Forecasting Patient Early Readmission from Irish Hospital Discharge Records Using Conventional Machine Learning Models

Minh-Khoi Pham *, Tai Tan Mai, Martin Crane, Malick Ebiele, Rob Brennan, Marie Ward, Una Geary, Nick McDonald, Marija Bezbradica

Posted Date: 18 October 2024

doi: 10.20944/preprints202410.1476.v1

Keywords: Electronic Patient Records; Multimodal Deep Learning; Explainable AI; Data Imbalance

Article

Forecasting Patient Early Readmission from Irish Hospital Discharge Records Using Conventional Machine Learning Models

Minh-Khoi Pham [1,2,*], Tai Tan Mai [1,2], Martin Crane [1,2], Malick Ebiele [1,3], Rob Brennan [1,3], Marie Ward [4], Una Geary [4], Nick McDonald [5] and Marija Bezbradica [1,2]

[1] Adapt Centre, Dublin, Ireland
[2] Dublin City University, Dublin, Ireland
[3] University College Dublin, Dublin, Ireland
[4] St James's Hospital, Dublin, Ireland
[5] Trinity College Dublin, Dublin, Ireland
* Correspondence: minhkhoi.pham@adaptcentre.ie

Abstract: Predicting patient readmission is an important task for healthcare risk management, as it can help prevent adverse events, reduce costs, and improve patient outcomes. In this paper, we compare various conventional machine learning models on a multimodal dataset of electronic discharge records from an Irish acute hospital. We evaluate the effectiveness of several widely-used Machine Learning models that leverage patient demographics, historical hospitalization records, and clinical diagnoses codes, to forecast future clinical risks. Our work focuses on addressing two key challenges in the medical fields: data imbalance and the variety of data types in order to boost the performance of Machine Learning algorithms. Through extensive benchmarking and the application of a variety of feature engineering techniques, we successfully improved the Area Under the Curve (AUROC) score from 0.628 to 0.7 across our models on the test dataset. Furthermore, we also employ Shapley Additive Explanations (SHAP) value visualization to interpret the model predictions and identify both the key data features and disease codes associated to readmission risks, identifying a specific set of diagnoses codes that are significant predictors of readmission within 30 days. Our study demonstrates how we effectively utilize the routinely collected hospital data to forecast patient readmission through the use of conventional machine learning while applying explainable AI techniques to explore the correlation between data features and patient readmission rate.

Keywords: electronic patient records; multimodal deep learning; explainable AI

1. Introduction

Effective risk management in healthcare is crucial as it safeguards patient safety, optimizes resource utilization, and ensures the overall efficiency of healthcare systems, thereby mitigating potential adverse events and enhancing the quality of patient care [1]. Healthcare-associated infections are an example of a key risk for modern healthcare. To be able to assess and control this risk it is essential for hospitals to have reliable mechanisms to anticipate future events, such as in-hospital infection outbreaks or transmission. As digital transformation of healthcare becomes more mature, there is a need for new ways to leverage the healthcare data available for prediction of infection, treatment outcomes and risk management. The readmission rate is recognized as one of the pivotal metrics for assessing the effectiveness of patient treatment [2]. Clancy et al. [2] define hospital readmission as occurring when a patient returns to the hospital after initial discharge within 30 days. By forecasting individual patients' readmission risks, interventions can be identified to reduce costs and propose innovative strategies for preventing readmissions [3]. This motivation drives our research effort to evaluate the suitability of typical healthcare data to forecast 30-day readmission rates within one of the national acute hospitals in Ireland as part of a larger research project to improve the Prevention and Control of Healthcare-associated Healthcare Infections (PCHCAI).

This particular interdisciplinary PCHCAI project, ARK-Virus, has brought together researchers in Human Factors, health systems, data governance, data analytics, infection control and microbiology [4].

 doi:10.20944/preprints202410.1476.v1

The ARK-Virus project took a systems approach to managing the risk of healthcare-acquired infection in an acute hospital setting, supported by the ARK (Access Risk Knowledge) Platform. This approach uses the Cube socio-technical systems analysis methodology which requires evidence-based analysis, like the predictive machine learning (ML) data and Deep Learning (DL) analysis methods described in this paper. The analysis techniques must be suitable for the data available and the deployment context of risk management.

Conventional machine learning (ML) models, characterized by interpretability, ease of implementation and well-established frameworks [5], offer practical advantages in forecasting hospital readmission, providing informative insights into risk factors and facilitating trust among healthcare professionals. Their simplicity, limited data requirements and alignment with clinical reasoning make them effective tools for enhancing risk management in healthcare.

This research explores the potential of the different conventional ML models on this dataset to predict patient outcomes. We also incorporate several well-known DL models into our experiments for comparison. Additionally, we emphasize model explainability, aiming to identify specific data attributes that are likely to impact patient readmission since explainability is one of the most important requirements in healthcare [6].

In summary, this paper addresses two key research questions:

1. How effectively can conventional ML models utilize diverse data types within Irish healthcare discharge records to predict patient outcomes, such as the 30-day readmission rate?
2. How can we generate clinically meaningful explanations for these predictive models using a popular explainable AI technique?

Our main contributions are as follows:

- We evaluate various conventional ML algorithms, as well as a set of DL methods, for predicting patient readmissions using routinely collected discharge records data from an Irish acute hospital that encompass patient demographics, their clinical information, diagnosis codes and previous hospitalization.
- We extensively assess the importance of each data feature in predicting patient readmission by analyzing the impact of individual group features on the classification score, while also testing with different data sampling methods to address the class imbalance..
- To gain deeper insights into the models' predictions, we use SHAP visualizations to examine how patient clinical information and diagnosis codes are related to 30-day readmissions. Finally, we interpret the significance of these visualizations..

Overall, the structure of our paper is as follows: We present an overview of prior research in Section 2. Section 3 provides an introduction to our data sources and the methodology we employ. In Section 4, we present the outcomes of our benchmarking models and accompany these with visualizations depicting the importance of the features. Lastly, in Section 5, we engage in a discussion concerning the results, and limitations, and outline possible avenues for future research.

2. Related Work

2.1. Multimodal Machine Learning in Patient Readmission Forecast

Various research studies have explored the task of predicting patient readmission with multiple different time frames, such as 30, 180, or 365 days depending on their focus. In our study, our primary focus lies in predicting readmissions within a 30-day window. The surge in data potentially available for clinical and safety decision making as a result of digital transformation is not only marked by a substantial increase in volume but also by a growing complexity and diversity in its forms. These challenges contribute to a notable rise in the publication of research papers, particularly in the multimodal machine learning topics since multimodal methods address the data diversity.

 doi:10.20944/preprints202410.1476.v1

Many researchers [7–11] have substantiated the advantages of amalgamating diverse data types extracted from patients' Electronic Health Records (EHRs). These encompass patient demographics, medical diagnoses, vital signs, lab test results, and historical hospital visit records, all of which contribute to the prognosis of future readmissions. These studies have adeptly incorporated multimodal configurations into a variety of machine learning models to cater to this task, yielding superior performance compared to single-modal methodologies.

Hence, we aim to introduce and address the unique characteristics of our discharge records data, which may differ from data in other research. For example, lab tests and vital signs are pivotal features in both [9,10] whereas they were not available in our study. Additional supplementary information can also be leveraged for readmission prediction, such as integrating clinical prescription data with weather and air quality records [12]. However, the effectiveness of these measures largely depends on their design, implementation, and the timing of data collection, and they are not always readily available. Consequently, in Section 3.1.1, we will provide an introduction to our electronic data, highlighting the associated characteristics, and thereby distinguishing our approach from these aforementioned studies.

Given these differences, we shift our emphasis to the clinical diagnostic aspects, particularly the utilization of International Classification of Diseases (ICD) codes, since these are commonly available to access in the global healthcare system. Many studies have employed diverse techniques to extensively harness the information encapsulated in ICD code names for tasks related to predicting patient outcomes. These approaches span from straightforward tree-based models [10,13] to more advanced deep language models, such as convolutional networks [14] and recurrent networks [7,8,15,16]. Studies found that instead of relying exclusively on medical code names, leveraging the contextual descriptions of clinical concepts can amplify the effectiveness of predictive models [14,17]. In our experiment, we seek to conduct a comparative analysis between the use of code names and descriptions, in conjunction with other variables, to reassess this claim. Our position asserts that for the conventional model we employ for this dataset, utilising clinical codes rather than textual descriptions is seen to offer more effectiveness to the models' accuracy.

It is worth noting, however, that these studies typically do not address the challenge of class imbalance[1], which is a prevalent and ubiquitous issue when utilizing many types of the electronic data for the healthcare domain for AI application. While previous studies [7,9] suggest the introduction of additional weights in their cost function to mitigate this problem, they report that this approach is not particularly effective [7]. In contrast, our approach directly tackles this issue by comparing a variety of data sampling techniques, from random sampling to KNN-based sampling, which significantly improved the overall performance metrics of our methods.

