Benchmarking four algorithms for improved classification of agricultural injury cases from free-text analysis of pre-hospital care reports

Laura E. Jones, Erika Scott, Nicole Krupa, Megan Kern, Cristina Silvia Hansen-Ruiz, Paul Jenkins

Submitted to: JMIR Public Health and Surveillance
on: August 21, 2024

Disclaimer: This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Benchmarking four algorithms for improved classification of agricultural injury cases from free-text analysis of pre-hospital care reports

Laura E. Jones[1] PhD, MS; Erika Scott[2] PhD; Nicole Krupa[1] BS; Megan Kern[1] BS; Cristina Silvia Hansen-Ruiz[2] PhD; Paul Jenkins[1]

[1]Center for Biostatistics Bassett Research Institute Cooperstown US
[2]Northeast Center for Occupational Health and Safety in Agriculture, Forestry and Fishing Bassett Medical Center Cooperstown US

Corresponding Author:
Laura E. Jones PhD, MS
Center for Biostatistics
Bassett Research Institute
One Atwell Road
Cooperstown
US

Abstract

Background: Fatality rates in Agriculture, Forestry, and Fishing (AgFF) industries are historically the highest of any US sector, with a combined rate of 18.6 deaths per 100,000 workers. Despite clear trends for fatal AgFF workplace injuries in federal data, challenges remain in capturing nonfatal agricultural injuries. The Northeast Center for Occupational Health and Safety (NEC) developed a naïve Bayes-based classification strategy to extract non-fatal injury cases from pre-hospital (EMS) free-text records.

Objective: The aim of this paper is to improve retrieval rates, in terms of false positive rate required to obtain a true positive rate of 0.90, by benchmarking naïve Bayes against three other algorithms: elastic net regression, Support Vector Machines, and boosted decision trees (XGBoost).

Methods: Using a labeled, fully one-hot coded gold-standard dataset (N=60,143) with substantial (24%) missing data, we benchmark these algorithms on complete case data (N=44,566) and imputed data from two imputation schemes: grouped hot-deck and recoding of missing units to the category “unknown,” using a 75:25 train/test split and stratified sampling.

Results: All models produced similarly accuracies (~0.98) on complete case data, though necessary False Positive Rates (FPR*) varied from 0.055 (XGBoost) to 0.20 (naïve Bayes) on training data, and on predictions, the range was 0.10 (elastic net) to 0.22 (naïve Bayes). On imputed data, accuracies ranged from 0.96 (Bayes) to 0.98 (XGBoost) for training data, yielding false positive rates from 0.095 (XGBoost) to 0.34 (Bayes). Predictions from imputed data showed FPR* ranging from 0.12 (XGBoost) to 0.41 (Bayes) depending on imputation scheme.

Conclusions: While all four models perform well on complete data, missing units are substantial and can result in misclassification and in omissions, both requiring human coding. Reliance on a machine learning method that is robust to missing data and imputation method, such as XGBoost, is a reasonable approach to improving classification rates without omitting data.

(JMIR Preprints 21/08/2024:65621)
DOI: https://doi.org/10.2196/preprints.65621

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Benchmarking four algorithms for improved classification of agricultural injury cases from free-text analysis of pre-hospital care reports

Laura E. Jones, PhD, MS, MS[1,3], Erika Scott, PhD[2], Nicole Krupa, BS[1] ,
Megan Kern, BS[1], Cristina S. Hansen-Ruiz, PhD[2], and Paul Jenkins PhD[1]

Affiliations:
[1]Center for Biostatistics, Bassett Research Institute, Bassett Medical Center
One Atwell Road, Cooperstown, New York
[2]Northeast Center for Occupational Health and Safety in Agriculture, Forestry, and Fishing, Bassett Medical Center, One Atwell Road, Cooperstown, New York
[3]Corresponding author: laura.jones@bassett.org

Abstract

Background. Fatality rates in Agriculture, Forestry, and Fishing (AgFF) industries are historically the highest of any US sector, with a combined rate of 18.6 deaths per 100,000 workers. Despite clear trends for fatal AgFF workplace injuries in federal data, challenges remain in capturing nonfatal agricultural injuries. The Northeast Center for Occupational Health and Safety (NEC) developed a naïve Bayes-based classification strategy to extract non-fatal injury cases from pre-hospital (EMS) free-text records.

Objective. The aim of this paper is to improve retrieval rates, in terms of false positive rate required to obtain a true positive rate of 0.90, by benchmarking the reference naïve Bayes against three other algorithms: elastic net regression, Support Vector Machines, and boosted decision trees (XGBoost).

Methods. Using a labeled, fully one-hot coded gold-standard dataset (N=60,143) with substantial (24%) missing data, we benchmark these algorithms on complete case data (N=44,566) and imputed data from two imputation schemes: grouped hot-deck and recoding of missing units to the category “unknown,” using a 75:25 train/test split and stratified sampling.

Results. All models produced similarly accuracies (~0.98) on complete case data, though necessary False Positive Rates (FPR*) varied from 0.055 (XGBoost) to 0.20 (naïve Bayes) on training data, and on predictions, the range was 0.10 (elastic net) to 0.22 (naïve Bayes). On imputed data, accuracies ranged from 0.96 (Bayes) to 0.98 (XGBoost) for training data, yielding false positive rates from 0.095 (XGBoost) to 0.34 (Bayes). Predictions from imputed data showed FPR* ranging from 0.12 (XGBoost) to 0.41 (Bayes) depending on imputation scheme.

Conclusions. While all four models perform well on complete data, missing units are substantial and can result in misclassification and in omissions, both requiring human coding. Reliance on a machine learning method that is robust to missing data and imputation method, such as XGBoost, is a reasonable approach to improving classification rates without omitting data.

Keywords:

surveillance, work-place injuries, agriculture, pre-hospital records, free text, machine learning, missing data, imputation

Introduction

Fatalities in Agriculture, Forestry and Fishing (AgFF) industries are historically the highest of any sector in the United States, with a combined rate of 18.6 deaths per 100,000 workers [1]. In the farming sector alone, agricultural workers die at rates five times greater than typical workers in other sectors [2]. Despite clear trends in fatal AgFF workplace injuries documented in federal data, challenges remain in capturing nonfatal agricultural injuries and illnesses. In the half century since the Occupational Health and Safety Act [3] was established, agricultural workers remain disproportionately disadvantaged, with regulatory exemptions directly contributing to the current inadequate state of injury reporting in this sector [4]. Given that common data sources like the Survey of Occupational Injuries and Illnesses (SOII) exclude self-employed agricultural workers and workers at farms with less than 11 employees [5, 6], under-reporting may be particularly true of northeast farming operations, many of which are small-scale or family-run. Accurate surveillance of nonfatal agricultural injuries using novel data sources is thus needed to more effectively capture injury burden in this sector.

There have been numerous attempts to close the gap in non-fatal agricultural injury surveillance. A series of agricultural injury surveys, including the Occupational Injury Surveillance of Production Agriculture (OISPA, since 2001), the Childhood Agricultural Injury Survey (CAIS, since 1998), and the National Agricultural Workers Survey (NAWS injury module, since 1999), once routinely identified farm injuries, yet due primarily to unsustainable cost, were discontinued in 2015 [7]. Similarly, a work-related supplement to the National Electronic Injury Surveillance System (NEISS-Work) identified workplace injuries from a nationally representative sample of 67 hospital emergency departments beginning in 1998 [8], but was discontinued in 2023 due to lack of funding (NIOSH, personal communication, Erika Scott. 2023). The above underscores the need for a surveillance system that 1) passively identifies injuries to avoid burdening workers, businesses or healthcare systems; 2) provides ongoing data; and 3) is detailed enough to be useful but inexpensive enough to be sustainable.

Some surveillance efforts have employed pre-hospital care records to capture otherwise undocumented incidences of non-fatal agricultural injury [9, 10]. Pre-hospital care records (PCRs) are particularly useful for occupational analysis as they yield a higher rate of injury records than other administrative data sources and are available at relatively low cost [11]. The use of PCRs lends itself to machine learning techniques due to algorithmic efficiency in processing large quantities of data, much of it textual; facility with large dimensional problems, and ability to extract patterns from data, allowing for interpretation and prediction [12]. Supporting this idea, Lehto et al. (2009) showed that naïve Bayes methods functioned well to classify injury narratives from administrative databases [13] and subsequently demonstrated the utility of naïve Bayes in a public health surveillance setting, where the algorithm reduced manual review of records by 68% [14].

The complexity of health data, comprising images, free text, genomic sequences, electronic health records and more, has led to application of machine learning applications for real-time patient monitoring, resource planning and crowd management in the emergency department, and predictive modeling of injuries and admissions [15]. A 2022 review of occupational injury prediction [16] found that the most common machine learning algorithms used for text-based analysis were K-means and K-nearest neighbor (KNN) clustering algorithms, Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), and Decision Tree-based ensemble methods such as Random Forests (RF) and Gradient-Boosted Decision Trees (XGBoost). These methods were all effective in performing tasks such as classifying accident type, determining patient eligibility for clinical trials, identifying causal factors of occupational accidents, and predicting occupational injury outcomes from prehospital records or injury free-text data [16-18].

Given the variability in terminology and abbreviations used in injury reports, and frequent concerns over quality and quantity of data sources, it is unsurprising that no single learning method has consistently proven most effective for health applications [15]. Logistic Regression/Elastic Net (LR/Net), Random Forest (RF), and Support Vector Machine (SVM) techniques, however, have proven utility in benchmarking tasks relative to Naïve Bayes (NB). Marrucci-Welman et al. found that a regularized (penalized) logistic regression model had superior performance in classifying narratives from a large occupational administrative dataset when compared to NB, SVM, and single word and bi-gram models [19], a finding supported by a comparison of NB and LR in coding exposures leading to injury in free text narratives obtained from worker’s compensation claims [20]. Decision Tree based ensemble classifiers have also had success in accurate prediction of injury outcomes and severity when compared to NB model performance [21, 22], and like SVM, will handle potential interactions, while NB assumes independence between features. Support Vector Machines outperformed a Naïve Bayes approach as demonstrated by performance in classification of construction accident and near-miss narratives and transcriptions of emergency calls [23, 24].

The Northeast Center for Occupational Health and Safety (NEC) adopted a naïve Bayes (NB) classification approach and in 2021 documented that use of a NB classification algorithm reduced the time associated with identifying agricultural injury cases from PCRs, decreasing manual record review by 69.5% [10]. However, significant reviewer time was still required despite this reduction, with 1.7 full-time staff equivalents needed to visually inspect and code each record returned by the algorithm. Alternative machine learning methods coupled with rigorous cleaning and feature reduction may thus optimize the existing surveillance system by decreasing misclassification and further decreasing necessary manual review hours.

Regularized logistic regression, Support Vector Machine techniques or boosted decision tree-based classifiers may each outperform the existing Naïve Bayes approach to identifying agricultural injury cases in PCRs. The aim of this paper is to document, benchmark and compare the performance of four common machine learning algorithms in classification of injury cases from PCR data, to design an improved, more efficient surveillance system for nonfatal agricultural injuries.

Materials and Methods

Data. We assembled a labeled gold-standard dataset for the benchmarking exercise as follows; briefly, EMS records from the state of Maine from 2008 through May 2023 were pulled, and duplicates omitted and records excluded as described in detail by Hirabayashi and others [25]. Cleaning involved first removing duplicate records and records of no interest. Duplicates were identified based on matching four variables: sex, admission date, ZIP code, and date of birth; once identified, one duplicate was retained at random and the others omitted. Before exclusions and deduping, Maine records from this period comprised 3,720,304 records, and after deduping and exclusions, 2,520,180 records remained, of which 60,143 had been tagged as to occupational injury status. Our dataset comprises these 60,143 labeled records. See Supplemental Tables 1-2 for a summaries of raw, deduped and tagged records by year, and dedup/exclusion criteria. The Institutional Review Board (IRB) of the Mary Imogene Bassett Hospital approved all protocols. In addition, approval was also granted by Maine Data Use Board.

Outcome/labels. Labels comprised a four-level case-class variable, as follows: 0 (non-agricultural, non-traumatic/acute, or both), 1 (confirmed agricultural, confirmed traumatic/acute = true case), 2 (confirmed traumatic/acute, suspected agricultural), or 3 (suspected traumatic/acute, confirmed agricultural). To simplify the classification problem, this labeling system was converted to a “nothing versus everything” binary variable with 0 indicating confirmed non-agricultural non-traumatic/acute injuries, and 1 including the remaining three confirmed and suspected agricultural and/or acute case types (i.e., case-class 0 versus case-class 1, 2, or 3). The aim of this recoding is to facilitate correct classification of non-agricultural non-traumatic/acute injury cases (0), which constitute most (over 97 %) of our records. Even this substantially reduces the number of records that must be reviewed by human coders, who will then only be reviewing case-class types 1, 2, or 3.

