Adverse Drug Reactions Studied

Across literature, predictive modeling of ADRs exhibits diverse strategies regarding the target outcomes and, even when studies use the same source of ADR labels, preprocessing and selection strategies differ. For example, [21] uses drug-ADR pair labels obtained from the SIDER 4.1 database, which contains information on marketed drugs and their reported ADRs collected from public documents. ADR terms are standardized using the MedDRA preferred term (PT) vocabulary to avoid semantic redundancy. In this study, only ADRs observed for at least one drug are included, ensuring that the dataset contains meaningful drug-ADR associations for predictive modeling. These curated ADRs are then combined with drug structural and gene expression features to train the BiMPADR model. Similarly, [22] also relies on SIDER 4.1 but applies additional filtering to handle overlapping drug-ADR and drug-indication associations, treating indications as negative examples to reduce class imbalance. Only ADRs that are consistently associated with drugs after filtering are included in the dataset for training the ADRP-VETA model, which uses a random forest classifier and voting ensemble to predict drug-ADR associations.

In contrast, [19] relies on the FAERS database, which contains voluntary clinical reports of ADRs along with patient demographics. Here, ADRs are also based on MedDRA preferred terms, but extensive preprocessing is required to handle inconsistencies in drug naming, including brand names, ingredient names, and typographical errors. Both demographic and non-clinical drug features are leveraged to train Random Forest (RF) and deep learning models for ADR prediction.

Study [23] integrates multiple sources, including SIDER, CTD, LINCS L1000, PubChem, and MeSH. ADRs are harmonized across vocabularies and included only if they appear in multiple databases, while ADRs that cannot be consistently mapped are excluded. This study also performs additional preprocessing to link ADRs with gene and chemical interaction data, creating a multi-source, high-quality dataset. The study then proposes DGANet, a CNN-based model that integrates chemical-gene interaction (CGI) and drug-perturbed gene expression (GE) into drug features, and gene-disease associations (GDA) into ADR features, to predict drug-ADR associations.

In study [24], the ADRs come directly from commercial electronic health record (EHR) claims data. Experts from the IQVIA clinical team extracted adverse reaction records for each

patient and linked them to the 12 months of EHR data preceding the event. Unlike studies that rely on public databases like SIDER or FAERS, the ADR labels in [24] are patient-specific clinical outcomes, reflecting real-world observed adverse events rather than literature- or report-based drug-ADR associations. Building on this dataset, the study proposes pADR, a personalized multi-sourced prediction network that combines chemical drug features with patient EHRs through a hierarchical multi-sourced transformer, addressing the challenge of personalized ADR prediction.

In addition to studies that model broad sets of ADRs across many drugs, some research narrows its scope to specific adverse events of clinical interest. For example, [25] uses the IBM MarketScan Explorys Claims-EMR dataset to define two cohorts, each focused on a single ADR. The first cohort predicts chronic opioid use as a harmful outcome of prescription opioid exposure, while the second predicts extrapyramidal symptoms (EPS) associated with antipsychotic initiation. Unlike studies that harmonize large ADR vocabularies such as MedDRA, this work relies on targeted cohort definitions and carefully engineered risk factors (e.g., demographic, clinical, behavioral, and pharmacological variables) to model well-specified ADR outcomes. Such focused approaches allow researchers to investigate risk factors and prediction strategies in depth for ADRs of high medical and societal relevance.

Likewise, Song et al. [26] concentrated on predicting the likelihood of two particular adverse events, myocardial infarction (MI) and acute kidney injury (AKI), in patients with acute coronary syndrome (ACS) who were treated with percutaneous coronary intervention (PCI). The authors implemented eight ML models, among which were XGBoost, random forest, and linear discriminant analysis (LDA), using data from 5,240 patients to predict death by any cause within a twelve-month period, MI during their stay in the hospital, and AKI. Shapley Additive Explanations (SHAP) scores were used to identify patient features, which included demographics, clinical assessments, lab results, and prescription use. The LDA model performed the best, showing good discriminative power with an AUC of 0.83 for overall mortality and 0.74 for AKI and MI.

These examples demonstrate that even when researchers collect data from the same ADR source, differences in preprocessing, filtering, and inclusion criteria can lead to substantially different final datasets. While some studies focus on specific ADRs, others, like the present work, consider a broader set of adverse events, with this variability highlighting the importance

of clearly describing ADR selection and preprocessing methods when comparing predictive modeling approaches in literature.

Approaches to Adverse Drug Reaction Prediction

Initially, studies on detecting ADEs mostly depended on drug-centric techniques that compiled post-market data, like the research conducted by Padhi et al. (2024) [27], which used FAERS spontaneous report data and drug similarity metrics along with a label propagation algorithm to detect ADEs in advance. Their results showed that by incorporating pre-clinical information with drug similarity metrics enhanced early ADE detection compared to standard algorithms, yet the approach focused on the patient population as a whole rather than treating each patient individually, which made its clinical relevance in tailored care limited, underscoring the need for more sophisticated, patient-centered approaches.

In a similar vein, other studies adopted a drug-centric approach by leveraging the molecular and interaction characteristics of drugs. One such work is [28], which developed a deep learning predictor that combined multiple drug features, including one- and two-dimensional molecular structure information, drug-protein interactions, and drug similarity. The model was able to achieve improved performance over conventional multi-label classifiers, with an AUPR increase from 64.02% to 66.19% on a benchmark dataset, and from 34.65% to 68.82% on LIU's dataset. While this approach advanced ADR prediction in drug discovery and demonstrated robustness across evaluation metrics, it still concentrated solely on the medications rather than patient-related information.

Combined, these drug-centric studies underscore the strengths of ADR signal detection using detailed drug information but also reveal the lack of information on patient-level, indicating the need for personalized ADR prediction approaches that incorporate individual demographics, comorbidities, and clinical histories to improve clinical applicability.

The analysis conducted by Yang et al. [29] demonstrates how advanced ADR detection is now possible throughout the entire drug development and monitoring lifecycle thanks to the adoption of ML approaches. ML makes it easier to combine data from various sources, including clinical data, genomics, and electronic health records, which is essential for complex problems such as ADR prediction. Additionaly, it makes it possible to detect safety signals from unstructured text and structured data, including social media and scientific publications, as well

as identify new, unknown ADRs. Regardless of these advancements, the authors also highlight three main challenges associated with ML, including the need for extensive data, problems with model interpretability, and legal restrictions. Moreover, they emphasize the use of genomic data for individualized ADR risk predictions which suggests a major change from one-size-fits-all models to precision pharmacovigilance.

Building on these advances, Dsouza et al.'s systematic review [3] summarizes the results of 13 studies that employ multimodal ML models in healthcare, which, in order to evaluate ADR risks, they frequently incorporate demographics, clinical laboratory results, comorbidities, and medical histories. When compared to models that utilize a limited set of features, the review discovered that incorporating such a wide range of patient-specific variables greatly enhances the accuracy and precision of ML-based ADR prediction. It also emphasized the importance of proper data integration across patient parameters for increasing AI-driven pharmacovigilance and patient safety, as well as the need for more robust model development and evaluation in order to ensure clinical applicability.

The afforementioned findings highlight the shift toward multimodal, patient-centered ADR prediction, like, for example, study [24], which proposes a personalized multi-sourced ADR prediction model with the named pADR. The proposed approach initially converts each individual data source into proper representations and then employs a hierarchical multimodal Transformer to model interactions and fuses them for the final prediction. In particular, the authors compose a new dataset by combining descriptions of drug pharmacodynamics, EHR from the 12 months prior to an adverse event, patient demographics, like age and gender, and chemical characteristics of medications in SMILES codes. Therefore, employing this multi-source dataset enables tailored ADR prediction by framing ADR prediction as a multi-label prediction problem, where each medicine may be linked to many negative side effects at the same time. The findings gathered via experiments conducted using this dataset show that pADR is superior to cutting-edge drug-based baseline approaches, while case and ablation studies also confirm the efficacy of the suggested fusion techniques, as well as the robustness of each module of the structure.

Extending these multimodal strategies, Bi et al. [5] provide a in-depth review of the increasing significance of genomic data in the medical domain, with an emphasis on the role that pharmacogenomics plays in ADR prediction. Since ADRs are a serious public health issue that

contribute to an increased rate of morbidity and mortality, the authors highlight how advancements in genomic technologies have deepened our understanding of the genetic bases of drug reactions. In addition to summarizing current use of pharmacogenomics in ADR prediction, diagnosis, prevention, control, and individualized treatment, the review explains the relationship between ADRs and genetic polymorphisms. Bi et al. also discuss how multimodal learning models, which combine genomic, phenotypic, electronic health record, and environmental data, can detect intricate drug response patterns, increase prediction accuracy, and facilitate tailored risk assessment by integrating diverse patient-specific information such as genome, epigenetic, transcriptome, clinical history, medication records, and biomarkers.

Similarly, Anastopoulos et al. [30] offer a novel approach that integrates drug-level structural data and patient-level data through the use of graph convolutional networks (GCNs). This study uses the FDA's FAERS dataset and the UK Biobank's dataset to train and evaluate their algorithm, and, according to their experiments, models that combine pharmacological, demographic, and genetic variables perform better than those that only use one data source. In particular, their proposed model outperforms standard ML methods at predicting hospitalization and death, with an R^2 of 0.37 for hospitalization and an AUC of 0.90 for death. However, despite the fact that the addition of genetic and demographic characteristics to the FAERS dataset greatly enhances the model's capacity to predict outcomes like hospitalization and death, issues like dataset bias and the requirement for calibration still exist. In summary, this study demonstrates how, in the setting of polypharmacy, integrating diverse patient features along with molecular drug characteristics might facilitate tailored ADR prediction.

Despite their significant contribution to the discipline, the previously mentioned studies typically don’t make use of the temporal structure of patient health records, although the time dimension in clinical settings can heavily influence the outcome. For this reason, a number of studies attempt to incorporate the temporal aspect of the data in their approaches, one of which is Shi et al. [31], which introduces an hierarchical, time-aware neural network called HTNNR, that mimics the sequential structure of longitudinal patient data, for ADE Risk prediction. HTNNR captures intra- and inter-visit interactions by leveraging semantic embeddings to encode diagnoses, procedures, and drugs, which are processed through a hierarchical bi-directional Recurrent Neural Network (RNN) architecture, with the model's ability to gain knowledge from the time intervals between visits being further enhanced by time-aware components. Using a

dataset of half a million patients, HTNNR was compared to several common methods for ADE prediction like Random Forest, XGBoost, LSTM, and CNN, and managed to outperform all other approaches, achieving an AUC of 0.86, precision of 0.80, recall of 0.48, and accuracy of 0.87. It is evident that, across all metrics, the suggested HTNNR approach consistently performs the best, with these results confirming the importance of incorporating patients' long-term and evolving health history when predicting ADE risk.

Another interesting study is [32] that uses LSTM models for assessing the temporal ADR data obtained by the Center for ADR Monitoring of Guangdong, with the authors converting the data into sequences and employing a LSTM model for data classification. The presented LSTM managed to outperform commonly used approaches, such as Logistic Regression, Random Forest, K-Nearest Neighbor, and MLP, by achieving an F1-score of 0.9706 and a sensitivity of 0.9516. The model's perfomance, as presented in this work, demonstrates the ability of LSTM architectures to identify sequential patterns in ADR data, and therefore highlights the significance of modeling temporal information for personalized ADE prediction.

An additional AI-based pharmacovigilance study conducted an evaluation on Gradient Boosting, CNNs, and LSTMs using both structured EHR data and unstructured social media data [33]. The results revealed that CNN performed best on social media data, obtaining an AUC-ROC of 0.91 and a sensitivity 90%, while Gradient Boosting achieved the highest performance on structured EHR data, obtaining an AUC-ROC 0.89. Moreover, feature importance analysis and SHAP explanations were employed to provide interpretability, revealing patient age, drug type, and dosage as key predictors for ADRs. The study also demonstrates how Natural Language Processing (NLP) can be valuable in extracting ADR signals from unstructured sources, which may further enhance traditional pharmacovigilance, and emphasizes the importance of balancing predictive performance with interpretability.

Study [34], expanding on the use of unstructured data for ADR prediction, applied a hybrid AI framework which integrated structured EHR data, lab results, and unstructured clinical notes from trials, with their proposed CNN models combined with NLP techniques, like BERT and GPT, reaching an accuracy of 85%, and outperforming traditional ML methods such as Logistic Regression and Support Vector Machine (SVM). Furthemore, feature analysis highlighted patient demographics along with clinical factors as key predictors for ADRs,

demonstrating that applying NLP can potentially improve pharmacovigilance and personalized ADR risk assessment.

We can observe some common trends in these studies, specifically that drug-centered, post-market detection tools have given way to patient-centered, tailored models that use a variety of clinical and genomic data. In order to improve the practicality and predictive performance of models, multiple data types, including genomics, clinical data, and patient history need to be combined into one collective structure, while it is important to note that time-based models make it easier to monitor how a patient's health progresses over time, especially in cases of chronic or complicated conditions. In addition, recent advances emphasize how integrating unstructured data sources, such as social media and scientific publications, and employing interpretable AI methods can enhance our understanding of the logic behind a model's prediction, as well as our trust in clinical outcomes. Ultimately, even if AI has a lot of potential, real-world applications still face significant challenges in clinical practice like data quality, interpretability, ethical monitoring, and class imbalance, that need to be tackled.