Furthermore, the existing research papers to date often lack adequately transparent explanations for their models' performance. This deficiency is primarily attributed to the fact that such explanations are typically tailored to specific models and cannot be readily applied to others [13], or they necessitate the integration of additional components like attention modules [15,17]. In our research, we have adopted a widely recognized and universally applicable Shapley Additive Explanations (SHAP) method, regardless of the specific architecture of the machine learning models employed. While SHAP is a popular tool for explaining feature importance in ML models, many studies report SHAP values for only a single model. In practice, different models may capture and represent feature relationships with the target variable in diverse ways, leading to variations in SHAP values across models. Therefore, we not only use SHAP to investigate the relevance of patient information and diagnostic code groups that may influence early readmission but also aggregate SHAP values across multiple models to ensure the stability and reliability of our findings.

---

[1] The imbalance of classes can be defined as when the number of one type of data points outnumbers the instances of another type

 doi:10.20944/preprints202410.1476.v1

2.2. Explainable AI in Healthcare domain

There has been a growing demand for transparency of AI models to be utilized in the healthcare sector [6]. In our research, we advocate the utilization of the SHAP [2] method [18], which is based on the Shapley value [19], to visualize the importance of model features. SHAP values possess the advantage of being model-agnostic, making them universally applicable in conjunction with any machine learning models. This represents a straightforward approach that offers additional interpretability without affecting the models' parameters, thus preventing compromising model accuracy.

A recent comprehensive survey [6] provides an extensive list of explainable methods for interpreting model predictions and investigating their underlying behaviours. The survey notes that in many previous studies utilizing EHRs as data, SHAP is often the preferred choice among explainability methods. However, it's worth noting that these studies commonly overlook the inclusion of patients' diagnosis codes and descriptions [20–22]. Thus, in our study, we illustrate how this approach can provide useful information on which factors that potentially influence patient readmissions based on their prior clinical diagnoses codes and its descriptive terms in our dataset.

3. Method

3.1. The Collected Data

3.1.1. Data Description

The data used for this retrospective study was an anonymized sample of typical operational Irish healthcare electronic data, collected daily from inpatients at St. James' Hospital, Dublin, Ireland. As such, it represents the diversity of interlinked data that is needed for this sort of analysis and highlights challenges with data consistency, sparseness and other quality factors that are often found in operational data. Demonstrating methods that can overcome these challenges is a major goal of this study.

In Irish healthcare settings, inpatient data is digitally collected through the Hospital Inpatient Enquiry (HIPE) system. HIPE plays a pivotal role in Irish healthcare data collection, generating approximately 1.7 million records annually [3]. A HIPE discharge record is created when a patient is discharged from (or dies in) a hospital. It encompasses patients' demographic, clinical, and administrative data for each bed day, including discharge codes noted by medical administrators, which reflect patient conditions for a discrete episode of care. An episode of care begins at admission to hospital, as a day or inpatient, and ends at discharge from (or death in) that hospital.

The HIPE data in our study is neatly organized in a table and includes both structured and semi-structured fields, covering a time period from 2018 to 2022 consisting of almost 1 million rows. Each row in this dataset represents one patient's stay at the hospital, recorded every night. Furthermore, the dataset includes patients' demographic information as well as their hospitalization, including the episode admission and discharge time, medical wards they were admitted to and discharged from and as well as the speciality of that ward. The data features are described in Appendix **??**, and the main statistical description of the dataset is reported in Table 1.

---

[2] We highly recommend looking at this extensive book for a more intuitive explanation of SHAP value https://christophm.github.io/interpretable-ml-book/shapley.html

[3] Information for HIPE can be found at https://hpo.ie/

 doi:10.20944/preprints202410.1476.v1

Table 1. Overall statistics of the data after processing. Clinical codes include International Classification of Diseases (ICD), Irish Coding Standards (ICS), Diagnosis Related Groups (DRG), Major Diagnostic Categories (MDC).

Feature name	Statistics
Number of patients	50,159
Number of episodes	82,713
Highest value of number of clinical visits	46
Number of unique clinical codes	9851
Number of unique generalized clinical codes	3771

Besides patient demographics such as gender and county of residence, the dataset also includes information on up to 29 diagnoses and 29 medical procedures for each hospital stay. These are entered as clinical codes by hospital discharge administrators based on the free-text medical notes from doctors. These clinical codes are categorized using the widely recognized guidelines within the HIPE system, namely Irish Coding Standards (ICS) [23]. The ICS are guidelines for the collection of data for all discharges from hospitals in Ireland using the HIPE Portal software and the 10th Edition ICD-10-AM/ACHI/ACS classification system. The ICD-10 system, published by the World Health Organization (WHO) [24], is a standardized method for categorizing and labelling diagnoses, symptoms, and procedures related to hospital care. Additionally, the dataset contains codes and descriptions for Diagnosis Related Groups (DRG) and Major Diagnostic Categories (MDC). These codes help categorize different diagnoses for the purpose of healthcare reimbursements. In total, the dataset encompasses **9,851** distinct diagnosis and procedure codes. Some illustrative examples are presented in Table 2.

Table 2. Example of a data instance where multiple codes might be assigned in an episode. These codes are shown at their most granular level.

Diagnosis Code	**Code Description**
J44	Chronic obstructive pulmonary
D351	Benign neoplasm of parathyroid gland
I20	Angina
Z22.3	Carrier of other specified bacterial diseases
Procedure Code	**Code Description**
9555009	Allied Health Intervention Pharmacy
3031500	Subtotal Parathyroidectomy
DRG Code	**Code Description**
K05A	Parathyroid Procedures
MDC Code	**Code Description**
10	Parathyroid Procedures

3.1.2. Data Cleaning and Feature Engineering

In this section, we present a comprehensive overview of our data processing pipeline. We begin by selectively retaining episodes that occurred between January 1^{st}, 2018, and February 28^{th}, 2022, stripping episodes that do not have a final discharge date or ones that are missing an admission date as they would likely generate noises in our data. In order to maintain data consistency, we exclude records associated with patients who are either deceased during hospitalization or were transferred to another healthcare facility. These cases are considered outliers since our primary objective is predicting patient readmission within 30 days, and these situations may not align with that goal [10].

For numerical data features, we apply standardization to ensure numerical stability, while categorical data features are converted into one-hot encoding matrices. These preprocessing steps are not necessary for LightGBM and CatBoost, as these frameworks handle them internally.

 doi:10.20944/preprints202410.1476.v1

The sequential data from patients' past hospital visits holds valuable insights for predicting their future hospitalizations, a concept well-supported by prior research [7,25,26]. Thus, we have manually integrated cumulative features based on the patients' historical admissions. The representative features for each data sample include cumulative metrics from the past *t episodes*, such as the *total length of stay since the last t episodes*, *the number of ICU admissions in the last t episodes*, and *the counts of diagnoses and procedures from prior admissions within the last t episodes*. We also present an ablation study in Appendix A3 to examine the impact of *t*, which appears to significantly influence model performance..

For the medical codes in our dataset, studies show that utilizing the most granular levels of medical codes may not be imperative for effective predictions [7,8,13,15,16]. Following their footsteps, we follow this approach by grouping the first three digits of ICD-10 diagnosis codes and five digits of ICS procedure codes. This grouping significantly reduces the number of unique codes (as shown in Table 1). These characters provide a high-level grouping of diseases which represents the category of the diagnosis , which provides sufficient information for machine learning models to learn in our study. We also report the ablation study for this setting in A1.

For the code's descriptions, we apply standard text preprocessing techniques, including converting text to lowercase, and eliminating stopwords, punctuation, and numerical characters. Missing values in these columns are uniformly filled with "None" to ensure consistency. Finally, to maintain data integrity, we remove duplicate rows and empty columns from the dataset. This step ensures that our dataset remains clean and well-structured for analysis.

3.2. Models

3.2.1. Data Sampling Method

Our dataset demonstrates a significant class imbalance between the number of readmitted and non-readmitted patients, as shown in Figure 1, with the positive class representing only 12% of the data.