Features. Narratives from the EMS records were used to create a list of stemmed keywords as shown in Table 1. The keyword list was developed by choosing agricultural terms from the National Occupational Research Agenda Agriculture, Forestry, and Fishing (NORA AFF) Dictionary [26]; identifying records containing these terms within a single year of data [2008], and reviewing a frequency table of all words contained within these records to identify additional agricultural keywords. The keyword list was finalized by continually assessing its ability to identify true cases [25]. EMS narratives were processed by removing punctuation, lowercasing and stemming all words using the Natural Language Toolkit's (NLTK) Snowball stemmer (v 3.5). For our purposes here, identified and obviously duplicate stems such as 'chain_saw' and 'chainsaw' were combined into a single binary variable. Additional features included sex, age, (determined from date of birth and incident date), zip code of residence (truncated to first three digits), incident location, mechanism of injury, dispatch reason, and primary impression. After age values were computed, ages less than zero or greater than 110 years are set to 'NA' and later omitted, hot-deck imputed or replaced by 'Unknown' depending on the strategy for missing units described below. All zip codes from Canada were assigned the level 'CAN', and American zip codes outside of the Northeastern states of Vermont, New Hampshire, Maine, Connecticut, and Massachusetts were assigned the level 'US.' Missing zip codes and those labeled 'unknown' were assigned to the level 'Unknown' (UNK).

Table 1. Stems. Keyword stems and distributions in the 60143 records comprising the gold standard dataset. A value of “1” indicates that the stem appears in a given record, and a value of “0” indicates that it does not appear. The 119 binary stems retained and shown below include at least one “stem = 1” value. All fully zero columns contribute no new information and were omitted. Table entries in italics are among the top 5 for variable importances for models in Tables 5AB.

Stem	Stem = 0	Stem = 1	Stem	Stem = 0	Stem = 1	Stem	Stem = 0	Stem = 1
wood	***53716***	***6427***	pig	59987	156	bobcat	60117	26
tree	53936	6207	dairi	59999	144	coveral	60117	26
limb	56357	3786	auger	60001	142	hoof	60118	25
yard	56710	3433	prune	60003	140	logger	60119	24
pen	57281	2862	farmer	60003	140	sprayer	60122	21
blade	57337	2806	ram	60007	136	agricultur	60123	20
feed	57430	2713	bind	60010	133	silo	60125	18
hors	*57549*	*2594*	implement	60010	133	amish	60127	16
farm	*57862*	*2281*	bunker	60012	131	chopp	60127	16
anim	58486	1657	*trough*	*60014*	*129*	cleanser	60127	16
cart	58630	1513	winch	60020	123	crop	60128	15
barn	*58656*	*1487*	forestri	60026	117	pipelin	60129	14
milk	58725	1418	spreader	60039	104	livestock	60130	13
tractor	*58740*	*1403*	splitter	60040	103	scraper	60131	12
chicken	58948	1195	corral	60041	102	*poultri*	*60132*	*11*
plow	59129	1014	pastur	60050	93	yearl	60132	11
irrig	59162	981	sanit	60054	89	beater	60133	10
gear	59241	902	hog	60056	87	methan	60133	10
fenc	59264	879	chute	60060	83	compost	60133	10
chain	59369	774	timber	60062	81	slaughter	60134	9
combin	59572	571	goat	60065	78	debark	60135	8
cabl	59646	497	defac	60066	77	udder	60136	7
hay	*59676*	*467*	gator	60068	75	kicker	60137	6
stall	59692	451	*greenhous*	*60071*	*72*	unhitch	60137	6
chainsaw	59697	446	fenc_post	60077	66	kicker	60137	6
buck	59710	433	buggi	60088	55	drive_line	60138	5
mower	59722	421	skidder	60095	48	*skidsteer*	*60138*	*5*
arch	59737	406	tie_down	60101	42	harrow	60138	5
turkey	59787	356	sheep	60102	41	fop	60139	4
wagon	59819	324	breed	60106	37	kickback	60139	4
bull	59823	320	manur	60108	35	forag	60140	3
calv	59864	279	pto	60108	35	guywir	60140	3
cow	*59885*	*258*	pesticid	60108	35	hoov	60140	3
hitch	59912	231	vacuum_pump	60108	35	sheav	60141	2
entangl	59926	217	fertil	60109	34	*silag*	*60141*	*2*
bale	59934	209	uncap	60109	34	slack_line	60141	2
straw	59944	199	skid_steer	60111	32	3pt_hitch	60141	2
shear	59970	173	bulldoz	60112	31	choker	60142	1
loader	59987	156	digger	60115	28	*post_hole_digger*	*60142*	*1*

For all categorical variables, any levels that included just one hit were combined with other similar levels to avoid level differences in test and training sets, and levels such as e.g., ‘blank value’, ‘other,’ ‘unknown’, ‘not applicable, ‘not available,’ ‘not recorded,’ ‘not reporting,’ ‘unspecified,’ were all recoded as ‘Unknown.’ For each stem, counts of 1 (stem/keyword appears in record) and 0 (stem/keyword does not appear) are shown in Table 1. Levels for multivariate categorical covariates are shown in Table 2.

Table 2. Features – Demographic/Other

Feature	Description	Levels
Sex	(3 levels)	Male/female/Unknown
age_cat	Age category (by quintile)	Q1, Q2, Q3 ,Q4, Q5
Dispatchreason	Reason for dispatch (38 levels)	Unknown, abdominal pain, Anaphylactic Reaction, Animal Bite, Assault, auto vs pedestrian, back pain (non-traumatic), breathing problem, burns, cardiac, chest pain, choking, CO poisoning / Hazmat, diabetic, drowning, electrocution, eye injury, fall victim, headache, heat/cold exposure, hemorrhage/laceration, medical alarm, MCI, overdose, poisoning, pregnancy-childbirth, respiratory arrest, sick person, Standby, industrial accident, stroke, stabbing-gunshot, psychiatric, traffic accident, traumatic injury, pain, seizure/convulsions, unconscious/fainting
Primaryimpression	Primary caregiver impression (40 levels)	Unknown, Anaphylaxis, Chest Pain, Cardiac, Death, Dehydration, Diabetic, Abdominal-GI, Electrocution, Fever, Hemorrhage, Hypotension, Hypothermia, COPD, CHF, Psychiatric, Transfer, Nausea, Pain-Nontraumatic, Overdose, Substance Abuse, Ophthalmological, OB-Gyn, Malaise, Fainting, Disoriented-confused, Respiratory, Inhalation injury, Shock, Stroke, Toxic exposure, Trauma-back, Trauma-burn, Trauma-head, Trauma-torso, Trauma-extremity, Trauma, Seizure/convulsions, Unconscious
mechinj	Mechanism of injury	Unknown, Blunt, Burn, Penetrating, None (no injury)
incident_loc	Incident location (13 levels)	Unknown, Airport, Business/service, Farm, HealthFacility, Home/Residence, Industrial, Lake/River/ Ocean, Mine/Quarry, Public Building, Residential institution, Street/Highway
zipcode	Zip code Truncated to first three digits (76 levels)	UNK, CA, US, 010 to 020, 021 to 030, 031 to 040, 041 to 050, 051 to 060, 061 to 070, 074, 076 to 080, 081 to 083, 085 to 089

Missing Data. Classification on incomplete data is common in data science, but it presents a challenge, with performance of the classifier affected most by missing units in the test set and declining rapidly as a function of test set missingness [27]. Missing data is typically classified into three types: MCAR, MAR, and MNAR. In type MCAR, data is 'missing completely at random,’ with missing units independent of both unobserved and observed features. When data are MAR, or ‘missing at random,’ missing units are associated with measured, but not unmeasured features [28]. If data are MNAR, or ‘missing not at random,’ the probability of missingness varies for unobserved reasons linked to the value of the missing observation itself [28, 29]. Here, standard approaches such as listwise deletion and complete case analysis may result in a biased classifier that systematically misses classes, or in omission of classes entirely. Incorrect imputation can also lead to misclassification, thus addressing MNAR or ‘nonignorable’ missingness may involve either creating a mechanistic missingness model, or using judgment and topical expertise to account for it [30]

Although our variables all include “Unknown” levels, they are otherwise complete except for two key variables: subject age and residence zip code. Due to missing values in the EMS record ‘date of birth’ (DOB) field, subject age (‘Age’), derived from DOB and incident date, has over 24% missingness (See Table 3), more than twice the acceptable level for list-wise deletion and complete case analysis. Zip code, with approximately 4% missingness, is less problematic, as we’ve assigned missing values to ‘Unknown’, an existing level as described above. The default choice for handling missing data in most statistical packages is list-wise deletion, omitting any rows with one or more missing units [31]. Provided the dataset is large, missingness is at least MAR and the percentage of missing units is not extreme, list-wise deletion can be a sensible choice. However when missing units are above 10-15%, even if confined to just a few key variables in a large data frame, this results in unreasonable loss of information [32], especially for classification applications on injury cases.

Table 3. Missing Data by Covariate

Covariate	Missing (%)
Age	14607 (24.3)
zip code	2521 (4.1)
incident location	322 (0.5)
dispatch reason	236 (0.4)

Multiple imputation (MI) is often the go-to solution for missing data in statistical analysis, but because this is a predictive study and we are not performing inference, MI is unnecessary, as well as difficult to implement in the context of machine learning [33]. MI is employed when we are interested in determining the uncertainty associated with imputed values, and is necessary when accurately determining variance for the purposes of inference. For our classification problem, our goal is empirical: we seek to replace missing units in a manner that returns stable, accurate predictions on as many samples as possible. We thus employ two strategies to retain samples that might otherwise be omitted due to missing units: first we impute age, incident location and dispatch reason using grouped hot-deck single imputation, where donors are randomly selected from existing distributions in specified columns grouped by zip code (first 3 digits). Donor-based methods such as grouped hot deck are fast, efficient and conservative methods that perform well for large datasets with mixed continuous and categorical values, and for predictive analytical methods that do not directly involve inference [32, 34, 35].

A second approach recodes all missing units as 'Unknown,' in the same way that small number of missing zip code data is handled above. The approach of capturing a formal 'missing for unknown reasons' category was suggested by the fact that it is more than twice as likely for a class = 1 label to have missing age data (5.4%) than a class = 0 label to have missing age data (2.1%, though note that this comprises many more counts since the label is zero-inflated), suggesting that missing age data may result from cases that are unresponsive at pick-up (i.e., due to injury severity), and thus missingness in age may be MNAR. Indeed, the full dataset includes N = 60143 samples, 1806 (3%) of which are case Label = 1 (that is, injury classes 1, 2, or 3), and 58337 are case Label = 0 (non-cases). In complete case data, N = 44566 samples contain just 891 (2%) Label = 1 samples. Thus greater than half (51%) of case Label = 1 samples have missing age data (DOB information) and are omitted in a listwise deletion complete case analysis scheme, which is necessary for methods such as penalized regression.

Both imputation methods yield completed datasets on 60143 samples with 156 features before one-hot encoding. The standard 'complete case' approach omits all records with missing age data, resulting in 44566 samples with 156 features prior to one-hot encoding. Analysis will be performed on 'complete case' data, on data with missing units recoded as 'Unknown,' and on hot-deck imputed data and the results compared.

Learning methods. Our features are comprised primarily of zero-inflated keyword stem variables (140 bivariate variables), with additional multilevel categorical variables capturing sex (bivariate), year (14 levels), injury mechanism (5 levels), location of the injury incident (13 levels), reason for emergency dispatch (41 levels), caregiver first impression of the victim (42 levels), zip code (64 levels), and lastly age (continuous, derived from birth date and incident date). Approximately 25% of the zip code levels are sparsely populated, including just one or two counts, which can present a challenge for training a model as levels may appear in the training set but not the test set, and vice versa. A solution is to combine or omit levels with single cases, categorize age by quintile, and then fully one hot code (dummy code) all categorical data for all model families. Several of our models require one-hot or dummy coding, so we will train and test all models on one-hot coded data to facilitate comparison between results. Without one-hot coding, data includes 156 features. Once the data are fully one-hot coded and all-zero columns are omitted, the dataset expands to approximately 314 (318 features for missings coded "Unknown") primarily sparse, zero-inflated binary categorical features, and this presents a challenge to many algorithms.

