International Journal of Information Management Data Insights 2 (2022) 100070

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Empirical evaluation of performance degradation of machine learning-based predictive models – A case study in healthcare information systems

Zachary Young [a], Robert Steele [b,*]

[a] Computer Science Lab, Capitol Technology University, 11301 Springfield Rd., Laurel, MD 20708, United States
[b] Department of Computer Science, Capitol Technology University, 11301 Springfield Rd., Laurel, MD 20708, United States

ARTICLE INFO

Keywords:
Health predictive model
Performance degradation
Machine learning
Health information systems

ABSTRACT

While there have been a very large number of academic studies of proposed machine learning-based health predictive models, it is widely recognized that machine learning-based models in all domains typically degrade in performance over time, post training. This known characteristic of machine learning-based models could present significant risks in the healthcare setting to patient quality of care or safety. Nevertheless, there has been little study of the performance degradation of such models on real-world data. In this article, we empirically measure performance degradation of predictive models that predict at time of admission, emergency patient mortality, drawing upon a large dataset of over 1.83 million patient discharge records. We demonstrate important empirical results including both relatively slow performance degradation over two and a half years, but also significant differences in the rate and extent of performance degradation between different machine learning model types and time period of the training set.

1. Introduction

A number of information technology and legislative developments, notably including the Health Information Technology for Economic and Clinical Health (HITECH) Act, part of the American Recovery and Reinvestment Act of 2009, spurred over the last decade, much wider adoption of electronic health records (EHR) and digitized health data by health providers in the US (Adler-Milstein & Jha, 2017). Digitized health data provided the pre-requisite step for associated health information systems including Patient Administration Systems (PAS), Computerized Provider Order Entry (CPOE) systems, and Radiology Information Systems (RIS) and thereby subsequent health information systems for the large-scale aggregation of patient health data such as Clinical Data Warehouses (CDW) amongst others. Such health data aggregation systems have then provided the foundation for the application of such technologies as machine learning (ML) and artificial intelligence (AI) to healthcare prediction and classification problems.

Machine learning (ML)-based predictive models appear to have many intrinsic advantages over traditional methods for healthcare related prediction, which we desribe further below, and can provide additional decision support to physicians. ML-based predictive models, are trained using ML algorithms on historical datasets, are relatively efficient to produce and evaluate, can provide models based on very recently available data and can be developed so as to be broadly applicable or also alternatively specific to a given hospital or other custom scenario.

Nevertheless, despite the promise of ML-based healthcare predictive models, an important outstanding area where there is a significant lack of research and understanding is the occurrence of and the rate of predictive model performance degradation post-training, in real-world settings as:

- such decline has important implications for model efficacy and patient quality of care and indeed patient safety,
- it has implications for if and when model retraining may be needed and what data should be used for potential retraining, and
- as identified in the recent 2021 systematic review by Stanford University researchers of clinical ML-based model performance maintenance "[t]here was limited research in preserving the performance of machine learning models in the presence of temporal dataset shift in clinical medicine" (Guo et al., 2021)

This current article, to further investigate ML-based healthcare model performance degradation, is one of the first to undertake to empirically evaluate the extent and characteristics of such health predictive model performance degradation on a large real-world dataset and clinical predictive problem. The specific predictive problem considered is the prediction of the in-hospital mortality of emergency patients, pre-

* Corresponding author.
E-mail addresses: zeyoung@captechu.edu (Z. Young), rjsteele@captechu.edu (R. Steele).

https://doi.org/10.1016/j.jjimei.2022.100070
Received 2 September 2021; Received in revised form 12 March 2022; Accepted 12 March 2022

Table 1
Number of admissions, survival and died count from 2016, 2017 and 2018 MD SID datasets.

	All Admissions	Emergency Admissions	Survivals	Deaths
2016 H1	312,314	170,861	165,910	4832
2016 H2	310,501	176,144	171,029	5017
2017 H1	308,402	180,069	174,753	5191
2017 H2	304,677	172,935	167,979	4858
2018 H1	301,476	175,808	170,485	5185
2018 H2	297,275	168,512	163,486	4899
Total	1834,645	1044,329	1013,642	29,982

dicted at the time of hospital admission. It has also been suggested that accurate mortality prediction by the doctor alone may not be sufficient (Yeun, Levine, Mantadilok & Kaysen, 2000). The information systems community is well-placed to make important contributions to the study of the performance and effectiveness of machine learning models when deployed into real-world settings and business and industry environments (Won, Kim & Ahn, 2018) (Min et al., 2019) (Kushwaha, Kar & Dwivedi, 2021) (Konovalenko & Ludwig, 2021) including contributions to model explainability (Meske, Bunde, Schneider & Gersch, 2021) (Sharma, Kumar & Chuah, 2021).

Being able to predict emergency patient mortality at the time of admission is an important factor in quality of care decision making. Traditionally such predictions upon admission have been made on a more ad-hoc basis based on the clinical expertise and experience of individual doctors. In recent decades, severity of illness scoring systems (Varghese, Kalaiselvan, Renuka & Arunkumar, 2017) for patients once already in an ICU setting but post their admission, have been developed, with such emergency patient scoring systems including Apache III (Knaus et al., 1991), SAPS II (Le Gall, Lemeshow & Saulnier, 1993), SOFA (Minne, Abu-Hanna & de Jonge, 2008) and Apache IV (Varghese, Kalaiselvan, Renuka & Arunkumar, 2017) scores. An advantage of such severity score systems over purely physician-based prediction is that they can provide more quantitative and consistent approaches to severity and mortality prediction.

However, ML can provide a potentially significantly more extensible, flexible and sophisticated approach to predicting patient outcomes. As an example when compared with severity scoring systems: while Apache II utilizes physiology, age and chronic conditions amongst a total of 12 physiological variables and Apache III utilizes a total of 17 physiological variables (Chatzicostas et al., 2002), neither use exogenous or non-hospital attributes. On the other hand ML-based models can utilize a potentially far greater number of input attributes, including all available at time of hospital admission, thereby allowing the building of models for earlier point in time prediction, prior to the collection of any significant in-hospital clinical data. Additionally, the relative ease of development and evaluation of ML-based models supports both the development of very widely applicable/generalizable predictive models and also the development of a larger number of domain-specific, custom predictive models.

To conduct this empirical performance degradation study of ML-based healthcare models we have made use of a number of large patient discharge data sets made available from the Agency for Healthcare Research and Quality's (AHRQ) Healthcare Cost & Utilization Project (HCUP) (Agency for Healthcare Research & Quality 2022). These include the Maryland 2016, 2017 and 2018 State Inpatient Database (SID) datasets (HCUP 2022), comprising in aggregate a total of over 1.83 million patient discharge records (see Table 1). These contain the majority of all patient admissions/discharges in the US state of Maryland for the years of 2016 through 2018. To evaluate performance degradation we have first trained and evaluated using 10-fold cross validation (CV) a significant number of candidate predictive models on the emergency admissions data from the first calendar half of 2016 (2016 H1). After selecting four of the best performing in terms of a widely used metric of evaluation, Area Under the Receiver Operator Characteristic Curve (AUC) we then evaluated their performance maintenance on data from subsequent calendar half years drawn from 2016, 2017 and 2018. We also evaluated the effect of choosing different calendar halves for the training dataset. The results show, at least for the best performing models, high initial discriminative performance and relatively strong performance maintenance over the subsequent years. The results also show interesting variation in the extent of performance decline between different model types and choice of period of training set and provide new insights into the nature of ML-based health predictive model behavior. We discuss those implications for ML-based model performance and degradation in the Discussion section of this article.

The remainder of the article is structured as follows Section 2. provides a detailed review of related works Section 3. describes in detail the methodology used for the study Section 4. provides the results of the study and Section 5 discusses these results and their place in the literature and implications for practice. This is followed by the conclusion of the article.

2. Background and related work

Prediction of mortality of emergency patients at the time of admission provides an important input to allow safe and improved quality of care. Machine learning has been used in numerous other healthcare prediction applications including in relation to emergency patient outcomes. However there has been little study of model performance degradation (Guo et al., 2021).

A wide range of the existing studies of ML-based inpatient mortality prediction relate to specific diagnoses or health conditions for example sepsis (Perng et al., 2019), traumatic brain injury (Rau et al., 2018), acute cardiovascular cases (Metsker, Sikorsky, Yakovlev & Kovalchuk, 2018) or trauma from motorcycle accidents (Kuo et al., 2018) amongst numerous other condition specific models. There are also other studies that have considered mortality prediction over a longer time span post-discharge, such as 30-day mortality (Blom et al., 2019) and 1 year mortality for acute coronary syndrome patients (Sherazi et al., 2020).

Studies that have considered all-condition emergency patient in-hospital mortality prediction include such studies as (Steele & Hillsgrove, 2019, April) that predicts mortality at time of admission, (Klug et al., 2020) that predicts short term (10 day post admission) mortality using gradient boosting machine learning using information available at time of triage, (Zhai et al., 2020) that considered such models as SVM and XGBoost on a dataset of just 1624 cases, (Tang, Xiao, Wang & Zhou, 2018) that considered recurrent neural networks to factor in physiological temporal patterns post-admission, (Meiring et al., 2018) using data collected from within the ICU during the first and second day since admission and (Hillsgrove & Steele, 2019, March) demonstrating the use of Random Forest on a large admissions dataset. It is to be noted that studies that have sought to compare the performance of ML-based models with traditional severity scoring systems have generally reported better performance in terms of such metrics as AUC achieved with the ML-models versus the severity scoring systems (Kim, Kim & Park, 2011, van Doorn et al., 2021).

Nevertheless, existing studies have not considered the performance maintenance or degradation of ML-based emergency patient mortality predictive models or indeed performance degradation of health predictive models in general. The first and only previous work we are aware of to consider performance degradation of mortality predictive models are (Young & Steele, 2021) in relation to emergency patients and (Brettle & Steele, 2021) (Brettle & Steele, 2021) in relation to elective patients. Other works on health predictive model maintenance deal with quite separate problems such as performance maintenance of Medicare fraud detection models (Leevy, Khoshgoftaar, Bauder & Seliya, 2020).

