Enhancing Predictive Modeling in Emergency Departments

Mojgan Kouhounestani[a], Long Song[b], Ling Luo[c] and Uwe Aickelin[d]
School of Computing and Information Systems, University of Melbourne, Grattan Street, Parkville, 3010, VIC, Australia

Keywords: Electronic Health Records, Hospitalisation, Target Encoding, Nominal Features, Predictive Data Mining.

Abstract: Increasing global Emergency Department (ED) visits, exacerbated by COVID-19, has presented multiple challenges in recent years. Electronic Health Records (EHRs) as comprehensive digital repositories of patient health information offer a pathway to construct prediction systems to address these issues. However, the heterogeneity of EHRs complicates accurate predictions. A notable challenge is the prevalence of high-cardinality nominal features (NFs) in EHRs. Due to their numerous distinct values, these features are often excluded from the analysis, risking information loss, reduced accuracy, and interpretability. This study proposes a framework, integrating a preprocessing technique with target encoding (TE-PrepNet) into machine learning (ML) models to address challenges of NFs from MIMIC-IV-ED. We evaluate performance of TE-PrepNet in two specific ED-based prediction tasks: triage-based hospital admissions and ED reattendance within 72 hours at discharge time. Incorporating three NFs, our approach demonstrates improvements compared to the baseline and outperforms previous research that overlooked NFs. Random forest model with TE-PrepNet in the prediction of hospitalisation achieved an AUROC of 0.8458, compared to the baseline AUROC of 0.7520. For the prediction of ED reattendance within 72 hours, the utilisation of XGBoost yielded an improvement, attaining an AUROC of 0.6975, outperforming the baseline AUROC of 0.6166.

1 INTRODUCTION

In recent years, the application of artificial intelligence has increased in various aspects of modern life, including in medicine. The speed, superior performance, and accuracy of machine learning (ML) models are motivators for their widespread use in the medical and health fields. ML models serve as decision-making aids to clinicians to enhance and support patient access to care. Furthermore, they can replicate medical expertise and workflows in repetitive tasks, allowing physicians to focus on higher-value jobs (Liu et al., 2022). Therefore, ML has enormous potential to improve the health and well-being of the healthcare industry. The Emergency Department (ED) is an important part of the healthcare system that provides immediate medical attention to patients. The demand for ED services has increased in recent years owing to an ageing population and limited access to primary care (Brownell et al., 2014), further compounded by the emergence of the COVID-19 pandemic. This escalating demand for emergency care leads to overcrowding in the ED, extended service delays, prolonged waiting times, and declined quality of care. This ultimately affects the overall satisfaction of patients (Kusumawati et al., 2019) and increased in-hospital mortality (Guttmann et al., 2011).

Electronic Health Record (EHR) is a comprehensive digital repository of a patient's health information generated through various encounters in different healthcare settings. It is intended to improve healthcare practitioners' efficiency and workflow by producing a complete record of a patient's clinical interaction and assisting in other care-related activities, such as providing evidence-based decision support, maintaining quality, and tracking outcomes (Häyrinen et al., 2008). Most Australian public hospitals now implement an EHR system, enabling healthcare providers to gain easier access to critical patient information (Mollart et al., 2020). EHRs have become indispensable tools in ED, allowing clinicians quick access to important patient information. Such accessibility holds the potential to increase the quality of care and minimise the likelihood of errors. EHR adoption also improves communication and coordination between EDs and other healthcare professionals,

[a] https://orcid.org/0000-0001-7935-6410
[b] https://orcid.org/0000-0001-8494-4364
[c] https://orcid.org/0000-0002-1363-8308
[d] https://orcid.org/0000-0002-2679-2275

Kouhounestani, M., Song, L., Luo, L. and Aickelin, U.
Enhancing Predictive Modeling in Emergency Departments.
DOI: 10.5220/0012568900003699
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 10th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2024)*, pages 37-46
ISBN: 978-989-758-700-9; ISSN: 2184-4984

ensuring patients' continuity of treatment even if they are moved to another facility.

EHRs encompass an expansive array of data types, spanning from, numerical data - such as blood pressure; categorical data like pain scale assessments; textual information including prescription details- to even temporal data, indicating the timing of measurements. This extensive variety of data types contributes to the heterogeneity of this dataset. On the other hand, most ML algorithms are primarily designed to handle numerical data and face difficulties when dealing with non-numerical types like categorical data, which can be categorised into nominal data (without any inherent order) and ordinal data (characterised by a specific order). Despite significance of this information in enhancing the interpretability of ML models, they pose challenges. Conventional techniques can convert these features into numerical variables; however, the increasing number of unique values results in high-dimensional feature matrix and computational challenges, especially when used with computationally demanding models.