Learning methods employed here include Naïve Bayes, regularized logistic regression (elastic net with L_2 and L_1 norm penalties), Support Vector Machine and XGBoost (gradient-boosted decision trees) classifiers. The aim of this benchmarking exercise is two-fold: to identify and rank the variables that provide each classifier with information critical to accurate classification (compute variable importance), and to select a classifier from these examples that accurately classifies at least 90% of non-cases (zero class members) with minimum false positives (that is, identifying a 1, 2, or 3 class as a zero class).

Naïve Bayes. Our reference algorithm is naïve Bayes method (Scott et al., 2021). Naïve Bayes models comprise a family of classifiers based on Bayes theorem, and operate under the assumption that pairs of features used to classify observations are conditionally independent. Naïve Bayes is robust, reliable, fast, and is one of the most popular algorithms used for text classification [13, 36]. Some of our categorical variables have a large number of possible levels such that certain combinations of level values and class labels can result in zero probabilities. We address this issue, which is known as the 'zero frequency' problem, by applying additive smoothing – essentially adding a minimal number of counts (1) for every feature level – class label combination. Introducing this smoothing also regularizes the model. Permutation variable importances for naïve bayes are computed using a model-free scheme that employs the ROC curve.

Penalized/Regularized Logistic Regression (elastic net). A penalized logistic regression model imposes penalties on the logistic model for having features that contribute little to the model fit, resulting in shrinkage of associated coefficients to zero, also known as regularization. We will fit an elastic net model, which applies both L_1 (Taxicab) and L_2 (Euclidian) norm based penalty functions to the model. Variable importance metrics for regularized Logistic Regression are computed by taking the absolute value of the t-statistic for each feature.

Support Vector Machine (SVM). Support vector machines have been highly effective with other surveillance applications [37] and can perform binary and multiclass classification, though the latter is performed by running successive binary classifications. They work by constructing optimal hyperplanes in feature space to separate two classes and are especially good choices when the number of features exceeds the number of samples. Significant downsides for this application are that they are designed to perform on continuous features, thus categorical data must be one-hot encoded, and that they predict classes rather than probabilities. SVM does not directly provide a measure of permutation variable importance, but one may be estimated using recursive feature elimination, or by extracting coefficients and comparing their absolute values relative to the maximum. We use the former approach as it is closest to permutation importance.

Gradient-Boosted Decision Trees (XGBoost). XGBoost, an ensemble learning method, is a popular algorithm for classification and regression applications because it handles large (wide and deep) datasets efficiently, and provides high accuracy. XGBoost iteratively trains ensembles of weak learners - “stumps” or shallow decision trees - and with each fitting round uses residuals from a prior round to fit the next ensemble. The algorithm effectively handles missing values, and is robust against overfitting. Variable importances for tree-based methods are computed via permutation importance, based on permuting out-of-bag (OOB) data subsets for each individual tree. Prediction accuracies on the OOB subset are computed for each permuted OOB dataset, then each predictor is permuted (shuffled) within the OOB dataset, and accuracy is computed again. For each predictor, the differences between the two accuracies are averaged over all trees, and normalized by standard error. Per-predictor importances are summed over each boosting iteration.

Chained Models. Model chaining or stacking can be an efficient way of expanding the accuracy and improving the speed of a machine learning workflow. Here, we break down our classification problem into two parts: first, apply a fast machine learning model (e.g., one of NB or Elastic Net) for feature reduction to reduce the number of sparse and zero-inflated features to only those that contribute information to the classifier, and selecting the top 100 features determined via variable importance. We then apply a second more accurate and computationally intensive classifier such as XGBoost on the reduced dataset. Based on similarity between variable importances for NB and XGBoost as described below, we'll use NB for our exploratory feature reduction stage. Since we are using one model primarily for feature reduction and the second for classification, we do not have to worry about error propagation between models.

Performance Metrics

For each method we will report variable importances as described above for training data, as well as accuracy (fraction of correct predictions), sensitivity, specificity, and area under the receiver-operator curve (AUC) for both training and testing sets. Variable importances from training data rank the value of each feature in predicting whether it has utility for a given classifier in predicting cases [38]. The trapezoidal rule is used to compute AUC from the Receiver-Operator Curves (ROC), which we also display for models fitted on training data, and predictions from test data. We also compute the False Positive Rate (FPR*) required to capture a critical 90% of non-cases (TPR*). Reliably identifying and segregating non-cases (class = 0) from true cases and potential cases (class =1, 2, 3) substantially reduces the labor by human coders that is required to identify true cases (class = 1).

Data preparation and initial cleaning was done in SAS; further tidying was performed in R tidyverse; imputation was performed in R (version 4.3.4) using the VIM [39] and devtools packages, and machine learning was performed in R using the caret [40], mlbench [41], and e1071 packages that call the naïvebayes [42], glmnet [43, 44], and XGBoost [45] packages. All models are trained and tested on fully one-hot coded complete case, imputed and "missings-replaced" data, on parameter grids with 5-fold cross-validation, repeated 10 times, using a 75:25 train-test split, and employing stratified random sampling (stratified on the label) to ensure that train and test splits included the same approximate proportion of cases and non-cases.

Results

Of 60,143 cleaned and labeled records from 2008 through May 2023, after listwise deletion of missing units, 44,566 records were complete and were used in the complete case study (Table 4).

Since deleting records also deletes cases that we hope to capture, we applied two strategies for retaining records with missing units: simply recoding missing units to “Unknown,” and applying hot-deck imputation. All records were used in the imputed and replaced studies. All

Table 4. Performance metrics, Complete Case.
Performance Metrics on training set (N = 33425 samples) and prediction accuracy on testing data (N = 11141), complete case study. Train-Test split is 75:25 and uses stratified sampling due to highly imbalanced label distribution. Performance metrics on training set (N = 45105 samples) and prediction accuracy on testing (N = 15035) data.

Model Family	Metric	Training	Testing
Naïve Bayes	Accuracy	0.98	0.98
	Sensitivity	0.98	0.99
	Specificity	0.43	0.43
	AUC	0.94	0.93
Regularized Logistic (Elastic Net)	Accuracy	0.98	0.98
	Sensitivity	0.99	0.98
	Specificity	0.69	0.68
	AUC	0.97	0.95
[1]Support Vector Machines (SVM)	Accuracy	0.98	0.98
	Sensitivity	0.98	0.98
	Specificity	0.84	1
	AUC	N/A	N/A
XGBoost	Accuracy	0.99	0.98
	Sensitivity	0.98	0.98
	Specificity	0.80	0.67
	AUC	0.99	0.95

[1]Classifies most non-cases (Label = 0) correctly, hence high specificity, but tends to misclassify cases (Label = 1). AUC is not available for SVM and must be estimated via kernel.

Table 5. Performance metrics, Hot deck imputed.
Performance metrics on training set (N = 45105 samples) and prediction accuracy on testing (N = 15035) data. Any missing units are imputed using hot-deck single imputation grouped on zipcode. Train-Test split is 75:25 and uses stratified sampling due to highly imbalanced label distribution**.**

Model Family	Metric	Training	Testing
Naïve Bayes	Accuracy	0.96	0.96
	Sensitivity	0.98	0.98
	Specificity	0.39	0.41
	AUC	0.90	0.88
Regularized Logistic (Elastic Net)	Accuracy	0.97	0.97
	Sensitivity	0.98	0.975
	Specificity	0.60	0.63

	AUC	0.92	0.92
[1]Support Vector Machines (SVM)	Accuracy	0.97	0.97
	Sensitivity	0.97	0.97
	Specificity	1.0	NA
	AUC	NA	NA
XGBoost	Accuracy	0.98	0.97
	Sensitivity	0.98	0.98
	Specificity	0.87	0.69
	AUC	0.97	0.95

[1]Classifies most non-cases (Label = 0) correctly, hence high specificity, but tends to misclassify cases (Label = 1). AUC is not available for SVM and must be estimated via kernel.

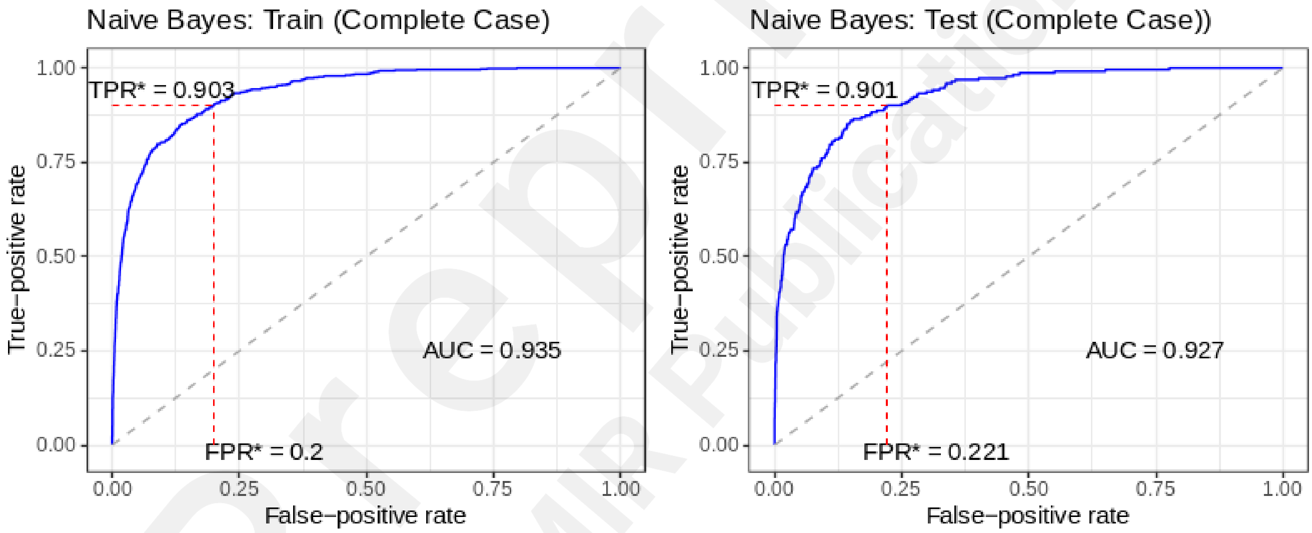
benchmarked methods performed well on the cleaned and labeled data, returning accuracies of at least 0.98 on complete case training and testing data, and accuracies between 0.96 and 0.98 on hot deck imputed data (Table 5).

As both speed and accuracy on large datasets are considered important here, there is no straightforward winner, given that we observe the usual trade-off between model complexity and computation speed. Simple models such as naïve Bayes and regularized regression (elastic net) run at light speed, but generally have slightly lower accuracy than much slower decision tree-based methods such as XGBoost, and often much lower AUC values, especially on imputed data. AUC is highest for tree-based methods across all missingness treatments, and elastic net and support vector machines (SVM) also showed improved accuracy over naïve Bayes (NB). NB, SVM and XGBoost all return qualitatively the same results for variable importance, especially for the top 5 to 10 variables, a mixture of keywords (stems) and other features, while logistic regression largely returns stems. Generally, NB yielded the poorest performance on training and testing data for specificities, followed by penalized logistic regression, and tree-based methods provided the best performance. However, chained models using NB for feature reduction followed by XGBoost for subsequent classification on a reduced feature set performed well.