There are emerging efforts to analytically consider ML-based model determination in terms of underlying dataset concept drift. While approaches such as concept drift detection (Barros & Santos, 2018) are

potentially relevant, they in themselves do not quantify the future performance of a given predictive model. As per (Barros & Santos, 2018) there is a wide range of concept drift detection algorithms, they perform differently for different datasets and do not themselves indicate how model performance will change in the future as per (Mauri & Damiani, 2021). An important goal for those engaged in information management is decision support (Rawat, Rawat, Kumar & Sabitha, 2021) and the analytic approaches developed to-date do not address many of the questions about when and at what rate ML-based models will degrade and support decisions around use and re-training.

It should be noted that general severity scoring systems are developed up to decades earlier than at the time their performance is being considered and evaluated, developed using significant time and effort by medical and clinical researchers. This current study makes a contribution to considering the time-based degradation of ML-based predictive models and hence also raises an important fundamental difference between how severity scoring systems and ML-based predictive models are evaluated. The ML models are being evaluated on data two to three years after training, not decades after development. This current article provides empirical findings based on large real-world datasets on characteristics of health predictive model performance maintenance.

3. Methodology

3.1. Methods

Broadly the study makes use of a data mining methodology, involving data preparation, model development and evaluation and model deployment on datasets in the future of the training data. First an appropriate, large dataset is obtained from the MD SID HCUP datasets (see Section 3.2) and specific selected attributes chosen so that only those that would be available at time of patient admission are included, and this dataset is then split into various chronological subsets used for training and testing of a range of candidate ML models.

By training an ML model on data from an earlier point in time, for example the first half of 2016, and then evaluating how accurate the predictions of such a trained model are when considered against chronologically later test sets, for example each subsequent half year, we can evaluate the characteristics of performance maintenance or degradation of such a model based on how it would have performed if trained from the first half of 2016 and then used in deployment for the next two and a half years.

We also then explore what impact on future performance results from training on later or more recent data, for example 2016 H2, 2017 H1 or each later half year. This allows us to consider the impact of frequently re-training a given model, for example each half year.

3.2. Data collection

The data for this study is derived from the HCUP Maryland (MD) SID datasets for years 2016 (AHRQ 2016), 2017 (AHRQ 2017) and 2018 (AHRQ 2018) in aggregate totalling more than 1.83 million admissions. The data is limited to only emergency admission patients (ATYPE = 1), limiting the data of interest to approximately 1.04 million admissions (see Table 1), and the “DIED” attribute is to be used as the target attribute. As such, all entries with a missing “DIED” attribute, a miniscule number, were removed from the datasets. The DIED attribute set to the affirmative refers to an in-hospital death of the admitted patient, and not death at some defined period post-discharge.

All three yearly datasets were split into half-year subsets Table 1. shows the numbers of patient discharge records for each half year, both total discharges and emergency patient discharges, and the respective numbers of surviving patients and in-hospital deaths amongst the emergency patients.

Fig. 1 shows the percentage of emergency patients, per each half year, who died. The ClassBalancer filter available via the widely used ML toolkit utilized for this study (Hall et al., 2009), is used to re-weight the instances in each half year dataset, to provide an equivalent effect of re-sampling with replacement to achieve balanced datasets.

Attributes are selected so as to be known at the time of admission and to be present in all three datasets. For this reason, only a limited number of attributes are included in the ML models (Table 2) and many other possible clinical attributes that would become available during the hospital stay are not included. The predictive problem we are addressing is mortality prediction at time of admission, not prediction at some point in time post-admission.

All attributes are converted so that their type is either numeric or nominal data type, depending on the appropriate semantics of the attribute. The set of such model attributes all known at admission, with their description is shown in Table 2. For more detailed descriptions of each attribute and allowable values see (HCUP Central Distributor SID Availability of Data Elements 2018).

3.3. Model development and performance degradation evaluation method

All candidate ML models were trained and evaluated using 10-fold evaluation using version 3.8.4 of a widely used Java-based open-source ML toolkit (AHRQ 2018).

The candidate models are initially trained and evaluated on the 2016 H1 dataset of emergency patient discharge records. The model types used in the 2016 H1-based model development were initially the range of base models present in (Young & Steele, 2021) and for each their default model parameters were utilized. Models with an AUC of at least 0.65 were then re-evaluated with the AdaBoostM1 (Freund & Schapire, 1996, July) and Bagging (Breiman, 1996) meta classifiers to consider if a higher performing meta-model version of that model type was readily obtainable.

The top four performing models (importantly not including more than one of the same base model type), in terms of Area Under the Receiver Operating Characteristic curve, from this initial 2016 H1 training and evaluation, were then selected for more detailed performance degradation evaluation. This means within the top four chosen models, we did not include more than one of the same base model type, so as to best allow exploration of possible differences in model degradation between models of different base types. The model degradation evaluation involved training of each of the top four models on each of the subsequent half year datasets and evaluation of the trained model on all of the half years subsequent in time, onwards from that half year used for training.

For example, when a model was trained on 2016 H1 its future performance was evaluated on each of subsequent half year datasets, which are: 2016 H2, 2017 H1, 2017 H2, 2018 H1 and 2018 H2. When the model was trained on 2016 H2, that model was evaluated for its subsequent half year datasets: 2017 H1, 2017 H2, 2018 H1 and 2018 H2. And so on, including training on 2018 H1 and performance evaluation on 2018 H2 alone, and finally training on 2018 H2 alone (the final half year in the three years of data used for this study).

In both the initial training and 10-fold CV evaluation and all subsequent half year evaluations, four performance metrics were captured: the Area Under the Receiver Operating Characteristic curve (AUC) which is considered to provide a threshold-independent measure of model discriminative performance (Hanle & McNeil, 1982); accuracy, or the percentage of correctly predicted mortality outcomes; weighted precision, or the weighted average of the fraction of instances which are predicted to be of a particular outcome (e.g. died or survived) which actually did have that outcome; and weighted recall, or the weighted average of the fraction of all of the instances of a particular outcome (e.g. died or survived) that were correctly predicted to have that outcome.

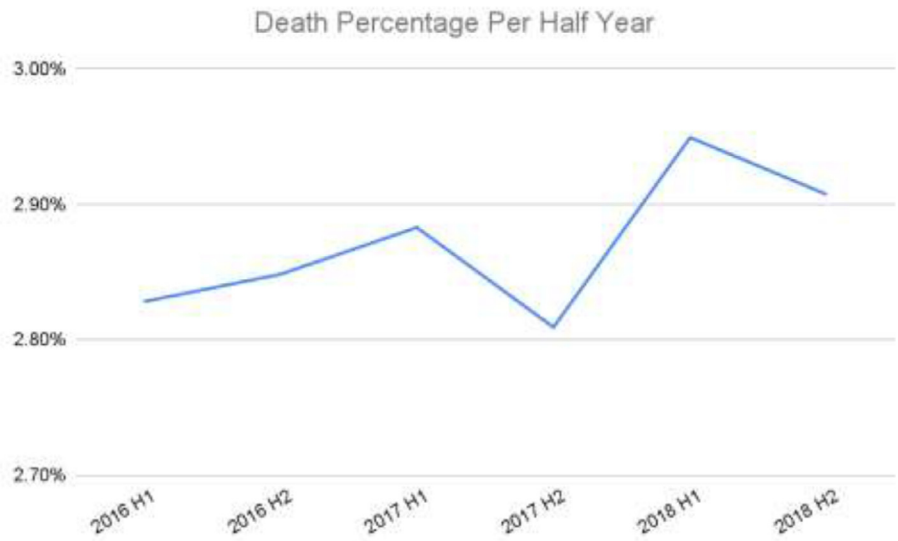

Fig. 1. Percentage of emergency patient admissions died per half year.

Table 2
Model Attributes (see (HCUP Central Distributor SID Availability of Data Elements 2018) for more detailed attribute descriptions).

Attribute #	Attribute Name	Description	Data Type
1	Age:	Age in years at admission	Scale
2	AType:	Admission Type	Nominal
3	AWeekend:	Admission day is on a weekend	Binary
4	DQTR:	Discharge Quarter	Scale
5	Female:	Indicator of sex	Binary
6	Hispanic:	Indicates hispanic ethnicity	Nominal
7	Homeless:	Indicator of Homelessness	Binary
8	HOSPST:	Hospital State Postal Code	Nominal
9	I10_DX1:	First listed diagnosis from I10_DXn. Diagnosis based on the ICD-10-CM coding	Nominal
10	I10_PR1:	First listed procedure from I10_PRn. Procedures based on the ICD-10-PCS coding system.	Nominal
11	MaritalStatus_X:	Patient's marital status, as received by the source	Nominal
12	MDNUM1_R:	Anonymized physician identifier	Nominal
13	MEDINCSTQ:	Quartile classification for patients income across state median	Scale
14	Pay1:	Expected Primary Payer, uniform	Nominal
15	Pay2:	Expected Secondary Payer, uniform	Nominal
16	PL_CBSA:	Metropolitan, Micropolitan, and Outside Core-Based Statistical areas (CBSA)	Nominal
17	PL_RUCC:	Rural-urban Continuum codes based on sub-county (1/10 county) population size	Nominal
18	PL_UIC:	Relationship of sub-county(1/9 county) to major metropolitan areas	Nominal
19	PL_UR_CAT4:	four-category urban-rural identifier.	Nominal
20	PSTATE:	Two-character state code	Nominal
21	PSTCO:	5-number county code	Nominal
22	RACE:	Race	Nominal
23	READMIT:	If this admission is a readmission	Binary
24	YEAR:	Calendar Year	Scale
25	ZIP3:	First three digits of Patient's zip code	Nominal
26	DIED:	Died during hospitalization	Binary

4. Results

The results from the study are summarized in the following figures Section 4.1. addresses initial model performance results including identification of the top four models selected, Section 4.2 provides the results of comparative model degradation over time of the models trained on the 2016 H1 dataset and Section 4.3 provides model performance degradation results related to the time period of the dataset used for model training.

4.1. Model 10-fold cross validation performance

Figs. 2 through 5 present the candidate model performances when trained upon 2016 H1 and evaluated using 10-fold CV for the metrics of AUC, accuracy, precision and recall respectively. The top four performing models in terms of AUC (see Fig. 2) where each is chosen to be of a *different base model*, were: Bagged Bayesian Network, Boosted Decision Table, Bagged Naive Bayes and Boosted Decision Stump.

Fig. 3 presents the accuracy of each evaluated candidate model. The models were evaluated using 10-fold CV on the 2016 H1 dataset and the accuracy is the weighted accuracy between the two labels of the target class.