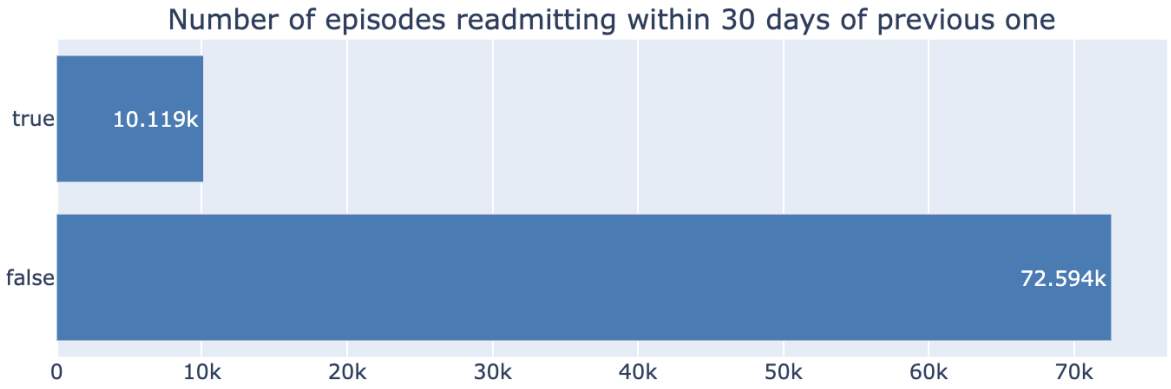

Figure 1. Readmission class distribution in the dataset, which is highly imbalanced toward the negative class.

We conducted a power analysis to estimate the required sample sizes for predicting 30-day patient readmission, considering varying effect sizes, as shown in Figure 2. Using a significance level of 0.05 and a desired power of 0.8, our analysis revealed that as the effect size decreases, the required sample size also increases. For smaller effect sizes, such as 0.2, a much larger sample size is necessary to achieve sufficient power, while larger effect sizes, such as 0.8, require fewer samples. Based on our analysis, for an effect size of 0.3, and 12% class imbalance ratio taken into account, we estimate that approximately 500-3,000 samples for each class would be needed to achieve the desired power.

 doi:10.20944/preprints202410.1476.v1

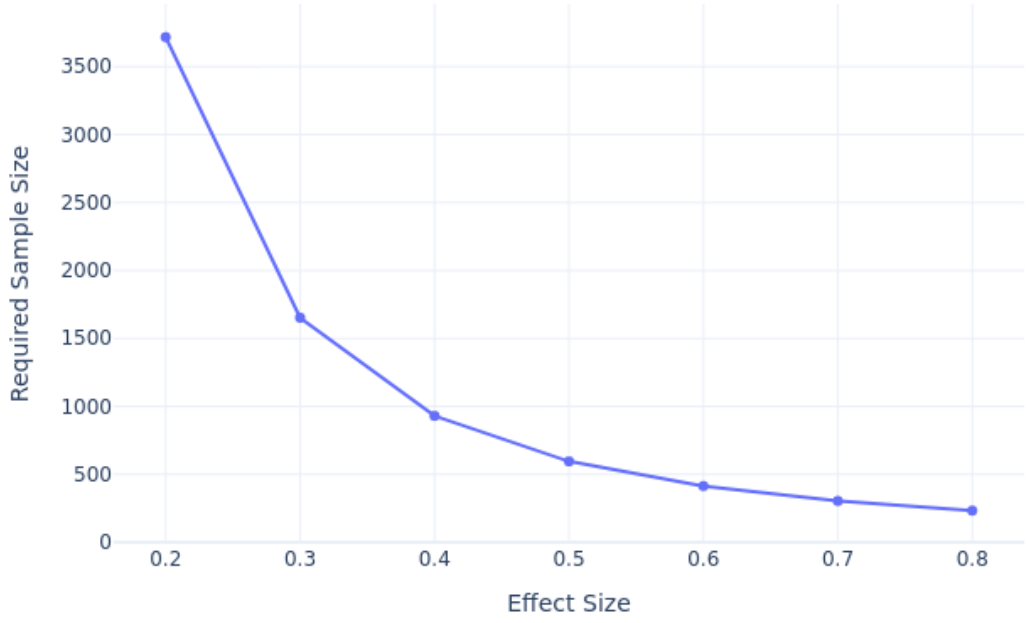

Figure 2. Power analysis with different effect size.

To tackle this issue, we conduct an extensive experiment to identify the optimal data sampling method, covering the five most commonly used techniques

- **Majority Undersampling**: Randomly removes instances from the majority class to balance the class distribution.
- **Minority Oversampling**: Randomly replicates instances from the minority class to balance the class distribution.
- **Tomek Link Undersampling** [27]: Identifies and removes majority class samples that form Tomek links (pairs of nearest neighbours from different classes) to clean the decision boundary.
- **Synthetic Minority Over-sampling Technique (SMOTE)** [28]: Generates synthetic samples for the minority class by interpolating between existing instances.
- **Adaptive Synthetic Sampling (ADASYN)** [29]: Generates synthetic samples by focusing more on the minority class instances that are near the decision boundary or in regions with more majority class samples.

3.2.2. Term Frequency-Inverse Document Frequency

To encode textual data for ML models, we employ a commonly-used technique namely Term Frequency-Inverse Document Frequency (TF-IDF) [30]. TF-IDF assesses the importance of individual terms within a text by calculating their frequency. We utilize these encoding techniques for experiments that employ clinical codes or code descriptions as training features for machine learning models. In medical texts, it is common for terms to appear as compound nouns. To try capturing potential associations between these terms, we generate n-gram tokens from the code descriptions. We retain only those terms that appear at least twice in the episode treatment. We also report the ablation study for choosing n in Appendix A2. It should be noted that when using the clinical diagnosis codenames rather than descriptions as the training features, these associations are reflected in the codenames due to their hierarchical structure design. Initially, TF-IDF is trained on the vocabulary of the training

 doi:10.20944/preprints202410.1476.v1

set and subsequently encodes entire medical paragraphs as single vectors for each episode in the validation set.

3.2.3. Machine Learning Models

In our investigation, we perform a comprehensive benchmarking comparison to identify the most optimal choice among nine models, encompassing both machine learning and deep learning approaches, with the results detailed in Table 3. These methods encompass some of the most well-known conventional machine learning methods: Logistic Regression, Random Forest, AdaBoost, XGBoost, CatBoost and LightGBM. Logistic Regression serves as a fundamental algorithm primarily utilized for binary classification tasks, recognized for its simplicity and interpretability [31]. In contrast, Random Forest operates as an ensemble learning technique, combining multiple decision trees to enhance predictive accuracy while mitigating overfitting risks [32]. AdaBoost, short for Adaptive Boosting, is tailored to enhance the performance of weak learners by assigning greater weight to misclassified instances [33]. Among the high-performing gradient boosting frameworks, we have selected XGBoost, CatBoost, and LightGBM. Extreme Gradient Boosting (XGBoost) [34] is recognized for its speed and performance due to features such as regularization, missing value handling, and parallelization, making it particularly suited for structured or tabular data. CatBoost [35] is engineered to manage categorical features natively, eliminating the need for extensive preprocessing like one-hot encoding. LightGBM [36] employs a leaf-wise tree growth algorithm, offering faster training times and improved performance on large datasets. Additionally, we leverage the Optuna tool [37] to optimize hyperparameters across all experimental models detailed in Table 3.

3.2.4. Deep Learning Models

We also incorporate representative deep learning models that are frequently utilized for natural language data: CNN, LSTM, and Transformer. Convolutional Neural Networks (CNN), primarily designed for image processing, can also effectively learn from natural language data by employing convolutional filters to capture local feature interactions. In the context of language data, CNNs can automatically identify patterns or dependencies among neighbouring language tokens, thereby facilitating the learning of structured or spatial relationships within the data [38]. Long Short-Term Memory Networks (LSTM), a variant of recurrent neural networks (RNN), excel at capturing sequential dependencies in data. When applied to tabular datasets, LSTMs are capable of modeling temporal or sequential relationships between features, particularly when the data has a time series or ordered structure. Transformers [39], characterized by their self-attention mechanism, can effectively manage tabular data by focusing on feature importance across the entire dataset. Unlike CNNs and LSTMs, Transformers are effective at capturing complicated feature interactions without being constrained by local or sequential structures, making them well-suited for high-dimensional tabular data with diverse feature relationships. To utilize these models with our feature sets without modifying their architecture, we only use clinical code names and descriptions as training features.

4. Results & Discussion

4.1. Validation Method and Performance Metrics

Researchers have extensively studied suitable evaluation metrics for classification tasks within the healthcare domain [40]. They have emphasized the importance of not relying solely on a subset of metrics, as doing so could potentially yield misleading results when implementing models in clinical settings. To comprehensively assess model performance, we employ multiple evaluation metrics, including Sensitivity, Specificity, F1 score and Area Under the Curve score (AUROC). The equations for the first three metrics are provided below:

 doi:10.20944/preprints202410.1476.v1

$$Specificity = \frac{TN}{TN + FP} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Here, TP, FP, TN, and FN denote True Positives, False Positives, True Negatives, and False Negatives, respectively. Specificity places a greater emphasis on correctly classifying negative samples, while sensitivity aims to minimize the misclassification of positive instances, making it a critical metric in medical studies [40].