In recent studies, particularly in the field of medicine, there is a growing trend of using a subset of values extracted from nominal features (NFs) to a harmonious balance between optimising data utility and managing the dimension of the dataset. Nonetheless, this approach has potential downsides, including the risk of losing valuable information and heavily relying on domain expertise to select the most relevant values. Therefore, in numerous applications, these features are often disregarded or considered to be leveraging domain knowledge, so only a subset of their distinctive values is considered. In this study, our contributions are to:

- Tackle the challenges associated with NFs in EHRs by employing the proposed target encoding preprocessing framework (TE-PrepNet).
- Optimise high-cardinality NFs handling by minimising dependency on domain experts, while maximising the integration of embedded values. This optimisation is accomplished through incorporating the TE-PrepNet.
- Assess two distinct ED-based prediction tasks: prediction of hospital admissions at the time of triage in the ED; prediction of reattendance to the ED within 72 hours after discharge.

We applied the target encoding approach on a set of chosen NFs (race, arrival transport mode, and chief complaint), encompassing both high and low-cardinality characteristics, which are extracted from the Medical Information Mart for Intensive Care IV Emergency Department (MIMIC-IV-ED) dataset (Johnson et al., 2021). The results highlighted the performance enhancements and effectiveness of using the TE-PrepNet on both of the aforementioned prediction tasks. In particular, the implementation of random forest with target encoding achieved an AUROC of 0.8458, outperforming the baseline AUROC of 0.7520. Furthermore, in predicting 72-hour reattendance, the use of XGBoost with target encoding achieved an AUROC of 0.6975, showing an improvement from the baseline's previous AUROC of 0.6166.

2 RELATED WORK

Given the continual influx of data into EHRs, the integration of ML holds promise in facilitating comprehensive analysis. By discerning trends, detecting patterns, and offering predictions pertaining to a patient's well-being, ML can play a pivotal role in enhancing healthcare. In recent years, ML models have capitalised on the potential offered by EHRs to undertake a spectrum of predictions pertinent to the ED. These efforts contain predictions related to hospital admission (Barak-Corren et al., 2017; Xie et al., 2022; Hong et al., 2018; Graham et al., 2018; Al Shalabi et al., 2006), early prediction of sepsis or septic shock in the ED (Wardi et al., 2021), predictions concerning the length of stay within the ED (Gurazada et al., 2022; Rahman et al., 2020), as well as forecasts regarding the length of stay for COVID-19 patients specifically within the ED (Etu et al., 2022). So, implementing early prediction models for patient admissions can be beneficial in addressing the problem of long boarding times and expediting resource allocation, and enhancing overall patient care efficiency.

Conventional techniques, such as one-hot encoding (or dummy encoding), have been employed in the handling of nominal variables with a limited number of distinct values (Hancock and Khoshgoftaar, 2020). These methods effectively convert a nominal variable with N unique values into N new variables (or $N-1$ variables in the case of dummy encoding) to capture its categorical nature. However, their effectiveness significantly decreases when dealing with NFs with many distinct values, primarily due to the inherent challenge of high dimensionality. The escalation in dimensionality poses computational and interpretational difficulties, limiting the applicability of these methods in scenarios where NFs exhibit a multitude of unique values.

Apart from employing one-hot encoding, proposing the use of clustering techniques is also an option. These techniques involve grouping individual values into K sets. Although this approach results in fewer introduced variables compared to one-hot encoding ($K << N$) (Micci-Barreca, 2001), in cases involving high-cardinality NFs, the challenge of high dimensionality persists because the number of clusters, K, remains relatively large.

Achieving the right balance between dimensionality reduction and information retention is essential, guided by domain-specific insights (Xie et al., 2022; Hong et al., 2018; Barak-Corren et al., 2017). In this approach, domain knowledge is leveraged to select a subset of distinctive values from NFs, which are then incorporated as binary features into ML models.

Target-based techniques, which leverage information pertaining to the target variable, often demonstrate superior performance compared to approaches that neglect such information, particularly when dealing with high-cardinality NFs (Pargent et al., 2022). The utilisation of score-based target encoding to predict unplanned hospitalisations among elderly patients has enhanced the performance of ML models (Nazyrova et al., 2022). In addition to demographic attributes, their analysis extended to encompass subsets of drug and disease categories. Notably, the highest cardinality among the features was linked to 31 distinct values, significantly lower than what we introduce as high-cardinality features with thousands of distinct values.

3 METHODOLOGY

The primary objective of this approach is to convert NFs into continuous scalar values, compatible with ML models. This conversion is achieved while preserving the original dimensionality of the dataset without introducing additional attributes. Fig. 1 provides an overview of the training phase of the aforementioned prediction tasks.

3.1 Target Encoding

The core concept involves assigning a probability ($\hat{X}_l$) estimate to each value (l) of the nominal variable (X), based on its association with the outcome attribute (y_j) on the training set, as depicted in Equation 1, where n_l represents the frequency of occurrence of level l in the training set (Pargent et al., 2022).

$$\hat{X}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} y_j \qquad (1)$$

By incorporating this probability estimate, nominal values are effectively transformed into a format that captures the likelihood of information pertaining to the target attribute (Micci-Barreca, 2001). This enables the utilisation of nominal data in ML algorithms, enhancing their ability to leverage the probabilistic characteristics of the data. The target variable in this context can be associated with either binary classification tasks or multi-class classification tasks.