Figs. 4 and 5 provide respectively the precision and recall of the candidate models, evaluated via 10-fold CV on the 2016 H1 dataset. Each of precision and recall are weighted precision and recall respectively.

4.2. Comparative model performance degradation

The top four selected models after being trained on 2016 H1, are compared against each other in terms of performance degradation from 2016 H1 through 2018 H2, based on AUC, accuracy, precision and recall, in Figs. 6 through 9 respectively.

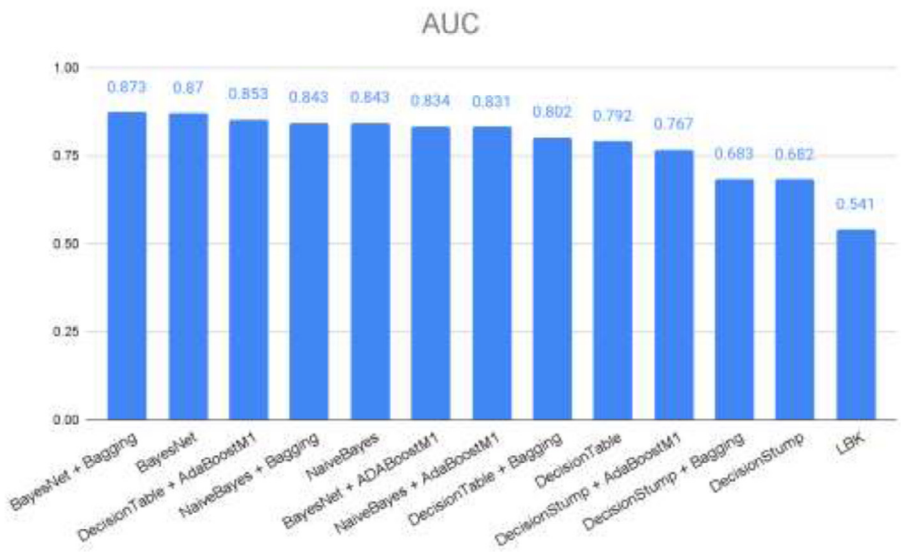

Fig. 2. Top performing models in terms of AUC based on 10-fold CV on 2016 H1.

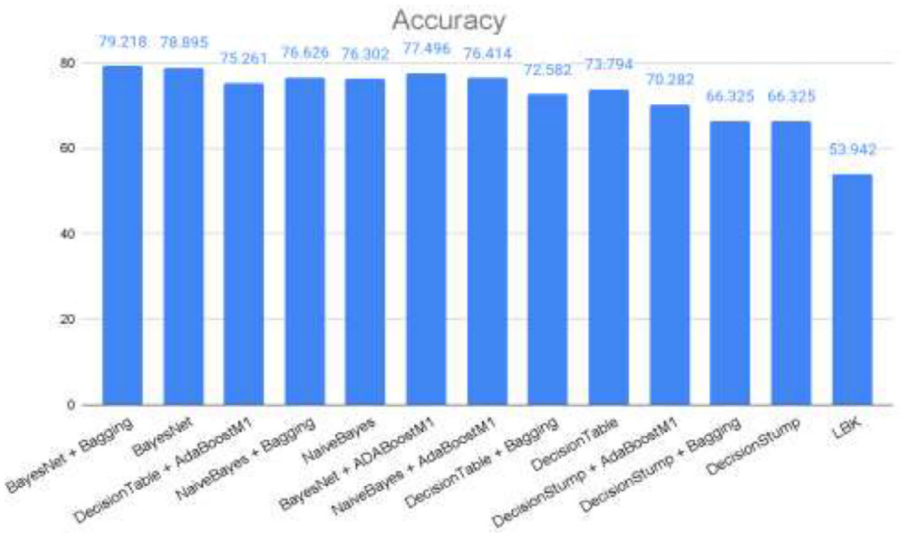

Fig. 3. Top performing models in terms of accuracy based on 10-fold CV on 2016 H1.

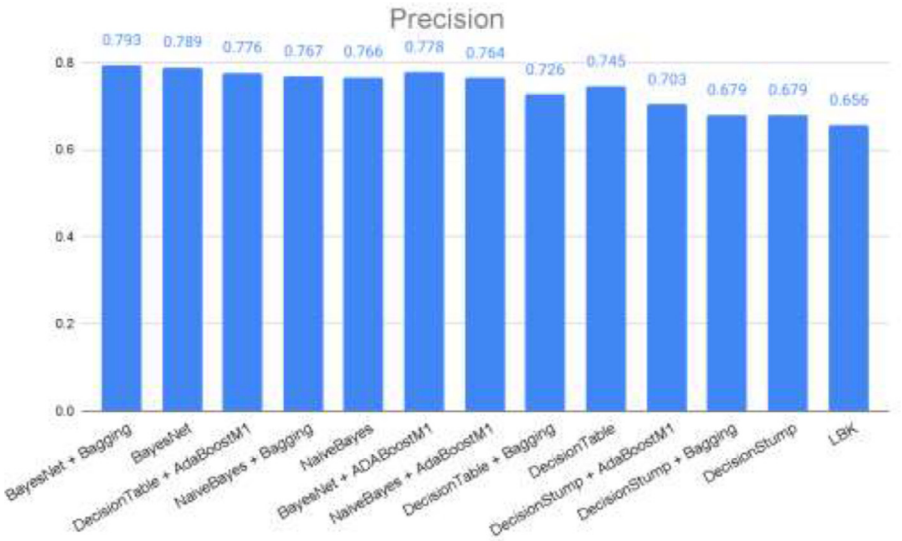

Fig. 4. Top performing models in terms of precision based on 10-fold CV on 2016 H1.

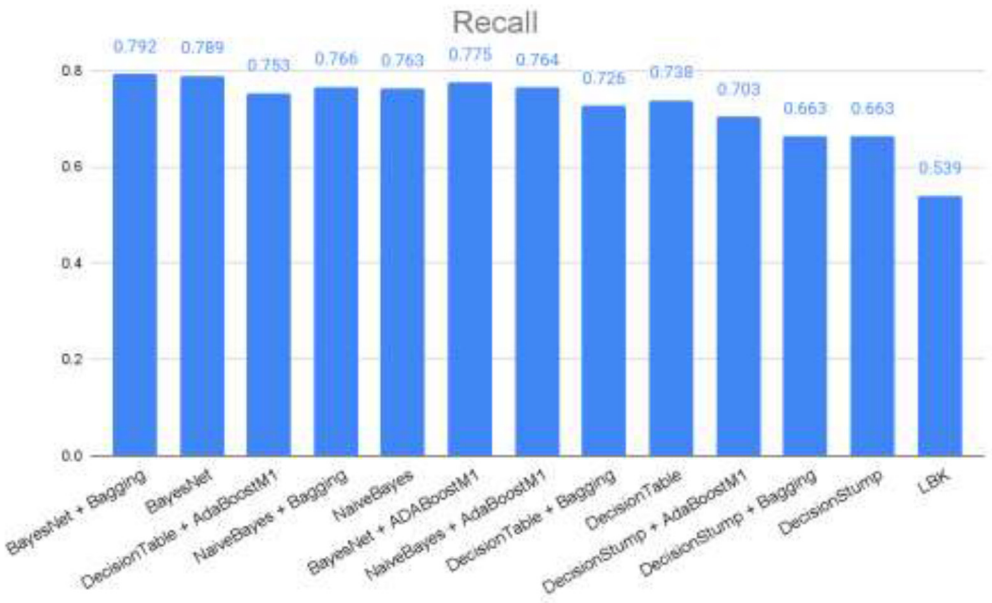

Fig. 5. Top performing models in terms of recall based on 10-fold CV on 2016 H1.

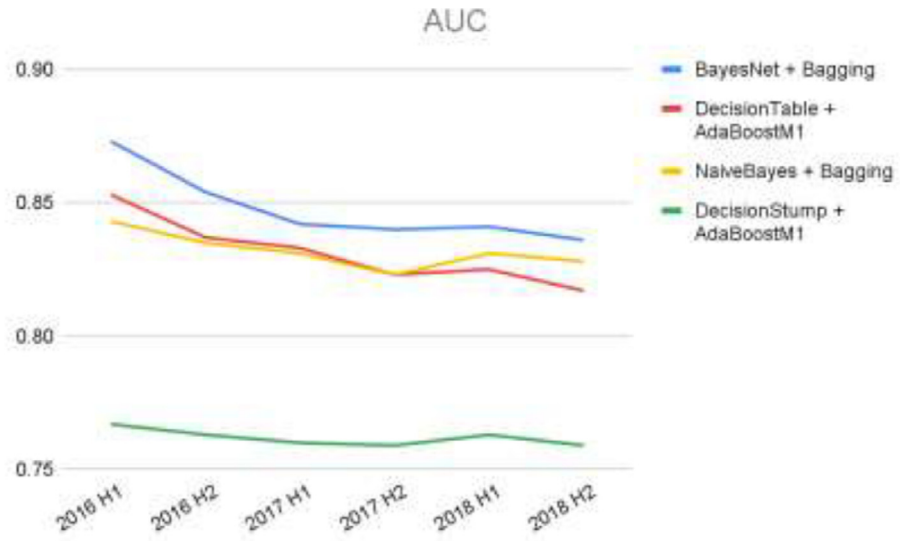

Fig. 6. Comparative AUC degradation 2016 H1 through 2018 H2.

Fig. 6 depicts the degradation of AUC for the four top models, starting from 2016 H1 through to 2018 H2. The AUC for 2016 H1 is that calculated via 10-fold CV when the model is trained on 2016 H1 and the AUC for each subsequent half year is the result of evaluating the performance of the 2016 H1-trained model on the data from those subsequent calendar half years.

Fig. 7 similarly provides a graphical representation of the degradation of the same top four models from 2016 H1 until 2018 H2, in this case showing accuracy degradation. The 2016 H1 value is from 10-fold CV training of the model, and the subsequent values from testing the 2016 H1-trained model on the data from the subsequent half years.

Fig. 8 demonstrates the degradation of model precision from 2016 H1 through 2018 H2. Again the 2016 H1 value is from the 10-fold CV training of the model and the subsequent half year values from testing the model on future half year datasets.