The F1 score represents the harmonic mean of precision and recall, providing a balanced measure that penalizes extreme values of either precision or recall. AUROC, on the other hand, has gained popularity in machine learning [41] due to its favourable properties when dealing with imbalanced classes. It is used to distinguish between positive and negative classes across different threshold settings regardless of class distribution, making it an effective balance measurement. As a result, we have selected AUROC as the balance metric for model comparison.

To assess the robustness of our models, we employ cross-validation techniques, particularly stratified 5-fold cross-validation. This approach ensures that each fold maintains reasonable class distribution, mitigating the effects of class imbalance problems when using cross-validation. Following the guidance of [40], we avoid sharing data from the same patients across different folds to prevent introducing bias during the parameter tuning phase. We report the mean and standard deviation of model performance across these folds, and predictions made on the test set are presented for all the experiments.

4.2. Model Performance

Table 3 presents a performance comparison among traditional ML and DL approaches, revealing that CatBoost stands out as the leading model in terms of AUROC score, while LightGBM achieves the highest F1 score. It is intriguing that a straightforward algorithm like Logistic Regression shows a commendable balance in metric scores, likely due to the effectiveness of our robust data engineering pipeline and meticulous hyperparameter tuning. In contrast, the advanced deep learning methods appear to be less suited for this task, as evidenced by their underwhelming AUROC scores. This may be attributed to the fact that these models are solely trained on textual features, potentially overlooking significant predictors and missing out on the advantages offered by feature engineering. Zarghani [42] also found that deeper models such as LSTM tends to overfit on the training set and not generalizable to the smaller validation subset; resulting in inferior performance than methods like XGBoost and LightGBM in predicting patient readmission in diabetic patients. Figure 3 illustrates the ROC curves for these models.

 doi:10.20944/preprints202410.1476.v1

Table 3. Comparison between conventional ML models and DL models on the test set using their best settings on the validation set found by the Optuna optimizer.

Model	Results on test set using cross-validation				
	Balanced Accuracy	**Specificity**	**Sensitivity**	**F1 score**	**AUROC score**
Logistic Regression [31]	**0.646 (0.003)**	0.57 (0.009)	0.72 (0.008)	0.37 (0.003)	0.706 (0.001)
Random Forest [32]	0.633 (0.002)	0.56 (0.01)	0.706 (0.005)	0.364 (0.002)	0.683 (0.013)
AdaBoost [33]	0.63 (0.007)	0.52 (0.0004)	0.75 (0.015)	0.364 (0.006)	0.69 (0.004)
XGBoost [34]	0.64 (0.005)	0.6 (0.009)	0.678 (0.015)	0.37 (0.003)	0.69 (0.009)
CatBoost [35]	0.641 (0.002)	0.51 (0.013)	**0.768 (0.011)**	0.369 (0.002)	**0.71 (0.002)**
LightGBM [36]	0.645 (0.001)	0.61 (0.015)	0.68 (0.016)	**0.378 (0.0013)**	0.704 (0.006)
CNN1D	0.614 (0.01)	0.57 (0.123)	0.655 (0.11)	0.355 (0.009)	0.663 (0.013)
LSTM	0.632 (0.009)	**0.633 (0.052)**	0.63 (0.068)	0.372 (0.008)	0.68 (0.01)
Transformers [39]	0.62 (0.008)	0.566 (0.087)	0.672 (0.078)	0.358 (0.007)	0.67 (0.008)

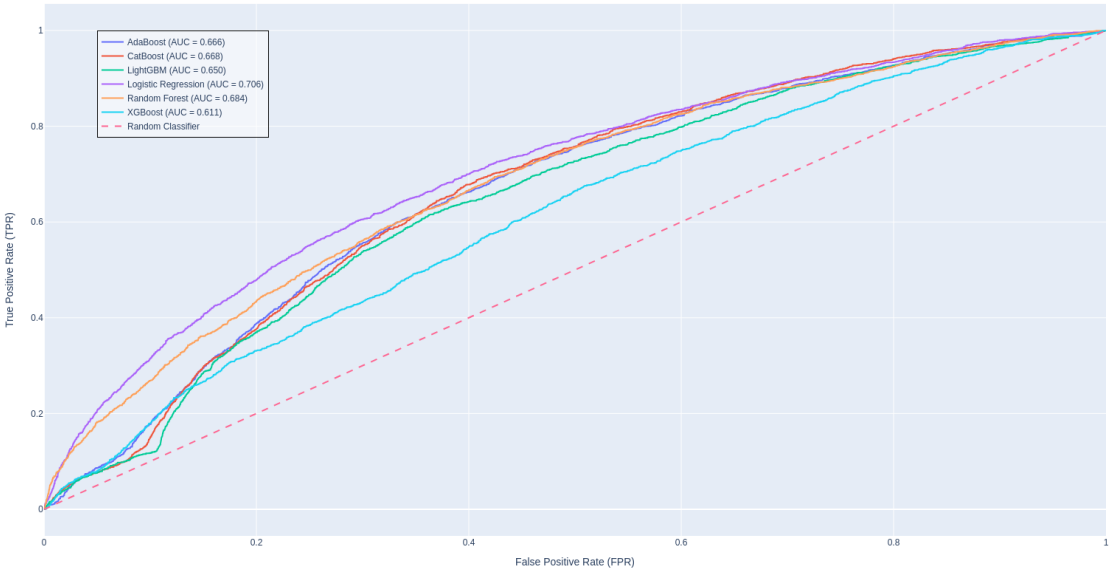

Figure 3. ROC Curves for all experiment models.

We also conduct a thorough evaluation of the individual features included in our models in order to assess their importance in Table 4. This evaluation involved systematically including each feature one at a time and examining whether there is an improvement in the evaluation score. We perform this setup for all ML models and aggregate the results for each setup. This ensures the consistency of the experiment and proves that the inclusion of newly engineered features is the main contribution to the performance. Otherwise, we include Figure 4 showing the impact of each group of features on the models individually.

Table 4. Comparison of effectiveness between features used in training the ML models. The results are aggregated across ML models and cross-validation folds and benchmarked on our data test set.

Exp. ID	**Patient Information**	**Code Description**	**Code names**	**Past Episode**	**Data Sampling**	**Specificity**	**Sensitivity**	**F1 score**	**AUROC score**
1	x					0.968 (0.037)	0.066 (0.086)	0.088 (0.096)	0.628 (0.034)
2	x	x				**0.972 (0.02)**	0.1 (0.053)	0.161 (0.075)	0.658 (0.014)
3	x		x			0.971 (0.021)	0.097 (0.051)	0.15 (0.073)	0.655 (0.013)
4	x		x	x		0.954 (0.037)	0.148 (0.088)	0.2 (0.097)	0.686 (0.02)
5	x		x	x	x	0.55 (0.069)	**0.726 (0.066)**	**0.369 (0.0098)**	**0.7 (0.01)**

 doi:10.20944/preprints202410.1476.v1

Figure 4. Effectiveness of each data features' inclusion and techniques applied across different models.

As presented in Table 4, it is crucial to note that the inclusion of clinical codes and descriptions has the most significant impact on the performance of the model, leading to an improvement of 3% in the AUROC score. Additionally, our findings agree with the conclusions drawn by Ma *et al.* [14] regarding the utility of code descriptions. In our case, the inclusion of code descriptions indeed led to performance improvement as opposed to using code names only (Experiment 2 and 3). (Experiment 2 and 3), For more information on this, we attach the detail evaluation table in Appendix A1.

Moreover, leveraging data from prior admissions introduces valuable prior knowledge for future decision-making. Experiment 4 received a substantial AUROC score improvement of approximately **3%** thanks to the incorporation of cumulative features from previous hospital episodes. As anticipated, we gain significant performance enhancement when all features from various modalities are incorporated. This observation strongly supports the efficacy of multimodal techniques in our predictive model.

It is important to highlight there is a significant imbalance between Specificity and Sensitivity in most of our reported experiments in Table 4. Experiments 1-4 exhibit extremely high Specificity scores, indicating a strong bias toward the negative class, which constitutes a substantial portion of the dataset. In contrast, Experiment 5, which integrated a sampling technique, achieve a more balanced distribution of scores between Specificity and Sensitivity, resulting in improved overall balance performance. We also report ablation study on different choices of data sampling methods in Appendix A4.

Overall, traditional models demonstrate a reasonable level of effectiveness in dealing with complex and diverse datasets like patient discharge records when appropriate data processing and engineering techniques are applied.

4.3. Feature Importance

In this section, we demonstrate the application of SHAP value visualization to illustrate feature importance for our models. The computation of SHAP values involves systematically perturbing input features and analyzing how these modifications impact the final model predictions. Figure 5 demonstrates the relative rankings of features based on their absolute SHAP value, aggregated across multiple models. In detail, for each feature (see the feature description in Appendix **??**) , we compute its absolute SHAP value, and then we rank the features based on this value. We do this for all ML models we mentioned, and then scale the ranking to 0-1 and average them. The higher the value of average ranking, the more influence it has on the patient readmission within the next 30 days.

 doi:10.20944/preprints202410.1476.v1

As can be seen, *CODENAMES* and *CODE_DESCRIPTION* is showing strong influence on patient readmission since they directly reflect the patient's condition. In contrast, other features display a wide variation in ranking distribution, indicating that different models prioritize different features as important.