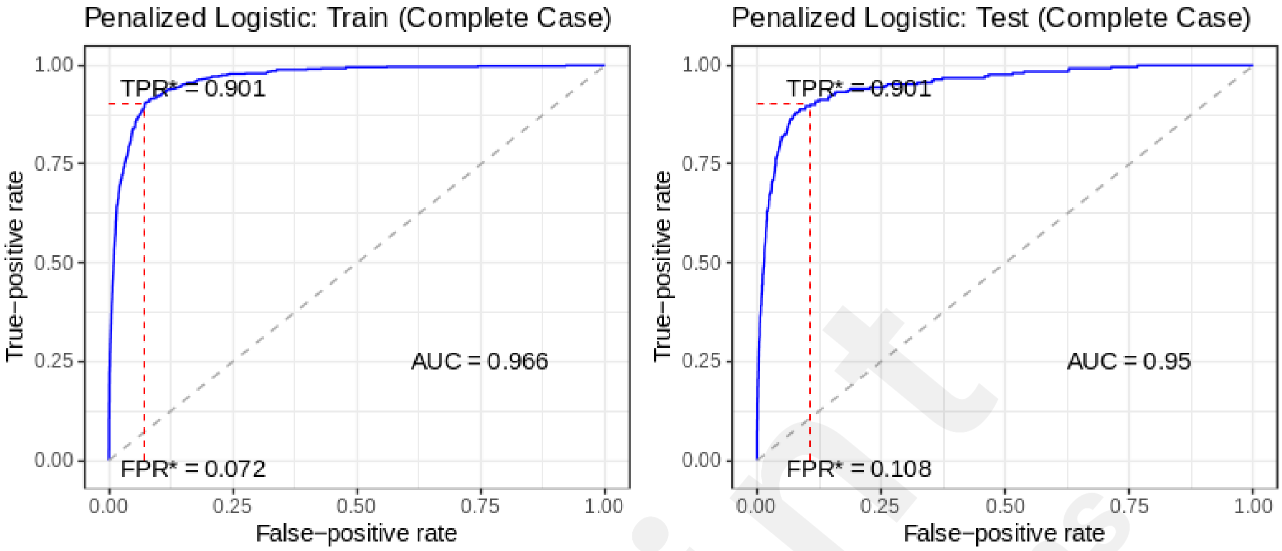
Training. Training and cross-validating a model on the naïve Bayes algorithm had the shortest run-time, concluding in minutes where tree-based methods took several hours.

Complete case data produced the best results (see Figure 1 for ROC curves, also Table 4); all models had similar accuracies (~0.98 or better), though because specificity was often much lower, ranging from 0.43 for NB to 0.84 for SVM, AUC varied from 0.94 (NB) to 0.99 (XGBoost). SVM is not a probability-based classifier, and we were not able to produce ROC curves from SVM models fit on this dataset. SVM classifies most non-cases correctly, hence high specificity, but tends to misclassify cases (Label = 1). We were not able to compute ROC curves for SVM on this dataset (Table 4).

Figure 1. Complete Case.
Receiver-Operator Characteristic (ROC) curves for training and testing sets, trained and tested on labeled Maine data (2008-2023), complete case. Area under the curve (AUC) is shown on each panel, as well as the False Positive Rate (FPR*) required to return a True Positive rate (TPR*) of 0.90, indicated in dashed red line.

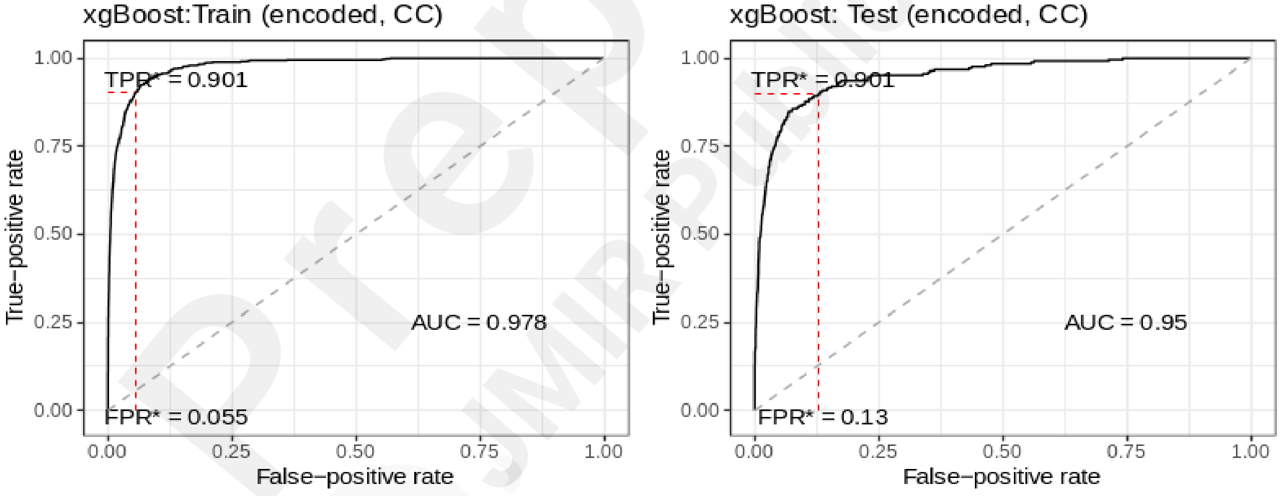


A. **Naïve Bayes Model (reference).**



B. Penalized Logistic (elastic net) Model.

XGBoost: Train (complete case) XGBoost: Test (complete case)

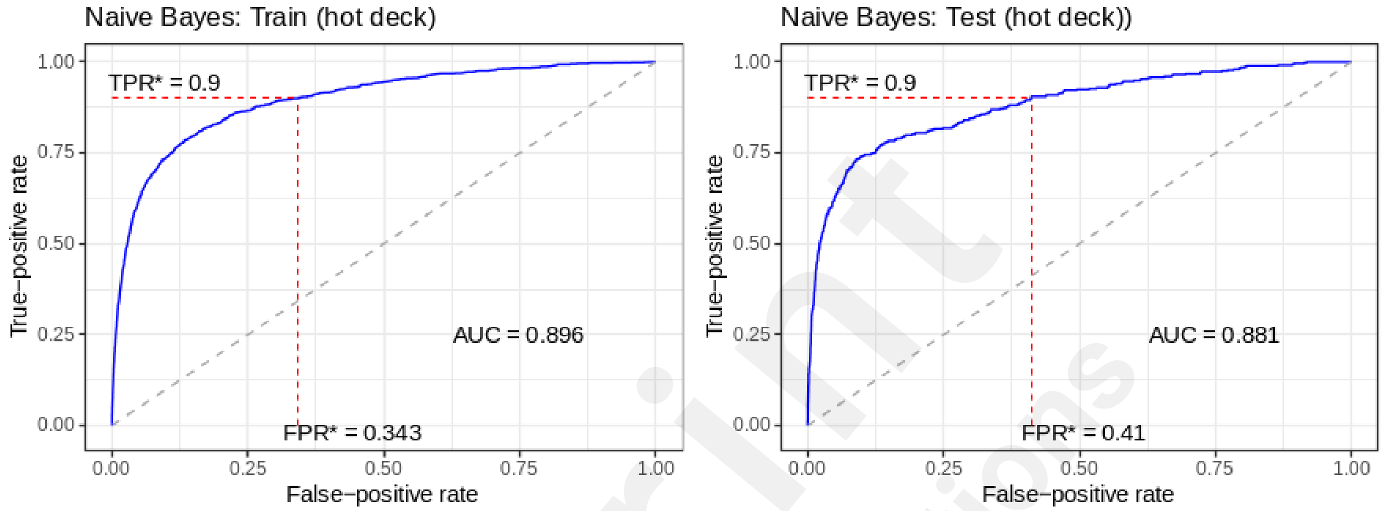


C. XGBoost model (one-hot encoded).

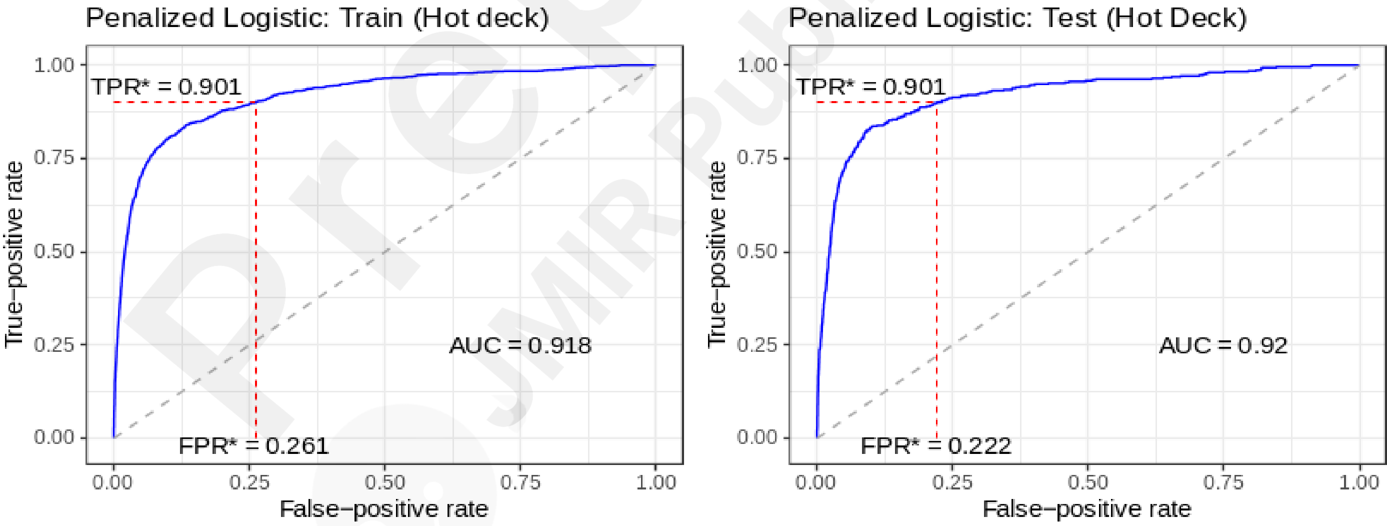
On *imputed data* (see Figure 2; Table 5), accuracies were lower for all methods, ranging from 0.96 for NB to 0.98 for XGBoost, with AUC values ranging from 0.89 (NB) to 0.96 (XGBoost). On data with *missing units recoded to "Unknown,"* accuracies range from 0.95 (NB) to 0.98 (XGBoost), with slightly higher AUC values, ranging from 0.90 (NB) to 0.97 (XGBoost) (Figure 1; Table 2, Supplemental Information).

Figure 2. Hot-deck imputed.

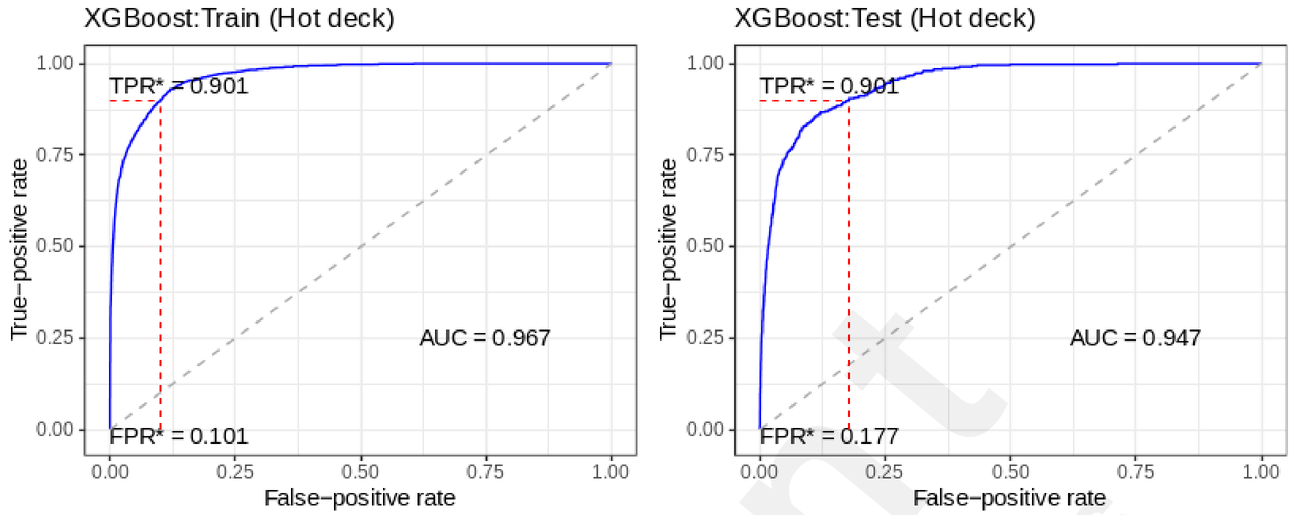
Receiver-Operator Characteristic (ROC) curves for training and testing sets, trained and tested on labeled Maine data (2008-2023), hot-deck single imputed. Area under the curve (AUC) is shown on each panel, as well as the False Positive Rate (FPR*) required to return a True Positive rate (TPR*) of 0.90, indicated in dashed red line.