3.2 Handling Nominal Features

This work centers on the incorporation of NFs, characterised by a substantial prevalence of distinct values. The first and straightforward method involves replacing each nominal value with corresponding scores generated through target encoding. However, this initial method is not robust enough to adequately address the complexities of our data. In practical scenarios, certain records in the dataset may exhibit multiple distinct values for specific nominal variables. Consider, for instance, the inclusion of chief complaints as an NF in the dataset, wherein a single record may encompass multiple distinct complaint names. Therefore, an important challenge arises due to the existence of diverse values for specific NF during particular instances. Another challenge is that, when dealing with high-cardinality nominal values, some values may be unseen in the training set. Consequently, addressing previously unseen values in the test dataset becomes imperative. To tackle these challenges, we design TE-PrepNet with following two phases:
During the training phase, we handle NFs in four steps:

I. Discern various expressions within each nominal value.

II. Apply target encoding to assign numerical values to distinct expressions associated with these features.

III. Establish a dictionary to facilitate the consolidation of these encoded values across all expressions within the training dataset.

IV. Compute cumulative sum of scores linked with each expression in instances where patients exhibit multiple complaints during their ED visits. These The diverse values are characterised by value separation via ‘,‘ resulting in the generation of unique expressions.

Then the testing phase is embarked upon:

- If expressions were previously seen during training, corresponding numeric values are directly assigned based on the established dictionary.
- If expressions were unseen during training, we calculate the target score in two steps by comparing the unseen new expression with the expression in our dictionary:

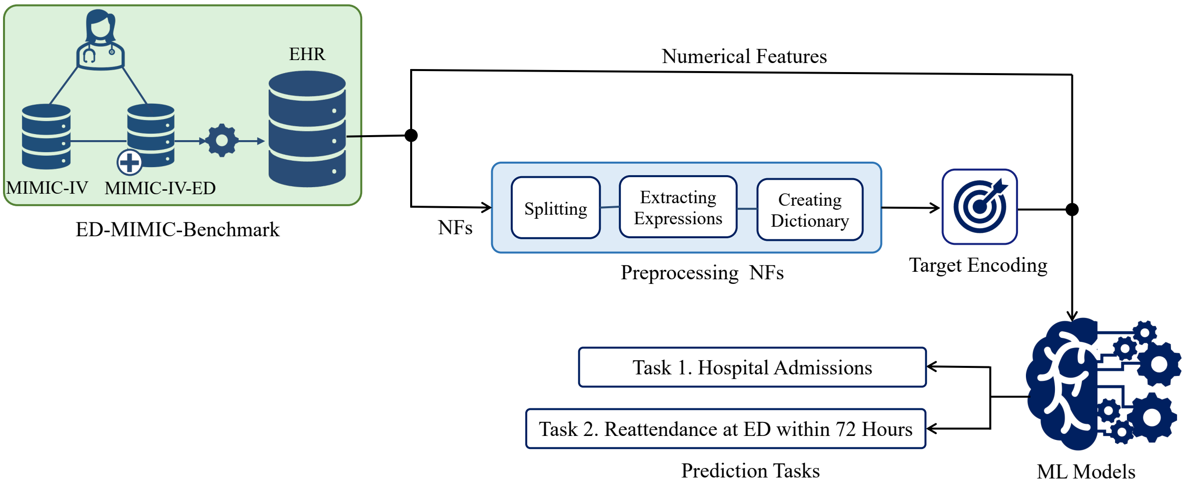


Figure 1: The schematic representation of the training phase. NFs refer to the nominal features. The preprocessing procedures are specifically applied to high-cardinality NFs.

i. Find five most similar expressions based on Jaccard similarity, which quantifies the similarity between two expressions (A and B) by assessing the fraction of the intersection of their sets divided by the union of all values belonging to the two expressions (Zahrotun, 2016).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

ii. Calculate the mean of target scores for the closest expressions.

These stages make a significant contribution to mitigating challenges associated with both low- and high-cardinality NFs, effectively addressing complexities posed by previously unseen expressions. Notably, our approach demonstrates universal applicability, proving effective for NFs with varying cardinalities, including those with thousands of unique values.

4 EXPERIMENTS

4.1 Experimental Setup

A combination of two datasets, MIMIC-IV and MIMIC-IV-ED, was employed in this study. Subsequent to this compound, various filters, detailed in the following subsection, were applied. Post-filtering, $440,285$ instances representing unique ED visits for the first task and $231,868$ instances corresponding to ED episodes involving reattendance within a 72-hour time frame were retained for further analysis.

We employed a diverse set of ML models, encompassing logistic regression (LR), gradient boosting (GB), random forest (RF), and an XGBoost (XGB), to make predictions for the two tasks under investigation. The selection of these four ML models was deliberate, with the aim of facilitating a comparative analysis with previous research conducted on the MIMIC-IV-ED dataset.

For assessing the model's performance, we employed 4 key metrics: Area Under the Receiver Operating Characteristic (AUROC) (Bradley, 1997), Area Under the Precision-Recall Curve (AUPRC), Sensitivity, and Specificity (Sofaer et al., 2019).