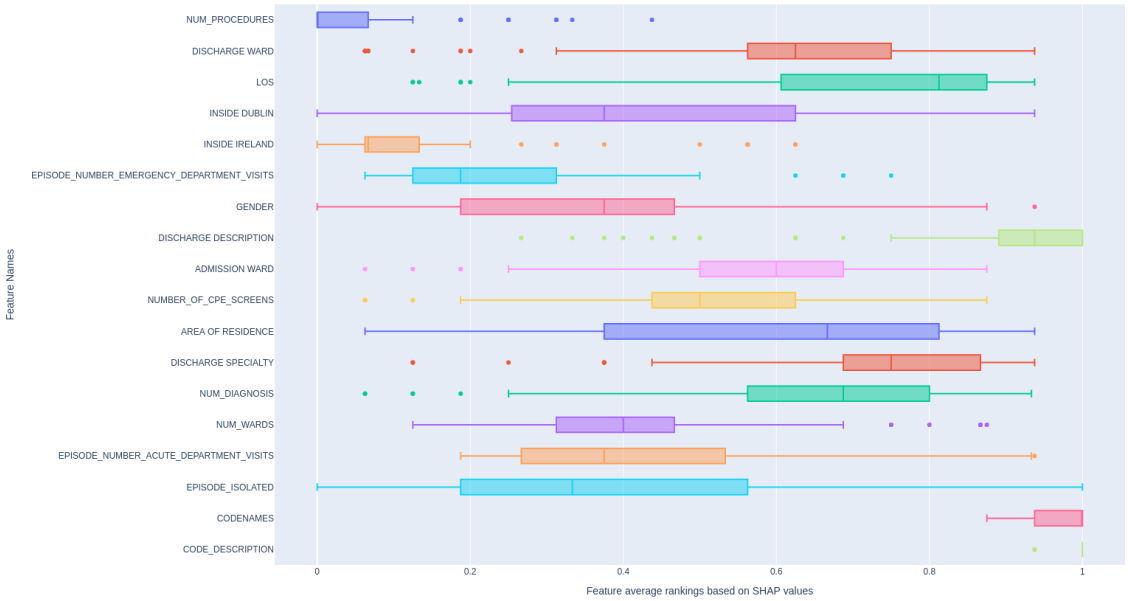

Figure 5. Overall feature Ranking based on SHAP absolute values across models.

In the following figures, the arrangement of features, from top to bottom, signifies their importance, with the most significant features occupying the upper positions. Additionally, each dot on the plot represents an individual data sample, with the colour intensity of each dot reflecting its numeric value when input into the model. The x-axis in the visualization displays the SHAP value for each plotted dot. Positive SHAP values indicate that the model is more inclined to classify the corresponding sample as positive (in our context, signifying that the patient is likely to be readmitted within the 30-day window) and negative values suggest the opposite. A SHAP value *"0.0"* is indicative of minimal impact. Within our focus, we exclusively examine the SHAP values associated with clinical codes and descriptions to assess their contribution to the model's predictions.

In Figure 6, we focus on what are the top influencing factors that are associated with patient readmission. Firstly, DISCHARGE_DESCRIPTION and DISCHARGE_SPECIALTY emerge as the most crucial feature in our dataset. We hypothesize that these two features reflect the clinical decision at the time of discharge, potentially indicating the guidance for diagnosis, treatment and destination received by patients upon discharge. Consequently, patients discharged from some specific departments (e.g. ones that often deal with chronic or long-term conditions) might have a higher likelihood of readmission. Additionally, the Length of Stay (LOS) feature and the total count of previous hospitalizations for a patient carry meaningful predictive values, suggesting that patients with frequent readmission history and longer hospital stays have an increased likelihood of readmission within a month. Lastly, we observed that a higher value in NUM_DIAGNOSIS is associated with an elevated impact on the prediction of patient readmission.

 doi:10.20944/preprints202410.1476.v1

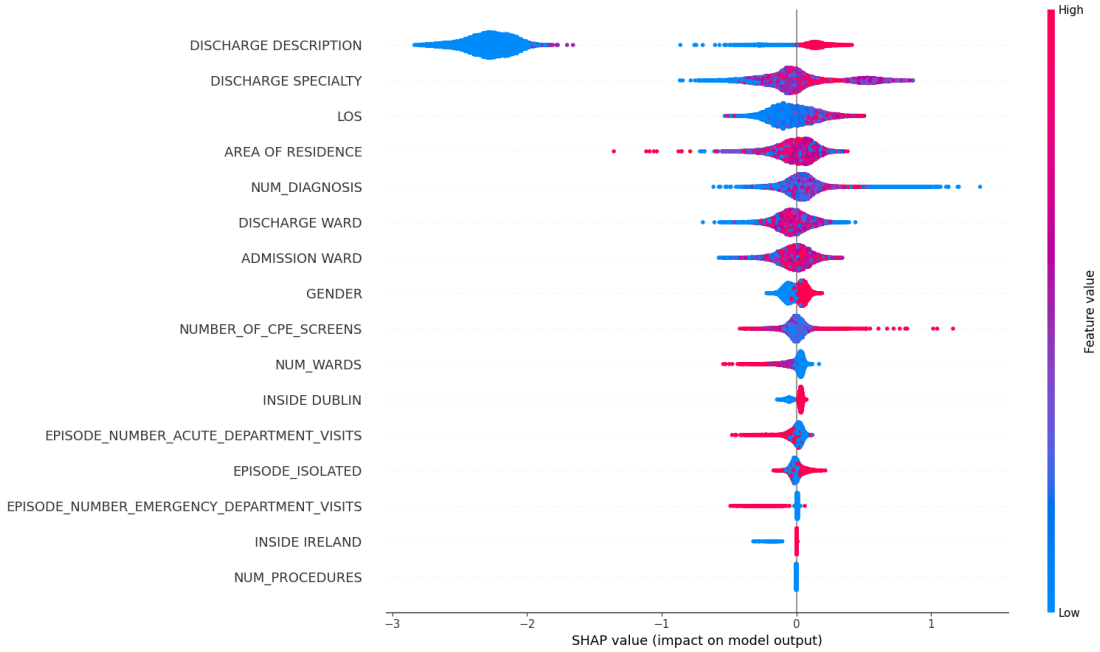

Figure 6. Feature Importance based on SHAP values, aggregated and averaged across models

Figure 7 provides an in-depth exploration of the most influential clinical codes on models, as outlined in Table 4. We have selectively chosen to present codes or terms that exhibit clearly distinguishable SHAP value measurements, meaning that their feature value does not vary along the x-axis. For those codes or terms that do not meet this criterion, despite having a significant impact, their interpretation is not straightforward in isolation and should be considered in conjunction with other terms or codes. Some of our findings in the feature importance align with medical insights from previous research.

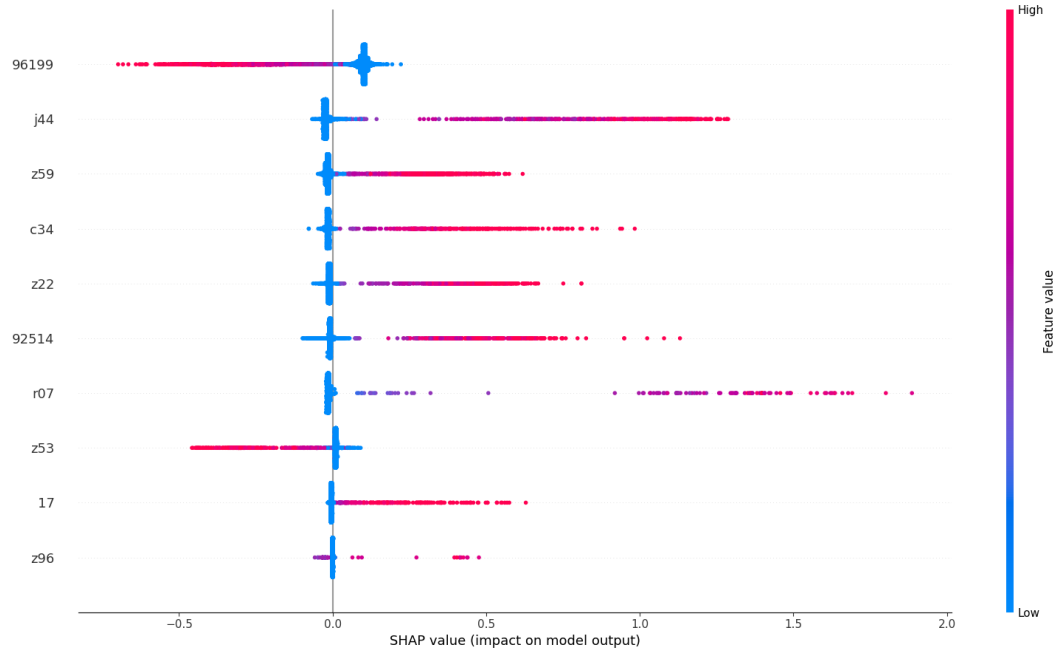

Figure 7. Clinical code Importance. We present generalized clinical codes that have a large impact on patient readmission prediction across models.

 doi:10.20944/preprints202410.1476.v1

Likewise, ICD-10 code *"Z59"* corresponds to patients facing *"Problems related to housing and economic circumstances,"* indicating a link between social factors and readmission rates. Previous studies have substantiated the association between social factors and all-cause unplanned readmissions, highlighting low socioeconomic status and housing instability as prominent contributing factors [43]. We also find "Z22" to be one of the minor causes for readmission; it presents a group of codes for "carrier of infectious diseases" which is of potential interest to the ARK project. In our interpretation, these clinical codes, indicating the diagnoses of patients' conditions and their social status, are important predictors of the patients' 30-day readmissions.