Fig. 9 graphs the degradation of model recall from 2016 H1 through 2018 H2 for the same top four models. The 2016 H1 value represents the result of 10-fold CV training on the 2016 H1 dataset and the subsequent values from testing that model on the subsequent half year datasets.

4.3. *Effect of time since training*

Each of the top four selected models are evaluated in terms of performance degradation of AUC, accuracy, precision and recall where different half year datasets are used for training, in Figs. 10 through 13 corresponding to each top four model respectively.

The four parts of Fig. 10 depict the performance for the Bagged Bayes Net model Fig. 10.a. shows the effect of the time period of the training set on AUC, Fig. 10b. shows the effect of time period of training set on accuracy, Fig. 10c. shows the effect of time period of the training time on precision and Fig. 10d. shows the effect of time period of the training set on recall.

Each of the different graph lines shows the future performance when the model is trained one half year later. For example the blue line in Fig. 10a. shows Bagged Bayes Net AUC when trained on 2016 H1 using 10-fold CV and then evaluated on each subsequent half year's dataset. The red line shows AUC degradation for a model trained on 2016 H2 using 10-fold CV and then evaluated on each subsequent calendar half year. The orange line shows the performance of Bagged Bayes Net when

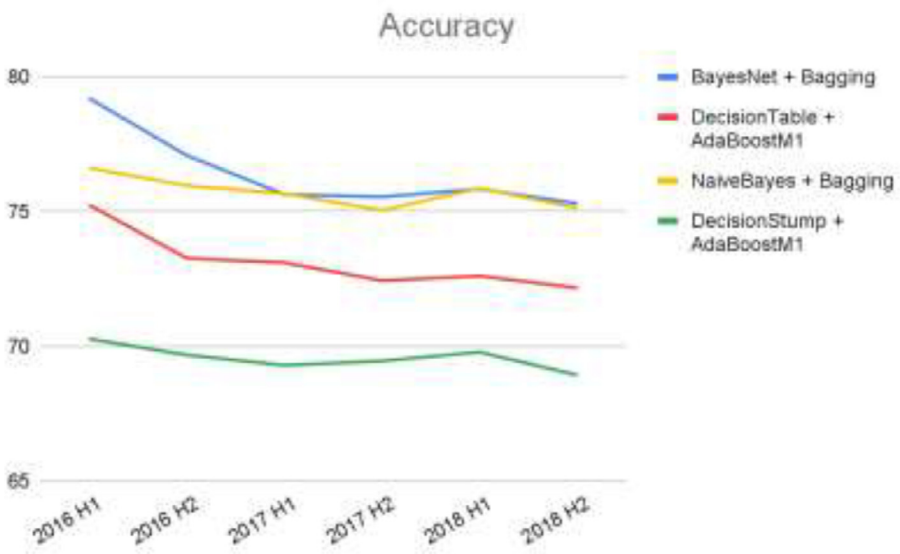

Fig. 7. Comparative accuracy degradation 2016 H1 through 2018 H2.

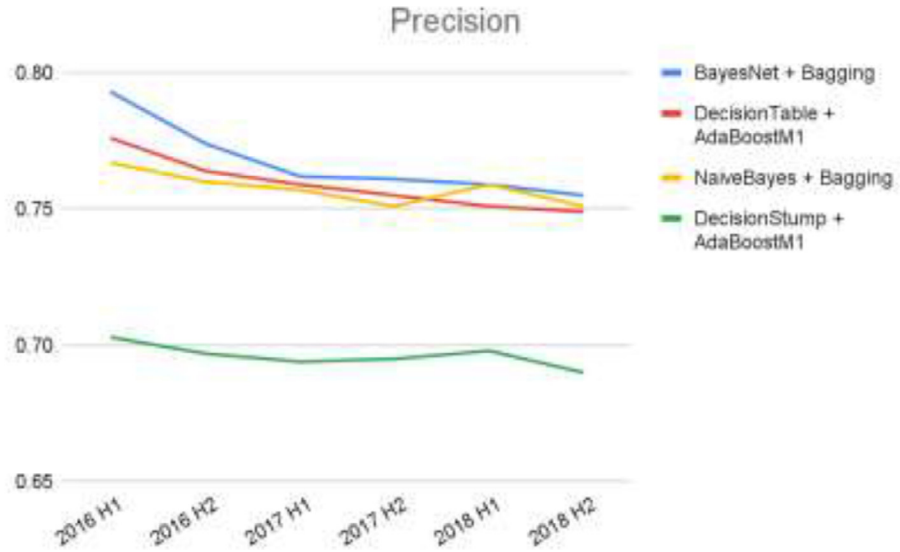

Fig. 8. Comparative precision degradation 2016 H1 through 2018 H2.

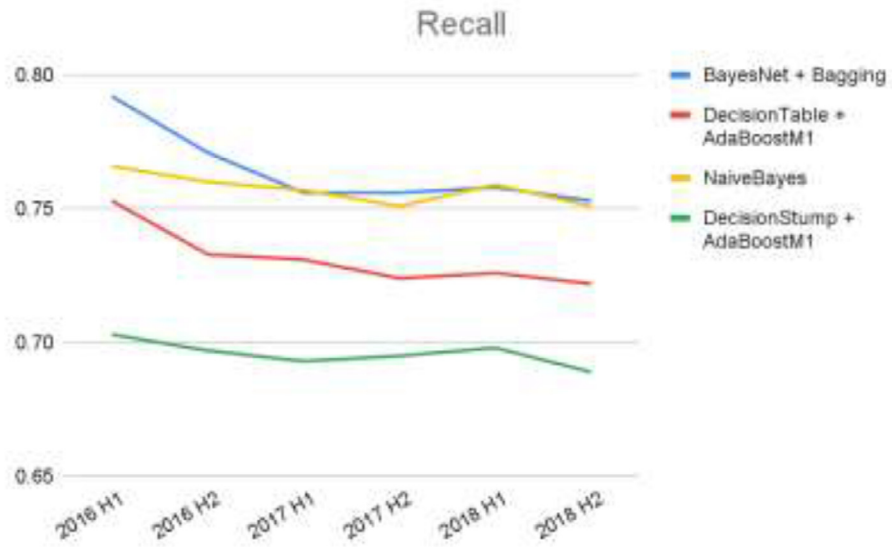

Fig. 9. Comparative recall degradation 2016 H1 through 2018 H2.

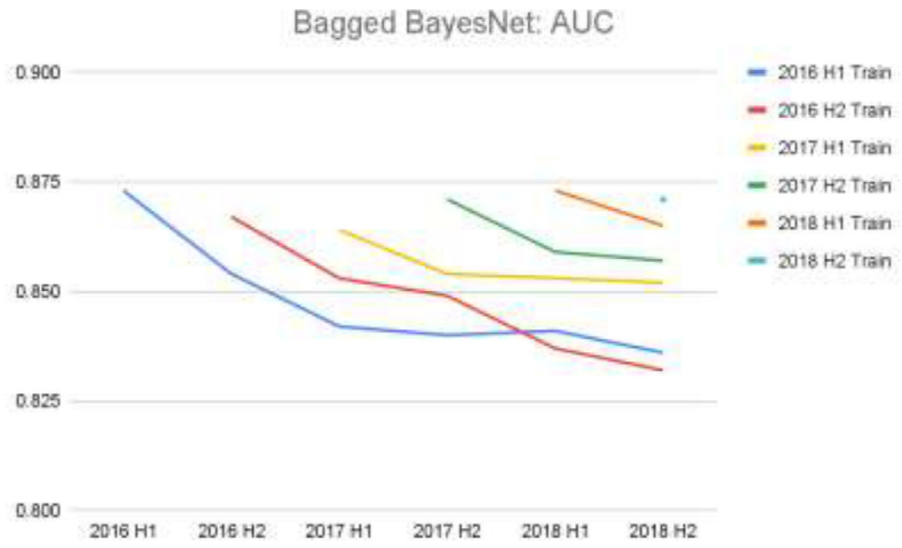

Fig. 10. a. Bagged BayesNet: Effect of time of training on AUC Fig. 10.b. Bagged BayesNet: Effect of time of training on accuracy Fig. 10.c. Bagged BayesNet: Effect of time of training on precision Fig. 10.d. Bagged BayesNet: Effect of time of training on recall.

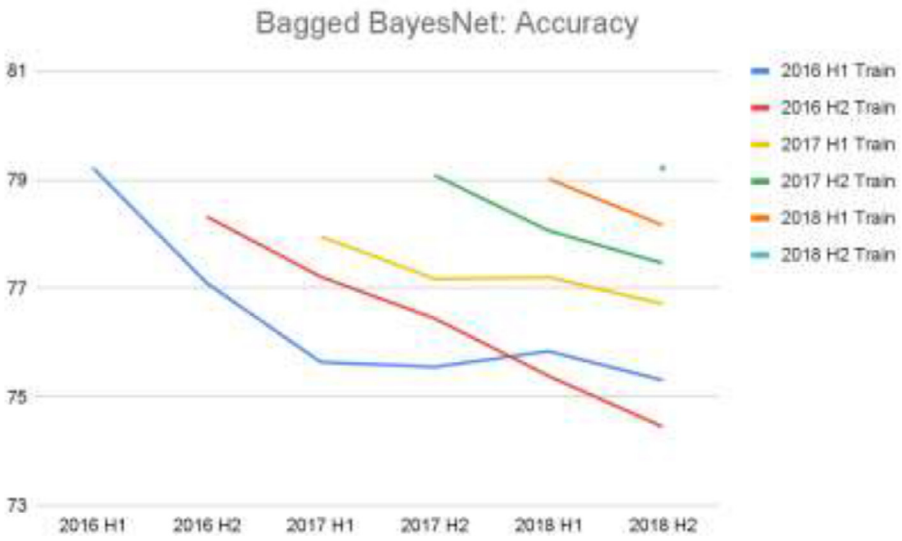

Fig. 10. Continued

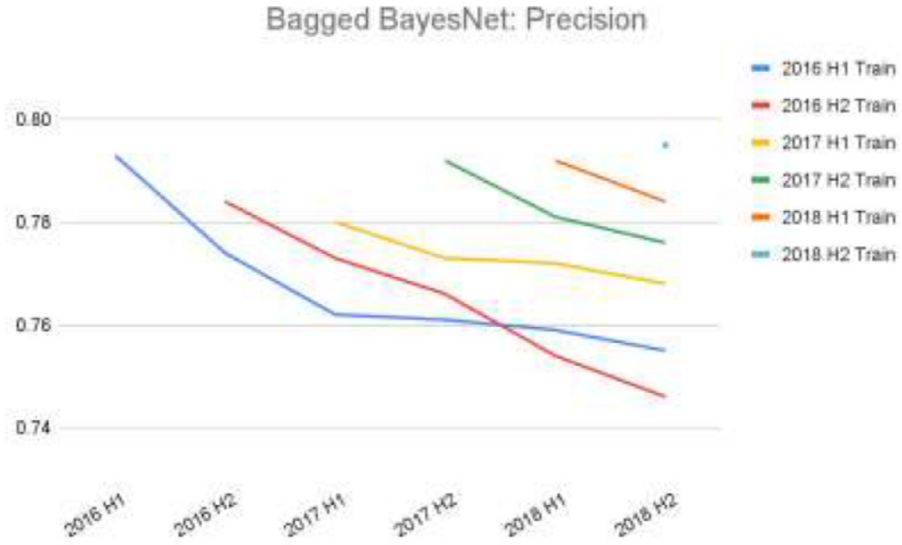

Fig. 10. Continued

Fig. 10. Continued

Fig. 11. a. Boosted Decision Table: Effect of time of training on AUC Fig. 11.b. Boosted Decision Table: Effect of time of training on accuracy Fig. 11.c. Boosted Decision Table: Effect of time of training on precision Fig. 11.d. Boosted Decision Table: Effect of time of training on recall.