Based on all of the aforementioned arguments, our study focuses on multimodal approaches for ADR prediction in polypharmacy patients by adopting methods that are considered appropriate for each type of data. In particular, we employ a multimodal, LSTM-based framework that integrates demographic information and genomic data as static features, while drugs and diagnoses are handled as temporal sequences, with this setup allowing the model to deal with the unique characteristics of each data source separately and capture how patient health evolves over time, which demonstrates an effort toward personalized ADR prediction.

Table 3. Approaches to ADR Prediction

Study	Data Sources	Methods	Key Contributions	Limitations
Padhi et al. (2024)	FAERS	Label propagation + drug similarity	Early ADE detection in post-market data	Not patient-centered
Ou et al. (2024)	Molecular data (drug structure, drug–protein interactions, drug similarity, benchmark & LIU datasets)	Deep learning fusion model (4 feature modules)	Improved predictive accuracy	Drug-centric; ignores patient heterogeneity and comorbidities
Luo et al. (2023)	EHR claims + drug features	Hierarchical transformer (pADR)	Personalized ADR prediction	Proprietary dataset; not public
Anastopoulos et al. (2021)	FAERS + UK Biobank	Graph Convolutional Networks	Integrates molecular + demographic + genetic data	Dataset bias; calibration required
Shi et al. (2020)	Half a million patients (longitudinal)	Hierarchical time-aware neural net	Temporal modeling with visit-level embeddings	Computationally heavy
Cao et al. (2025)	Temporal ADR monitoring data	LSTM	Strong sequential ADR prediction (F1 = 0.97)	Limited to regional dataset

Pramanik et al. (2024)	Structured EHR + unstructured social media	Gradient Boosting, CNN, LSTM + NLP	CNN best on social media (AUC 0.91); Gradient Boosting best on EHR (AUC 0.89); interpretable via SHAP	Accuracy-interpretability trade-off; complex data integration
Khemani et al. (2025)	Structured EHR, lab results, clinical notes	Hybrid AI: CNN + NLP (BERT, GPT)	Combines structured & unstructured data; interpretable; CNN accuracy 85%	Limited to clinical trial datasets; unclear generalizability

Interpretability

As ML models for ADR prediction become increasingly complex, interpretability has emerged as a critical consideration, because, while deep learning and ensemble methods can achieve high predictive performance, their "black-box" nature often limits the understanding of how individual features contribute to a model's decisions. This limitation poses a challenge for the adoption of such models in real-life clinical applications, since healthcare providers need transparent and explainable models in order to trust their predictions and make decisions based on them.

One study [19] illustrates this trade-off by comparing RF and deep learning (DL) models which are trained on demographic and non-clinical drug features, including chemical, molecular, and biological characteristics. In this research, the DL model managed to achieve higher predictive performance (AUC 0.674) but was inherently more difficult to interpret, while, on the other hand, the RF model, though slightly lower in predictive accuracy (AUC 0.611), was able to detect the top contributing features, which included five demographic and 15 non-clinical factors. It is also important to note that a simple RF model which used only these top 20 features was able to achieve comparable performance to a model that used the complete set of 2,315 features, demonstrating that interpretable, feature-focused models can make accurate predictions while also being transparent.

Expanding on interpretability, He et al. propose ToxBERT and demonstrate how transformer-based models can identify drug substructures that are associated with specific ADRs [35]. For example, ToxBERT achieved AUROC scores of 0.839, 0.759, and 0.664 for predicting drug-induced QT prolongation, rhabdomyolysis, and liver injury, respectively, while highlighting mechanistic links between chemical structures and ADRs. These results show that advanced deep learning models can provide insights regarding how biological mechanisms work while demonstrating high predictive accuracy, allowing black-box models to support applicable interpretability.

Additionally, the review study conducted by Nurzhan et al. underlines the importance of interpretability when integrating patient-specific factors beyond clinical and molecular data, based on the fact that environmental and lifestyle variables, such as diet, physical activity, smoking, and alcohol use, can significantly influence how an individual responds to a drug [36]. Such heterogeneous data types, including structured clinical data and non-clinical patient information can be handled using AI models, which in turn can provide more personalized predictions while maintaining transparency, with methods such as SHAP values, LIME, and feature importance rankings, that are crucial to explain model outputs, ensure regulatory compliance, and facilitate clinician trust, enabling safer adoption of AI in pharmacovigilance.

All of the above show how important it is to balance performance with interpretability in the context of ADR prediction. For patient-centered pharmacovigilance, understanding the drivers of model predictions, such as patient age, sex, or molecular drug properties, can inform safer prescribing practices, identify high-risk patient subgroups, and support regulatory reporting requirements. Integrating interpretability tools, such as feature importance analysis, SHAP, LIME, or attention mechanisms, remains an active area of research to make predictive models actionable in real-world clinical settings.

Data Imbalance

ADRs are extremely scarce in EHRs, with positive cases making up a very small portion of the overall patient population. ADR occurrences are significantly less common than non-ADR instances in many datasets, which leads to a very unbalanced class distribution. Because misidentifying a tiny proportion of infrequent events has little effect on the overall loss function, traditional ML algorithms, which are intended to optimize overall loss rates, are prone to be

biased toward the majority category, which is patients who are not suffering from ADRs. As a result, models may seem to function well based on conventional measures like accuracy, but they may be unable to recognize the minority ADR cases that are most clinically meaningful.

A number of studies have emphasized the challenges posed by class imbalance in ADR prediction, such as [28], [37], [38], [39], with these studies showing that unbalanced data can lead to misleading performance evaluations, reduced clinical relevance, and poor detection of rare but potentially serious ADRs. For instance, a model that predicts as "no ADR" every instance in a dataset dominated by non-ADR case, although it could achieve high overall accuracy, it would not be able to learn any patterns leading to ADR, which would potentially lead to putting patients' health in danger and decreasing the confidence of the predictive system.

Addressing class imbalance can be achieved by both data-centric strategies, such as resampling techniques, including oversampling the minority class or undersampling the majority class, which can rebalance the dataset, though they may introduce overfitting or result in the loss of information, as well as algorithm-level approaches, like cost-sensitive learning, weighted loss functions, ensemble methods, or anomaly detection, which can emphasize correct classification of rare events while maintaining model stability.

Anomaly detection, sometimes referred to as outlier detection [40], discovers observations, events, or data points that substantially differentiate from the dataset's typical behavior, which makes it suitable for identifying rare ADRs. Ensemble learning is a technique that combines two or more base learners, like neural networks or regression models, to make more accurate predictions than an individual model separately [41]. In ADR prediction, ensemble learning has been employed for anomaly detection, enhancing the ability to identify rare drug events, as shown in [42] for post-marketing ADR monitoring. In order to identify safety signals, this study combines many clustering or anomaly detection techniques in an ensemble framework, treating adverse occurrences with disproportionately high reporting rates as anomalous, with their combined approach demonstrating how ensemble anomaly detection can enhance the identification of rare yet clinically significant drug events by successfully identifying known ADRs included in the FAERS database.

Additionally, the selected evaluation metrics for a model's performance can be chosen in such a way that focuses on the minority class, by using metrics such as precision-recall curves, or F beta scores, that can provide a more accurate assessment of a model's ability to detect rare

occurrences of ADRs. Therefore, it is important to recognize and mitigate the effects of extreme class imbalance in order to develop more robust, and clinically relevant ADR prediction models and to ensure that rare but serious ADRs are not overlooked due to the abundance of non-ADR cases.

Time-to-Onset of ADRs

ADRs are characterized by not only being highly heterogeneous in how they are expressed in medical settings but also in their timing, which means that some reactions, such as immediate hypersensitivity, can occur within minutes of exposure, while others, including immune-mediated or neurological effects, may take days, weeks, or even months to emerge [9], [43].

The inherent heterogeneous nature of ADRs can complicate prediction, since individual differences in metabolism, comorbidities, genetics, and concomitant treatments, can make a single medication be linked to several ADRs, each with a unique onset pattern, while also the same ADR may appear at different time points across various patients. In addition to the afforementioned influential factors, predictive efforts are further complicated by the fact that environmental factors, including dietary habits, lifestyle, and exposure to other substances or medications, can further affect the likelihood and initiation of ADRs. Moreover, the occurrence and severity of ADRs can be also influenced by temporal patterns of drug administration, including dosage, frequency, and duration.

In summary, in order to accurately assess the clinical causality of ADRs and build predictive models we need to pay close attention to temporal dynamics, longitudinal patient monitoring, and integration of multi-source clinical, demographic, and molecular data. Through deeply understanding these temporal and mechanistic complexities, personalized ADR risk assessment tools can be developed, and patient safety can be imporoved in both clinical trials and post-marketing surveillance.

Chapter 3 - Data

Data Sources

The UK Biobank served as the source of this study's data, which is a British biomedical database that contains genetic and health data for approximately half a million UK participants. It is used as a resource for scientific research that improves human health. The participants in the cohort provided by UK Biobank are aged 52 to 89. To enhance heterogeneity of the cohort, 22 assessment centers were located around the UK and the participants' origins were either England (89%), Scotland (7%) or Wales (4%).

After acquiring the data access license, we downloaded genomic and clinical information that were considered relevant to the study's purpose for 502,357 participants from the DNANexus platform.

Demographic Data

For this project, the most appropriate patient-related variables linked to ADEs were chosen. [44] demonstrates how age has a major impact on how ADR develops. ADRs are more likely to occur in elderly people with a history of ADRs, various medical conditions, multiple medications, and who are unable to reduce medications. It additionally supports the case that because of biological variations between the male and female, gender can have an impact on ADRs. By changing the pharmacokinetics and pharmacodynamics of the pharmaceuticals, including the way they are absorbed, distributed, metabolized, and eliminated, these variations may have an impact on how the body responds to medications.

Social factors were also collected as relevant to ADEs. Alcohol drinking affects the metabolism of many drugs, and it facilitates the development of ADRs [45]. Ethnicity is controlled by genetic factors, that can result in differences on individual-level and polymorphisms in genes encoding drug metabolizing enzymes, drug transporters, and receptors. Smoking is identified as one of the risk factors for peptic ulcer, cardiovascular diseases and cancer. Smoking induces liver cytochrome 450 iso-enzymes CYP1A1, 1A2 and possibly 2E2 leading to the increase in metabolism of drugs that are substrate to this group of enzymes and decrease their pharmacological action.

Demographic and lifestyle information was obtained from several UK Biobank data files. The features used in this study include participant ID (participant.eid), birth year (p34), sex (p31), body mass index (p21001_i0), smoking status (p20116_i0), alcohol consumption status (p20117_i0), alcohol intake frequency (p1558_i0), ethnic background (p21000_i0), and educational qualifications (p6138_i0). Mortality-related information was collected in 'death_record.csv', which provides the date of death, and 'death_cause_record.csv', which contains the cause of death coded using ICD-10.

All of these files are linked via the participant ID, enabling the creation of a combined dataset that captures demographic, lifestyle, and health outcomes across multiple time points. For variables with multiple assessment instances (e.g., BMI, smoking, alcohol consumption, ethnic background, or qualifications), only the first available measure was retained in our dataset to ensure consistency across participants and to reduce missingness due to incomplete follow-up. The final dataset used in this study contains the selected features described in Table 4.

Table 4. Demoraphic data obtained from UKB Biobank

demographic.csv

Column	Data Type
participant.eid	INT64 (ex. '1000074')
participant.p34 (Birth Year)	INT64 - years (ex. 1946)
participant.p31 (Gender)	INT64, Categorical (1-male or 0-female)
participant.p21002 (Weight)	INT64 - kg (ex. 54)
participant.p50 (Standing Height)	INT64 - cm (ex. 63.78)
participant.p21001_i0 (BMI)	FLOAT64, Continuous - kg/m2 (ex. 33.6129)
participant.p20116_i0 (Smoking status)	INT64, Categorical (see Table 5)
participant.p20117_i0 (Alcohol Drinker Status)	INT64, Categorical (see Table 5)
participant.p1558_i0 (Alcohol Intake Frequency)	INT64, Categorical (see Table 6)
participant.p21000_i0 (Ethnic Background Visit)	INT64, Categorical (see Table 7)
participant.p6138_i0 (Qualifications)	INT64, Categorical (see Table 8)
Date of Death	DATE (ex. '2012-12-21')
Cause of death - ICD-10	STRING (ex. 'I38 Endocarditis, valve unspecified')

The UK Biobank provides several categorical variables in an encoded numeric format, including smoking status, alcohol status and style, ethnic background, and educational qualifications. The mappings between the numeric values and their corresponding categories are provided by UK Biobank and are summarized in Tables 5-8.