A. Naïve Bayes Model (reference).



B. Penalized Logistic (elastic net) Model



C. XGBoost model

Predictions. Performance metrics for predictions on test data were similarly optimal for *complete case data* (Figure 1; Table 4). Accuracies were 0.98 across all models, and AUC values varied from a low of 0.93 (NB) to a high of 0.95 (XGBoost). On *hot-deck imputed data* (Figure 2; Table 5), accuracies range narrowly between 0.96 (NB, Elastic Net) and 0.97 (SVM, XGBoost), with AUC values ranging from 0.91 (NB) to 0.94 (XGBoost) (Table 5). On *data with recoded unknowns*, AUC ranges from 0.89 (NB) to 0.94 (XGBoost) (Figure 1 and Table 2, Supplemental Information). While all models show drops in specificity between training results and results predicted on test data, the largest drops in specificity between training and testing are for XGBoost, suggesting some amount of overfitting for this method, which may still produce the best results. While other methods show some variability in accuracy and AUC, depending on treatment of missing units, XGBoost produces the most robust result across missingness treatments.

Required False Positive Rates. A summary of False Positive Rates (FPR*) required to return a True Positive rate (TPR*) of 0.90 for the training and testing scenarios on complete case and all missing data treatments is shown in Table 6. As above, complete case data produces the best results on both training and test data. For the complete case study on training data, the XGBoost model returns an FPR* of only 0.055 (5.5%), with elastic net returning an FPR* of 0.07 (7%), much improved over the FPR* of 0.19 returned by NB.

Table 6. Summary of False Positive Rate (**FPR***) required given a desired True Positive Rate (**TPR***) of 0.90 for classification of zero cases for each trial model and missing data strategy. Area under the ROC curve for each scenario is also shown. Imputed and recoded/replaced datasets both include N=60,143 samples. Models trained on 5-fold cross-validation, with 10 repeats. Chained model consists of feature reduction to 100 most important variables using naïve Bayes, then the reduced dataset pushed through the XGBoost classifier.

Strategy	Model	**Training**			**Testing**		
		AUC	Required TPR*	Necessary FPR*	AUC	Required TPR*	Necessary FPR*
Complete Case N=44,566	Naïve Bayes	0.935	0.90	0.20	0.93	0.90	0.22
	Elastic Net	*0.97*	*0.90*	*0.07*	*0.95*	*0.90*	*0.11*
	XGBoost	0.99	0.90	0.055	0.95	0.90	0.12
	Chained models	*0.98*	*0.90*	*0.05*	*0.96*	*0.90*	*0.11*
Missing recoded “Unknown”	Naïve Bayes	0.90	0.90	0.29	0.89	0.90	0.38
	Elastic Net	0.93	0.90	0.21	0.92	0.90	0.26
	XGBoost	*0.96*	*0.90*	*0.13*	*0.93*	*0.90*	*0.22*
	Chained Models	*0.98*	*0.90*	*0.092*	*0.94*	*0.90*	*0.21*
Hot deck imputed	Naïve Bayes	0.90	0.90	0.34	0.88	0.90	0.41
	Elastic Net	0.93	0.90	0.23	0.92	0.90	0.21
	XGBoost	0.98	0.90	0.10	0.95	0.90	0.18
	Chained models	*0.98*	*0.90*	*0.095*	*0.97*	*0.90*	*0.12*

On testing data, the rankings are reversed, with the elastic net model returning an FPR* of 0.10 (10%) and XGBoost returning 0.12 (12%). For the missingness treatments, results from hot-deck imputed data are superior to those from “missing recoded as unknown” across all model families (Table 6). Here, XGBoost performs best, returning an FPR* of 0.125 (12.5%) on training data and 0.18 (18%) on the test set, with elastic net a close second at 0.23 on the training set and, surprisingly, 0.21 (21%) on the test set. Note that XGBoost performs well in the presence of missing data [46], and results are robust to the choice of imputation method [47].

Variable Importance. Visualizations of the 30 most important features for each method on complete case training data, scaled to 100 for comparison, are shown in **Figure 3**. Tree-based methods tend to be more discriminatory, relying on fewer features more heavily; note the narrow “trunk” of the visualization for XGBoost, compared with other methods (Figure 3D), and e.g., note complete case importances ranging from 100 to 3.6 for the first 20 ranked variables for XGBoost, while naïve Bayes, elastic net and SVM have importance values ranging from 100 to ~20 - 30 for the top 20 features (Figure 3). Examining the top five features, NB and SVM return the same features (Figure 3), and in the same order: ‘stem_tractor,’ ‘incidentlocation: farm,’ stem_barn,' 'mechinj: blunt', sex: male for imputed data (Supplemental Figure 2). The order of features returned XGBoost varies by data preprocessing/imputation, but 'stem_tractor' is first, 'incidentlocation: farm,' is in the top three, and 'stem_barn' in the top four (see Figure 3, Supplemental Figure 2). Of note is that NB, SVM, and XGBoost tend to identify as important features those stem words with relatively high 'Stem = 1' counts (counts of how often that stem appears in records) such as 'Stem_farm', 'Stem_tractor', 'Stem_barn' whereas the elastic net algorithm prefers relatively low counts, such as 'Stem_post_hole_digger', 'Stem_silag' and 'Stem_skidsteer' (see Table 1 for stem counts). This may be due to the fact that elastic net is the only method applied here that operates under assumptions of linearity (it is a generalized linear model), and the metric used for variable importance may be sensitive to outliers.

Model chaining. A summary of results from chaining a naïve Bayes model (feature selection at top 100 features for this algorithm) followed by the XGBoost classifier are shown in Table 6. As above, accuracies and AUC values are higher for the complete case chained model study on the 100 complete case features in the training phase (Figure 4), though for testing phase, we obtain nearly equivalent AUC and FPR* values from running chained models on hot-deck imputed data (Figure 4).

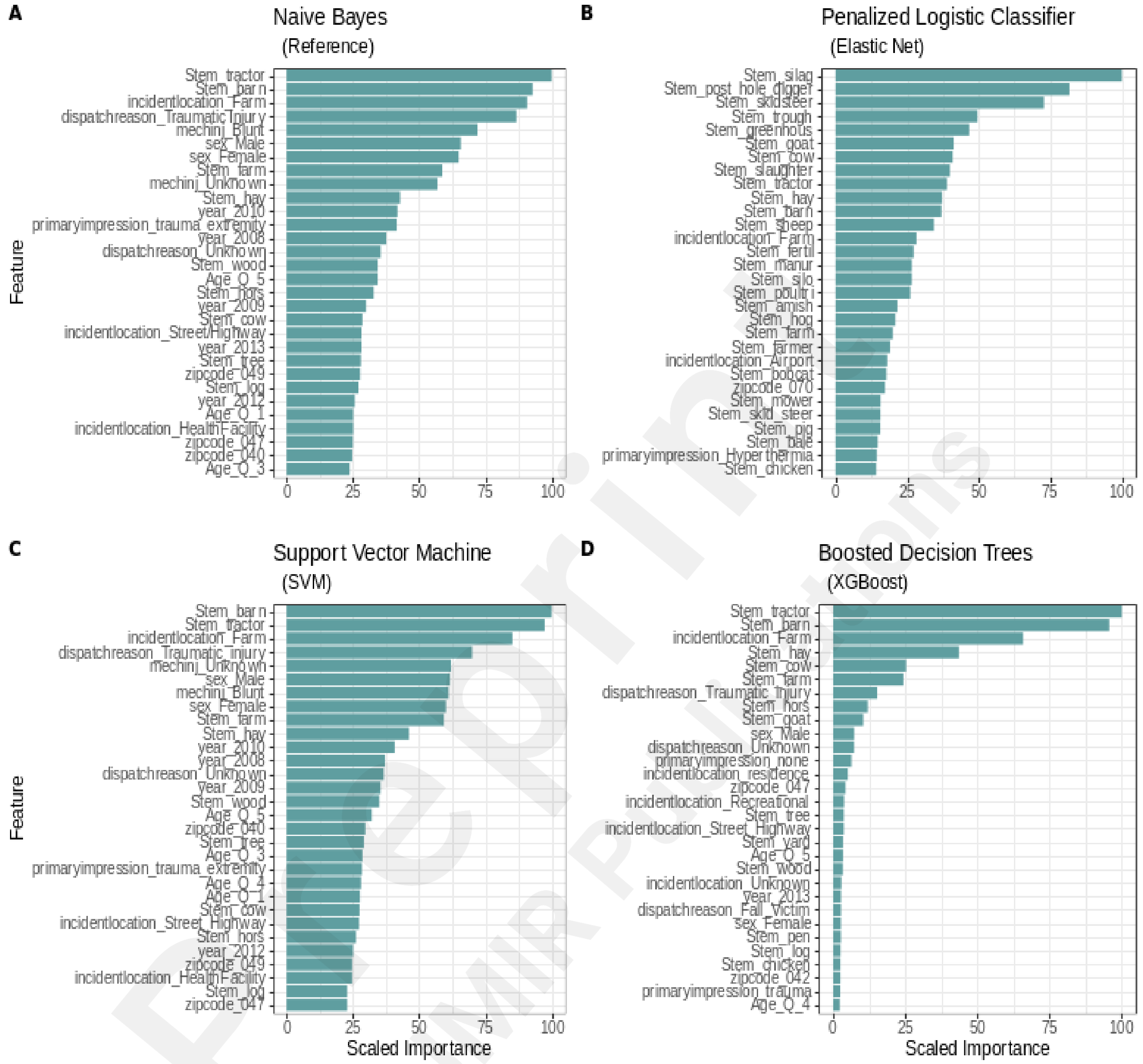
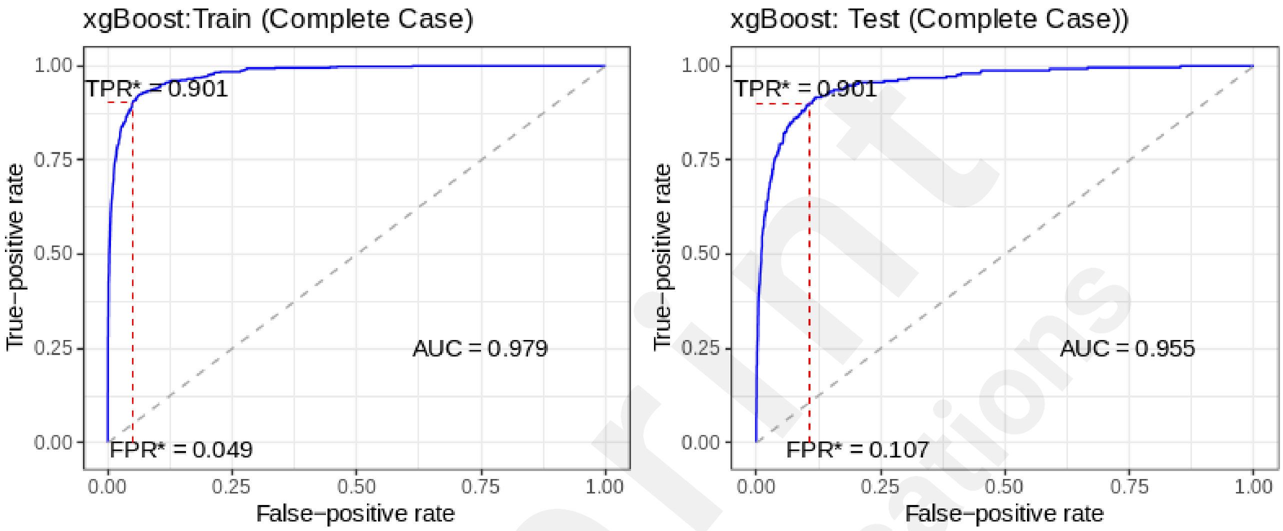


Figure 3.