trained on 2018 H1 data, and finally the single point represents the model's performance when trained on 2018 H2 there being no future half years to evaluate it upon Fig. 10.b. uses a similar approach to show the impact upon accuracy degradation resulting from choosing different training times/half years for model training.

Fig. 10c. similarly shows precision degradation resulting when different calendar half year datasets are chosen for initial training. Again, the blue line represents 10-fold CV training upon 2016 H1, the red line represents training upon 2016 H2 and so forth until finally the training is done upon the 2018 H2 dataset.

Fig. 10d. shows the effect of the time period of the training set upon recall degradation. For the first (blue) line, training using 10-fold CV has been done using the 2016 H1 dataset, and for each subsequent line a later calendar half year has been used for model training.

Figs. 11a., 11b., 11c. and 11d present the effect of different training times on AUC, accuracy, precision and recall degradation respectively, for the Boosted Decision Table model. The model shows various performance degradation aspects as impacted by time of training, that differ from that seen for Bagged Bayes Net as per Fig. 10.

Fig. 11b. shows the effect of the different choice of calendar half year for training upon model accuracy degradation. As with Fig. 10. the later the training half year, the shorter the period of future evaluation and hence length of graph line.

Fig. 11c. for the Boosted Decision Table model, provides a graphical representation of the effect of the time of training upon precision degradation. Here the selection of the calendar half for training, shows a significant impact on the rate of performance degradation.

Fig. 11d. again for the Boosted Decision Table model, represents the effect of time of training on recall degradation over time post-training. This illustrates that a later time of training does not necessarily lead to better performance than models trained on earlier data.

Figs. 12a., 12b., 12c. and 12d present performance degradation results for the Bagged Naïve Bayes model in terms of AUC, accuracy, precision and recall respectively.

Fig. 12b. shows the effect of time of training upon accuracy degradation. Each graph line shows where training using 10-fold CV has been carried out upon each of the 2016 H1 through 2018 H2 datasets respectively.

Fig. 12c. presents the effect of time of training upon precision degradation. Again, each graph line shows the case of training using 10-fold CV carried out upon each of the 2016 H1 through 2018 H2 datasets respectively.

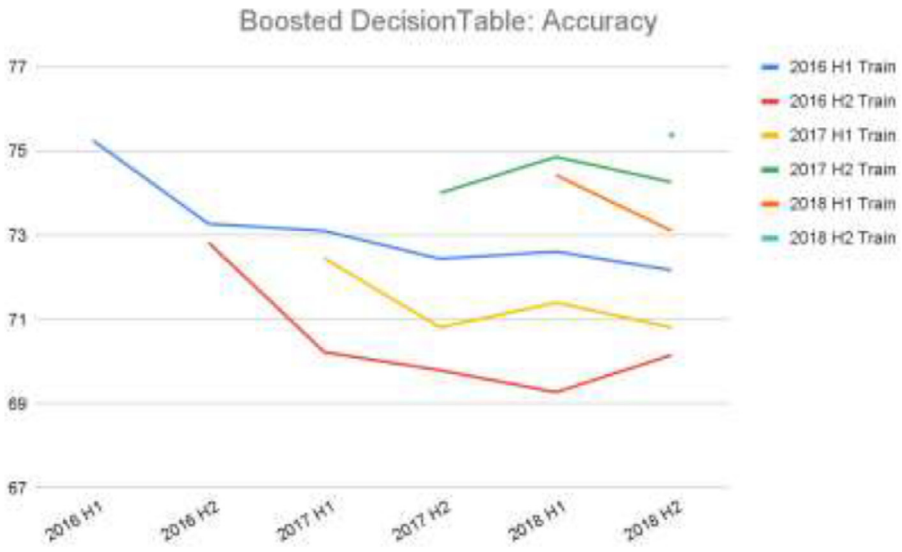

Fig. 11. Continued

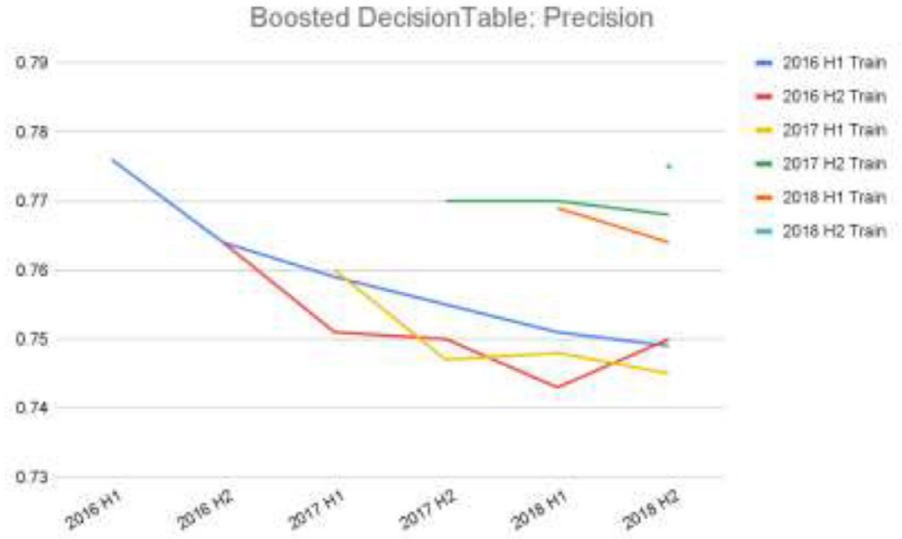

Fig. 11. Continued

Fig. 11. Continued

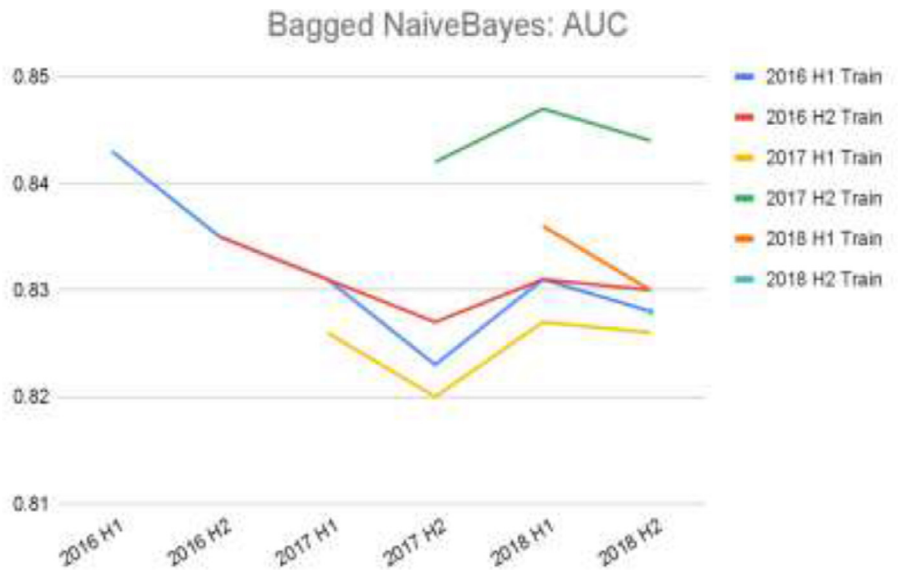

Fig. 12. a. Bagged Naive Bayes: Effect of time of training on AUC Fig. 12.b. Bagged Naive Bayes: Effect of time of training on accuracy Fig. 12.c. Bagged Naive Bayes: Effect of time of training on precision Fig. 12.d. Bagged Naive Bayes: Effect of time of training on recall.