Table 5. UK Biobank Encoded Values for Smoking and Alcohol Status

Smoking and Alcohol Status

Code	Name
-3	Prefer not to answer
0	Never
1	Previous
2	Current

Table 6. UK Biobank Encoded Values for Alcohol Style

Alcohol Style

Code	Name
-3	Prefer not to answer
1	Daily or almost daily
2	Three or four times a week
3	Once or twice a week
4	One to three times a month
5	Special occasions only
6	Never

Table 7. UK Biobank Encoded Values for Ethnic Background

Ethnic Background

Code	Name
1	White
1001	British
2001	White and Black Caribbean
3001	Indian
4001	Caribbean
2	Mixed
1002	Irish
2002	White and Black African

3002	Pakistani
4002	African
3	Asian or Asian British
1003	Any other white background
2003	White and Asian
3003	Bangladeshi
4003	Any other Black background
4	Black or Black British
2004	Any other mixed background
3004	Any other Asian background
5	Chinese
6	Other ethnic group
-1	Do not know
-3	Prefer not to answer

Table 8. UK Biobank Encoded Values for Qualifications

Qualifications

Code	Name
1	College or University degree
2	A levels/AS levels or equivalent
3	O levels/GCSEs or equivalent
4	CSEs or equivalent
5	NVQ or HND or HNC or equivalent
6	Other professional qualifications eg: nursing, teaching

Prescription Data

Prescription information for participants was obtained from the UK Biobank GP prescriptions dataset (gp_scripts), which contains over 56 million prescription records across 8 columns. The dataset provides detailed information on general practitioner (GP) prescriptions, that includes medication name, date, and dosage, and covers participants from August 2019

(version March 2023), but it is important to note that prescription records are available for only about 45% of the UK Biobank cohort, which is approximately 230,000 participants, meaning that a large portion of participants lack drug-related information, with this limitation being particularly important since medication data is critical for predicting ADRs. Therefore, the incomplete coverage of prescription data underscores one of the challenges faced in the field of ADR risk assessment and in modeling multimodal, temporal patient data, due to the fact that the absence of prescription history restricts the ability of the model to fully capture individual drug exposure patterns.

For drug-related information, we downloaded a file from the UK Biobank, with the name 'ukbb_prescription.csv' (see Table 9), by selecting data fields related to drug prescriptions and contains records of drugs prescribed to participants, including the participant ID, drug name, quantity issued, and date of prescription. Drugs are provided as drug names (e.g., Bisoprolol 2.5 mg tablets), and ATC codes are not included in the original file. To match drug names with their corresponding ATC classifications we used the publicly available ATC code-name table provided by the World Health Organization (WHO). This dataset provides a longitudinal record of medications dispensed to each participant and was linked to other data files via the participant ID for integration into the final dataset.

Table 9. Prescription Data Extracted from UK Biobank

ukbb_prescription.csv

Column	Data Type	Unique Values
Participant ID	INT64 (ex. '1000074')	222,073
Drug Name	STRING (ex. 'Bisoprolol 2.5mg tablets')	80,684
Quantity issued	STRING (ex. '84 tablet(s)')	94,066
Date prescription was issued	DATE (ex. '2012-12-21')	11,328

The initial UK Biobank prescription dataset ('ukbb_prescription.csv') contained 56,216,350 rows, covering 222,073 participants. The initial table did not include ATC codes, and no direct mapping from Readv2, DM+D, or BNF codes to ATC was available in the literature.

We processed the prescription data in three steps to assign ATC codes to drug names:

1. Active Substance Matching: The first word of the drug name, usually representing the active substance, was matched against a USAAR-provided list of 6,791 active

substances with corresponding ATC codes. If an active substance corresponded to multiple ATC codes (e.g., ibuprofen), only the first ATC code alphabetically was retained. This step left 22,245,463 rows unmatched (NaN).

2. Manual Mapping for Prevalent Drug Names: A manual dictionary of 35 common drug names (including package names, vaccines, creams, eye drops, supplements, etc.) was used to map additional unmatched entries. After this step, 14,928,646 NaNs remained, achieving approximately 73% coverage.
3. BNF Code Matching: For remaining unmatched rows, BNF codes were used to map drugs to ATC codes, increasing coverage slightly to 75% of the original dataset.

After concluding the processing phase, the final prescription dataset included 1,021 unique ATC codes, which represented a standardized classification of medications in the UK Biobank cohort, but, although the original dataset contained tens of thousands of distinct drug names, many of these names corresponded to the same pharmacological substance or therapeutic class. Mapping drug names to ATC codes therefore reduced the apparent diversity, consolidating multiple synonymous or related drug names into a smaller, manageable set of standardized codes. Furthermore, several drug names could not find any match to any ATC code, which led to excluding some rows from the final dataset, highlighting that while the ATC mapping standardizes the dataset, it does not preserve all original prescription records.

The resulting processed prescription dataset, which is named ‘prescription.csv’, contained only the most relevant information for longitudinal and sequence-based analyses, that is the participant identifier, the date of prescription, and the assigned ATC code, and the structure of this final dataset, including the data types and number of unique values in each column, is presented in Table 10.

Table 10. Prescription data with ATC codes

prescription.csv

Column	Data Type	Unique Values
Participant ID	INT64 (ex. '1000074')	201,117
Date prescription was issued	DATE (ex. '2012-12-21')	10904
atc_code	STRING (ex. 'R03AK06')	1,021

As shown in Table 10, the 'Participant ID' column allows the connection of multiple different UK Biobank datasets, which enables the integration of prescription data with phenotypic, clinical, and genomic data. More thoroughly, the 'Date prescription was issued' column provides information regarding when the prescription was issued and is necessary for constructing medication histories and extracting windows of prescriptions for modeling, while the 'atc_code' column represents the correcponding drug that was prescribed as an ATC code. By keeping only these three columns, the final file contains only the necessary information required for our modeling task.

Diagnoses Data

In order to collect information on diagnoses, we initially downloaded ICD-10 coded diagnosis reports from the UK Biobank, with the primary file, 'ICD10.csv' (see Table 11), containing the participant ID and the corresponding diagnosis (Data-Field p41270). Dates of first in-patient diagnoses are recorded across 259 separate files ('ICD10_dates_a0.csv' to 'ICD10_dates_a258.csv'), each containing the participant ID and an array representing the interpolated year of first diagnosis for a subset of conditions (Data-Fields p41280_a0 through p41280_a258), as described in Table 12. To consolidate these data for analysis, the files were merged, producing 'diagnoses.csv' (see Table 13), which contains the participant ID, ICD-10 diagnoses, and the corresponding first in-patient diagnosis dates.

Table 11. ICD-10 Diagnoses Extracted from UK Biobank

ICD10.csv

Column	Data Type	Counts
Participant ID	INT64 (ex. ‘1000074’)	502,364
participant.p41270 (Diagnoses - ICD10)	STRING - ex. ["M201", "Z139"]	404,991

Table 12. Dates of First In-Patient Diagnoses Extracted from UK Biobank

ICD10_dates_a0.csv - ICD10_dates_a258.csv

Column	Data Type
Participant ID	INT64 (ex. ‘1000074’)
participant.p41280_a0 (Date of first in-patient diagnosis - ICD10)	DATE - ex. ‘2012-12-21’
…	…
participant.p41280_a258 (Date of first in-patient diagnosis - ICD10)	DATE - ex. ‘2012-12-21’

Table 13. Combined ICD-10 Diagnoses with First Diagnosis Dates from UK Biobank

diagnoses.csv

Column	Data Type
Participant ID	INT64 (ex. ‘1000074’)
ICD-10	STRING (ex. ‘J111’)
Date	DATE - ex. ‘2012-12-21’

The initial UK Biobank disease dataset (see Table 11), containing ICD-10 codes and diagnosis dates, had a non-standard format where multiple dates corresponded to a list of ICD-10 codes. To convert this into a usable format we transformed the dataset into time series events. Participants whose ICD-10 lists did not match the corresponding dates were excluded.

This processing ensured that both prescription and disease datasets were structured consistently, with ATC-coded prescriptions and ICD-10-coded diagnoses suitable for integration into model analysis.

Genomic Data

Our ADE risk prediction model incorporates HLA-A and HLA-B genotypes as features because of their proven correlations with immune-mediated ADRs. Lamotrigine-induced cutaneous responses have been associated with specific alleles, including HLA-A*24:02, with meta-analyses demonstrating a significantly higher risk in comparison to tolerant controls [46]. In a similar manner, certain HLA-B alleles, such as HLA-B*15:02 and HLA-B*58:01, have been proven to be highly linked to severe adverse reactions like Stevens-Johnson syndrome and toxic epidermal necrolysis induced by carbamazepine and allopurinol [47]. Based on the previous, it becomes apparent how HLA genotypes may contribute to an individual's predisposition to ADEs, and subsequently how important it is to combine these genetic markers with clinical and demographic information in order to increase the predictive power of tailored ADE risk models.

A dataset containing data for the HLA-A and HLA-B genes were obtained from the UK Biobank (Data-Field 22182), which provides four-digit HLA alleles for each participant, and therefore there is no need to perform imputation. These data were the result of an analysis conducted by UK Biobank, as described in [48], which followed the HLA*IMP:02 method proposed by Dilthey et al. (2023), with a diverse reference panel being utilized for the imputation, that included datasets from African, Asian, Latino, and European (CEU+58) populations, along with the Type 1 Diabetes Genetics Consortium (T1DGC) and the 1000 Genomes Project. From the SNPs that were detected on the UK Biobank Axiom array, those that were within 1 Mb of each HLA locus were the only ones considered for the final imputation step.

In particular, the dataset we obtained includes the HLA types that are most likely to appear per individual, which means that for each patient there is one row that contains a single imputed HLA allele for each of the two chromosomes and the selected 2 loci (HLA-A and HLA-B). Additionally, the posterior probability, Q, is provided to measure the confidence of the imputation, with low-confidence calls being suggested to be filtered out using a recommended Q threshold of 0.7. The final HLA-A and HLA-B alleles used in our study are reported as combinations of star alleles (e.g., *1/*1 or *1/*2), which represent each participant’s diploid HLA genotype in a format suitable for downstream analyses.

Harmonization

This study is part of the larger SafePolyMed project, which attempts to develop predictive models for ADE risk assessment across multiple healthcare databases using federated learning because it allows models to be trained collaboratively on distributed datasets without the need to share patient-level data, meaning that privacy is preserved while the diversity and scale of multiple cohorts is exploited. In this context, it is important to harmonize each dataset to a common standard, such as the Observational Medical Outcomes Partnership (OMOP) data model, so that features obtained from different sources are comparable and can be used collectivelly as inputs for model training. Therefore, althought this thesis focuses on analyses that were performed using solely the UK Biobank dataset, the harmonization step performed converts it to a format that aligns it with other databases utilized by the SafePolyMed project for federated model development.

The following tables provide details on the resulting harmonized datasets showing the name of the columns after being converted to OMOP codes, the corresponding feature names, a brief description of the features, and, in some cases, the predefined sets of values in OMOP codes that specific categorical features accept. Specifically, Table 14 describes the demographic features, Table 15 the genotype features, Table 16 the diagnosis-related features, and Table 17 the prescription features, and, in total, these tables demonstrate how each original UK Biobank column has been standardized to a common coding system in order to ensure compatibility and make integration easier with other datasets.

Table 14. Demographic Data Features Description

Column (in OMOP Code)	Feature	Data Description	Values (OMOP Codes)
1008215	Participant ID	The participant ID (i.e. 'EGV1')	-
35810093	Year of birth	4-digit float or NaN (i.e. 1948.0)	-
35810091	Sex	OMOP code integers	35818840 – Female, 35820554 – Male 763013 – Not provided
35811394	BMI	Float number with inconsistent decimal points or NaN	-
35811013	Smoking status	OMOP code integers	35821337 – Never, 35822845 – Previous, 35817596 – Current, 763013 – Not provided
35811014	Alcohol status	OMOP code integers	35821337 – Never, 35822845 – Previous, 35817596 – Current 763013 – Not provided

35811393	Ethnicity	OMOP code integers	38003614 – White, 8515 – Asian, 700391 – Mixed background, 38003600 – African, 8522 – Other race, 763013 – Not provided
35810746	Qualifications	OMOP code integers	1620732 – Primary school education, 35821787 – GCSEs, 35821953 – e.g. nursing or teaching, 35817386 – College Degree, 45878298 – MSc 763013 – Not provided
35811395	Weight	Float number with inconsistent digits or NaN	-
35810116	Height	OMOP code integers as floats with 1 decimal point (e.g., 184.0 instead of 184) or NaN	-

35810226	Alcohol intake frequency	OMOP code integers	35821337 – Never, 35822845 – Previous, 35817596 – Current, 35821830 – One to three times a month, 35824290 – Three or four times a week, 35817657 – Daily or almost daily 763013 – Not provided

Table 15. Genotypes Data Features Description

Column (in OMOP Code)	**Feature**	**Data Description**
1008215	Participant ID	The participant ID (i.e. 'EGV1')
40792628	HLA-A	Allelic combination as string (i.e. '*201/*201') or NaN
40779425	HLA-B	Allelic combination as string (i.e. '*4001/*4402') or NaN

Table 16. Diagnosis Data Features Description

Column (in OMOP Code)	Feature	Data Description
1008215	Participant ID	The participant ID (i.e. 'EGV1')
44810321	Diagnosis	ICD-10 codes without dots as strings (i.e. I251 instead of I25.1) (see 'Diagnosis Preprocessing' section for handling -, +, and space characters)
4160852	Date of diagnosis	Date value as string (i.e. '1996-12-03')

Table 17. Prescription Data Features Description

Column (in OMOP Code)	Feature	Data Description
1008215	Participant ID	The participant ID (i.e. 'EGV1')
42528920	Date of prescription	Date value as string (i.e. '2012-04-12')
1031609	Prescription	ATC codes as strings (i.e. 'R03AK06')

Data Preprocessing

The preprocessing phase of the input data is an essential stage in model training, since the quality and structure of the data greatly influence how a model performs, especially in applications like risk assessment for ADEs which is a health-related problem that could affect an individual's well-being. Due to the reason that ADE risk assessment is a complicated process because of how multiple factors such as patient demographics, medication histories, diagnostic codes, and genetic factors can impact the outcome, it is required to apply suitable preprocessing

in order to ensure that the data can accurately depict the real-world circumstances under which an ADR may or may not occur, which could subsequently enhance reliable risk assessment.