In conclusion, our SHAP value visualization analysis has provided valuable insights into the global observation of how a patient's diagnosis contributes to the likelihood of their readmission within 30 days. Notably, diseases such as cancer, indicated by clinical codes like '96199' and terms like 'malignant neoplasm,' emerge as key contributors to increased readmission risks. Furthermore, chronic conditions like 'J44' (COPD) and social factors represented by 'Z59' are identified as significant determinants. However, being bacteria carriers, as indicated by code 'Z22' appear to have minimal impact on the readmission likelihood of patients, likely due to the scarcity of such cases during the study period. As one of our objectives is to study clinical risks, we leave the investigation of the effects of bacteria for our future research.

5. Conclusion & Future Works

Patient risk management is crucial for healthcare providers because it ensures patient safety, prevents adverse events, and reduces legal and financial risks. Forecasting patient readmission automatically using AI helps hospitals achieve these goals by allowing timely interventions, better post-discharge planning, and more efficient resource allocation.

In this study, we carry out a performance comparison of multiple conventional ML models on the extensive incorporation of multimodal discharge records data from an Irish acute hospital. Despite the complex and diverse nature of data within the HIPE system, gradient boosting algorithms come out as the top algorithms demonstrating their strong ability to simultaneously capture patient demographics, historical hospitalization records, and clinical diagnoses codes. Overall, after integrating all these features into the algorithm, we successfully improve the results from baseline AUROC score of **0.628 to 0.7** averaging over 9 models. Although an AUROC score at around 0.7 seems to be moderate, it falls within the typical range observed in an overview study for readmission prediction conducted by Yu and Son [44], which is between 0.51 and 0.93. This also highlights the utility of routinely collected data in hospitals to further improve the outcome of patients in this digital era.

The successful application of the conventional SHAP value technique to dissect the ML models has allowed us to provide visually interpretable figures that illustrate the predictions of the model without compromising accuracy on complex multimodal healthcare data. Notably, through this analysis, we have revealed from patients' diagnosis codes, establishing a robust correlation with their likelihood of future readmission. This understanding has the potential to empower hospital experts and organizations in anticipating future risks and costs more effectively, ultimately leading to improvements in hospital processes and enhanced patient outcomes. By recognizing that certain common illnesses within Irish hospitals significantly contribute to readmission risks, healthcare providers can tailor interventions and care plans to address the specific needs of patients with these diagnosis codes. This proactive approach has the potential not only to reduce the likelihood of readmissions but also optimize resource allocation, thus fostering a more efficient and patient-centric healthcare system.

Nevertheless, there are several limitations to our study. One challenge when working with clinical codes is their high dimensionality due to their complicated and sparsely populated nature. As evidenced by the approximately **10,000 distinct codes** reported in Table 1, sparsity is a significant concern. Conversely, utilizing clinical terms results in a less sparse word embedding matrix, enabling the capture of the hierarchical structure inherent in these clinical codes. Despite our efforts to address

 doi:10.20944/preprints202410.1476.v1

this issue through code generalization, it proved to be an ineffective solution - as evidenced in Appendix A1 - possibly due to loss of information. It is also possible that missing data may contain concealed information that could serve as potential predictors as has been reported in [45]. We have noticed that other authors have recommended other methods for missing data and its impact on ML classifiers in similar problems [46,47]. In addition to these, we have conducted comprehensive benchmarks to assess different data sampling methods. However, the same configurations have yet to be applied to deep learning models, largely due to the complexities associated with the data. Future research should aim to investigate more sophisticated strategies to effectively address both of these challenges.

In the realm of healthcare, future research may need to delve deeper into the architectures of deep learning models to adapt them to this specific domain for better performance, but it will also require additional efforts to enhance transparency in dealing with their black-box architectures, as well as seeking domain expert's validation and verification. For instance, language models like CNN [48], Transformer [39], or process mining techniques could be explored further to leverage sequential patient visits and improve predictive capabilities while maintaining interpretability.

Author Contributions: Conceptualization, MK.P, M.C. and M.B.; methodology, MK.P.; software, MK.P.; validation, MK.P.; formal analysis, MK.P.; investigation, MK.P.; resources, M.C.; data curation, MK.P, M.E., R.B., U.G., N.M and M.W.; writing—original draft preparation, MK.P.; writing—review and editing, everyone; visualization, MK.P.; supervision, M.C., M.B., T.T.M.; project administration, M.C. M.B. ; funding acquisition, M.B., M.C.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Institutional Review Board Statement: The study was conducted using real-world anonymised data set that was shared as part of the ARK risk management research project that had ethics approval from the hospital (TUH/SJH Rec Project ID: 0291) and Dublin City University (DCUREC/2021/118).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AI: Artificial intelligence; DL: Machine Learning; DL: Deep Learning; HCAI: Healthcare-Associated Infection; PCHCAI: Prevention and Control of Healthcare Associated Healthcare Infections; EPRs: Electronic Patient Records; AUROC: Average area under the receiver operating characteristic curve; HIPE: Hospital Inpatient Enquiry; ICD-10: International Classification of Diseases Codes version 10;

 doi:10.20944/preprints202410.1476.v1

Appendix A. Description of the Used Data Features

Patient demographics

GENDER
Patient's sex. Male, Female or Other

AREA OF RESIDENCE
The patient's place of living, includes Dublin Districts, other Ireland counties, and a small amount of other European countries.

Patient episode information

EPISODE_ISOLATED
Boolean value whether the patient was in an isolation room in the current episode

LOS
Length of stay of the current episode

DISCHARGE DESCRIPTION
Episode's outcome

DISCHARGE SPECIALTY
Episode's discharge ward specialty

Clinical code information

NUM_DIAGNOSIS
Number of diagnoses that patient received during the episode

NUM_PROCEDURES
Number of procedures that patient received during the episode

CODENAMES

Ward transition information

ADMISSION WARD
The codename of the ward to which the patient was admitted

DISCHARGE WARD
The codename of the ward on which the patient was discharged

EPISODE_NUMBER_ACUTE_DEPARTMENT_VISITS
Number of acute departments that the patient visited during the current episode. In our study, we refer to wards that admit patients directly from the Emergency Department as Acute Departments.

EPISODE_NUMBER_EMERGENCY_DEPARTMENT_VISITS
Number of emergency departments that patient visited during the current episode

NUM_WARDS
Number of ward transitions during the episode

Appendix B. Ablation Study

Figure A1 illustrates that the implementation of code generalization results in a broader range of performance outcomes, while also exhibiting a higher median performance compared to the non-generalized approach.