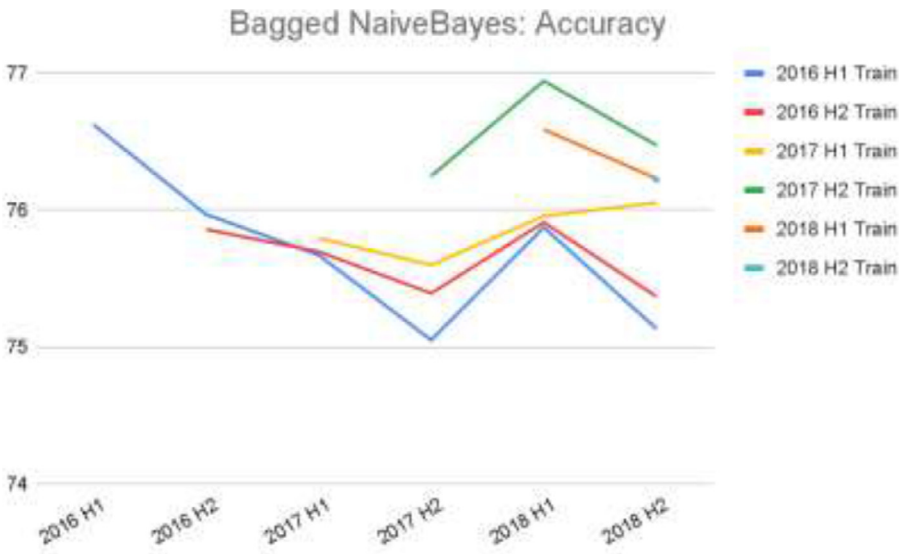

Fig. 12. Continued

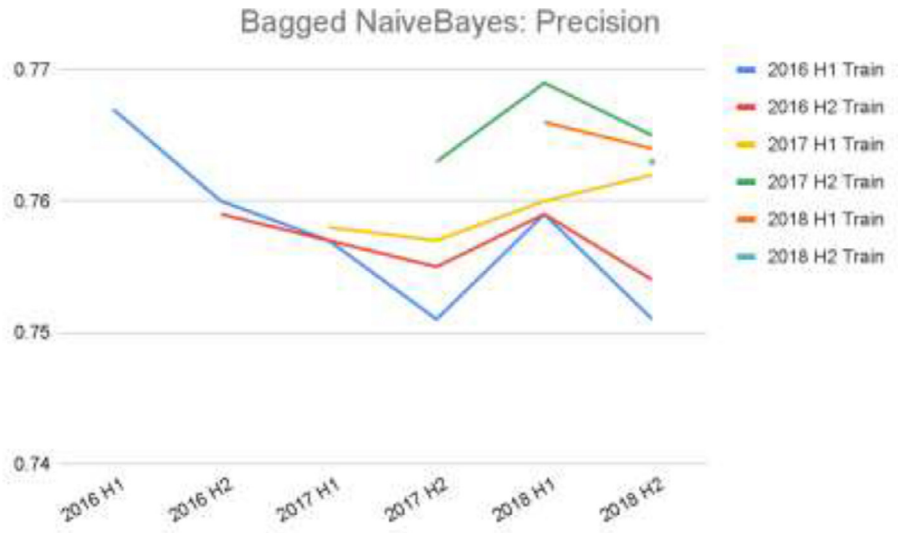

Fig. 12. Continued

Fig. 12. Continued

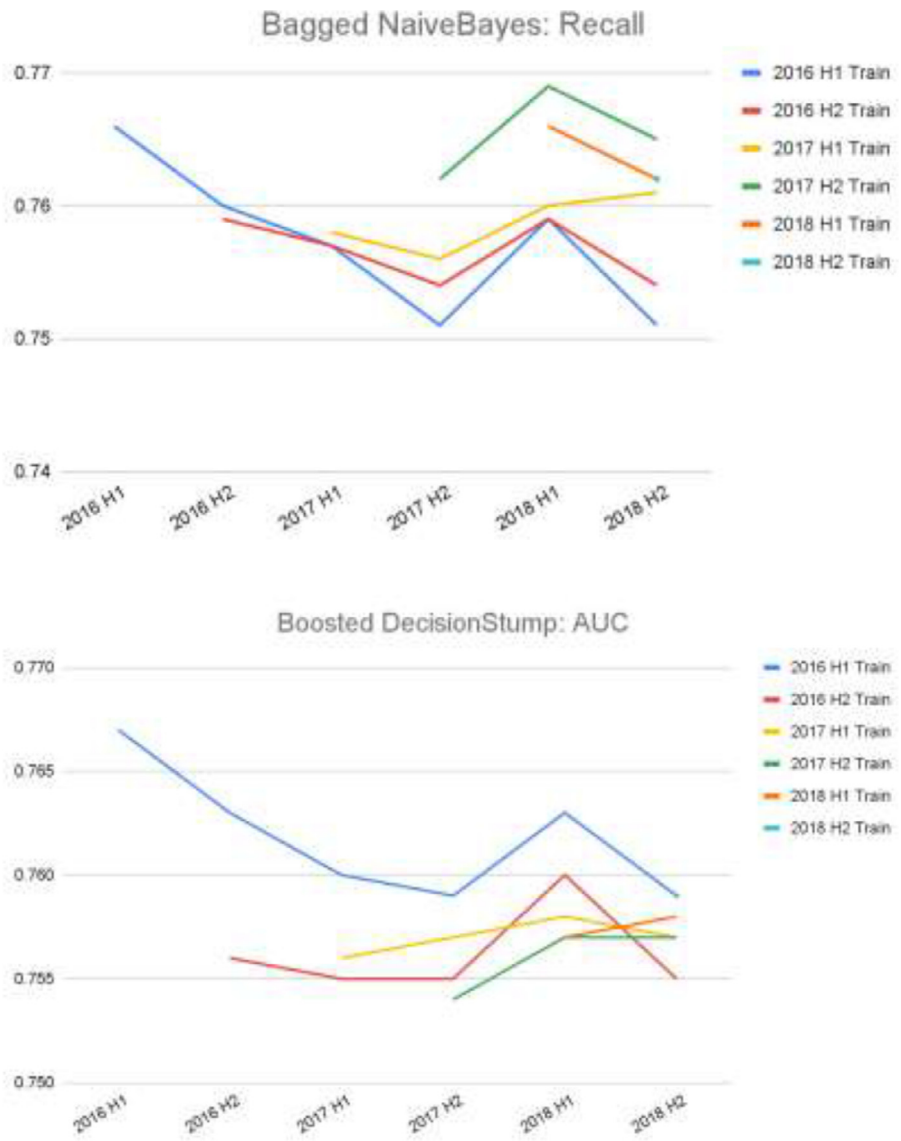

Fig. 13. a. Boosted Decision Stump: Effect of time of training on AUC Fig. 13.b. Boosted Decision Stump: Effect of time of training on accuracy Fig. 13.c. Boosted Decision Stump: Effect of time of training on precision Fig. 13.d. Boosted Decision Stump: Effect of time of training on recall.

Fig. 12d. shows the effect of changing the half year dataset selected for training upon model recall degradation. 10-fold CV-based training is again carried out on each of the half year datasets 2016 H1 through 2018 H2 and the performance maintenance of each model graphed.

Figs. 13a., 13b., 13c. and 13d depict for the Boosted Decision Stump model, the effect of the time of training upon AUC, accuracy, precision and recall degradation respectively Fig. 13.a. depicts AUC degradation where 10-fold CV model training is carried out using the datasets from each of 2016 H1 through 2018 H2 respectively.

Fig. 13b. depicts for Boosted Decision Stump the effect of time of training upon accuracy degradation. The graph lines depict the training of the model using 10-fold CV upon the 2016 H1 through 2018 H2 datasets respectively.

Fig. 13c. depicts for Boosted Decision Stump the effect of time of training upon precision degradation. The graph lines depict the training of the model using 10-fold CV upon the 2016 H1 through 2018 H2 datasets respectively.

Fig. 13d. depicts for Boosted Decision Stump the effect of time of training upon recall degradation. The graph lines depict the training of the model using 10-fold CV upon each of the 2016 H1 through 2018 H2 datasets respectively.

5. Discussion

5.1. Model predictive performance

Upon initial evaluation of candidate base models such models as Bayesian Network, Naive Bayes and Decision Table are shown to demonstrate high performance (Figs. 2 through 5). For these models, meta-variants in a number of cases improved upon the base model performance and hence were selected for consideration of future performance degradation characteristics within the set of top four models. The best performing model based on 10-fold CV on 2016 H1 was Bagged Bayesian Net, achieving an AUC of 0.873 on the 2016 H1 data. An AUC of over 0.9 is considered 'excellent' discriminative performance and the 0.873 can be considered very good discriminative performance. Bagged Bayes Net also provided the highest initial performance in terms of accuracy (79.2%) (Fig. 3), precision (0.793) (Fig. 4) and recall (0.792) (Fig. 5). As has been found in previous HCUP data based predictive model studies, Bayesian-based models such as Bayesian Network and Naive Bayes can perform well on such HCUP patient discharge datasets (Steele & Hillsgrove, 2019; Young & Steele, 2021).