Diagnoses Preprocessing

Each row in the diagnoses file is meant to represent a single diagnosis code linked to a particular patient report, however multiple diagnoses are frequently included in a single report. Therefore, the data structure is adjusted so that each ICD-10 code is displayed in a distinct row by splitting reports that contain multiple codes instead of just one. This means that if a single report has more than one diagnosis code, this specific report will be represented by several rows in the dataset, each of which will have the same report identification but a different ICD-10 code.

In greater depth, these sets of numerous diagnosis codes that corresponded to a single report were frequently separated by characters like "-", "+", or even spaces (for example, "I78-O67", "I78+O67", or "I78 O67"), which led us to manage different characters as distinct scenarios by setting specific guidelines to guarantee that every row complied with the one-code-per-row requirement. The two primary cases of these guidelines are the following:

1. Hyphen-separated codes ("-"): This case means that numerous unique diagnoses were recorded in a single patient report, which means that two or more diagnoses, such as "I78-O67", that are reported in a single report indicate that all codes (I78 and O67 in this example) were reported on the same date. In that scenario, we divide the combined diagnosis codes into separate reports while maintaining the original report identifier for each of them.
2. Plus-sign or space-separated codes ("+" or " "): Codes separated by plus-signs or spaces may indicate that it is not clear which code is the correct diagnosis or that only one of the codes is applicable. For example, "I78+O67" or "I78 O67" mean that there is confusion regarding which of the several diagnoses was the actual one, which led us to put off handling these cases until the windows extraction phase. Instead of attempting to resolve this during preprocessing, we put off handling these occurrences until the windows extraction phase. At the windows extraction phase, distinct windows are generated for every code. That means that if a diagnosis report falls within this case by combining multiple codes using plus-signs or spaces, such as "K296+B980", then the report is divided into as many

windows as are the codes, in this scenario two windows, each of which will contain one of the codes until all codes are used. Therefore, in our example, one window will contain the “K296” code while the other one will contain the “B980” code at the same time-point, guaranteeing that each one of the codes is taken into account independently.

Several other formatting guidelines were used in order to standardize the ICD-10 codes. Codes were restricted to a minimum of three characters and a maximum of four characters. Since the trailing “X” in ICD-10 codes serves as a placeholder and does not indicate important diagnostic information, it was dropped from codes that concluded with the letter “X” (for instance, “H71.X”).

Demographics Preprocessing

The demographic data was preprocessed in order to ensure that the data are complete, consistent, and compliant with this study’s objectives and, to achieve that, we initially reindexed the dataset so that it included the required demographic variables that were selected as features relevant to ADRs and followed a predetermined column order. Patients without any information on age, gender, or BMI were removed entirely from the study due to the importance of the features for the specific task, without any imputation technique being applied to fill empty fields that could increase model bias [19]. For categorical features that were encoded with OMOP standards, we replaced any empty values with the concept ID 763013 which indicates that the data was not provided ("Not provided") without requiring imputation.

Body Mass Index (BMI), if not provided by the participant, was calculated using height and weight measurements, if available, but before applying the BMI formula, height was translated from centimeters to meters. Once calculated, the original height and weight columns were removed since BMI is a more informative feature for ADR prediction than height or weight alone. Additionally, the final BMI value was rounded and set to -3 if it was not provided or could not be calculated due to the absence of height and weight. Moreover, instead of keeping the date of birth, we extracted and used from it only the year of birth, which makes time-based modeling simpler by calculating the age of the participant within each window solely based on the year of birth.

After the preprocessing was concluded, the output dataset was filtered so that it would include only the following columns: participant ID, year of birth, gender, BMI, smoking status, alcohol status, ethnicity, qualifications, and alcohol intake frequency.

ADRs List

For this study, the list of ADRs used to flag windows as ADR or non-ADR were based on on a comprehensive set of ICD-10 codes synthesized by study [49], which systematically identified ICD-10 codes typically used to identify ADEs in large healthcare databases. Clinical professionals, that we collaborated with, reviewed and refined the code list to preserve only clinically relevant codes by following the approach of a validation study [50]. To provide additional information, they conducted a strict and broad analysis, where they included all A1 and A2 level codes while excluding those that have low PPV, while also flagging codes that appeared to be odd.

The results that they collected were then shared with us, and by utilizing the additional information that the experts supplied for each code, we managed to eliminate broad or vague codes, while preserving ADR related codes that were clearly relevant. This additional preprocessing of the initial ADR list produced a set of 131 high-confidence ADE-related ICD-10 codes and ensured that the final list contained only codes with strong clinical relevance, reducing the risk of including ambiguous diagnoses that could dilute model accuracy. This list was then used to characterize patient windows as ADE or non-ADE, as described in section “Windows Extraction”, with the full list of selected ADR-related ICD-10 codes being available in the supplementary file ‘complete_ADEs_list.xlsx’.

Windows Extraction

Since ADRs are events that happen due the interaction of multiple drugs that were consumed within a time period that they were active, it is important to construct the data in such a way that these interactions between medication exposure and ADE occurrences are captured. For this reason, we constructed windows that define the problem based on its clinical context, in a manner that positive windows illustrate real-life ADE cases, as well as real-life interactions between drugs and diagnoses, while negative windows accurately represent non-ADE occurrences, with the aim to reduce dataset noise resulting from keeping information that is irrelevant to ADEs. By labeling each window as ADE or non-ADE, we treat this problem as a

binary classification task, with the objective being to determine whether a set of events within a window relates to an ADE.

The windows extraction procedure for a patient, that is illustrated in Figure 1, starts with the identification of a medication ($medication_0$), which is known as the target medication, and acts as the starting point of the window, and the date on which this drug is issued is t_0. It is important to note that each patient's medications are sorted from earliest to oldest based on the prescription date, since we do not have detailed information regarding the medication plan, such as the duration, or whether the medication was actually consumed, and the windows extraction process starts with the earliest prescribed medication as $medication_0$. Due to the lack of more detailed information on the medication plan we can only assume that the drug might have been consumed in the days following the prescription, but there is no way to confirm that it was actually consumed. Having the prescription date t_0 of the $medication_0$ as the starting point for the current window, a 60-day period after date t_0 (duration: $[t_0+1, t_0+60]$) is evaluated to see if an ADE diagnosis has been reported.

In the case that no ADE is reported during this 60-day period, the window is labeled as a non-ADE window and includes all prescription and diagnosis events that happened from 30 days prior to t_0 until the actual prescription day (duration: $[t_0-30, t_0]$), while the search for the next focus medication starts the day after t_0.

If an ADE does occur during the 60-day period, the window is labeled as an ADE window, and, in that scenario, the window includes any event that was reported from 30 days before t_0 to one day before the ADE date ($t_{ADE}-1$). In this case, the search for the next focus medication starts from the day after the ADE ($t_{ADE}+1$) occurred, because drugs that were prescribed on t_{ADE} may or may not be related to the ADE, since we don't know if they were consumed right before the ADE, making it difficult to decide between one of the two scenarios.

The time durations used for assessing whether an ADE occurred or which events are considered relevant to an ADE or a non-ADE window is based on findings in the literature, which show that ADEs can have different onset times, like immediate reactions within hours or delayed effects that are triggered over weeks or even months. Given the heterogeneity that characterizes ADEs, a period of 2 to 3 months after the consumption of a drug provides a reasonable balance by being long enough to capture delayed-onset ADEs but short enough to maintain only clinically relevant events, while the 30-day period before the prescription of the

focus medication provides additional information regarding the patient's health before the drug prescription, and can help distinguish conditions that already existed from events that may actually be related to the medication, like the diagnosis that led to the prescription of the specific medication.

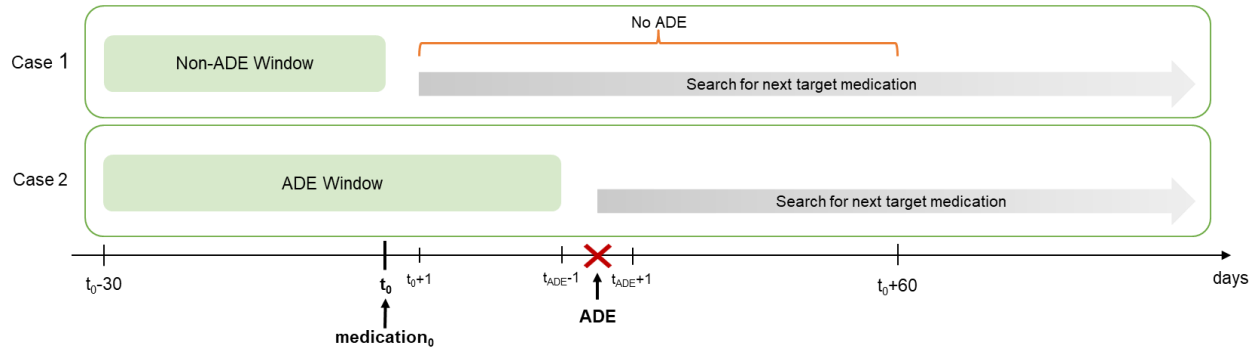


Figure 1. Standard approach on windows extraction

Although the standard approach constructs windows that reflect quite well how ADEs occur, there are some scenarios in which special handling is necessary to more precisely capture events related to ADE occurrences, and in order to prevent adding bias or noise to the dataset, these scenarios are addressed separately based on each scenario's unique requirements.

For instance, since ADEs are characterized by specific diagnosis codes, the absence of a patient's diagnosis data automatically suggests that there are no ADEs. Consequently, all extracted windows for these individuals are marked as non-ADE, as shown in Figure 2.

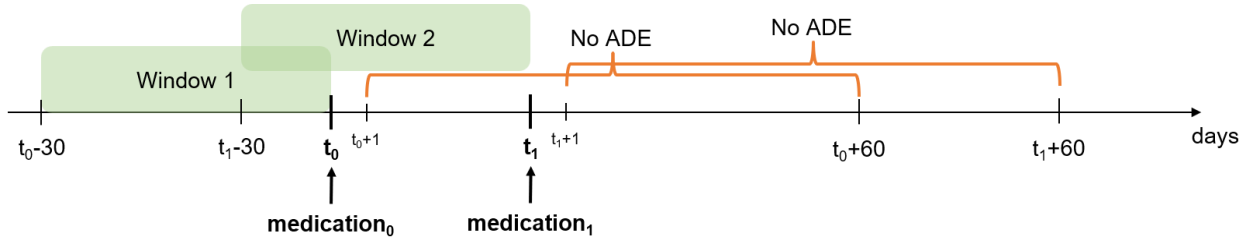


Figure 2. Windows extraction scenario where no diagnoses are reported at all

When an ADE diagnosis is discovered during the [t_0-30, t_0] interval that was utilized to create a candidate non-ADE window, as indicated by the red "X" in Figure 3, another unique scenario occurs. The existence of an ADE diagnosis within the actual window period renders it inappropriate as a clean non-ADE example, even if no ADE is found within 60 days after the prescription date t_0. To prevent uncertainty in these situations, the window is rejected.

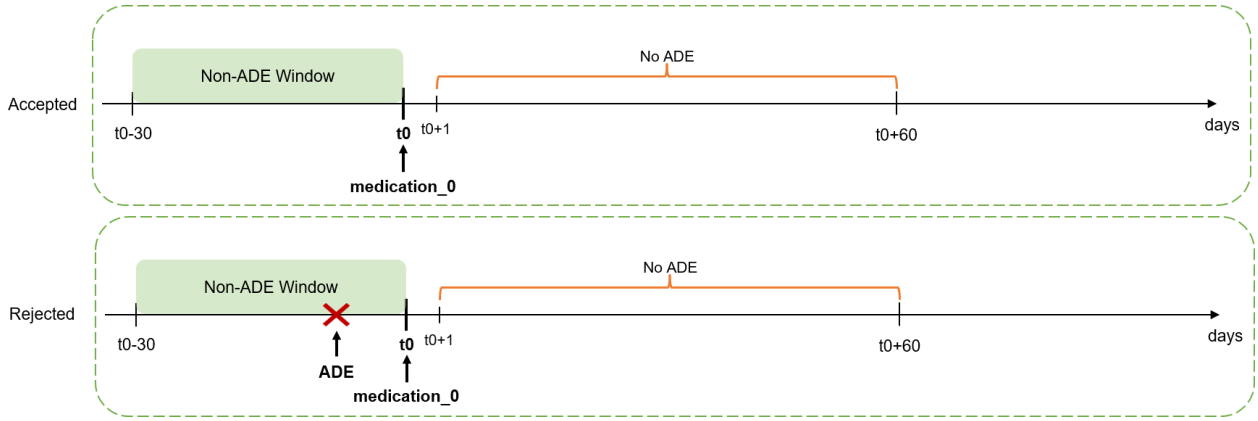


Figure 3. Windows extraction scenario where an ADE diagnosis is reported within the candidate non-ADE window period

A separate scenario discusses window extraction's approach to handle chronic diseases. Similar to other diagnoses, a chronic disease is included in the associated timeframe when it is discovered for the first time in the data. Even if the chronic disease is not specifically mentioned again, it is automatically carried forward to all following windows for that patient. It is anticipated that chronic conditions will continue to exist in every subsequent window following their initial occurrence since they persist throughout time.