To ensure the relevance and quality of the dataset, a filtering method was applied to eliminate ED visits made by patients below 18 years of age and those lacking primary emergency triage category assignments. Additionally, the MIMIC-EXTRACT (Wang et al., 2020) was utilised for outlier detection. Each patient visit to the ED is denoted by a unique $subject_{id}$ linked to a corresponding $stay_{id}$. In cases where an ED visit is followed by an inpatient stay, the $stay_{id}$ can be associated with an inpatient admission identified as $hadm_{id}$ in the *edstays* table.

4.2 Baseline

This study leveraged a standardised reference framework denoted as the ”ED-MIMIC Benchmark” (Xie et al., 2022). The benchmark incorporates a comprehensive dataset, encompassing variables derived from the MIMIC-IV-ED (Johnson et al., 2021) and MIMIC-IV (Johnson et al., 2020). Previous work (Xie et al., 2022) utilised 64 numeric features, and our endeavor involved the reproduction of its results. In augmenting this benchmark, we introduced previously overlooked three NFs and implemented the conventional technique of one-hot encoding to address the integration of these new NFs. Thus the baseline was updated to ensure a consistent number of selected features. This updated baseline serves as a foundational reference point for our study, providing a standardised framework for comparison and evaluation.

It is noteworthy to mention that, in the process of updating the benchmark encompassing 67 features, various methods were explored, including the bitmap technique. Initially, an attempt was made to transform all extant features into a binary representation, leading to a training dataset characterised by $25,586$ columns. Despite utilising a high-performance computing system with 256 gigabytes of memory, the chosen configuration proved to be computationally inoperable. Consequently, in order to address the computational challenges, the decision was made to adopt one-hot encoding as the baseline technique to handle NFs. This approach not only facilitates the handling of high cardinality NFs but also remains computationally feasible for the entire dataset.

4.3 Feature Overview

In our dataset, we incorporate three distinct sets of selected NFs: race, arrival transport mode, and chief complaint. These features encompass both low- and high-cardinality characteristics. Specifically, we designate race and arrival transport modes as demographic features. Notably, in the training dataset, the arrival transport mode feature comprises only five distinct values, while the race feature exhibits greater diversity with 34 unique values, enriching the dataset's variability. The chief complaint, with $52,478$ distinct values, represents a high-cardinality NF integral to our study.

Table 1 provides a comprehensive overview of the features employed in the prediction tasks, encompassing demographic, medical characteristics, arrival transport mode, primary concerns (10 binary features), and history of patients' visits to the ED, ICU and hospitals. Due to the extensive number of distinct values in the chief complaint and the race variables, it is impractical to present all values in the table. In alignment with the features previously employed in the ED-MIMIC Benchmark, our ML models also incorporate an set of 35 binary features, serving as indicators for patients' comorbidities. By including these features, we aim to maintain consistency with the benchmark's established framework and leverage the same set of attributes to ensure comparability and coherence in our modeling approach.

Medical features, such as vital signs, may be recorded multiple times during a single ED visit. For the first prediction task, the ML models used the values from the first set of measurements. Conversely, for the prediction of reattendance to the ED within 72 hours, the most recent measurements are taken into consideration. Moreover, in the second task, three additional features become available which were not present at the time of triage, dedicated as 72-Hour Reattendance Predictors in Table 1. These features, namely length of stay at the ED, the number of medications, and the count of medication reconciliations, are specifically utilised for the second task.

5 RESULTS

Table 2 shown in the TE-PrepNet yields substantially better results in the realm of predicting patient hospitalisation within the ED at the time of triage. The outcomes distinctly reflect that the integration of nominal attributes, particularly those with high-cardinality through the utilisation of TE-PrepNet, substantially enhances the performance of all models. This table effectively underscores the pronounced effectiveness of our proposed methodology in contrast to prior investigations involving the MIMIC-IV-ED dataset, which inadvertently neglected these attributes, and even with updated baseline that use traditional method of one-hot encoding for handle NFs. The results show that RF model achieved notable performance in predicting hospitalisation during the triage process, boosting AUROC and AUPRC values to 0.8459 and 0.8148, respectively.

Table 3 presents the results of TE-PrepNet, considering three additional NFs, in comparison to the performance of the baseline for the prediction of reattendance at ED within 72 hours. As delineated earlier, this predictive task benefits from an extended feature set, culminating in a total of 70 input features. The XGBoost model exhibits superior predictive performance for reattendance within 72 hours among discharged patients, achieving an AUROC of 0.6975. This performance surpasses other models and even demonstrates an improvement compared to the baseline.

6 DISCUSSION

We exploited the ED-MIMIC-Benchmark pipeline to analyse the newly released MIMIC-IV-ED database, focusing on NFs often overlooked in ML tasks. Our findings demonstrate that the incorporation of both low and high-cardinality NFs substantially enhances the performance of ML models in predicting hospi-

Table 1: Basic characteristics of the dataset. Mean (SD) values are presented for the continuous variables; and count (%) is presented for the binary or categorical variables. ED LOS: Length of stay at ED in minutes.