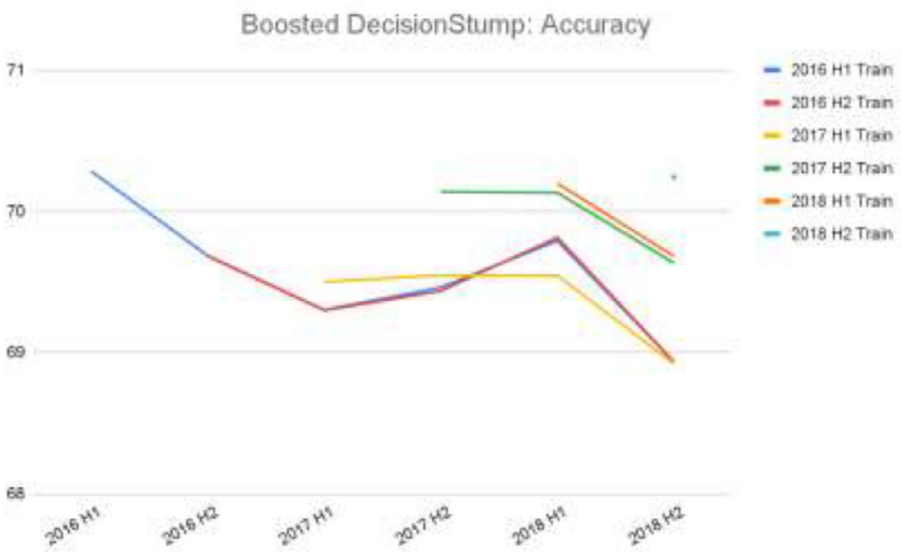

Fig. 13. Continued

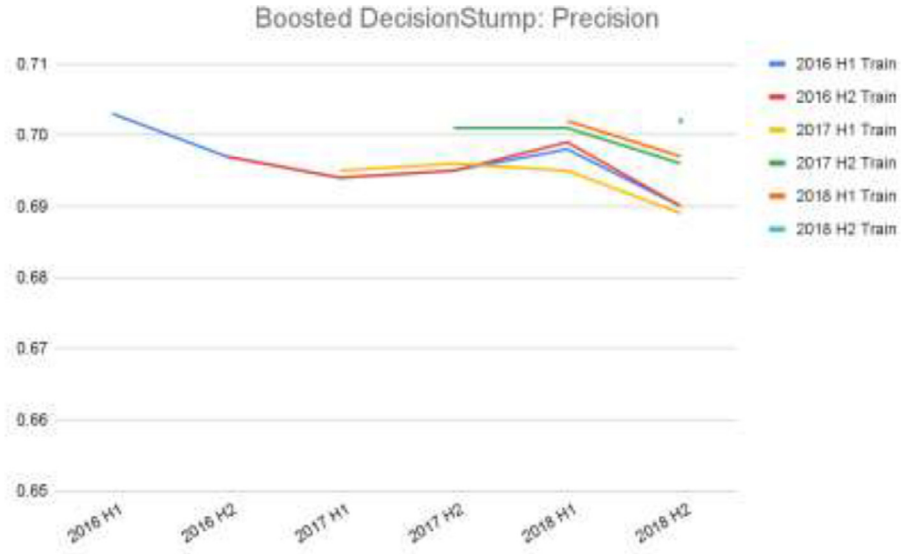

Fig. 13. Continued

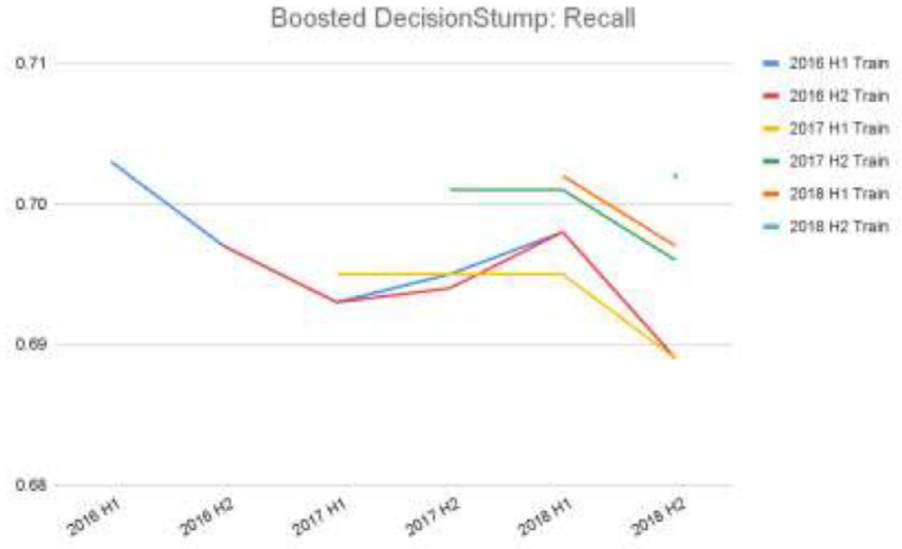

Fig. 13. Continued

5.2. Comparative model performance degradation

Of note is the high performance maintenance over time of the best performing models. Over a period of 2.5 years post-training, the best performing model Bagged Bayes Net drops from an initial AUC of 0.873 to an AUC of 0.836 for 2018 H2 (Fig. 6). We also note, that ranking of model performance based on 10-fold CV evaluation is not maintained over time: the second ranked model in terms of AUC upon training, Boosted Decision Table begins with an AUC of 0.853 but drops to third ranked with an AUC of 0.817 by 2018 H2 (Fig. 6). This initially suggests that exploration of model performance degradation is an important question, with such standard generalization techniques as 10-fold CV, not providing an accurate measure on its own of relative performance maintenance into the future. By 2018 H2 Bagged Naive Bayes is the second best performing model in terms of AUC with an AUC of 0.826.

Other observations on model AUC performance degradation include: (1) faster degradation in the first half year post-training with a trend towards smaller drops in performance each additional subsequent calendar half; (2) non-monotonic performance change, for example all four models increasing in AUC slightly from 2017 H2 to 2018 H1; but (3) a general trend towards decreased model performance in each further half year from time of model training. A possible factor in the performance increase from 2017 H2 to 2018 H1 could be suggested to be some greater potential similarity of the 2018 H1 dataset to the H1 dataset from 2016 used for training given the same first calendar halves of the year, but contrary to this, no increase in AUC is seen going from 2016 H2 to 2017 H1. Despite the creation of the test sets by calendar half years there is no clear seasonal trend (Zulkarnain & Rutledge, 2018) seen in the performance maintenance and degradation. For the purposes of studying the predictive models in this article we are not attempting to take into account seasonal factors but rather consider relative performance maintenance between differing initially high-performing models.

Model accuracy degradation (Fig. 7) shows similarities to AUC degradation. We do see a changing of rank of models during the period of model degradation evaluation, with Bagged Naive Bayes slightly outperforming Bagged Bayes Net in terms of accuracy in 2017 H1 and 2018 H1. This again provides some support for 10-fold CV not necessarily capturing future performance maintenance when evaluated empirically.

Comparative performance maintenance in terms of precision (Fig. 8) and recall (Fig. 9) show a number of similar characteristics including: non-maintenance of model rank order over time, non-monotonic performance decrease and an initially more rapid decline in performance with the rate of decline leveling off for later time intervals post-training.

5.3. Effect of time of training

Fig. 10 graphs the effect of time of training for the best performing model, Bagged Bayes Net, on performance in terms of AUC, accuracy, precision and recall degradation in Figs. 10a through 10d respectively. We see such AUC characteristics as: (1) sharper initial decline followed by a leveling off of rate of decline; (2) different training periods can lead to different rates and extents of decline e.g. Bagged Bayes Net trained on 2016 H2 deteriorates to be worse than the same model trained on 2016 H1 for the whole of 2018 (Fig. 10a); and (3) examples of non-monotonic performance decline. While AUC for Bagged Bayes Net generally declines more with more time since training, this is not necessarily the case, and it appears that training on some dataset periods can lead to worse performance on future data even if more recently trained.

In terms of accuracy (Fig. 10b), precision (Fig. 10c) and recall (Fig. 10d) we again see that training on 2016 H2 leads to an overall greater rate and magnitude of performance decline and also apparently no leveling off of decline as seen for other training sets, at least as of the end of 2018. This again suggests that neither initial 10-fold CV performance nor time since training provide deterministic predictors of the relative performance degradation of Bagged Bayes Net for this predictive model application area.

For Boosted Decision Table (Fig. 11) we see more erratic impact due to choice of training period. We see the model trained on 2016 H2 consistently has a lower AUC into the future than the model trained on 2016 H1 for the same evaluation periods. We also see the model trained on 2017 H2 increase in performance into the future and little AUC decline for the model trained on 2018 H1. While 10-fold CV is considered to evaluate for a generalizable model, these results suggest that Boosted Decision Table for this health predictive model use case, is more sensitive to the specific training data used as to how its performance will deteriorate in the future.

In terms of accuracy (Fig. 11b), precision (Fig. 11c) and recall (Fig. 11d) we see similar future increase of performance for the 2017 H2 trained model. We see particularly strong performance maintenance in terms of accuracy and recall for the model trained on 2016 H1.

For Bagged Naive Bayes (Fig. 12) we see more erratic performance maintenance behavior, more similar to Boosted Decision Table than Bagged Bayes Net. We again see the model trained on 2017 H2 increase in performance in subsequent halves. We see the model trained on 2017 H1 perform substantially worse than both models trained on earlier half years. We also see the model trained on 2018 H2 begin with an AUC (0.828) no higher than the model trained on 2016 H1 (two and a half years earlier) (Fig. 12a). This provides an interesting demonstration that at least for this model, more recent training does not provide better predictive performance.

Similarly for Boosted Decision Stump (Fig. 13) no more recently trained model exceeds the model trained earliest in 2016 H1 in terms of AUC (Fig. 13a) for the same test period. Even upon training using 2018 H2 an AUC of 0.759 is achieved, equal to the performance of the 2016 H1 trained model evaluated at 2018 H2.

5.4. Contributions to the literature

This is one of the first studies of ML-based model performance maintenance and degradation over time post-training based on a large real-world dataset. Other existing examples of such studies include (Leevy, Khoshgoftaar, Bauder & Seliya, 2020), (Leevy, Khoshgoftaar, Bauder & Seliya, 2019), (Galen & Steele, 2021, January), (Galen & Steele, 2021, April) and (Thabtah, Hammoud, Kamalov & Gonsalves, 2020). This is suggestive of an opportunity for the Information Systems community to contribute to the performance evaluation of ML algorithms when applied to a given real-world business predictive problems and to identify empirical findings that can help bridge the gap between new ML algorithm development and practical applicability including in health applications (Galetsi, Katsaliaki & Kumar, 2020) (Aggarwal, Mittal & Battineni, 2021). This is in general an under addressed research area, with much of the research focus on the development of novel ML algorithms, not on their empirically measured generalizability in real-world evolving data environments.

In relation to model evaluation, 10-fold CV performance is typically considered a measure of the generalizability of a ML model or how well the model will perform on future, previously unseen data, and is the basis to choose a model for deployment. The results in this study suggest that 10-fold CV does not fully capture how well performance will be maintained on future data at least for some predictive models. The study also provides insights in relation to model retraining a not well understood problem (Schelter et al., 2018). It is demonstrated that just using the most recent historical data for training, does not necessarily lead to the highest performing ML model, see Fig. 13a. for a clear example of this

Further, the current article describes an approach to assess model degradation empirically based on real-world datasets for given ML models, to complement methods for assessing data drift and model degradation seen in such works as (Leevy, Khoshgoftaar, Bauder & Seliya, 2020), (Barros & Santos, 2018) and (Mauri & Damiani, 2021).

In relation to health information systems, this is the first detailed study of the performance degradation over time of a ML-based predictive model on a large, real-world dataset. In relation to the specific predictive problem of emergency patient mortality prediction at time of admission, the article demonstrates that performance degradation over a span of years is relatively slow. No previous study has considered performance maintenance or degradation for such health ML models and none has demonstrated this empirical finding and it has important implications for practice. It suggests that such models once trained can be used with likely high performance for a significant period into the future. It also provides an important data point in considering the possible performance maintenance of other similar health predictive models.

5.5. Implications for practice

An important implication is that such high performing predictive models as developed in this work can potentially be deployed in real-world settings with confidence that they will remain relatively high performing for at least a coming year or more. This makes such models developed in the work of real-world applicability, as they were trained from data drawn from all Maryland hospitals and could be considered applicable to all such hospitals. Certainly this would immediately be the case within the context of the Maryland health setting, but as similar HCUP data is available for a wide range of US states, it is very likely that such performance maintenance characteristics would be seen for similar models trained for use for other US states.