 doi:10.20944/preprints202410.1476.v1

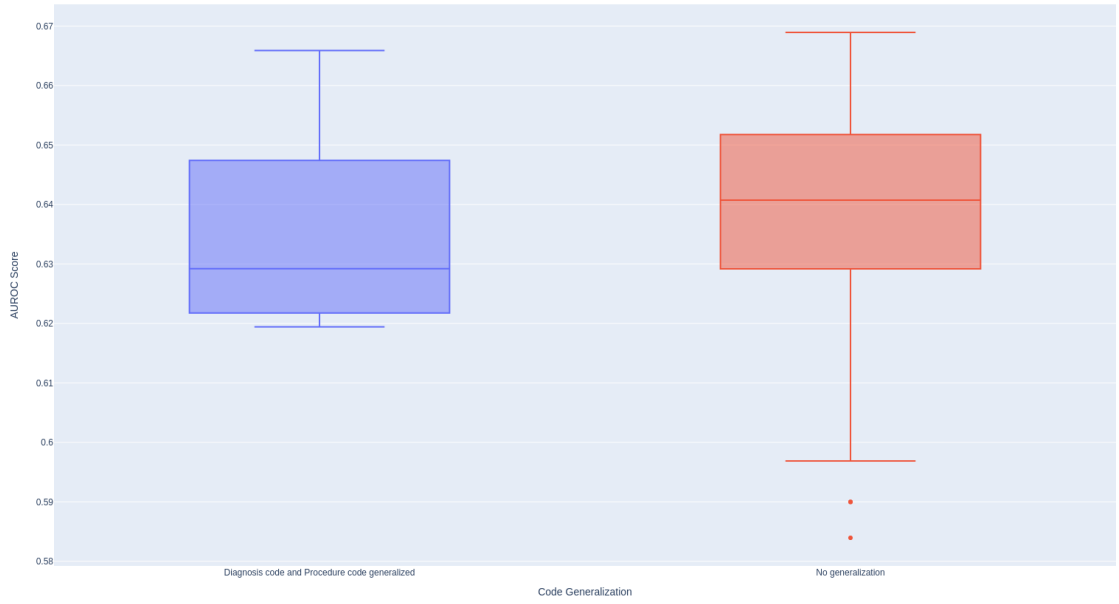

Figure A1. Effectiveness of Generalized Codes. In the generalized setting, all diagnosis codes are truncated to the first three letters, while all procedure codes are truncated to the first five letters.

As demonstrated in Figure A2, the utilization of the 1-gram model provides the most stable distribution of performance. Conversely, employing larger n-grams tends to result in diminished performance.

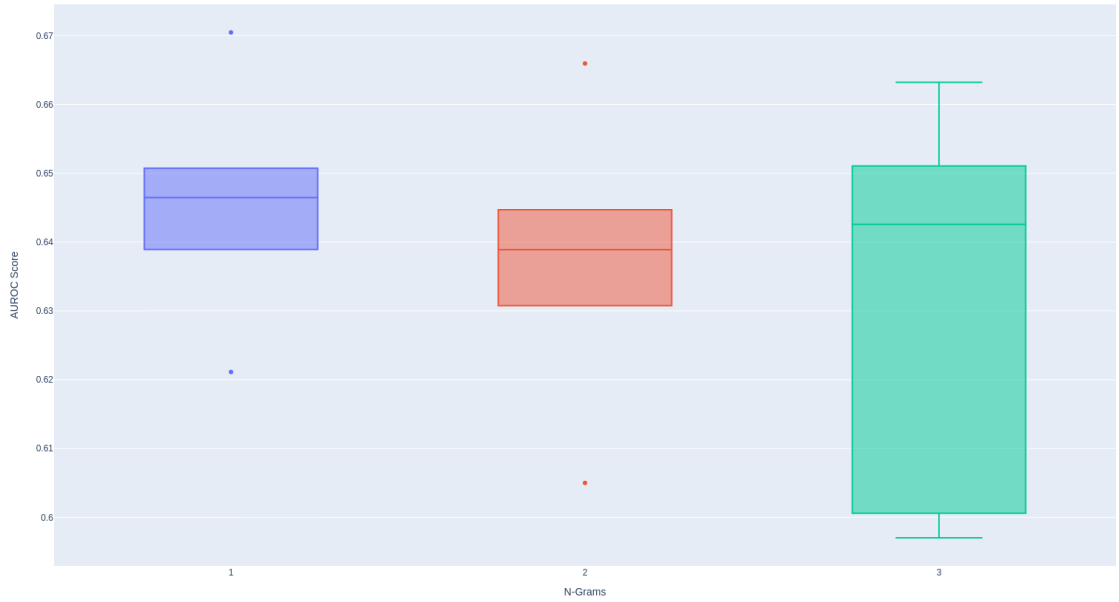

Figure A2. Effectiveness of N-gram Models When Encoding Textual Data Using TF-IDF.

Figure A3 indicates that incorporating information from past episodes enhances the performance of all models, underscoring the significance of historical data in predicting patient readmissions.

 doi:10.20944/preprints202410.1476.v1

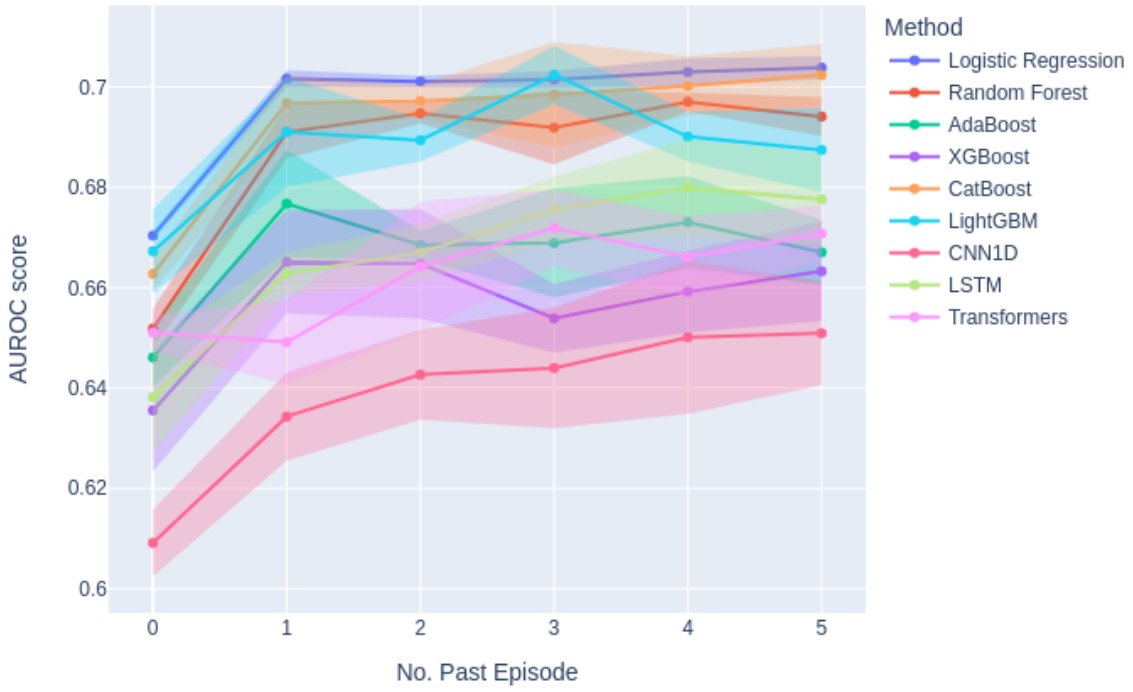

Figure A3. Effectiveness of the Number of Past Episodes Considered for Prediction Across Different Models.

Figure A4 illustrates that all sampling methods employed effectively address the class imbalance issue, demonstrating improvements in F1 scores across the models.

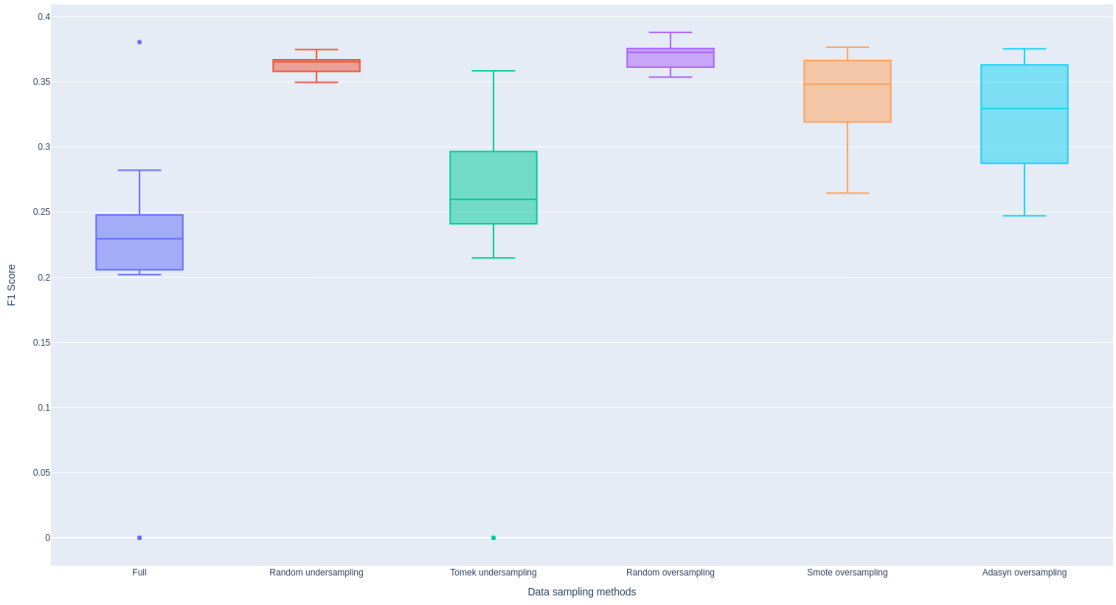

Figure A4. Effectiveness of Various Data Sampling Methods Aggregated Across Models.

 doi:10.20944/preprints202410.1476.v1

Table A1. Comparative Effectiveness of Code Names versus Code Descriptions in Machine Learning and Deep Learning Models.