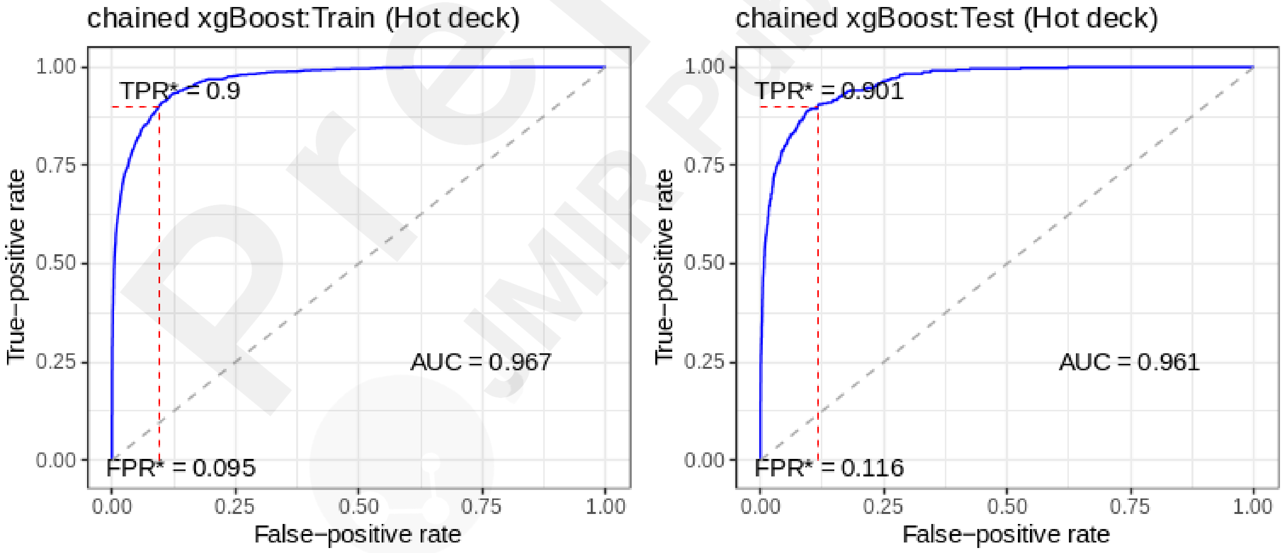
Top 30 complete case feature importances compared for (A) Naïve Bayes (reference model); (B), Penalized Logistic; (C) Support Vector Machine, and (D) Boosted Decision Tree (XGBoost) classifiers.

Figure 4. Chained Models

Receiver-Operator Characteristic (ROC) curves for training and testing sets, trained and tested on labeled Maine data (2008-2023), chained models: feature reduction via naïve Bayes followed by classification by XGBoost. Train-Test split is 75:25 (Train N = 45105 samples; Test N = 15035) using stratified sampling due to imbalanced label distribution. Area under the curve (AUC) is shown on each panel, as well as the False Positive Rate (FPR*) required to return a True Positive rate (TPR*) of 0.90, indicated in dashed red line. **A.** complete case data (N=44566); **B.** Hot-deck single imputed (N=60143).



A. Chained XGBoost classifier, complete case data



B. Chained XGBoost classifier, hot deck imputed data

Discussion

We have demonstrated, on labeled and clean data, that we may improve case retrieval and further reduce the burden of identifying agricultural injury cases from PCRs by switching the algorithm(s) used. Our results directly contribute to the Centers for Disease Control and Prevention – National Institute for Occupational Safety and Health (CDC-NIOSH) mission of improved leveraging of existing data for agricultural injury surveillance [48]. Furthermore, given the most recent National Academies of Sciences report on occupational health and safety surveillance emphasized coordinated, cost-effective approaches [49] our approach, using no-cost data, is responsive to this recommendation, especially in comparison to costlier past surveillance systems such as routine surveys.

In order to provide a level field of competition for the four algorithms, we competed on "square" data frames only, thus on complete case and imputed or replaced data. In theory, running (for example) a generative classifier such as naïve Bayes on data with high levels of missing data in one or more features should be possible without additional processing, since omissions are internal: the algorithm assumes (correctly or not) that all features are independent, and computes conditional probabilities independently, using only complete rows for each feature. Naive Bayes is said to be "missing data-resistant" (Farhangfar et al., 2008), producing accurate classification in the presence of missing data, provided feature independence assumptions are met, and the implementation platform supports automatic by unit or by row omission of missing units. However, high levels of missingness, non-independence between features, and MAR (where missingness in one feature is systematically related to observed features) or non-ignorable missingness can however result in reduced classification accuracy with this algorithm.

While all the algorithms competed well in terms of accuracy on complete case data, row-wise omission of missing data results in a loss of almost 25% and, for this dataset, over half of cases. Thus a reliable imputation strategy is essential for retaining data in the classifier and accurately classifying it. Of the strategies tested, grouped hot-deck imputation performs best for all models in the testing and prediction exercise, yielding overall the highest AUC values and lowest necessary False Positive Rates across the tested methods for imputed or replaced data.

The winning algorithm for this particular dataset is clearly XGBoost, followed closely by the penalized logistic/elastic net model, the latter with added attraction of computation speed. However XGBoost has the benefit of accommodating interactions or other nonlinearities, being robust in the presence of missing data and also relatively robust to choice of imputation method. The primary drawback to XGBoost on a wide (315 features) and relatively sparse dataset has been lengthy computation time. However, the similarity between variable importances returned by several of the methods (Naïve Bayes, SVM, and XGBoost) provided the final and more efficient approach of model chaining: start with a fast model (e.g., NB) to perform feature reduction by selecting the top 100 features for this method, followed by training the reduced data set on XGBoost, a slower method that is robust to imputation methods. A chained model produces equivalent results to running XGBoost on the full (imputed) dataset, and with reduced computation time.

Limitations

Our surveillance method identifies injuries where EMS were called and where the resulting records contained features, including keywords, related to agriculture. We thus omit cases where medical treatment was sought without EMS involvement, such as injuries transported to the hospital in a private vehicle. Missing data are a large and inescapable part of our challenge, especially since injured persons may be unresponsive at pickup, and records from these cases may not include information such as a reliable birth date. We rely on a simple and efficient donor-based grouped hot-deck single imputation method that has performed well for other prediction tasks, but another as yet determined method may be more robust.

Comparisons with Prior Work

As described in Scott et al., 2021, we successfully employed a naïve Bayes model trained on tagged data from Maine and New Hampshire years 2008 – 2010 to predict true cases in data from Maine (2011 – 2016) and New Hampshire (2011 – 2015), reducing the volume of records requiring visual inspection by two-thirds over a prior keyword search-based strategy. This study did not however explicitly summarize or address missing data or a strategy for managing it, and employed only one algorithm (NB) with relatively lower accuracy than the remaining four strategies benchmarked here on tagged and cleaned data.

Conclusions

Reliance on a machine learning method that is robust to missingness and the imputation method used to address it is the best approach to accurately capturing and classifying sparse and zero-inflated surveillance data from free-text records.

Abbreviations

AgFF – Agriculture, Forestry, and Fishing

AUC – area under the receiver-operator curve

CAIS – Childhood Agricultural Injury Survey

CDC – Centers for Disease Control and Prevention

CPU – central processing unit

DOB – date of birth

EMS – pre-hospital free-text records

FPR* – False Positive Rates

GPU – graphics processing unit

IRB – Institutional Review Board

KNN – K-means and K-nearest neighbor

LR – Logistic Regression

LR/Net – Logistic Regression/Elastic Net

MAR – missing at random

MI – multiple imputation

MNAR – missing not at random

NAWS – National Agricultural Workers Survey

NB - Naïve Bayes

NEC – Northeast Center for Occupational Health and Safety

NEISS-Work – National Electronic Injury Surveillance System

NIOSH – National Institute for Occupational Safety and Health

NLTK – Natural Language Toolkit

NORA AFF – National Occupational Research Agenda Agriculture, Forestry, and Fishing

OOB – out-of-bag

OISPA – Occupational Injury Surveillance of Production Agriculture

OSHA – Occupational Health and Safety Act

PCRs – Pre-hospital care records

RF – Random Forests

ROC – receiver operator characteristic curve

SAS – Statistics Analysis System

SOII - Survey of Occupational Injuries and Illnesses

SVM – Support Vector Machine

TPR* – True Positive Rate

UNK – Unknown

VIM – Visualization and Imputation of Missing Values

XGBoost – Gradient-Boosted Decision Trees

Author contributions

LEJ: conceptualization, data curation and cleaning, analytical plan, software, analysis, writing (original draft), manuscript development and revision; ES: funding, project management, manuscript development and revision; NK: data selection, curation and cleaning; MK: conceptualization, data curation, manuscript development and revision; CH-R: project management, manuscript development and revision, and PJ: conceptualization, manuscript development and revision. All authors read and approved the final manuscript.

Availability of data and material

The data analyzed here are available in raw form from the Maine Bureau of Emergency Medical Services, but restrictions apply and it is used under license for the current study. Those interested in applying may contact the Maine EMS Bureau.

Code availability

The workflow and code developed by the authors is available by a written request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Ethical approval

This study was approved by the Institutional Review Board of the Mary Imogene Bassett Hospital (Bassett Medical Center).

Acknowledgements.

We thank Alex Levenstein for helpful discussions during development of this phase of the project. Funding for this study was provided by the Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health (CDC-NIOSH) Grant No. 2U54OH007542.

References

[1] (2023). *National Census of Fatal Occupational Injuries in 2022* . [Online] Available: https://www.bls.gov/news.release/pdf/cfoi.pdf

[2] (2023). *Civilian occupations with high fatal work injury rates, 2022*. [Online] Available: https://www.bls.gov/charts/census-of-fatal-occupational-injuries/civilian-occupations-with-high-fatal-work-injury-rates.htm

[3] (1970). *Occupational Safety and Health Act of 1970. U S Public Law 91-596.* [Online] Available: https://www.osha.gov/laws-regs/oshact/completeoshact

[4] T. W. Kelsey, "The agrarian myth and policy responses to farm safety," (in eng), *Am J Public Health*, vol. 84, no. 7, pp. 1171-7, Jul 1994, doi: 10.2105/ajph.84.7.1171.

[5] J. P. Leigh, J. P. Marcin, and T. R. Miller, "An estimate of the U.S. Government's undercount of nonfatal occupational injuries," *J Occup Environ Med*, vol. 46, no. 1, pp. 10-8, Jan 2004, doi: 10.1097/01.jom.0000105909.66435.53.

[6] J. Leigh, J. Du, and S. McCurdy, "An estimate of the U.S. government's undercount of nonfatal occupational injuries and illnesses in agriculture," *Annals of Epidemiology*, vol. 24, no. 4, 2014, doi: https://doi.org/10.1016/j.annepidem.2014.01.006.

[7] (2015). *Looking to the future for agriculture injury surveillance at NIOSH - Agriculture, Forestry and Fishing Resources.* . [Online] Available: http://www.cdc/gov/niosh/agforfish/aginjurysurv.html

[8] (2018). *Consumer Product Safety Commission NEISS Hospitals*. [Online] Available: https://www.cpsc.gov/s3fs-public/NEISS_Hospital_Map_2018.pdf?6gAfTlFla.YEZWTkBH5hF6zcHm.1eweZ

[9] K. Patel, S. Watanabe-Galloway, R. Gofin, G. Haynatzki, and R. Rautiainen, "Non-fatal agricultural injury surveillance in the United States: A review of national-level survey-based systems," *Am J Ind Med*, vol. 60, no. 7, pp. 599-620, Jul 2017, doi: 10.1002/ajim.22720.

[10] E. Scott, L. Hirabayashi, A. Levenstein, N. Krupa, and P. Jenkins, "The development of a machine learning algorithm to identify occupational injuries in agriculture using pre-hospital care reports," *Health Inf Sci Syst,* vol. 9, no. 1, p. 31, Dec 2021, doi: 10.1007/s13755-021-00161-9.

[11] E. Scott, E. Bell, L. Hirabayashi, N. Krupa, and P. Jenkins, "Trends in Nonfatal Agricultural Injury in Maine and New Hampshire: Results From a Low-Cost Passive Surveillance System," *J Agromedicine,* vol. 22, no. 2, pp. 109-117, 2017, doi: 10.1080/1059924X.2017.1282908.

[12] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized machine learning techniques for prediction of occupational accidents," *Computers & Operations Research,* vol. 106, pp. 210-224, 2019, doi: https://doi.org/10.1016/j.cor.2018.02.021.

[13] M. Lehto, H. Marucci-Wellman, and H. Corns, "Bayesian methods: a useful tool for classifying injury narratives into cause groups," *Inj Prev,* vol. 15, no. 4, pp. 259-65, Aug 2009, doi: 10.1136/ip.2008.021337.

[14] H. R. Marucci-Wellman, M. R. Lehto, and H. L. Corns, "A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Naive Bayes algorithms," *Accid Anal Prev,* vol. 84, pp. 165-76, Nov 2015, doi: 10.1016/j.aap.2015.06.014.

[15] G. Chenais, E. Lagarde, and C. Gil-Jardine, "Artificial Intelligence in Emergency Medicine: Viewpoint of Current Applications and Foreseeable Opportunities and Challenges," *J Med Internet Res,* vol. 25, p. e40031, May 23 2023, doi: 10.2196/40031.

[16] M. Z. F. Khairuddin *et al.*, "Predicting occupational injury causal factors using text-based analytics: A systematic review," *Front Public Health,* vol. 10, p. 984099, 2022, doi: 10.3389/fpubh.2022.984099.