Feature name	Overall	Discharge	Hospitalised	72-hour ED Reattendance
ED Visits	440285	231868	208417	15791
Age	52.8 (20.6)	46.3 (19.4)	60.0 (19.5)	50.5 (18.7)
Gender				
Female	239305 (54.4 %)	133573 (57.6%)	105732 (50.7%)	7386 (46.8%)
Male	200980 (45.6%)	98295 (42.4%)	102685 (49.3%)	8405 (53.2%)
Arrival Transport Mode				
Ambulance	158304 (36.0%)	52333 (22.6%)	105971 (50.8%)	5460 (34.6%)
Helicopter	560 (0.1%)	32 (0.0%)	528 (0.3%)	3 (0.0%)
Other	1351 (0.3%)	726 (0.3%)	625 (0.3%)	41 (0.3%)
Unknown	15180 (3.4%)	7861 (3.4%)	7319 (3.5%)	1008 (6.4%)
Walk	264890 (60.2%)	170916 (73.7%)	93974 (45.1%)	9279 (58.8%)
Triage Acuity				
Level 1	25249 (5.7%)	5338 (2.3%)	19911 (9.6%)	478 (3.0%)
Level 2	146837 (33.4%)	45332 (19.5%)	101505 (48.7%)	3947 (25.0%)
Level 3	236958 (53.8%)	151458 (65.3%)	85500 (41.0%)	10183 (64.5%)
Level 4	30074 (6.8%)	28624 (12.3%)	1450 (0.7%)	1124 (7.1%)
Level 5	1167 (0.3%)	1116 (0.5%)	51 (0.0%)	59 (0.4%)
Pain Scale	4.2 (3.6)	4.7 (3.6)	3.6 (3.5)	4.8 (3.8)
Vital Signs				
Temperature (Celsius)	36.7 (0.5)	36.7 (0.5)	36.7 (0.6)	36.7 (0.4)
Heart Rate (bpm)	85.0 (17.5)	83.9 (16.3)	86.3 (18.6)	79.9 (13.9)
Respiratory Rate (bpm)	17.6 (2.5)	17.3 (2.1)	17.9 (2.8)	17.0 (1.9)
Oxygen Saturation (%)	98.4 (2.4)	98.8 (2.0)	97.9 (2.7)	98.2 (2.9)
Systolic BP (mmHg)	134.8 (22.1)	135.1 (20.7)	134.5 (23.7)	128.8 (19.5)
Diastolic BP (mmHg)	77.5 (14.7)	78.8 (13.8)	76.0 (15.6)	76.0 (13.5)
Previous Visits				
30-Day ED Visits	0.2 (0.8)	0.2 (0.8)	0.3 (0.8)	1.1 (2.3)
90-Day ED Visits	0.5 (1.6)	0.5 (1.6)	0.6 (1.6)	2.3 (4.8)
360-Day ED Visits	1.4 (4.2)	1.2 (4.1)	1.6 (4.2)	6.0 (12.6)
30-Day Hospitalisation	0.2 (0.5)	0.1 (0.4)	0.2 (0.6)	0.6 (1.3)
90-Day Hospitalisation	0.4 (1.0)	0.2 (0.8)	0.5 (1.2)	1.2 (2.7)
360-Day Hospitalisation	1.0 (2.7)	0.6 (2.2)	1.4 (3.1)	3.3 (7.6)
30-Day ICU Stays	0.0 (0.2)	0.0 (0.1)	0.0 (0.2)	0.0 (0.2)
90-Day ICU Stays	0.0 (0.3)	0.0 (0.2)	0.1 (0.3)	0.1 (0.3)
360-Day ICU Stays	0.1 (0.5)	0.0 (0.3)	0.2 (0.6)	0.2 (0.6)
Chest Pain				
False	409599 (93.0)	218112 (94.1)	191487 (91.9)	14842 (94.0)
True	30686 (7.0)	13756 (5.9)	16930 (8.1)	949 (6.0)
Abdominal Pain				
False	389515 (88.5)	206134 (88.9)	183381 (88.0)	13746 (87.0)
True	50770 (11.5)	25734 (11.1)	25036 (12.0)	2045 (13.0)
Headache				
False	423730 (96.2)	219938 (94.9)	203792 (97.8)	15130 (95.8)
True	16555 (3.8)	11930 (5.1)	4625 (2.2)	661 (4.2)
Shortness of Breath				
False	439002 (99.7)	231468 (99.8)	207534 (99.6)	15765 (99.8)
True	1283 (0.3)	400 (0.2)	883 (0.4)	26 (0.2)
Back Pain				
False	422691 (96.0)	219524 (94.7)	203167 (97.5)	15144 (95.9)
True	17594 (4.0)	12344 (5.3)	5250 (2.5)	647 (4.1)
Cough				
False	431030 (97.9)	226582 (97.7)	204448 (98.1)	15533 (98.4)
True	9255 (2.1)	5286 (2.3)	3969 (1.9)	258 (1.6)

Table 1: Basic characteristics of the dataset. Mean (SD) values are presented for the continuous variables; and count (%) is presented for the binary or categorical variables. ED LOS: Length of stay at ED in minutes (cont.).