An important consideration arises when a chronic disease also qualifies as an ADE. In this setting, the initial occurrence of the chronic ADE is handled similarly to a typical ADE diagnosis: an ADE window is established if it happens within 60 days of the target medication's prescription. Once the chronic ADE has been discovered, it is no longer regarded as a trigger for future ADE windows. By default, it is incorporated into all upcoming windows and is handled as a chronic, ongoing condition. However, if the first instance of the chronic ADE occurs during a potential non-ADE window's period of events (that is, between t_0–30 and t_0), the window is dismissed.

Each time window is restricted to a maximum of 15 medication events and 15 diagnosis events in order to guarantee consistency in the model input and preserve a balance among informativeness and computational performance. This number is a reasonable number of events that are expected to occur within the given timeframe and was established based on clinical relevance, while also, it enables the model to get a realistic view of the clinical status of the patient without adding too much information that may be redundant or unnecessary. It is worth mentioning that, within each window, the medication events are ordered from oldest to most

recent, with medication_1 being the oldest while medication_15 being the most recent, while the same ordering is being applied to diagnosis events (diagnosis_1 to diagnosis_15) as well.

When it comes to prescription drugs, we start by retaining the ones that were prescribed earlier in that time period until we reach the maximum number of events, which is set to 15, due to fact that the prescription date, which is not the actual intake date, is the one recorded in the data that we use. This makes it difficult to determine if a drug that was given close to the end of the window was taken early enough to have an impact. Our goal in giving priority to earlier prescriptions is to incorporate drugs that are more likely to have been taken and, as a result, have affected the patient's clinical condition. By doing this, the noise from medications that might not have been consumed yet is minimized.

Diagnostic data, on the other hand, are handled in a different way. A reported diagnosis is a verified clinical finding at a certain moment in time and is not dependent on any subsequent action to be deemed legitimate. A diagnosis becomes an irrefutable part of the patient's record once it is documented. Hence, we choose diagnoses beginning at the end of the window and working backward in time when the number of diagnoses in a window surpasses the acceptable limit. By following this methodology, we are more certain that the most recent and possibly most relevant diagnostic data is preserved, while chronic diseases, that are continuous by nature, are present in all following windows and featured within the initial diagnoses of the windows, to ensure that the input of the model constantly contains those long-term health problems.

When the windows extraction process is finished, the data is combined with patient-specific information, like demographics, lifestyle factors, and genetic data, with this integration producing a comprehensive dataset in which each row corresponds to a single patient window and contains all the information necessary for model prediction. The resulting dataset, that is summarized in Table 18, contains 43 columns in total among which are temporal patterns of drug exposure, diagnoses, chronic conditions, the ADE target label, and patient-specific characteristics.

Table 18. Patient Window Dataset

#	Column	Data Type	Description	Example
1	patient_id	INTEGER	Patient unique identifier	1000299
2	medication_1	STRING	A medication ATC code or NULL	'D06AX01'
…	…	…	…	…
16	medication_15	STRING	A medication ATC code or NULL	'D06AX01'
17	diagnosis_1	STRING	A diagnosis ICD-10 code or NULL	'J80'
..	…	…	…	…
31	diagnosis_15	STRING	A diagnosis ICD-10 code or NULL	'J80'
32	35811394	FLOAT	BMI (not NULL)	74.6
33	age	INTEGER	Age of the patient in the current window (not NULL)	54
34	35810091	INTEGER	Gender in OMOP codes	35818840 – Female, 35820554 – Male, 763013 – Not provided
35	35811013	INTEGER	Smoking status in OMOP codes	35821337 – Never, 35822845 – Previous, 35817596 – Current, 763013 – Not provided

36	35811014	INTEGER	Alcohol drinker status in OMOP codes	35821337 – Never, 35822845 – Previous, 35817596 – Current, 763013 – Not provided
37	35811393	INTEGER	Ethnic background in OMOP codes	38003614 – White, 8515 – Asian, 700391 – Mixed background, 38003600 – African, 8522 – Other race, 763013 – Not provided
38	35810746	INTEGER	Qualifications in OMOP codes	1620732 – Primary school education, 35821787 – GCSEs, 35821953 – e.g. nursing or teaching, 35817386 – College Degree, 45878298 – MSc 763013 – Not provided

39	35810226	INTEGER	Alcohol intake frequency in OMOP codes	35821337 – Never, 35822845 – Previous, 35817596 – Current, 35821830 – One to three times a month, 35824290 – Three or four times a week, 35817657 – Daily or almost daily 763013 – Not provided
40	40792628	STRING	HLA-A allele pairs or NULL	'*101/*2601'
41	40779425	STRING	HLA-B allele pairs or NULL	'*801/*3801'
42	target	INTEGER	Target class	0 – Non-ADE, 1 – ADE
43	target_ADE	STRING	The ICD-10 code of the ADE responsible for classifying a window as ADE, else NULL	'L51.0'

An additional step was the filtering of the windows dataset so that only those windows that contain at least five different drugs are included in the final dataset, due to the fact that our study concentrates on polypharmacy cases.

The final dataset is described in Table 19, which presents important statistics on the window distribution, with only 10,604 (0.39%) of the 2,729,165 windows being classified as

ADE, whereas 2,718,561 windows, representing 99.61%, being classified as non-ADE, suggesting a highly imbalanced dataset. Such imbalance can cause issues in model learning and performance, so it is necessary to handle the imbalance by applying suitable methods like class weighting, resampling, or specific evaluation metrics.

Table 19. Class Distribution of Windows in the Full Dataset

	Number of Windows	Percentage (%)
Negative (non-ADE)	2718561	99.61
Positive (ADE)	10604	0.39
Total	2729165	100

Table 20 provides information on the study population and presents the gender distribution of the patients in the dataset, with 54.40% (49,804) being female and 45.59% (41,735) being male, which means that the distribution is relatively balanced with a good representation of both genders in our cohort, since almost half of the participants are women and the other half are men.

Table 20. Gender Distribution of Patients in the Dataset

	Number of Patients	Percentage (%)
Female	49804	54.40
Male	41735	45.59

More gender specific statistics in relation to ADE occurrences are shown in Table 21, which separates patients with at least one ADE from those without any. By observing Table 21 we can see that, while a significant number of patients in both groups did not suffer from any ADE, 6.04% (3,012) of female patients had at least one ADE, in comparison to 5.78% (2,416) of male patients, with the percentages fluctuating at similar levels between the two genders.

Table 21. ADE Occurrence by Gender (Patients With at Least One ADE vs. None)

	Female	Percentage (%)	Male	Percentage (%)
At least one ADE	3012	6.04	2416	5.78
No ADE	46792	93.06	39319	94.22

Age-related statistics on the dataset's participants can be observed in Table 22, where we can see a mean age of 63.33 years and a median age of 65 years, along with a standard deviation of 7.89 years, and participants' ages ranging from 23 to 80 years, which indicates that even

though the majority of the participants were close to the average age, there was still a lot of diversity within the cohort.

Table 22. Descriptive Statistics of Participant Ages

	Value
Mean age	63.33
Median age	65.00
Standard deviation	7.89
Minimum age	23
Maximum age	80

Chapter 4 - Methodology

Model Input Preparation

Since we have a diverse set of data with different data types, we utilize a multimodal model architecture that handles demographic and lifestyle factors, clinical code sequences (ATC and ICD-10), and genetic data separately so that each module handles each data type in accordance with the specific data type's needs.

Firsly, with regard to the demographic and lifestyle factors, age and BMI were standardized using min-max normalization so that values are translated between 0 and 1, with the age range used for the normalization being from 18 years old up to 90 years old, while the corresponding BMI range was from 10 up to 50. One-hot encoding was used to encode binary categorical features, such as gender and ethnicity, since their categories are unordered and mutually exclusive; each category was transformed into a distinct binary column, with 0 denoting absence and 1 denoting presence. In the case of a male patient, for instance, the "male" column would be set to 1 and the "female" column to 0. To encode ordinal categorical data, such as qualifications, smoking status, alcohol status, and alcohol style, integers representing their natural order were assigned. Valid categories were assigned a value of 1, whereas missing or empty values were assigned a value of 0. Subsequently, these ordinal encoded values were adjusted to fall within the [0, 1] range.

For the clinical data, two vocabularies were constructed: one for diagnosis codes (ICD-10) and another for drug codes (ATC). These vocabularies were produced by using the complete lists of codes from reliable sources: the complete list of ATC codes was retrieved from the research institute of AOK (WidO) (available at https://www.wido.de/fileadmin/Dateien/Dokumente/Publikationen_Produkte/Arzneimittel-Klassifikation/ATC_2025/atc_gkv-ai_2025.zip), and the ICD-10 codes list was retrieved from the U.S. Centers for Disease Control and Prevention (CDC) (available at https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/April-1-2023-Update/). In these vocabularies, every code was given a distinct integer, beginning with 1. Using these vocabularies, every sequence was encoded by substituting the corresponding integer from the vocabulary for each code and padding tokens were encoded as 0.

In a similar vein, 2 distinct vocabularies were constructed for the genomic data, one for each of the genes in Table 15. Clinical specialists involved in the project contributed lists of potential allele combinations for each gene, which served as the basis for the creation of these vocabularies. Every possible allele combination in a gene's vocabulary was given a unique integer, beginning with 1. Each allele combination in the dataset was replaced with its appropriate integer value by utilizing these vocabularies to encode genotype data; missing or unknown values were encoded as 0.

The preprocessed dataset was then divided into training and testing subsets in such a way that maintained the class distribution demonstrated in the real-world data, by assigning 90% of the data from both the ADE and non-ADE classes as the training set and keeping the remaining 10% for testing. By using a stratified split, the model was guaranteed to be trained and assessed using data that accurately reflected the class disparity that exists in clinical settings. The training set is further split into training and validation set in the same manner as the train-test split (90-10%). Table 23 summarizes the number of windows in each subset.

Table 23. Distribution of Negative and Positive Windows Across Train, Validation, and Test Sets

Set	Negative windows	Positive Windows	Total Number of Windows
Train	2202035	8590	2210625
Validation	244670	954	245624
Test	271856	1060	272916

Model Architecture

The model is designed as a multimodal Attention LSTM architecture in order to effectively manage the many forms of medical data needed to assess the risk of ADEs. Since demographic, genetic, diagnostic, and medication data are all distinct types of information, each category must be processed and represented appropriately to capture its distinct qualities. This multimodal design allows the model to take full advantage of the wide range of clinical information for accurate ADE risk prediction by ensuring that all pertinent and meaningful data from each modality is properly processed and incorporated.

The architecture suggested in [51], which was first created for textual data, served as the model's basis for our LSTM. In our adaptation, sequential inputs are processed by two parallel multi-layer LSTM networks, one for diagnostic codes and the other for medication prescriptions. Instead of depending on pre-trained text embeddings, which are made for natural language and fail to identify connection between clinical codes, the embedding layers in both branches change the categorical inputs by capturing significant interactions between codes. The embedded sequences are subsequently processed through the LSTMs to comprehend temporal dependencies. Each branch's outputs are forwarded to a dot-product attention mechanism, which generates a context vector that highlights the most informative time steps. This computationally efficient choice of attention enables the model to concentrate on clinically significant events and is especially appropriate given the size of our dataset (~2 million sequences). To lessen overfitting and promote the model's learning of solid representations across patients' medical and diagnostic records, we also included a dropout layer following the attention context vector.

For every gene, genetic characteristics are mapped to embeddings. These embeddings are then concatenated and put through an MLP comprising fully linked layers with batch normalization, dropout, and nonlinear activations. The batch normalization module normalizes layer inputs, which in turn reduces internal covariate shift and enables the use of greater learning rates that enhances training robustness and speed, while also serving as a regularizer that helps to avoid overfitting in cases of extremely unbalanced data [52]. LeakyReLU is utilized to address the "dying ReLU" issue, which occurs when neurons that receive only negative input would normally output zero and cease updating after training [53]. Due to the fact that uncommon genetic variations can yield sparse or negative signals in the data we possess, LeakyReLU ensures activity in all neurons by permitting a small gradient for negative inputs. This guarantees that all genes provide useful information for prediction, and batch normalization stabilizes training and serves as a regularizer to minimize overfitting.

In contrast, static patient features do not require embeddings. Instead, they are projected directly through an MLP, as done with genetic data. This results in a representation that is consistent with the other modalities.

A single feature vector is created by combining the representations from sequential, genetic, and static data. The model may dynamically select the most useful sources by assigning learnable weights to each category of input, with the final prediction being then produced by

processing this combined representation through fully connected layers that include dropout, LeakyReLU activations, and batch normalization. The model architecture can be seen in Figure 4.