[17] M. Z. F. Khairuddin *et al.*, "Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance," (in eng), *Int J Environ Res Public Health,* vol. 19, no. 21, Oct 27 2022, doi: 10.3390/ijerph192113962.

[18] R. Stemerman, T. Bunning, J. Grover, R. Kitzmiller, and M. D. Patel, "Identifying Patient Phenotype Cohorts Using Prehospital Electronic Health Record Data," (in eng), *Prehosp Emerg Care,* pp. 1-14, Jan 25 2021, doi: 10.1080/10903127.2020.1859658.

[19] H. R. Marucci-Wellman, H. L. Corns, and M. R. Lehto, "Classifying injury narratives of large administrative databases for surveillance-A practical approach combining machine learning ensembles and human review," *Accid Anal Prev,* vol. 98, pp. 359-371, Jan 2017, doi: 10.1016/j.aap.2016.10.014.

[20] S. J. Bertke, A. R. Meyers, S. J. Wurzelbacher, A. Measure, M. P. Lampl, and D. Robins, "Comparison of methods for auto-coding causation of injury narratives," *Accid Anal Prev,* vol. 88, pp. 117-23, Mar 2016, doi: 10.1016/j.aap.2015.12.006.

[21] A. Yedla, F. D. Kakhki, and A. Jannesari, "Predictive Modeling for Occupational Safety Outcomes and Days Away from Work Analysis in Mining Operations," *Int J Environ Res Public Health,* vol. 17, no. 19, Sep 27 2020, doi: 10.3390/ijerph17197054.

[22] S. T. Sarkar, A. P. M.; Maiti, J.; Reneirs, G., "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Safety Science,* vol. 125, 12 February 2020 2020, doi: 104616.

[23] Y. M. Goh and C. U. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques," *Accid Anal Prev,* vol. 108, pp. 122-130, Nov 2017, doi: 10.1016/j.aap.2017.08.026.

[24] T. Anthony, A. K. Mishra, W. Stassen, and J. Son, "The Feasibility of Using Machine Learning to Classify Calls to South African Emergency Dispatch Centres According to Prehospital Diagnosis, by Utilising Caller Descriptions of the Incident," *Healthcare (Basel)*, vol. 9, no. 9, Aug 27 2021, doi: 10.3390/healthcare9091107.

[25] L. Hirabayashi, E. Scott, P. Jenkins, and N. Krupa, "Occupational injury surveillance methods using free text data and machine learning: creating a gold standard data set.," 2020, doi: https://doi.org/10. 4135/9781529720488.

[26] (2008). *National Occupational Research Agenda for Agriculture, Forestry, and Fishing. Second Decade. Appendix 2: Dictionary of Terms for Agricultural, Forestry and Fishing Safety and Health Professionals.*

[27] T. Shadbahr *et al.*, "The impact of imputation quality on machine learning classifiers for datasets with missing values," (in eng), *Commun Med (Lond)*, vol. 3, no. 1, p. 139, Oct 06 2023, doi: 10.1038/s43856-023-00356-z.

[28] D. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, 3, pp. 87-94, 1976.

[29] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, 1987, 2004, p. 258.

[30] Y. Zhou, S. Aryal, and M. Bouadjenek, "A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms," *ArXiv*, 4/9/2024

[31] M. Soley-Bori, *Dealing with missing data: key assumptions and methods for applied analysis*. Boston: Boston University, 2013.

[32] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," (in eng), *J Big Data*, vol. 8, no. 1, p. 140, 2021, doi: 10.1186/s40537-021-00516-9.

[33] M. Kuhn and K. Johnson, "Handling Missing Data , in Feature Engineering and Selection: A Practical Approach for Predictive Models," ed: Chapman and Hall/CRC Press, 2019.

[34] J. McAdam, L. E. Jones, X. X. Romeiko, and E. M. Bell, "Exogenous factors associated with legacy PFAS concentrations in the general U.S. population: NHANES 1999-2018," *Water Emerg Contam & Nanoplastics*, vol. 3, no. 15, 2024, doi: https://dx.doi.org/10.20517/wecn.2024.05.

[35] S. Otaru, L. E. Jones, and D. O. Carpenter, "Associations between urine glyphosate levels and metabolic health risks: insights from a large cross-sectional population-based study," (in eng), *Environ Health*, vol. 23, no. 1, p. 58, Jun 27 2024, doi: 10.1186/s12940-024-01098-8.

[36] G. Nanda, K. M. Grattan, M. T. Chu, L. K. Davis, and M. R. Lehto, "Bayesian decision support for coding occupational injury data," (in eng), *J Safety Res*, vol. 57, pp. 71-82, Jun 2016, doi: 10.1016/j.jsr.2016.03.001.

[37] M. L. Chee *et al.*, "Artificial intelligence and machine learning in prehospital emergency care: A scoping review," (in eng), *iScience*, vol. 26, no. 8, p. 107407, Aug 18 2023, doi: 10.1016/j.isci.2023.107407.

[38] N. Al-Dury *et al.*, "Identifying the relative importance of predictors of survival in out of hospital cardiac arrest: a machine learning study," (in eng), *Scand J Trauma Resusc Emerg Med*, vol. 28, no. 1, p. 60, Jun 25 2020, doi: 10.1186/s13049-020-00742-9.

[39] A. Kowarik and M. Templ, "VIM : Visualization and Imputation of Missing Values," vol. 74, no. 7, p. 15, October 2016,

[40] M. Kuhn, "Building Predictive Models in R Using the caret Package," *Journal of*

Statistical Software, vol. 28, no. 5, pp. 1-26, 2008, doi: doi:10.18637/jss.v028.i05.

[41] F. Leisch and E. Dimitriadou, "*mlbench: Machine Learning Benchmark Problems*," vol. R package version 2.1-5, ed, 2024.

[42] *High Performance Implementation of the Naive Bayes Algorithm*. (2024). https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf. [Online]. Available: https://github.com/majkamichal/naivebayes

[43] J. Friedman, R. Tibshirani, and T. Hastie, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, 1, pp. 1-22, 2010, doi: doi:10.18637/jss.v033.i01 .

[44] J. Tay, B. Narasimhan, and T. Hastie, "Elastic Net Regularization Paths for All Generalized Linear Models," *Journal of Statistical Software*, vol. 106, 1, pp. 1-31, 2023, doi: doi:10.18637/jss.v106.i01 .

[45] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presented at the The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016.

[46] Z. Aydin and Z. Ozturk, "Performance Analysis of XGBoost Classifier with Missing Data," presented at the The 1st International Conference on Computing and Machine Intelligence (ICMI 2021), 2021.

[47] S. Stokanović, D. Đukić, and N. Miljković, "The Robustness of XGBoost Algorithm to Missing Features for Binary Classification of Medical Data," presented at the 23rd International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2024.

[48] E. Scott, B. Weichelt, and J. Lincoln, "The Future of U.S. Agricultural Injury Surveillance Needs Collaboration," (in eng), *J Agromedicine*, vol. 28, no. 1, pp. 11-13, Jan 2023, doi: 10.1080/1059924X.2022.2148032.

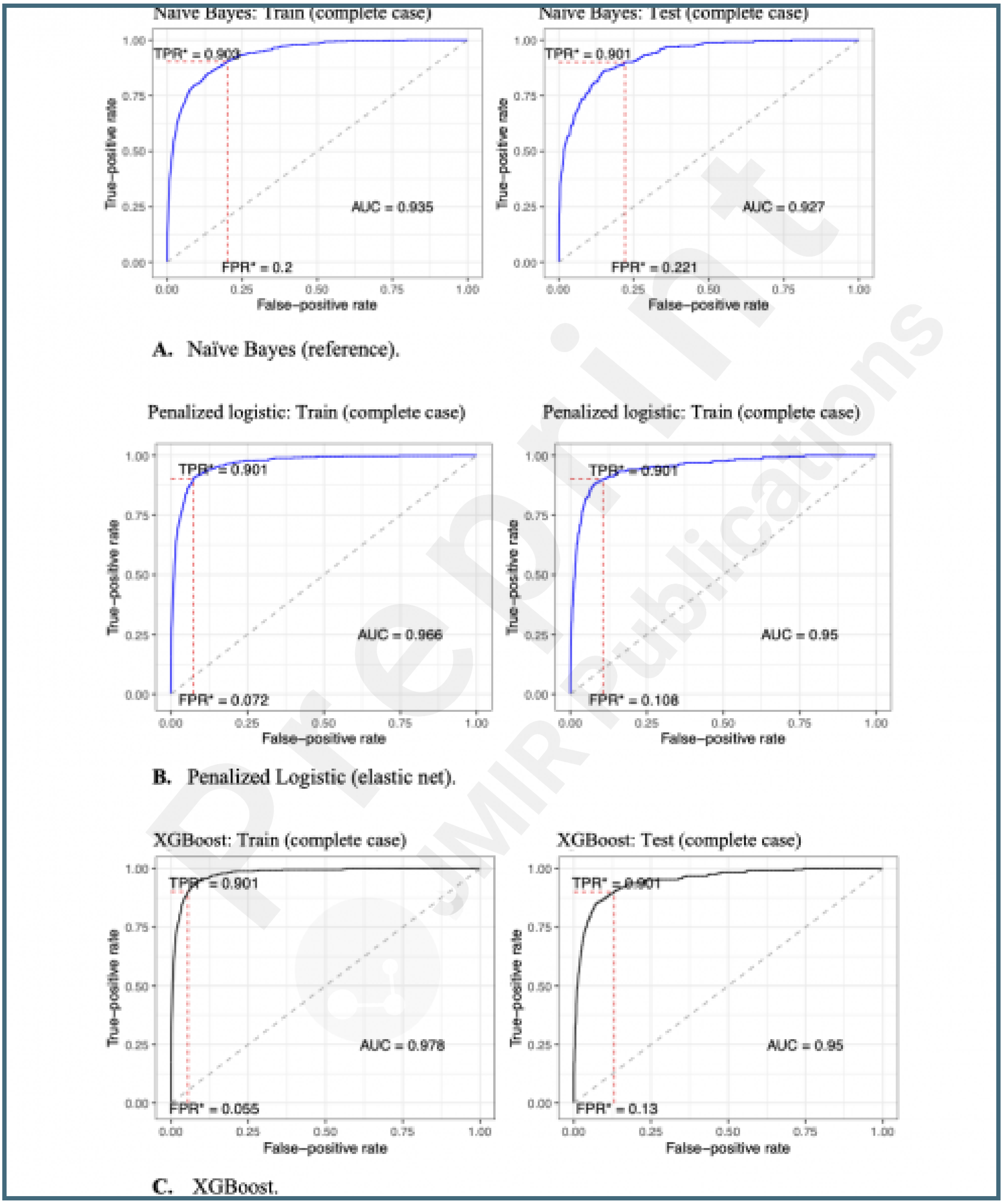
[49] E. National Academies of Sciences, and Medicine, *A Smarter National Surveillance System for Occupational Safety and Health in the 21st Century*. Washington, DC: The National Academies Press (in English), 2018, p. 318.

Supplementary Files

This is the current draft. Please replace the other draft with this and send it to reviewers.
URL: http://asset.jmir.pub/assets/b0dd420f7a5702a840962f01689dbbcf.docx

Figures

Complete Case. Receiver-Operator Characteristic (ROC) curves for training and testing sets, trained and tested on labeled Maine data (2008-2023), complete case. Area under the curve (AUC) is shown on each panel, as well as the False Positive Rate (FPR*) required to return a True Positive rate (TPR*) of 0.90, indicated in dashed red line.