Feature Name	Overall	Discharge	Hospitalised	72-hour ED Reattendance
Nausea Vomiting				
False	429639 (97.6)	226273 (97.6)	203366 (97.6) d	15379 (97.4)
True	10646 (2.4)	5595 (2.4)	5051 (2.4)	412 (2.6)
Fever Chills				
False	425051 (96.5)	227233 (98.0)	197818 (94.9)	15377 (97.4)
True	15234 (3.5)	4635 (2.0)	10599 (5.1)	414 (2.6)
Syncope				
False	432098 (98.1)	227467 (98.1)	204631 (98.2)	15615 (98.9)
True	8187 (1.9)	4401 (1.9)	3786 (1.8)	176 (1.1)
Dizziness				
False	429377 (97.5)	225542 (97.3)	203835 (97.8)	15487 (98.1)
True	10908 (2.5)	6326 (2.7)	4582 (2.2)	304 (1.9)
72-Hour Reattendance Predictors				
ED LOS (minutes)	385.4 (264.2)	-	-	407.6 (283.3)
# Medication	2.9 (3.3)	-	-	2.7 (3.2)
# Medication Reconcilation	6.1 (6.8)	-	-	5.2 (6.6)

Table 2: Performance comparison of TE-PrepNet approach and baseline across different ML models for hospitalisation prediction at triage in the ED. Sens: Sensitivity. Spec: Specificity.

Baseline (67 Features)				
ML Model	AUROC	AUPRC	Sens.	Spec.
LR	0.7985	0.7751	0.7113	0.7274
RF	0.7520	0.7168	0.6908	0.7242
GB	0.7232	0.6755	0.6279	0.7147
XG Boost	0.7577	0.7183	0.6677	0.7162

TE-PrepNet (67 Features)				
ML Model	AUROC	AUPRC	Sens.	Spec.
LR	0.8353	0.8030	0.7729	0.7440
RF	0.8458	0.8149	0.7736	0.7630
GB	0.8383	0.8103	0.7670	0.7560
XG Boost	0.8458	0.8142	0.7761	0.7624

Table 3: Performance comparison of TE-PrepNet and baseline across different ML models for predicting ED reattendance within 72 hours post triage. Sens: Sensitivity. Spec: Specificity.

Baseline (70 Features)				
ML Model	AUROC	AUPRC	Sens.	Spec.
LR	0.6267	0.0869	0.5816	0.5960
RF	0.5687	0.0679	0.5038	0.6026
GB	0.5730	0.0757	0.3822	0.7222
XGBoost	0.6166	0.0886	0.5812	0.5756

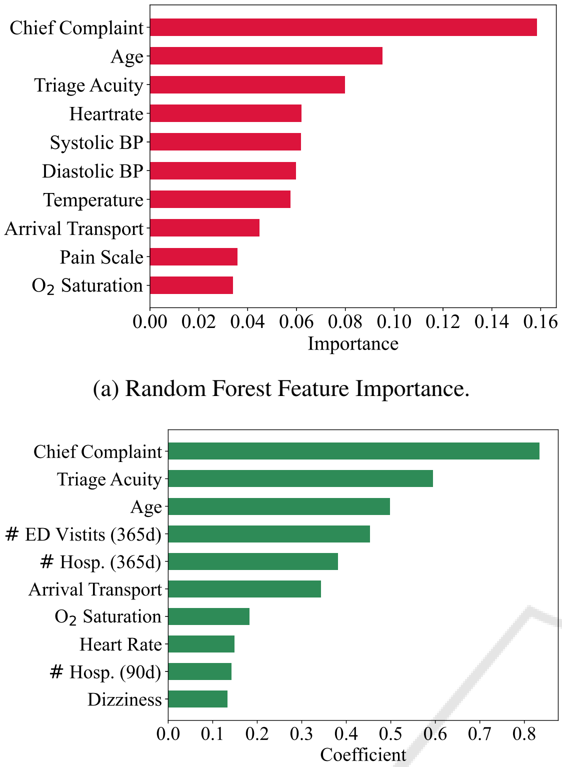
TE-PrepNet (70 Features)				
ML Model	AUROC	AUPRC	Sens.	Spec.
LR	0.6888	0.1404	0.6137	0.6611
RF	0.6833	0.1441	0.6237	0.6447
GB	0.6905	0.1474	0.5997	0.6837
XGBoost	0.6975	0.1382	0.6224	0.6729

tal admissions and reattendance at the ED within 72 hours following discharge. We successfully addressed the challenge of handling these features with thousands of unique values, all without relying on clinical expertise to manually select important values for ML tasks.