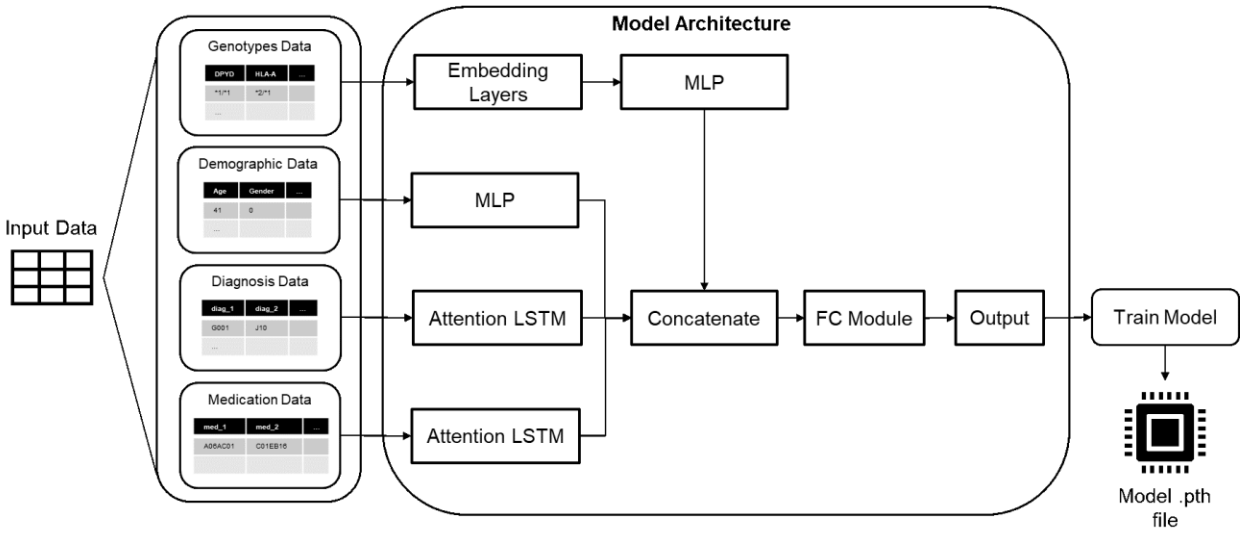


Figure 4. Multimodal Attention LSTM Pipeline for ADE Risk Prediction

Model Training

Due to the huge amount of training data, the model was trained in batches of 256, since by using mini-batches the model is able to update weights incrementally, which reduces memory needs and speeds up the training process, that is particularly important when handling large datasets. The batch size of 256 was selected via experimentation by aiming to balance training stability with computational efficiency.

Since our dataset was extremely unbalanced, we tried to resolve imbalance with common techniques used in such scenarios, that is random oversampling and undersampling, with oversampling duplicating samples from the minority class to increase the representation of both classes, and undersampling removing samples from the majority class until the dataset becomes more balanced [54]. Although these techniques can improve balance, they often fail to enhance model generalization, especially when it comes to high-dimensional feature spaces, while, in our experiments, oversampling actually increased the risk of overfitting to rare positive cases, whereas undersampling resulted in the loss of valuable negative patterns.

Another widely used technique for handling data imbalance is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic data points instead of merely

duplicating existing ones, by interpolating between minority class samples in the feature space to create new, artificial samples that better represent the minority class distribution [55]. However, earlier research, such as [56], found that applying methods like random oversampling, undersampling, or SMOTE does not always enhance discrimination in terms of AUROC, and in some cases, can produce biased probability estimates. Subsequently, the afforementioned findings highlight the limitations of traditional imbalance correction methods in predictive modeling, especially in the medical field, where a lot of diseases are rare in the population.

The final model weights and decision threshold were determined using the F3-score, which gives recall a higher weight than precision [57], in order to ensure that the model focused on prioritizing to identify as many positive cases as possible. By evaluating the model on the validation set across a range of thresholds during each training epoch, with up to a maximum of 50 epochs, and recording the threshold that maximized the F3-score for the weights of that epoch the optimal threshold for every candidate model was determined, since model weights vary from epoch to epoch, while the final model was chosen based on the weights and threshold combination that achieved the highest F3-score over all epochs. Early stopping with patience of 5 epochs was employed to prevent overfitting, halting training when no improvement in F3-score was observed for 5 consecutive epochs.

Focal Loss, although commonly applied in multi-label classification, is also effective for binary classification problems. This loss function improves performance on difficult-to-classify examples by down-weighting the contribution of easily classified samples, thereby directing the model's attention toward instances where it tends to make errors [58]. As a result, Focal Loss improves the recall-focused F3-score by enhancing the model's capacity to accurately detect rare positive cases. This loss is parameterized by α and γ, where α controls the relative weighting of positive and negative classes, and γ determines the degree to which easy-to-classify examples are down-weighted. We selected $\alpha = 0.99$ to heavily emphasize the minority positive class, and $\gamma = 2$ to ensure that the model focuses more on misclassified or difficult examples, complementing the use of the recall-focused F3-score to improve detection of rare positive cases. The values of the parameters were determined empirically by assessing the performance of the model using different combinations.

The trained model's predicted ability on unseen data was then evaluated using the independent test set that contained the remaining 10% of each class and, in order to calculate the

risk of an ADE event, because a model's predicted probabilities may differ from the actual likelihood of an event, as described in [59], calibration using isotonic regression was used to perform any necessary adjustments on the model's probabilities. Isotonic regression can be particularly suitable in our case because it has been shown to perform well in imbalanced scenarios, by providing better overall calibration and handling rare positive events effectively, although it can perform unstably in extremely imbalanced scenarios and therefore when interpreting the results we must be careful [60]. The calibration was performed using the validation set due to practical constraints, instead of creating a separate calibration set, because when we attempted to further split the data it reduced the number of ADE cases in each subset and resulted in insufficient ADE samples for the model to be efficiently trained and validated on ADE patterns. When tested, this imbalance led the optimization process to favor a very low decision threshold, which in turn led to a considerable proportion of non-ADE samples being mistakenly classified as ADE while also decreasing the model's predictive performance. Finally, the calibrated probabilities were used to assess the model and to compare their performance with the performance achieved using raw probabilities.

Chapter 5 - Results

During the initial stages of our study, every patient's clinical data was combined into a single record, with each sample containing the patient's complete medical history, including drug prescriptions and diagnoses, as well as demographic and genetic data, and the classification into ADE or non-ADE was done on patient-level by considering as ADE-positive cases patients who had at least one reported ADE diagnosis across all of their diagnoses, while all other patients were considered to be controls. By using this dataset, we first experimented with several machine learning techniques like Random Forest, Logistic Regression, Gradient Boosting, and SVM, but this approach on how the data were organized did not reflect the actual occurrence of ADEs in the real world, due to the fast that an ADE is not something a patient has or does not have. Rather, it usually occurs as an outcome of certain medications administered during a specific period of time, along with other health problems or therapies that may raise the risk. Through the integration of a patient's complete medical history into a single record, the models combined information from various time periods, including incidents that had nothing to do with ADEs. As a result, it was challenging for these models, which do not consider the sequence of events, to determine which medications or diseases actually contributed to the ADE and failed to detect the time-dependent patterns required for precise ADE prediction.

The difficulties encountered by models trained on combined patient records made it clear that the problem needed to be redefined in order to capture time-dependent patterns. Therefore, instead of handling a patient's whole medical history as a single example, we concentrated on organizing the data in order to demonstrate the timeframe in which ADEs take place. By constructing multiple time windows per patient, we were able to identify periods during which prescription drugs, and medical diagnoses could significantly affect the likelihood of an ADE, with the steps involved in creating these windows being described in Section “Windows Extraction” in more detail. The reconstruction of the data in a way that accurately reflects the time-dependent nature of ADEs, we focused on sequence-based models and decided to use LSTM networks as one of our main approaches, mostly due to its effective application in earlier ADE prediction research, as well as its suitability for handling sequential data, the LSTM was selected. Having the LSTM networks as our basis, we optimized the model by separating different feature types into tailored modules, by employing a MLP that processed static features

like demographic and genetic data and a LSTM network to handle sequential data like drug prescriptions and diagnoses. In order to determine the optimal configuration, we methodically adjusted hyperparameters during this procedure.

Given the rarity of ADE samples, a significant obstacle that we came across was the severe class imbalance, which we initially tried to tackle by using data-driven resampling techniques, such as undersampling the majority class and oversampling the minority class, however undersampling introduced the risk of underfitting by removing important negative cases, while oversampling resulted in overfitting on rare positive samples. Another approach applied was the optimization of the threshold based on recall, but in this case the number of ADE instances tended to be overestimated, while the application of weighted loss did not accomplish to handle the imbalance, with the high number of negatives still prevailing and making the model underestimate ADE instances. As a result, we shifted to algorithm-level methods by using focal loss with adjusted parameters in a way that it reduced the impact of multiple easy to predict non-ADE examples and concentrated on more challenging to identify ADE cases, while also utilizing the F3 score for threshold optimization, with the final model being able to compensate for the extremely low number of ADE samples.

The final model structure was established by incorporating techniques for handling class imbalance while also processing sequential data and static features via modules specifically designed for each data type, with this strategy improving the prediction of ADEs in comparison to the original methods that were applied, since it can more effectively identify informative patterns in the data and detect rare ADE cases despite their extreme rarity.

The final model was trained and evaluated in a process that took approximately 1.7 hours, with early stopping halting training at epoch 36 out of a maximum of 50 epochs. The optimal decision threshold, determined during validation, was 0.275 and was used to calculate all performance metrics. Several criteria were used to assess the model's prediction ability, as shown in Table 24, each of which shed light on a different facet of classification, especially in light of the severe imbalance in ADE cases.

Table 24. Model performance before and after calibration

	Before Calibration	After Calibration
Accuracy	0.996	0.997 ↑
Precision	0.533	0.84 ↑
Recall	0.446	0.406 ↓
F1 Score	0.486	0.547 ↑
F3 Score	0.454	0.428 ↓
PR AUC	0.434	0.427 ↓
ROC AUC	0.763	0.763

Performance before and after calibration is shown in Table 24, enabling evaluation of how isotonic regression improved the predicted probabilities. Accuracy quantifies the proportion of positive and negative observations that were properly classified [61]. Although accuracy was very high in both cases (0.996 before, 0.997 after), this metric alone is not useful in highly imbalanced datasets, as the large number of non-ADE instances increases the accuracy score by simply classifying all samples as non-ADE.

The precision, or the percentage of predicted ADE cases that were actual positives [62], increased from 0.533 prior to calibration to 0.84 following calibration. Because it lowers false positives and prevents unnecessary procedures for patients who are not affected, high precision is crucial in clinical settings. In our case, where recall decreased from 0.446 to 0.406 after performing calibration, recall quantifies the proportion of real ADE instances that it has managed to successfully identify and shows what is the model's ability to identify rare positive cases, which is crucial for prompt medical intervention when dealing with heavily unbalanced data [62].

The F1 score, that is a single statistic that is used to assess the balance between false positives and false negatives by providing a harmonic mean of precision and recall [61], improved from 0.486 to 0.547, while the F3 score declined from 0.454 to 0.428, with F3 score emphasizing the importance of identifying as many instances from the rare ADE cases as possible by weighting recall more highly than precision [63].

The model's overall performance is summarized by the precision-recall area under the curve (PR AUC) and is especially useful for imbalanced datasets where we are aiming for the positive class [64]. As shown in Table 7, PR AUC decreased slightly from 0.434 to 0.427. The

Receiver Operating Characteristic Area Under the Curve (ROC AUC) score, which measures the degree of separability among positive and negative classes, with higher values suggesting better distinction and less overlap between their predictions [65], remained stable at 0.763.

Chapter 6 - Discussion

The accuracy of the model was very high both before and after calibration. However, this metric is misleading because of the significant class imbalance, where only 0.39% of the over 2 million samples were positive. Due to the fact that an overwhelming majority of instances are non-ADE, a high accuracy can be accomplished by just predicting all instances as negative, which means that the model does not capture any meaningful patterns for ADEs, making this statistic entirely useless. In contrast, measures like precision, recall, F1/F3 scores, and PR AUC, that take performance on the minority class into consideration, can provide a more insightful assessment on model performance in scenarios of severe imbalance.

Our model managed to significantly improve precision after calibration, rising from 0.533 to 0.84, which suggests that the predicted probabilities that follow isotonic regression can more accurately represent the probability of actual ADE cases, while false positives are reduced by this increase in precision, which is crucial in clinical settings in order to prevent needless interventions for patients who are not experiencing any ADE. However, recall remained generally low, with 0.446 before calibration and 0.406 after calibration, which reflects the model's difficulty in identifying extremely rare positive ADE instances, due to the severe class imbalance as well as the incomplete prescription data, since prescription records only provide the dates of prescriptions, without any information on actual drug usage or adherence. As a result, many medications could be lost when constructing the input sequences for the model, and some windows may have missed active therapies, which is something that could lead to the loss of drug combinations that would more accurately represent positive cases. Overall, our low recall and high precision illustrate the inherent trade-off between detecting uncommon events and preventing unnecessary interventions in healthcare applications.

The balance between precision and recall improved with calibration, as shown by the F1 score increasing from 0.486 to 0.547, while the PR AUC and F3 score slightly decreased from 0.434 to 0.427 and from 0.454 to 0.428, correspondingly. These changes reflect subtle compromises that are introduced by calibration, as shown by the significant improvement in precision, but the slight reduction of the model's ability to capture all rare positive ADE cases. Finally, ROC AUC remained stable at 0.763 and showed that the overall discriminative ability between ADE and non-ADE cases was not improved, but neither did it worsen.