Exp.	AUROC score results on test set using cross-validation								
	XGBoost	AdaBoost	LightGBM	CatBoost	Logistic Regression	Random Forest	CNN1D	LSTM	Transformers
Code Names	0.60426 (0.0106)	0.638 (0.0037)	**0.6317 (0.006)**	0.64555 (0.0048)	0.666 (0.0016)	**0.6513 (0.004)**	0.61 (0.0066)	0.638 (0.01)	**0.651 (0.0035)**
Code Description	**0.616 (0.0066)**	**0.642 (0.0046)**	0.631 (0.008)	**0.649 (0.0037)**	**0.669 (0.0014)**	0.638 (0.0075)	**0.66 (0.001)**	**0.655 (0.0016)**	0.643 (0.004)

References

1. McGowan, J.; Wojahn, A.; Nicolini, J.R. Risk management event evaluation and responsibilities **2020**.
2. Clancy, C.; Shine, C.; Hennessy, M. Spending Review 2022 Hospital Performance: An Analysis of HSE Key Performance Indicators **2023**.
3. Kripalani, S.; Theobald, C.N.; Anctil, B.; Vasilevskis, E.E. Reducing hospital readmission rates: current strategies and future directions. *Annual review of medicine* **2014**, *65*, 471–485.
4. McDonald, N.; McKenna, L.; Vining, R.; Doyle, B.; Liang, J.; Ward, M.E.; Ulfvengren, P.; Geary, U.; Guilfoyle, J.; Shuhaiber, A.; others. Evaluation of an access-risk-knowledge (ARK) platform for governance of risk and change in complex socio-technical systems. *International Journal of Environmental Research and Public Health* **2021**, *18*, 12572.
5. Chauhan, N.K.; Singh, K. A Review on Conventional Machine Learning vs Deep Learning. 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 347–352.
6. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of explainable AI techniques in healthcare. *Sensors* **2023**, *23*, 634.
7. Ashfaq, A.; Sant'Anna, A.; Lingman, M.; Nowaczyk, S. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics* **2019**, *97*, 103256. https://doi.org/10.1016/j.jbi.2019.103256.
8. Wang, W.W.; Li, H.; Cui, L.; Hong, X.; Yan, Z. Predicting Clinical Visits Using Recurrent Neural Networks and Demographic Information. 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)), 2018, pp. 353–358. doi:10.1109/CSCWD.2018.8465194.
9. Wang, H.; Cui, Z.; Chen, Y.; Avidan, M.; Abdallah, A.B.; Kronzer, A. Predicting Hospital Readmission via Cost-Sensitive Deep Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2018**, *15*, 1968–1978. doi:10.1109/TCBB.2018.2827029.
10. Davis, S.; Zhang, J.; Lee, I.; Rezaei, M.; Greiner, R.; McAlister, F.A.; Padwal, R. Effective hospital readmission prediction models using machine-learned features. *BMC Health Services Research* **2022**, *22*, 1415.
11. Michailidis, P.; Dimitriadou, A.; Papadimitriou, T.; Gogas, P. Forecasting hospital readmissions with machine learning. Healthcare. MDPI, 2022, Vol. 10, p. 981.
12. Ryu, B.; Yoo, S.; Kim, S.; Choi, J. Thirty-day hospital readmission prediction model based on common data model with weather and air quality data. *Scientific Reports* **2021**, *11*, 23313.
13. Deschepper, M.; Eeckloo, K.; Vogelaers, D.; Waegeman, W. A hospital wide predictive model for unplanned readmission using hierarchical ICD data. *Computer Methods and Programs in Biomedicine* **2019**, *173*, 177–183. doi:https://doi.org/10.1016/j.cmpb.2019.02.007.
14. Ma, F.; Wang, Y.; Xiao, H.; Yuan, Y.; Chitta, R.; Zhou, J.; Gao, J. Incorporating medical code descriptions for diagnosis prediction in healthcare. *BMC medical informatics and decision making* **2019**, *19*, 1–13.
15. Choi, E.; Bahadori, M.T.; Sun, J.; Kulas, J.; Schuetz, A.; Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* **2016**, *29*.
16. Choi, E.; Bahadori, M.T.; Schuetz, A.; Stewart, W.F.; Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. Machine learning for healthcare conference. PMLR, 2016, pp. 301–318.
17. Feucht, M.; Wu, Z.; Althammer, S.; Tresp, V. Description-based Label Attention Classifier for Explainable ICD-9 Classification. Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), 2021, pp. 62–66.
18. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
19. Shapley, L.S.; others. A value for n-person games **1953**.

20. Duell, J.; Fan, X.; Burnett, B.; Aarts, G.; Zhou, S.M. A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2021, pp. 1–4.
21. Du, Y.; Rafferty, A.R.; McAuliffe, F.M.; Wei, L.; Mooney, C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Scientific Reports* **2022**, *12*, 1170.
22. Alsinglawi, B.; Alshari, O.; Alorjani, M.; Mubin, O.; Alnajjar, F.; Novoa, M.; Darwish, O. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific reports* **2022**, *12*, 607.
23. (HPO), H.P.O. Irish Coding Standards (ICS) 2023 (V1) 10th Edition ICD-10-AM/ACHI/ACS, 2023.
24. Organization, W.H.; others. ICD-10. International Statistical Classification of Diseases and Related Health Problems: Tenth Revision 1992, Volume 1= CIM-10. Classification statistique internationale des maladies et des problèmes de santé connexes: Dixième Révision 1992, Volume 1 **1992**.
25. Choi, E.; Schuetz, A.; Stewart, W.F.; Sun, J. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association* **2017**, *24*, 361–370.
26. Lipton, Z.C.; Kale, D.C.; Elkan, C.; Wetzel, R. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677* **2015**.
27. Elhassan, T.; Aljurf, M. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global J Technol Optim S* **2016**, *1*, 2016.
28. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **2002**, *16*, 321–357.
29. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). Ieee, 2008, pp. 1322–1328.
30. Sparck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* **1972**, *28*, 11–21.
31. Wright, R.E. Logistic regression. **1995**.
32. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
33. Freund, Y.; Schapire, R.E.; others. Experiments with a new boosting algorithm. icml. Citeseer, 1996, Vol. 96, pp. 148–156.
34. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
35. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* **2018**, *31*.
36. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **2017**, *30*.
37. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.
38. Kim, Y. Convolutional Neural Networks for Sentence Classification, 2014, [arXiv:cs.CL/1408.5882].
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
40. Hicks, S.A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M.A.; Halvorsen, P.; Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports* **2022**, *12*, 5979.
41. Gaudreault, J.G.; Branco, P.; Gama, J. An analysis of performance metrics for imbalanced classification. International Conference on Discovery Science. Springer, 2021, pp. 67–77.
42. Zarghani, A. Comparative Analysis of LSTM Neural Networks and Traditional Machine Learning Models for Predicting Diabetes Patient Readmission. *arXiv preprint arXiv:2406.19980* **2024**.
43. Navathe, A.S.; Zhong, F.; Lei, V.J.; Chang, F.Y.; Sordo, M.; Topaz, M.; Navathe, S.B.; Rocha, R.A.; Zhou, L. Hospital readmission and social risk factors identified from physician notes. *Health services research* **2018**, *53*, 1110–1136.
44. Yu, M.Y.; Son, Y.J. Machine learning–based 30-day readmission prediction models for patients with heart failure: a systematic review. *European Journal of Cardiovascular Nursing* **2024**, p. zvae031.
45. Sharafoddini, A.; Dubin, J.A.; Maslove, D.M.; Lee, J.; others. A new insight into missing data in intensive care unit patient profiles: observational study. *JMIR medical informatics* **2019**, *7*, e11605.

 doi:10.20944/preprints202410.1476.v1

46. Zhang, X.; Yan, C.; Gao, C.; Malin, B.A.; Chen, Y. Predicting missing values in medical data via XGBoost regression. *Journal of healthcare informatics research* **2020**, *4*, 383–394.
47. Masud, J.H.B.; Kuo, C.C.; Yeh, C.Y.; Yang, H.C.; Lin, M.C. Applying deep learning model to predict diagnosis code of medical records. *Diagnostics* **2023**, *13*, 2297.
48. Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; Eisenstein, J. Explainable Prediction of Medical Codes from Clinical Text. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers); Walker, M.A.; Ji, H.; Stent, A., Eds. Association for Computational Linguistics, 2018, pp. 1101–1111. doi:10.18653/v1/n18-1100.