The establishment of a baseline through one-hot encoding and the unsuccessful attempt using bitmap highlight the inadequacy of solely adding features for improved performance. This is especially notable for NFs with numerous unique values. TE-PrepNet excels in addressing these challenges, efficiently managing high-cardinality NFs to optimise performance and mitigate the impact of unseen values in the test set.

In high-cardinality NFs, such as the chief complaint, it is greatly possible that during the testing phase, numerous values emerge that did not appear in the training process. For instance, in both the first and second prediction tasks, the chief complaint exhibits $9,125$ and $5,092$ previously unseen values in the respective test sets. Methods like one-hot encoding tend to ignore these unseen values during prediction. This situation implies that a substantial amount of potentially valuable information may be overlooked. Addressing the handling of high-cardinality NFs during both the training and testing phases is crucial for improving the model's robustness and effectiveness in capturing diverse and previously unseen data. Our TE-PrepNet approach efficiently addresses this, maximising the use of all available information.

Variable Importance. Ten most predictive variables for both prediction tasks (based on RF feature importance and LR coefficients) are shown in Figures 2 and 3, respectively. As depicted in Figure 2 for the



(a) Random Forest Feature Importance.

(b) Logistic Regression Coefficients.

Figure 2: Ten top important variables in the hospitalisation prediction task based on random forest variable importance and logistic regression coefficients values. # ED visits (365d): Number of ED visits within last year. # Hosp. (365d): Number of hospitalisation within last year. # Hosp. (90d): Number of hospitalisation within last 90 days.

prediction of hospitalisation at the time of the triage, NFs — particularly chief complaint, encompassing a substantial number of unique values, 52,478 — consistently rank at the top predictive variables. Importantly, these feature surpass the importance of triage acuity and age. Moreover, the arrival transport mode stands out as one of the ten most predictive variables, as indicated by both RF feature importance and LR coefficients. This signifies the great contribution of the NFs to the predictive outcome. In Figure 2b it is evident that the historical data pertaining to a patient's ED visits over the past year, along with the frequency of hospitalisations within the last three months and the last year, significantly contribute to the hospitalisation outcomes.

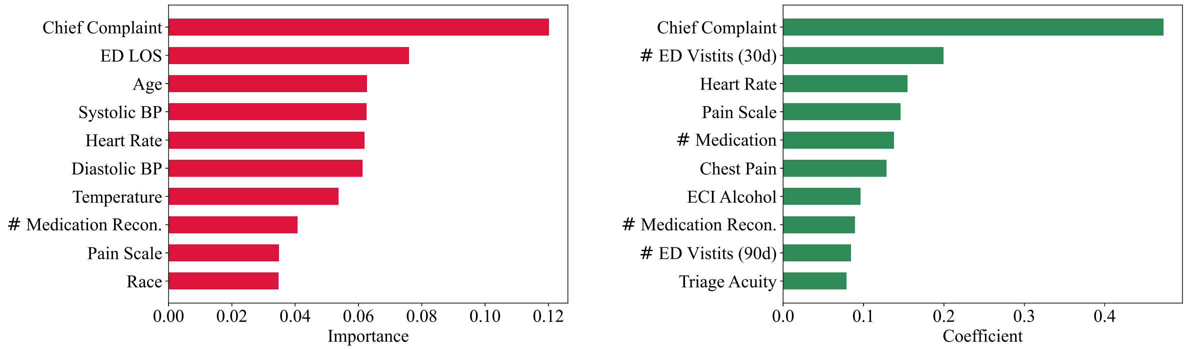
As previously mentioned, for the second prediction task, emphasis is placed on utilising the most recent values of all features for predictive analysis. Figure 3 shows that for the prediction of ED reattendance within 72 hours, chief complaint again is the top most important feature. The utilisation of this feature became possible due to the implementation of the TE-PrepNet, which allowed us to take advantage of their entire valuable information. It is worth mentioning that conventional methods encounter significant challenges when dealing with these features. These features are incorporated into ML models, with no dependence on clinician input, while preserving the original dimensionality of the dataset. In Figure 3a, the length of stay at the ED (ED LOS) stands out as the second most influential predictive variable, as indicated by the RF feature importance. ED LOS is exclusively applicable to the second prediction task.

Imbalanced Dataset. In the context of the second prediction task, there were a total of 15,791 distinct instances of reattendance at the ED within 72 hours. This figure constitutes approximately 3.5% of the total episode count for the cohort. The dataset exhibited a significant class imbalance, with instances of reattendance being notably underrepresented. This imbalance substantially contributed to a low AUPRC for the task. Despite an enhancement in AUPRC compared to the benchmark, the metric retained a relatively modest value.

7 CONCLUSION

In this study, we highlight the efficacy of our target encoding preprocessing framework (TE-PrepNet) in effectively managing nominal features inherent to electronic health records in the context of two emergency department based prediction tasks. While low-cardinality nominal features can be managed using various techniques, the handling of high-cardinality nominal features, those encompassing thousands of unique values, presents distinct challenges. These challenges directly arise from two main issues: the problem of high dimensionality, which becomes particularly problematic for particularly for computationally expensive models, and the reliance on domain-specific knowledge. The latter often results in either overlooking valuable features or considering only a subset of their potential values.