The optimal decision threshold, 0.275, was selected during the training process based on the F3 score, with small variations in the threshold during testing showing that the model was sensitive to threshold changes, leading to the conclusion that careful threshold selection is a crucial stage that depends on whether the clinical goal prioritizes identifying all positive cases or minimizing false positives.

Despite calibration affecting the outcome, the model's performance remains limited, because althought it can detect a small subset of rare positive ADE cases, it misses many true instances, which is something that demonstrates that even with careful calibration, the approach still struggles to capture all relevant to ADE patterns. Nevertheless, the choice of the modeling strategy should be aligned with the clinical priorities of the task, with some approaches aiming to detect all positive cases by tolerating false positives in order to ensure negatives are reliably ruled out, while others prioritizing high-confidence positive predictions by accepting that some true positives may be missed. Both strategies can be valuable, but, in our case, none of the two strategies is fully satisfied, despite our attempt to find as many positives as possible without over-predicting positives.

Moreover, improving recall may require richer patient-level data or inclusion of medication adherence proxies. Richer patient- and drug-focused data may offer more clinical context to better understand the problem of ADR prediction, while more detailed adherence proxies can decrease information loss by distinguishing prescription medications from those that are truly consumed. Furthermore, reframing ADE prediction as an anomaly detection problem, combined with ensemble learning techniques, provides complimentary ways to increase sensitivity to rare events, which makes these approaches intriguing directions for further research.

This study represents an initial attempt at using a multi-modal LSTM for ADR risk prediction, and no interpretability methods were applied. Although the model incorporates temporal sequences of drugs, diagnoses, and patient features, the lack of transparency further limits clinical trust. For practical application, explainable AI techniques are necessary to clarify which features drive predictions. Clinicians need insight into which drug combinations, comorbidities, or patient characteristics influence the predicted risk to confidently act on model outputs, which means that integrating interpretability would make predictions actionable and increase their value in actual clinical settings.

In summary, the findings indicate that the current approach is conservative, because it effectively limits false positives but still misses many actual ADE instances, since our attempts to simultaneously improve both recall and precision led to trade-offs, with the improvement of one metric often causing the decline of the other. Moreover, the incorporation of multiple data modalities that are tailored to each type, althought conceptually ideal, did increase the complexity of the model but did not reach satisfactory model performance mainly due to the severe class imbalance of the dataset.

In total, this study's findings underscore the difficulty of predicting rare ADRs in real-world datasets, since severe class imbalance, incomplete temporal medication data, and missing patient-level features can limit model effectivenes. Future work could focus on improving recall, integrating additional patient and drug information, and handling ADE prediction as an anomaly detection problem by employing ensemble methods, while also applying explainable AI techniques to enhance clinical trust in the model's predictions and support the translation of predictive models into actionable risk assessments.

Chapter 7 - Limitations

One of the most important limitations that we encountered during the course of our study is the severe class imbalance in the dataset, with only 0.39% of the more than 2 million samples being ADR cases. Due to the tendency of the model to expect the negative class, this significant imbalance made it difficult for the model to learn patterns that are linked to ADR outcomes, and, although several tactics were used to address this problem, the stability and generalizability of the model's predictions were still constrained by the insufficient number of positive samples.

An additional limitation refers to the temporal drug data, because prescription records do not include any details regarding the actual drug usage or adherence. Instead, they simply provide the dates of prescription, which means that the actual drug exposure is uncertain, and therefore many medications are lost during the extraction process of the windows. Subsequently, it is possible that many windows are going to miss ongoing medication treatments, resulting in the loss of some drug combinations that would more accurately represent one of the two classes, which in turn could restrict our model's capacity to learn complex temporal patterns and generate confident predictions.

Another limitation arises from the intrinsic complexity of ADRs, which involves multiple mechanisms, patient-specific factors, and heterogeneous clinical outcomes, while ADRs can also vary in time-to-onset, ranging from immediate reactions to effects that develop over weeks or months. This variability, combined with drug properties, dosage, drug interactions, drug-side effect relationships, comorbidities, and genetic predispositions, makes it difficult to clearly formulate the ADRs prediction problem and adds uncertainty to model development that could potentially limiti the generalization ability of the model to unseen patients.

Furthermore, the study was initially designed to include 16 additional genetic features as input, but difficulties in converting the imputed genotype data into their corresponding allele representations prevented us from including them, and, as a result, these potentially informative genetic features could not be leveraged.

Finally, no interpretability methods were applied. Our multi-modal LSTM model was designed solely to maximize predictive performance, and we did not implement any approach to identify which features had the greatest impact on the predictions. Models of this complexity, that are integrating temporal sequences of drugs, diagnoses, demographics, and genomic data,

are inherently difficult to interpret. In clinical settings, interpretability is essential for trust and adoption, as healthcare professionals need to understand the reasoning behind the predicitons of a model. Without any explanation, clinicians and patients may be hesitant to rely on the model's predictions, which limits the model's practical applicability even when its performance is strong.

Overall, the combined effect of severe class imbalance, insufficient patient- and drug-level data, missing genetic features, and the inherent complexity of ADRs makes it difficult to identify uncommon patterns, capture long-term dependencies, and define the problem in a way that fully reflects real-world clinical scenarios, potentially limiting the model's capacity to generalize to patients who have not yet been seen.

Chapter 8 - Future Work

Despite the increasing availability of massive clinical datasets, our study shows that having a lot of data by itself does not ensure highly accurate predictive modeling. Model performance is heavily influenced by the distribution, completeness, and quality of the data, especially when predicting rare events like ADRs, but also by how the prediction problem is formulated.

Future extensions of this work should focus on integrating all clinically relevant information for the given task. For instance, access to detailed treatment plans, as opposed to just prescription dates, would enhance model performance and give a more realistic picture of patient care in the context of drug exposure. In a similar manner, more detailed pharmacological information, such as active ingredients, doses, known interactions, and side effect associations between drugs that can be obtained from databases like SIDER, FAERS, and DrugBank, may be able to reveal patterns that can not be found solely from the utilization of ATC codes, and, therefore, it could lead to improvements in the model's capacity to identify drug combinations that cause ADRs.

Moreover, the exploitation of additional patient-level characteristics like laboratory results, and genomic information that are directly linked to effects and functionality of medications, could potentially improve the performance of future models by enabling tailored risk assessments and enhancing the detection of rare adverse events. The integration of such information could additionally boost models' interpretability and, therefore, make it easier to translate model predictions into applicable to real-life medical decisions.

This study mainly highlights the challenge produced by the extreme class imbalance in the context of ADR prediction, since, even when data-level balancing strategies are used to tackle it like resampling or synthetic data generation, traditional binary classification algorithms stll find it difficult to accurately model such exceptionally rare positive cases. As a result, future research should consider reframing ADR prediction as an anomaly detection problem that treats rare ADRs as deviations from the normal, non-ADR patterns, with this approach enabling models to concentrate on finding rare patterns, instead of just trying to artificially balance the dataset and lose information related to real-life data distribution.

Furthermore, anomaly detection techniques that are combined with ensemble learning may further increase the model's ability to detect rare ADR detection by combining several models that complete one another, in order to identify patterns that individual models might not notice. Collectively, these techniques provide a suitable approach based on the ADR problem and a methodologically robust framework for detecting the extremely rare ADRs while increasing the prediction accuracy, clinical relevance, and interpretability of the model.

Future research can overcome the limitations identified in this work by implementing the strategies that were suggested and include the integration of richer task-related data with an emphasis on pharmacological information, as well as the usage of anomaly detection and ensemble learning techniques in order to achieve more reliable and personalized ADR risk prediction, that would ultimately contribute to safer and more accurate clinical decision-making.

Chapter 9 - Conclusion

In this study, we investigated the problem of ADR prediction and risk assessment by using a variety of datasets obtained from the UK Biobank. Although our model achieved very high accuracy, this metric was misleading due to extreme class imbalance of our dataset, with only 0.39% of over 2 million samples representing positive ADR cases, which led us to focus on using evaluation metrics that weight the minority class higher in comparison to the majority class, such as precision, recall, F1/F3 scores, and PR AUC, with these metrics actually providing a clearer picture of the model's performance in detecting rare adverse events.

Calibration did improve the reliability of predicted probabilities by substantially increasing precision and, as a result, reducing false positives, which is critical in clinical applications in order to avoid unnecessary interventions, but recall remained at a low level, which is something that highlights how difficult it is to detect rare ADR instances and is a challenge further complicated by incomplete prescription data and the absence of some genetic features.

Our findings suggest that simply increasing data modalities or tailoring complex models for each data type does not necessarily improve performance in highly imbalanced datasets. Model performance is greatly influenced by the problem's formulation, model selection, completeness and quality of the data, as well as strategies for handling imbalanced data. Clinical priorities should also be taken into consideration when choosing a modeling technique. Approaches that aim to identify all positive cases may tolerate more false positives, whereas strategies prioritizing high-confidence predictions may accept missing some true positives.

Limitations of this study include severe class imbalance, incomplete patient- and drug-data, and the complexity of ADR mechanisms. Despite the limitations we came across, the model was able to detect a small subset of rare ADR cases without misclassifying a huge number of negative samples as positive in order to find as many ADR cases as possible, which highlights that even when incomplete temporal and genetic information are used, they can still provide some predictive signal, although this is not enough for reliable real-world applications.

Future work on ADR prediction should focus on integrating comprehensive clinical, pharmacological, and genomic information in order to improve the detection of rare ADR, with detailed treatment plans, drug-related information, and patient-specific characteristics being able

to reveal patterns that can not captured by coded data alone. Additionally, due to the severe class imbalance of the dataset that derives from the rare nature of ADRs in real-life, future research could explore alternative modeling approaches, such as anomaly detection and ensemble learning, which are approaches more suitable for rare events.

Overall, this study highlights the challenges of predicting rare ADRs in real-world healthcare datasets, since careful model calibration, feature selection, and consideration of clinical context helped the model detect some rare ADRs, but limitations such as severe class imbalance, incomplete temporal, genetic, and pharmacological data, along with the lack of model interpretability still restricted the performance and usability of the model in real-life applications. Although our model's predictive power is limited, addressing these challenges could enhance its usefulness for rare ADR detection in clinical practice, and therefore we suggest future research to combine more comprehensive clinical, temporal, genomic, and drug-related data with methods that are ideal for rare-event prediction, which could potentially enable more accurate, personalized, and clinically meaningful ADR risk assessment.

References

[1] 'Safety of medicines – adverse drug reactions'. World Health Organization, https://www.who.int/docs/default-source/medicines/safety-of-medicines--adverse-drug-reactions-jun18.pdf?sfvrsn=4fcaf40_2 (accessed 10 Dec. 2024)

[2] Ahmed A. Nafea, et al., 'An Ensemble Model for Detection of Adverse Drug Reactions', ARO-The Scientific Journal of Koya University, vol. 12, no. 1, pp. 41–47, Feb. 2024. doi: https://doi.org/10.14500/aro.11403

[3] Dsouza VS, Leyens L, Kurian JR, Brand A, Brand H. Artificial intelligence (AI) in pharmacovigilance: A systematic review on predicting adverse drug reactions (ADR) in hospitalized patients. Res Social Adm Pharm. 2025 Jun;21(6):453-462. doi: 10.1016/j.sapharm.2025.02.008

[4] de Souza IKC, Rosa-Souza FJ, de Lucena Alves CP, Duhamel TA, Waters DL, Martins RR, Costa EC. 'Polypharmacy, physical activity, and sedentary time in older adults: A scoping review'. Exp Gerontol. 2023 Nov; 183:112317, doi: 10.1016/j.exger.2023.112317

[5] Hongyu B., et al., 'Review on application and development of pharmacogenomics of adverse drug reactions', Journal of Holistic Integrative Pharmacy, vol. 6, no. 1, pp. 105–116, Mar. 2025, https://doi.org/10.1016/j.jhip.2025.03.005

[6] Bodas, S. and Pathak, S. S. 'Polypharmacy, Drug Interaction, and Their Consequences in Terms of Adverse Drug Reaction among Adults Receiving Cancer Chemotherapy in a Tertiary Care Hospital in Central India', Journal of Datta Meghe Institute of Medical Sciences University, vol. 19, no. 3, pp. 595–599, Sep. 2024. doi: 10.4103/jdmimsu.jdmimsu_51_23

[7] Zazzara MB, Palmer K, Vetrano DL, Carfì A, Onder G., 'Adverse drug reactions in older adults: a narrative review of the literature', European Geriatric Medicine, vol. 12, No. 3, pp. 463–473, Mar. 2021. doi: 10.1007/s41999-021-00481-9

[8] Afdhal, D., Ananta, K. W. and Hartono, W. S., "Adverse Drug Reactions Prediction Using Multi-label Linear Discriminant Analysis and Multi-label Learning," 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2020, pp. 69-76, doi: 10.1109/ICACSIS51025.2020.9263166

[9] Kommu S., Carter C., and Whitfield P., 'Adverse Drug Reactions', StatPearls, https://www.ncbi.nlm.nih.gov/books/NBK599521/ (accessed 17 Dec. 2024)'About our data', UK Biobank, https://www.ukbiobank.ac.uk/about-our-data/ (accessed 9 Mar. 2025)

[10] Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). doi: https://doi.org/10.1038/sdata.2016.35

[11] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. RRID:SCR_007345. [Online]. Available: https://doi.org/10.13026/C2XW26

[12] Mahendran D, McInnes BT. 'Extracting Adverse Drug Events from Clinical Notes', AMIA Jt Summits Transl Sci Proc., vol. 2021, pp. 420–429, May 2021.