Naive Bayes: Train (complete case)

Naive Bayes: Test (complete case)

A. Naïve Bayes (reference).

Penalized logistic: Train (complete case)

Penalized logistic: Train (complete case)

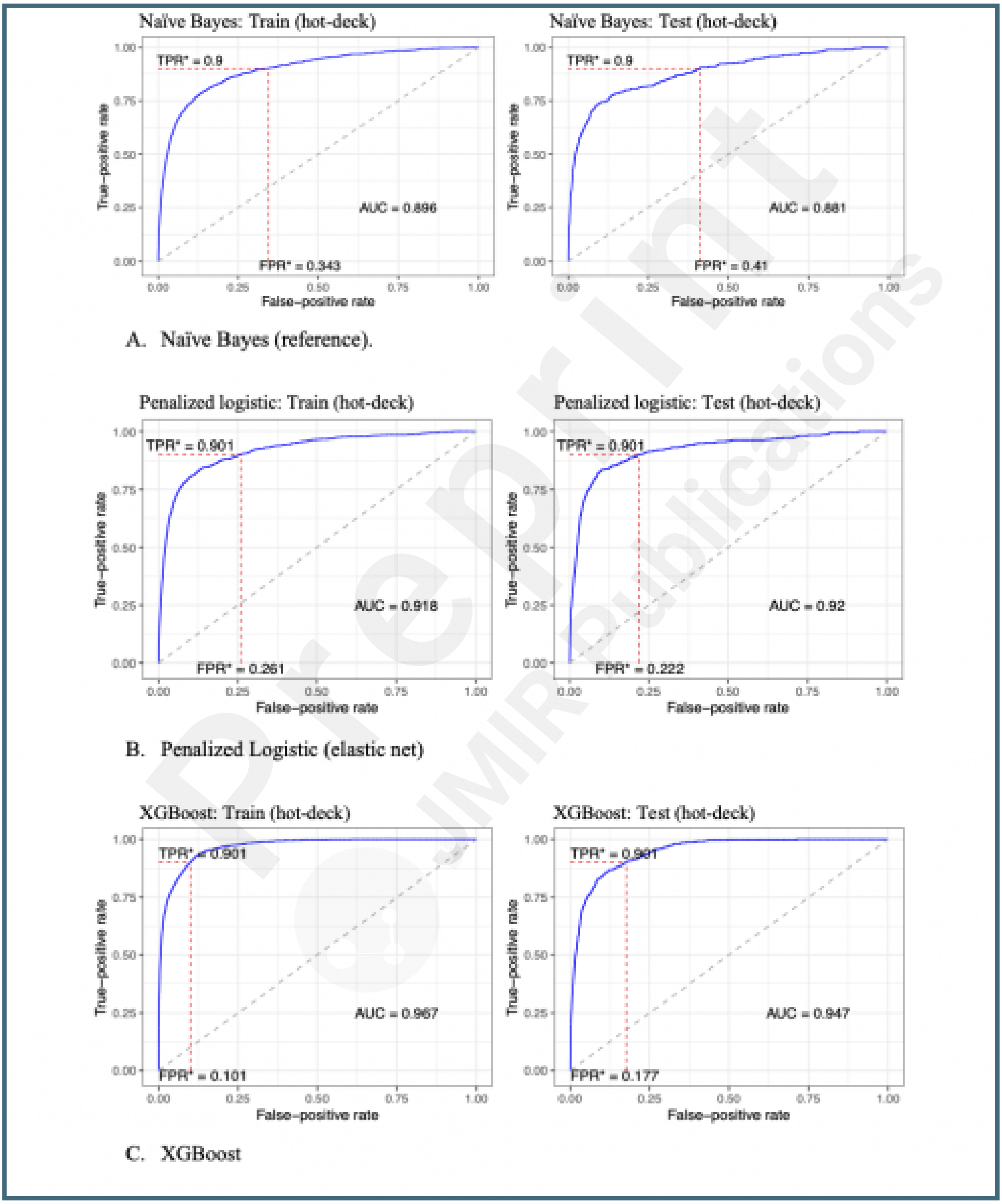
B. Penalized Logistic (elastic net).

XGBoost: Train (complete case)

XGBoost: Test (complete case)

C. XGBoost.

Hot-deck imputed. Receiver-Operator Characteristic (ROC) curves for training and testing sets, trained and tested on labeled Maine data (2008-2023), hot-deck single imputed. Train-Test split is 75:25 (N=60143, Train: N = 45105 samples; Test: N = 15035) using stratified sampling due to imbalanced label distribution. Area under the curve (AUC) is shown on each panel, as well as the False Positive Rate (FPR*) required to return a True Positive rate (TPR*) of 0.90, indicated in dashed red line.

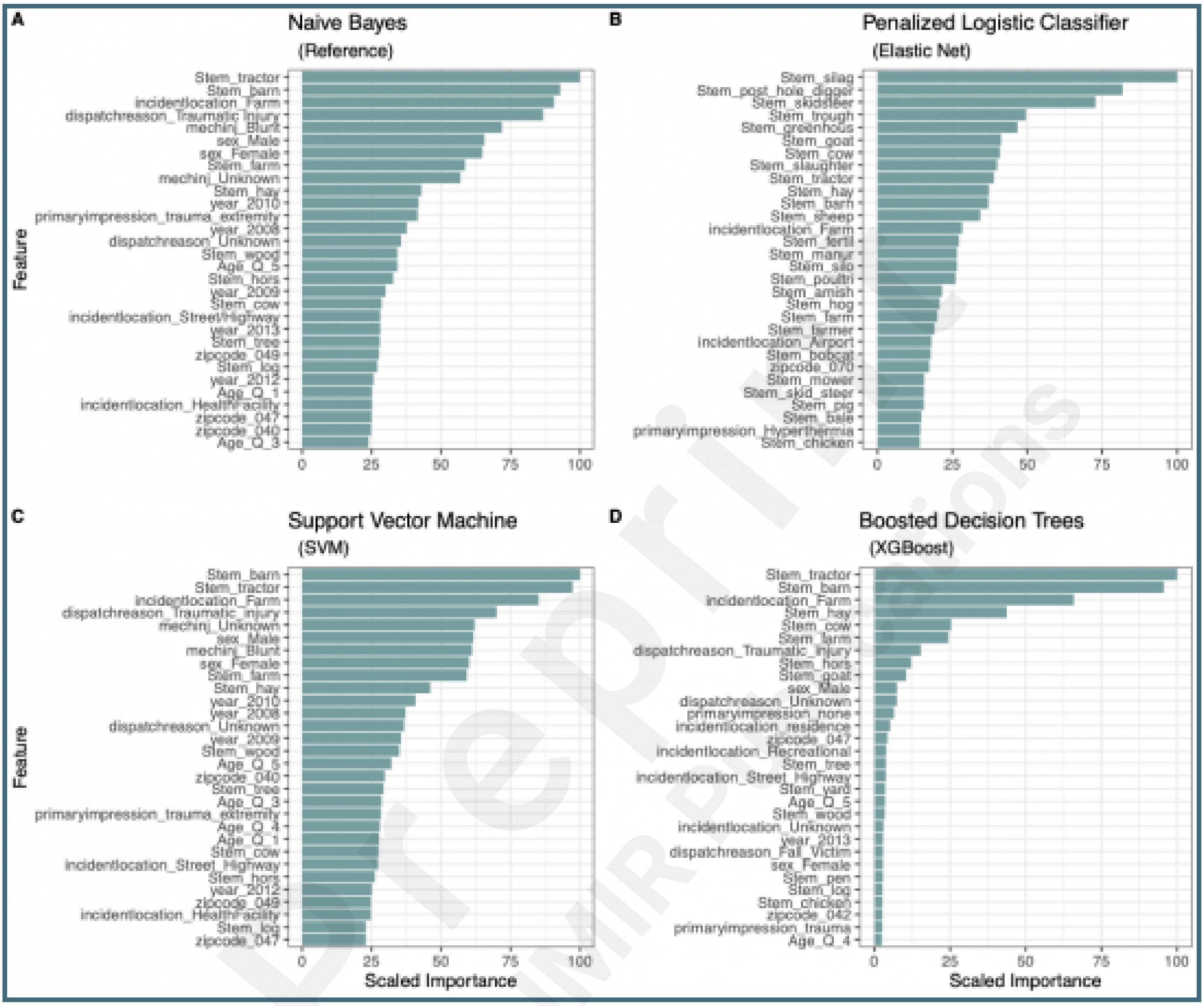


A. Naïve Bayes (reference).

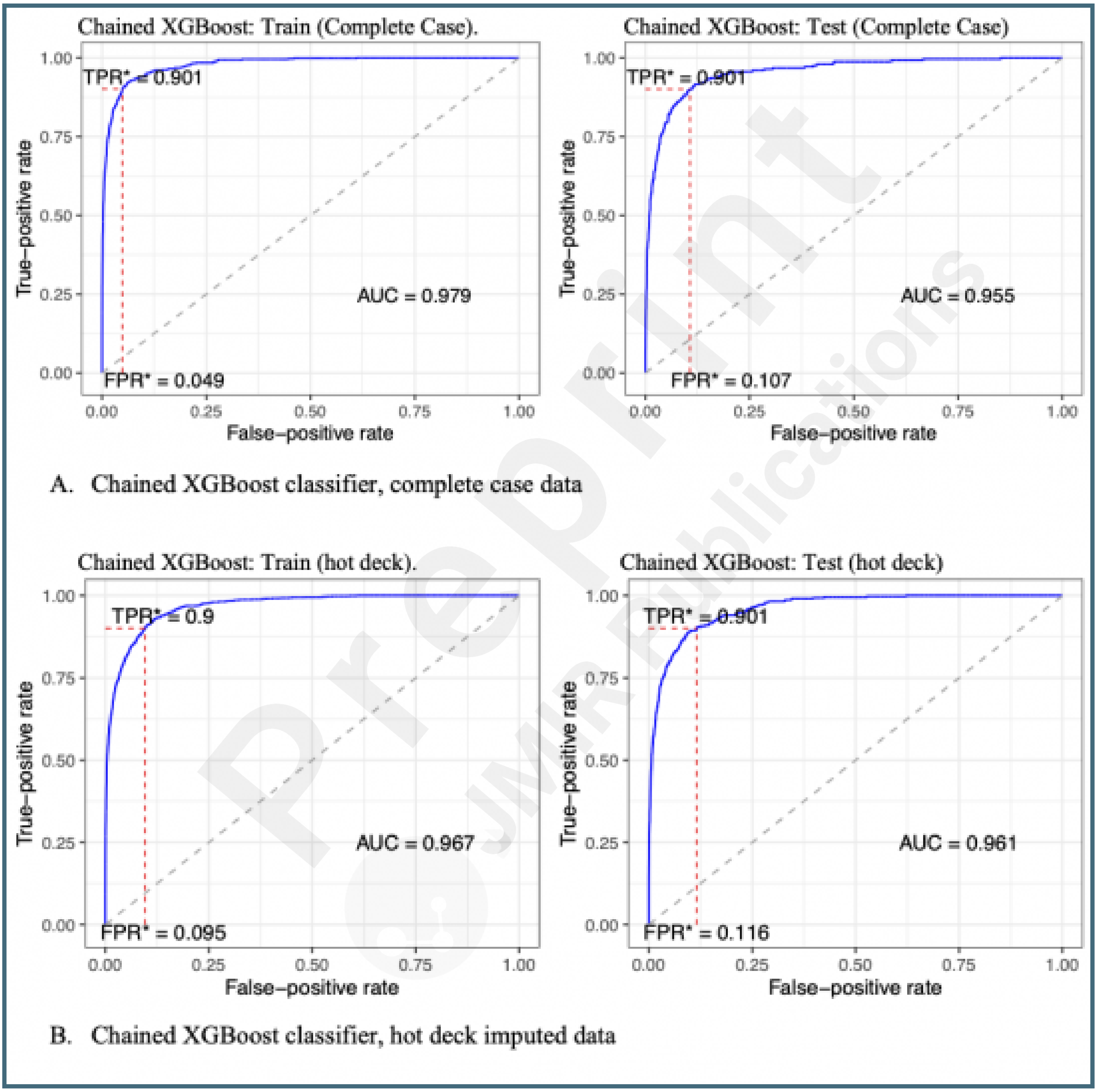
B. Penalized Logistic (elastic net)

C. XGBoost

Top 30 complete case feature importances compared for (A) Naïve Bayes (reference model); (B), Penalized Logistic; (C) Support Vector Machine, and (D) Boosted Decision Tree (XGBoost) classifiers.



Receiver-Operator Characteristic (ROC) curves for training and testing sets, trained and tested on labeled Maine data (2008-2023), chained models: feature reduction via naïve Bayes followed by classification by XGBoost. Train-Test split is 75:25 (Train N = 45105 samples; Test N = 15035) using stratified sampling due to imbalanced label distribution. Area under the curve (AUC) is shown on each panel, as well as the False Positive Rate (FPR*) required to return a True Positive rate (TPR*) of 0.90, indicated in dashed red line. A. complete case data (N=44566); B. Hot-deck single imputed (N=60143).



A. Chained XGBoost classifier, complete case data

B. Chained XGBoost classifier, hot deck imputed data

Multimedia Appendixes

Supplemental Information.
URL: http://asset.jmir.pub/assets/4b1532d118f5b9d2502b4bcf2cbe979c.pdf