An advantage of applicability of the findings arising from these models trained on the nationally collected HCUP dataset attributes, is the models make use of attributes universally captured or available at admission across all US hospitals. As such, models and the related performance degradation results may potentially be readily or approximately applicable nationally. As such, a non-Maryland version of the model could remain the same in terms of model type and input attributes but would need to be retrained on the other states' data. Similar models are also likely to be readily trainable for non-US hospital jurisdictions and the results of the current study suggest that such models may also demonstrate good performance maintenance. In all settings, this means that such models can be deployed with greater confidence that they will have high performance over time periods post-deployment and hence are a more valuable clinical tool.

6. Conclusion

This article provides a first detailed academic study of the empirical performance degradation of ML-based healthcare predictive models based on large real-world datasets, in this case drawing upon a dataset of more than 1.83 million patient discharge records from the US state of Maryland for the years 2016 through 2018. The results show it is possible and relatively simple to train high performing ML-based predictive models for predicting at time of admission, patient in-hospital mortality, the best model considered in this study having an initial AUC of 0.873, that are able to still maintain high performance for over 2.5 years after training. The study also shows that time passed since model training is not the only factor in future performance degradation and suggests the value of greater empirical evaluation and further analysis of such models in increasing the understanding of the expected future behavior of such ML-based health predictive models as it relates to both their performance maintenance and related safety in real-world use.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Adler-Milstein, J., & Jha, A. K. (2017). HITECH Act drove large gains in hospital electronic health record adoption. *Health Affairs, 36*(8), 1416–1422.

Agency for Healthcare Research and Quality. Healthcare Cost and Utilization Project. https://www.hcup-us.ahrq.gov/overview.jsp

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, Article 100004.

AHRQ, "Maryland State Inpatient Database File Structure". https://www.hcup-us.ahrq.gov/db/state/sidc/tools/filespecs/MD_SID_2018_CORE.loc

AHRQ, "Maryland State Inpatient Database 2017 File Structure". https://www.hcup-us.ahrq.gov/db/state/sidc/tools/filespecs/MD_SID_2017_CORE.loc

AHRQ, "Maryland State Inpatient Database 2016 File Structure". https://www.hcup-us.ahrq.gov/db/state/sidc/tools/filespecs/MD_SID_2016_CORE.loc

Barros, R., & Santos, S. (2018). A large-scale comparison of concept drift detectors. *Information Sciences, 451*, 348–370.

Blom, M. C., Ashfaq, A., Sant'Anna, A., Anderson, P. D., & Lingman, M. (2019). Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: A retrospective, population-based registry study. *BMJ open, 9*(8), Article e028015.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Brettle, C., & Steele, R. (2021,. March). Advance prediction of Maryland elective admission fatalities using machine learning. In 2021 *7th International Conference on Information Management (ICIM)* (pp. 107–112). IEEE.

Brettle, C., & Steele, R. (2021, December). Do predictive models always deteriorate in performance with time? A case study in elective mortality predictive model performance. In *Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET)*.

Chatzicostas, C., Roussomoustakaki, M., Vlachonikolis, I. G., Notas, G., Mouzas, I., Samonakis, D., et al. (2002). Comparison of Ranson, APACHE II and APACHE III scoring systems in acute pancreatitis. *Pancreas, 25*(4), 331–335.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. *International Conference on Machine Learning, 96*, 148–156.

Galen, C., & Steele, R. (2021,). Empirical measurement of performance maintenance of gradient boosted decision tree models for malware detection. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 193–198).

Galen, C., & Steele, R. (2021, January). The effect of training set timeframe on the future performance of machine learning-based malware detection models. In *HICSS* (pp. 1–10).

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management, 50*, 206–216.

Guo, L. L., Pfohl, S. R., Fries, J., Posada, J., Fleming, S. L., Aftandilian, C., et al. (2021). Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied Clinical Informatics, 12*(04), 808–815.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18 2009.

Hanle, A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

HCUP. State Inpatient Databases (SID) Overview. https://www.hcup-us.ahrq.gov/sidoverview.jsp

HCUP Central Distributor SID Availability of Data Elements. (2018). *Healthcare cost and utilization project (HCUP). may 2020*. Rockville, MD: Agency for Healthcare Research and Quality www.hcup-us.ahrq.gov/db/state/siddist/siddistvarnote2018.jsp.

Hillsgrove, T., & Steele, R. (2019, March). Utilization of data mining for generalizable, all-admission prediction of inpatient mortality. In *2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT)* (pp. 71–75). IEEE.

Kim, S., Kim, W., & Park, R. W. (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare Informatics Research, 17*(4), 232–243.

Klug, M., Barash, Y., Bechler, S., Resheff, Y. S., Tron, T., Ironi, A., et al. (2020). A gradient boosting machine learning model for predicting early mortality in the emergency department triage: Devising a nine-point triage score. *Journal of General Internal Medicine, 35*(1), 220–227.

Knaus, W. A., Wagner, D. P., Draper, E. A., Zimmerman, J. E., Bergner, M., & Bastos, P. G. (1991). The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults. *Chest, 100*(6), 1619–1636 et al.

Konovalenko, I., & Ludwig, A. (2021). Comparison of machine learning classifiers: A case study of temperature alarms in a pharmaceutical supply chain. *Information Systems, 100*, Article 101759.

Kuo, P. J., Wu, S. C., Chien, P. C., Rau, C. S., Chen, Y. C., & Hsieh, H. Y. (2018). Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: A cross-sectional retrospective study in southern Taiwan. *BMJ open, 8*(1), Article e018252 et al.

Kushwaha, A. K., Kar, A. K., & Dwivedi, Y. K. (2021). Applications of big data in emerging management disciplines: A literature review using text mining. *International Journal of Information Management Data Insights, 1*(2), Article 100017.

Le Gall, J. R., Lemeshow, S., & Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA, 270*(24), 2957–2963.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2019, December). The effect of time on the maintenance of a predictive model. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)* (pp. 1891–1896).

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2020). Investigating the relationship between time and predictive model maintenance. *Journal of Big Data, 7*, 1–19.

Mauri, L., & Damiani, E. (2021). Estimating degradation of machine learning data assets. *ACM Journal of Data and Information Quality (JDIQ), 14*(2), 1–15.

Meiring, C., Dixit, A., Harris, S., MacCallum, N. S., Brealey, D. A., & Watkinson, P. J. (2018). Optimal intensive care outcome prediction over time using machine learning. *PloS one, 13*(11), Article e0206862 et al.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2021). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, 1–11.

Metsker, O., Sikorsky, S., Yakovlev, A., & Kovalchuk, S. (2018). Dynamic mortality prediction using machine learning techniques for acute cardiovascular cases. *Procedia Computer Science, 136*, 351–358.

Min, Q., Lu, Y., Liu, Z., Su, C., & Wang, B. (2019). Machine learning based digital twin framework for production optimization in petrochemical industry. *International Journal of Information Management, 49*, 502–519.

Minne, L., Abu-Hanna, A., & de Jonge, E. (2008). Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. *Critical Care, 12*(6), 1–13.

Perng, J. W., Kao, I. H., Kung, C. T., Hung, S. C., Lai, Y. H., & Su, C. M. (2019). Mortality prediction of septic patients in the emergency department based on machine learning. *Journal of Clinical Medicine, 8*(11), 1906.

Rau, C. S., Kuo, P. J., Chien, P. C., Huang, C. Y., Hsieh, H. Y., & Hsieh, C. H. (2018). Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PloS one, 13*(11), Article e0207192.

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights, 1*(2), Article 100012.

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., & Szarvas, G. (2018). On challenges in machine learning model management. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Sharma, R., Kumar, A., & Chuah, C. (2021). Turning the blackbox into a glassbox: An explainable machine learning approach for understanding hospitality customer. *International Journal of Information Management Data Insights, 1*(2), Article 100050.

Sherazi, S. W. A., Jeong, Y. J., Jae, M. H., Bae, J. W., & Lee, J. Y. (2020). A machine learning–based 1-year mortality prediction model after hospital discharge for clinical patients with acute coronary syndrome. *Health Informatics Journal, 26*(2), 1289–1304.

Steele, R., & Hillsgrove, T. (2019,). Predicting all-condition, in-hospital mortality of elective patients at time of scheduling. In *Proc. Of 2019 SoutheastCon* (pp. 1–5).

Steele, R., & Hillsgrove, T. (2019,). Data mined models for predicting in-hospital mortality of emergency admissions at time of hospital admission. In *Proc. 2019 SoutheastCon* (pp. 1–5). IEEE.

Tang, F., Xiao, C., Wang, F., & Zhou, J. (2018). Predictive modeling in urgent care: A comparative study of machine learning approaches. *Jamia Open, 1*(1), 87–98.

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences, 513*, 429–441.

van Doorn, W.P., .Stassen, P.M., .Borggreve, H.F., .Schalkwijk, M.J., .Stoffers, J., Bekers, O. et al. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PloS one*, 16(1), e0245157.

Varghese, Y. E., Kalaiselvan, M. S., Renuka, M. K., & Arunkumar, A. S. (2017). Comparison of acute physiology and chronic health evaluation II (APACHE II) and acute physiology and chronic health evaluation IV (APACHE IV) severity of illness scoring systems, in a multidisciplinary ICU. *Journal of Anaesthesiology, Clinical Pharmacology, 33*(2), 248.

Won, H. R., Kim, M. J., & Ahn, H. (2018). A machine learning-based customer classification model for effective online free sample promotions. *The Journal of Information Systems, 27*(3), 63–80.

Yeun, J. Y., Levine, R. A., Mantadilok, V., & Kaysen, G. A. (2000). C-reactive protein predicts all-cause and cardiovascular mortality in hemodialysis patients. *American Journal of Kidney Diseases, 35*(3), 469–476.

Young, Z., & Steele, R. (2021). Performance maintenance of machine learning-based emergency patient mortality predictive models. In *Proc. of 4th International Conf on Computer and Informatics Engineering*.

Zhai, Q., Lin, Z., Ge, H., Liang, Y., Li, N., Ma, Q., et al. (2020). Using machine learning tools to predict outcomes for emergency department intensive care unit patients. *Scientific Reports, 10*(1), 1–10.

Zulkarnain, A., & Rutledge, M. S. (2018). *How does delayed retirement affect mortality and health?* (p. 11). Center for Retirement Research at Boston College, CRR WP.