[13] Johnson, A.E.W., Bulgarelli, L., Shen, L. et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10, 1 (2023). doi: https://doi.org/10.1038/s41597-022-01899-x

[14] Langenberger, B. 'Machine learning as a tool to identify inpatients who are not at risk of adverse drug events in a large dataset of a tertiary care hospital in the USA', British

Journal of Clinical Pharmacology, vol. 89, no. 12, pp. 3523–3538, Jul. 2023, doi: https://doi.org/10.1111/bcp.15846

[15] Sadeghi S, Hempel L, Rodemund N, and Kirsten T, 'Salzburg Intensive Care database (SICdb): a detailed exploration and comparative analysis with MIMIC-IV', *Scientific Reports*, vol. 14, May 2024, doi: 10.1038/s41598-024-61380-0

[16] Kuhn M, Letunic I, Jensen LJ, and Bork P, 'The SIDER database of drugs and side effects', *Nucleic Acids Res.*, vol. 44, pp. D1075–D1079, Oct. 2015, doi: 10.1093/nar/gkv1075

[17] Li, Y., *et al.*, 'Research on Adverse Drug Reaction Prediction Model Combining Knowledge Graph Embedding and Deep Learning', in *2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, Zhuhai, China, 2024, pp. 322–329. doi: 10.1109/MLISE62164.2024.10674360

[18] Farnoush A, Sedighi-Maman Z, Rasoolian B, Heath JJ, and Fallah B, 'Prediction of adverse drug reactions using demographic and non-clinical drug characteristics in FAERS data', *Sci Rep*, vol. 14, no. 23636, Oct. 2024, doi: https://doi.org/10.1038/s41598-024-74505-2

[19] 'FDA Adverse Event Reporting System (FAERS)', openFDA, https://open.fda.gov/data/faers/ (accessed 12 Dec. 2024)

[20] Li S., *et al.*, 'BiMPADR: A Deep Learning Framework for Predicting Adverse Drug Reactions in New Drugs', *Molecules*, vol. 29, no. 8, Apr. 2024, doi: 10.3390/molecules29081784

[21] Besharatifarda M, Ghorbanalia Z, and Zare-Mirakabad F, 'Adverse drug reaction prediction using voting ensemble training approach', *AUT J. Math. Comput.*, vol. 5, no. 1, pp. 45–59, 2024, doi: 10.22060/AJMC.2023.21538.1091

[22] He M, Shi Y, Han F, and Cai Y, 'Prediction of adverse drug reactions based on pharmacogenomics combination features: a preliminary study', *Front. Pharmacol.*, vol. 16, Mar. 2025, doi: 10.3389/fphar.2025.1448106

[23] Luo J, Qian C, Wang X, Glass L, and Ma F, 'pADR: Towards Personalized Adverse Drug Reaction Prediction by Modeling Multi-sourced Data', in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, Oct. 2023, pp. 4724–4730. doi: https://doi.org/10.1145/3583780.36154

[24] Choudhury O, Park Y, Salonidis T, Gkoulalas-Divanis A, Sylla I, and Das AK, 'Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning', in *AMIA Annu Symp Proc.*, Mar. 2020, pp. 313–322.

[25] Song, L., *et al.*, 'Using machine learning to predict adverse events in acute coronary syndrome: A retrospective study', *Clin Cardiol.*, vol. 46, no. 12, pp. 1594–1602, Aug. 2023, doi: 10.1002/clc.24127

[26] Padhi, B., *et al.*, 'Using multiple drug similarity networks to promote adverse drug event detection', *Heliyon*, vol. 10, no. 22, Nov. 2024, doi: 10.1016/j.heliyon.2024.e39728

[27] Ou, Q., *et al.*, 'A Fusion Deep Learning Model for Predicting Adverse Drug Reactions Based on Multiple Drug Characteristics', *Life*, vol. 15, no. 3, p. 436, 2025, doi: https://doi.org/10.3390/life15030436

[28] Yang S and Kar S, 'Application of artificial intelligence and machine learning in early detection of adverse drug reactions (ADRs) and drug-induced toxicity', *Artificial Intelligence Chemistry*, vol. 1, no. 2, Dec. 2023, doi: 10.1016/j.aichem.2023.100011

[29] Anastopoulos IN, Herczeg CK, Davis KN, and Dixit AC, 'Multi-Drug Featurization and Deep Learning Improve Patient-Specific Predictions of Adverse Events', *Int J Environ Res Public Health*, vol. 18, no. 5, p. 2600, Mar. 2021, doi: 10.3390/ijerph18052600

[30] Shi J, Gao X, Ha C, Wang Y, Gao G, and Chen Y, 'Patient ADE Risk Prediction through Hierarchical Time-Aware Neural Network Using Claim Codes', presented at the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 1388–1393. doi: 10.1109/BigData50022.2020.9378336

[31] Cao M, Chi Y, Yu J, Yang Y, Meng R, and Jia J, 'Adverse drug reaction signal detection via the long short-term memory model', *Front. Pharmacol.*, vol. 16, Jun. 2025, doi: https://doi.org/10.3389/fphar.2025.1554650

[32] Dr. Asutosh Pramanik, Dr. Gunaseelan.C, Dr. Shakeel Ahmad, Dr. Hari Narayan Singh, and Dr. Sukanta Bandyopadhyay, 'The role of AI in predicting adverse durg reactions: enhancing patient safety in pharmaceutical practice', *Journal of Population Therapeutics and Clinical Pharmacology*, vol. 31, no. 11, 2024, doi: 10.53555/nf2m4h73

[33] Dr. Bharti Khemani, Dr. Sachin Malave, Samyukta Shinde, Mandvi Shukla, Razzaq Shikalgar, and Harshita Talwar, 'AI-driven pharmacovigilance: Enhancing adverse drug reaction detection with deep learning and NLP', *MethodsX*, vol. 15, Dec. 2025, doi: 10.1016/j.mex.2025.103460.

[34] He Y, Lv X, Long W, Zhai S, Li M, and Wen Z, 'ToxBERT: an explainable AI framework for enhancing prediction of adverse drug reactions and structural insights', *Journal of Pharmaceutical Analysis*, vol. 15, no. 8, Aug. 2025, doi: 10.1016/j.jpha.2025.101387

[35] Nurzhan T, Aliya Z, Baurzhan Z, Almas Y, and Hameed A, 'Personalized Drug Response Prediction Using AI', *ResearchGate*, Jul. 2025.

[36] Santiso S, Casillas A, and Pérez A, 'The class imbalance problem detecting adverse drug reactions in electronic health records', *Health Informatics Journal*, vol. 25, no. 4, pp. 1768–1778, Sep. 2018, doi: 10.1177/1460458218799470

[37] Cartus AR, Samuels EA, Cerdá M, and Marshall BDL, 'Outcome class imbalance and rare events: An underappreciated complication for overdose risk prediction modeling', *Addiction*, vol. 118, no. 6, pp. 1167–1176, Feb. 2023, doi: 10.1111/add.16133

[38] Zhao YX, Yuan H, and Wu Y, 'Prediction of Adverse Drug Reaction using Machine Learning and Deep Learning Based on an Imbalanced Electronic Medical Records Dataset', *Association for Computing Machinery*, pp. 17–21, 2021, doi: 10.1145/3472813.3472817

[39] 'What is Anomaly Detection?', GeeksforGeeks, https://www.geeksforgeeks.org/computer-networks/what-is-anomaly-detection/ (accessed 20 Aug. 2025)

[40] Murel J, and Kavlakoglu E, 'What is ensemble learning?', IBM, https://www.ibm.com/think/topics/ensemble-learning (accessed 20 Aug. 2025)

[41] Chakraborty S, and Tiwari R, 'A Clustering Ensemble Method for Drug Safety Signal Detection in Post-Marketing Surveillance', *Ther Innov Regul Sci*, vol. 59, no. 1, pp. 89–101, Jan. 2025, doi: 10.1007/s43441-024-00705-7

[42] 'Assessing causality of adverse effects', Medicines Learning Portal, https://www.medicineslearningportal.org/2015/07/assessing-causality.html (accessed 5 Jan. 2025)

[43] Alomar MJ, 'Factors affecting the development of adverse drug reactions (Review article)', vol 22, no. 2, pp. 83–94, Apr. 2014, doi: 10.1016/j.jsps.2013.02.003
[44] Abubakar AR, Simbak NB, and Haque M, 'Adverse Drug Reactions: Predisposing Factors, Modern Classifications and Causality Assessment', *Research Journal of Pharmacy and Technology*, vol. 7, no. 9, pp. 1091–1098, Sep. 2014
[45] Li W, Wang J, Lin H, and Shen G, 'HLA-A∗24:02 associated with lamotrigine-induced cutaneous adverse drug reactions', *Medicine (Baltimore)*, no. 99, p. 52, Dec. 2020, doi: 10.1097/MD.0000000000023929
[46] Mushiroda T, 'Pharmacogenetic biomarkers associated with risk of developing severe drug eruptions and clinical implementation of HLA genetic testing', *Allergology International*, vol. 74, no. 3, pp. 380–385, Jul. 2025, doi: 10.1016/j.alit.2025.03.002
[47] 'Resource 182', biobank, https://biobank.ctsu.ox.ac.uk/ukb/refer.cgi?id=182 (accessed 14 Apr. 2025)
[48] Hohl CM, Karpov A, Reddekopp L, and Stausberg J, 'ICD-10 codes used to identify adverse drug events in administrative data: a systematic review', *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 547–557, May 2014, doi: https://doi.org/10.1136/amiajnl-2013-002116
[49] Nohner M, De Lima B, and Drago K, 'Validating ICD-10 codes for adverse drug events in hospitalised older adults: protocol for a cross-sectional study', *BMJ Open*, vol. 12, no. 11, Nov. 2022, doi: 10.1136/bmjopen-2022-062853
[50] Ed-drissiya El-allaly, Mourad Sarrouti, Noureddine En-Nahnahi, and Said Ouatik El Alaoui, 'A LSTM-Based Method with Attention Mechanism for Adverse Drug Reaction Sentences Detection', in *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019)*, Springer, Cham, Feb. 2020, pp. 17–26. doi: https://doi.org/10.1007/978-3-030-36664-3_3
[51] Ashutosh, 'Unleashing the Power of Batch Normalization in Multilayer Perceptrons (MLPs)', Medium, https://ashutoshkriiest.medium.com/unleashing-the-power-of-batch-normalization-in-multilayer-perceptrons-mlps-b339ba01152c (accessed 29 Aug. 2025)
[52] Juan C Olamendy, 'Understanding ReLU, LeakyReLU, and PReLU: A Comprehensive Guide', Medium, https://medium.com/@juanc.olamendy/understanding-relu-leakyrelu-and-prelu-a-comprehensive-guide-20f2775d3d64 (accessed 2 Sep. 2025)
[53] Baladram S, 'Oversampling and Undersampling, Explained: A Visual Guide with Mini 2D Dataset', Towards Data Science, https://towardsdatascience.com/oversampling-and-undersampling-explained-a-visual-guide-with-mini-2d-dataset-1155577d3091/ (accessed 8 Feb. 2025)
[54] 'SMOTE for Imbalanced Classification with Python', GeeksforGeeks, https://www.geeksforgeeks.org/machine-learning/smote-for-imbalanced-classification-with-python/ (accessed 8 Feb. 2025)
[55] van den Goorbergh R, van Smeden M, Timmerman D, and Van Calster B, 'The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression', *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, Sep. 2022, doi: https://doi.org/10.1093/jamia/ocac093
[56] Wood T, 'What is the F-score?', https://deepai.org/machine-learning-glossary-and-terms/f-score (accessed 12 Mar. 2025)

[57] Niyaz U, 'Focal loss for handling the issue of class imbalance', https://medium.com/data-science-ecom-express/focal-loss-for-handling-the-issue-of-class-imbalance-be7addebd856 (accessed 12 Mar. 2025)
[58] Bansal S, 'Understanding Model Calibration in Machine Learning', https://medium.com/@sahilbansal480/understanding-model-calibration-in-machine-learning-6701814dbb3a (accessed 25 Jul. 2025)
[59] Huang L, Zhao J, Zhu B, Chen H, and Vanden Broucke S, 'An Experimental Investigation of Calibration Techniques for Imbalanced Data', *IEEE Access*, vol. 8, pp. 127343–127352, 2020, doi: 10.1109/ACCESS.2020.3008150
[60] Czakon J, 'F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?', neptune.ai, https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc (accessed 7 Feb. 2025)
[61] Kashyap P, 'Understanding Precision, Recall, and F1 Score Metrics', Medium, https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093 (accessed 7 Feb. 2025)
[62] 'What is F-Beta Score?', GeeksforGeeks, https://www.geeksforgeeks.org/machine-learning/what-is-f-beta-score/ (accessed 14 Feb. 2025)
[63] 'What Is PR AUC?', arize, https://arize.com/blog/what-is-pr-auc/ (accessed 14 Feb, 2025)
[64] Dembla G, 'Intuition behind ROC-AUC score', Towards Data Science, https://towardsdatascience.com/intuition-behind-roc-auc-score-1456439d1f30/ (accessed 15 16 Aug. 2025)