Our results reveal promising outcomes, with the best predictive model for hospitalisation, random forest, achieving an AUROC of 0.8458. This performance notably surpasses the baseline model that used a conventional technique to handle these features and achieved an AUROC of 0.7520 for this specific prediction task. These findings highlight the contribution of the nominal features in enhancing the predictive accuracy of hospitalisation. In the prediction task for reattendance at the emergency department within 72



(a) Random Forest Feature Importance.

(b) Logistic Regression Coefficients.

Figure 3: Ten top important variables for the prediction of ED reattendance within 72-hours. # Medication Recon. : Counts of Medication Reconciliation. # ED Visits (30d): Number of ED visits within last month. ED LOS: Length of stay at ED. # ED Visits (90d): Number of ED visits within last 90 days.

hours, it is noteworthy that the XGBoost emerged as the top-performing model, achieving a AUROC score of 0.6975. This represents an enhancement in predictive performance compared to the baseline, where an AUROC score of 0.6166 was reported for this model.

Future Work. In the next stage, we aim to tailor target encoding techniques specifically for ED prediction. This customisation will involve optimising encoding methods to effectively capture unique patterns in emergency data. We plan to integrate this tailored target encoding seamlessly with our proposed preprocessing steps, streamlining the data preparation process and enhancing predictive accuracy.

REFERENCES

Al Shalabi, L., Shaaban, Z., and Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, 2(9):735–739.

Barak-Corren, Y., Israelit, S. H., and Reis, B. Y. (2017). Progressive prediction of hospitalisation in the emergency department: uncovering hidden patterns to improve patient flow. *Emergency Medicine Journal*, 34(5):308–314.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Brownell, J., Wang, J., Smith, A., Stephens, C., and Hsia, R. Y. (2014). Trends in emergency department visits for ambulatory care sensitive conditions by elderly nursing home residents, 2001 to 2010. *JAMA internal medicine*, 174(1):156–158.

Etu, E.-E., Monplaisir, L., Arslanturk, S., Masoud, S., Aguwa, C., Markevych, I., and Miller, J. (2022). Prediction of length of stay in the emergency department for covid-19 patients: A machine learning approach. *IEEE Access*, 10:42243–42251.

Graham, B., Bond, R., Quinn, M., and Mulvenna, M. (2018). Using data mining to predict hospital admissions from the emergency department. *IEEE Access*, 6:10458–10469.

Gurazada, S. G., Gao, S., Burstein, F., and Buntine, P. (2022). Predicting patient length of stay in australian emergency departments using data mining. *Sensors*, 22(13):4968.

Guttmann, A., Schull, M. J., Vermeulen, M. J., and Stukel, T. A. (2011). Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from ontario, canada. *Bmj*, 342.

Hancock, J. T. and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(1):1–41.

Häyrinen, K., Saranto, K., and Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Hong, W. S., Haimovich, A. D., and Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7):e0201016.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng IV, S. (2021). Mimic-iv-ed. *PhysioNet*.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2020). Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*.

Kusumawati, H. I., Magarey, J., and Rasmussen, P. (2019). Analysis of factors influencing length of stay in the emergency department in public hospital, yogyakarta, indonesia. *Australasian emergency care*, 22(3):174–179.

Liu, C., Tan, Z., and He, M. (2022). Overview of artificial intelligence in medicine. In *Artificial Intelligence in Medicine: Applications, Limitations and Future Directions*, pages 23–34. Springer.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1):27–32.

Mollart, L., Newell, R., Geale, S. K., Noble, D., Norton, C., and O'Brien, A. P. (2020). Introduction of patient electronic medical records (emr) into undergraduate nursing education: an integrated literature review. *Nurse Education Today*, 94:104517.

Nazyrova, N., Chaussalet, T. J., and Chahed, S. (2022). Machine learning models for predicting 30-day readmission of elderly patients using custom target encoding approach. In *Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part III*, pages 122–136. Springer.

Pargent, F., Pfisterer, F., Thomas, J., and Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, 37(5):2671–2692.

Rahman, M. A., Honan, B., Glanville, T., Hough, P., and Walker, K. (2020). Using data mining to predict emergency department length of stay greater than 4 hours: Derivation and single-site validation of a decision tree algorithm. *Emergency Medicine Australasia*, 32(3):416–421.

Sofaer, H. R., Hoeting, J. A., and Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10(4):565–577.

Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., and Naumann, T. (2020). Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235.

Wardi, G., Carlile, M., Holder, A., Shashikumar, S., Hayden, S. R., and Nemati, S. (2021). Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Annals of emergency medicine*, 77(4):395–406.

Xie, F., Zhou, J., Lee, J. W., Tan, M., Li, S., Rajnthern, L. S., Chee, M. L., Chakraborty, B., Wong, A.-K. I., Dagan, A., et al. (2022). Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658.

Zahrotun, L. (2016). Comparison jaccard similarity, cosine similarity and combined both of the data clustering with shared nearest neighbor method. *Computer Engineering and Applications Journal*, 5(1):11.