

1 **Naïve Bayes is an interpretable and**
2 **predictive machine learning algorithm**
3 **in predicting osteoporotic hip fracture**
4 **in-hospital mortality compared to other**
5 **machine learning algorithms.**

6 Jo-Wai Douglas Wang^{1,2*}.

7 1. Department of Geriatric Medicine, The Canberra Hospital, ACT Health, Canberra ACT

8 2. The Australian National University Medical School;, Canberra ACT.

9 * Corresponding author

10 E-mail: jo-waidouglas.wang@anu.edu.au

11

12 **0. Abstract**

13 Osteoporotic hip fractures (HFs) in the elderly are a pertinent issue in healthcare, particularly
14 in developed countries such as Australia. Estimating prognosis following admission remains
15 a key challenge. Current predictive tools require numerous patient input features including
16 those unavailable early in admission. Moreover, attempts to explain machine learning [ML]-
17 based predictions are lacking. We developed 7 ML prognostication models to predict in-
18 hospital mortality following minimal trauma HF in those aged ≥ 65 years of age, requiring
19 only sociodemographic and comorbidity data as input. Hyperparameter tuning was
20 performed via fractional factorial design of experiments combined with grid search; models
21 were evaluated with 5-fold cross-validation and area under the receiver operating
22 characteristic curve (AUROC). For explainability, ML models were directly interpreted as well
23 as analyzed with SHAP values. Top performing models were random forests, naïve Bayes
24 [NB], extreme gradient boosting, and logistic regression (AUROCs ranging 0.682 – 0.696,
25 $p > 0.05$). Interpretation of models found the most important features were chronic kidney
26 disease, cardiovascular comorbidities and markers of bone metabolism; NB also offers direct
27 intuitive interpretation. Overall, we conclude that NB has much potential as an algorithm, due
28 to its simplicity and interpretability whilst maintaining competitive predictive performance.

29 **Author Summary**

30 Osteoporotic hip fractures are a critical health issue in developed countries. Preventative
31 measures have ameliorated this issue somewhat, but the problem is expected to remain in
32 main due to the aging population. Moreover, the mortality rate of patients in-hospital remains
33 unacceptably high, with estimates ranging from 5- 10%. Thus, a risk stratification tool would
34 play a critical in optimizing care by facilitating the identification of the susceptible elderly in
35 the community for prevention measures and the prioritisation of such patients early during
36 their hospital admission. Unfortunately, such a tool has thus far remained elusive, despite
37 forays into relatively exotic algorithms in machine learning. There are three major drawbacks
38 (1) most tools all rely on information typically unavailable in the community and early during

39 admission (for example, intra-operative data), limiting their potential use in practice, (2) few
40 studies compare their trained models with other potential algorithms and (3) machine
41 learning models are commonly cited as being 'black boxes' and uninterpretable. Here we
42 show that a Naïve Bayes model, trained using only sociodemographic and comorbidity data
43 of patients, performs on par with the more popular methods lauded in literature. The model is
44 interpretable through direct analysis; the comorbidities of chronic kidney disease,
45 cardiovascular, and bone metabolism were identified as being important features
46 contributing to the likelihood of deaths. We also showcase an algorithm-agnostic approach
47 to machine learning model interpretation. Our study shows the potential for Naïve Bayes in
48 predicting elderly patients at risk of death during an admission for hip fracture.

49 **1. Introduction**

50 The osteoporotic hip fracture (HF) is a global issue with an estimated financial burden of 17
51 billion USD for the United States in 2002 and projected burden of £3.62 in 2023 for the
52 United Kingdom [1, 2]. Estimates for short-term (in-hospital) mortality following HF have
53 been placed in the vicinity of 2 – 10%, with an estimated mortality rate of 2.7% for HF
54 hospitalisation in Australia [3-5]. In developed countries, though preventative measures
55 (targeting reduction of hip fracture risk factors such as osteoporosis and falls) have reduced
56 the age-standardized incidence rate of hip fractures, the absolute rate is increasing due to
57 the ageing population [6]. In Australia, for instance, hospitalisations for HF in the elderly
58 increased by almost 20% between 2006-07 and 2015-16 from 15 900 to 18 700 respectively
59 [5]. With the trend towards an aged population expected to continue, including in Australia,
60 HFs in the elderly will remain a relevant, and increasingly pressing challenge in healthcare.

61

62 One key aspect in HF assessment and management challenge is the prognostication of poor
63 short-term outcomes. There exists a substantial amount of analysis from traditional statistical
64 methods (such as logistical regression, LR) in identifying key risk factors for predicting poor

65 outcomes, notably mortality, following HF and scoring tools that have risen to prominence
66 are the Nottingham Hip Fracture Score (NHFS) and the orthopaedic- Physiological and
67 Operative Severity Score for the enUmeration of Mortality and Morbidity (O-POSSUM) [7-
68 10]. Most of these tools require a combination of both clinical, laboratory and intra-operative
69 data; and the lack of laboratory and intra-operative data early during admission limits the use
70 of such tools in early risk stratification.

71 Non-traditional mathematical algorithms, especially those associated with artificial
72 intelligence (AI) and machine learning (ML), have become increasingly utilized in healthcare.
73 A variety of ML algorithms, including regression-based methods, decision-tree based
74 methods (i.e. decision trees [DT], Random Forests [RF], eXtreme Gradient Boosting [XGB]
75 implementation), neural networks (NN), Naïve Bayes and support vector machines (SVM)
76 have been used in the prognostication of patients in the general peri-operative [11-16] and
77 peri-HF [17-20] period with varying degrees of success. However, most of these tools
78 require data that is not readily available on admission (such as intra-operative data and
79 laboratory data), much like the tools developed from traditional statistical methods and most
80 do not predict short-term in-hospital mortality following HF.

81 Moreover, few studies compared multiple machine learning algorithms of different classes
82 with each other. An exception to this was work performed by Forssten et al., who trained
83 SVM, NB, and as well as LR (to be used as a baseline) models to predict the 1-year
84 mortality post-HF, and found that LR outperformed all other models [17]. To our knowledge
85 no study has trained a wider array of machine learning algorithms for prediction of short-term
86 outcome following HFs. Tree-based methods have received the majority of attention.

87 Another algorithm that has remarkable potential is naïve Bayes which is based on Bayes
88 theorem, with the additional 'naïve' assumption that features are conditionally independent. It
89 has been applied successfully across a wide variety of tasks in natural language processing
90 (e.g. detection of spam email [21], text sentiment analysis, text/document classification) as
91 well as in the medical field (e.g. the prognostication in cirrhotic patients following TIPS [22],

92 prediction of 30-day mortality following HF [23], prediction of osteonecrosis of femoral head
93 with cannulated screw fixation [24] and prediction of mortality in post-surgical intensive care
94 unit patients [25]).

95 While predictive ability is an important characteristic of any prognostic tool, it is increasingly
96 recognized that a desirable attribute of machine learning algorithm is that they are
97 interpretable (or ‘explainable’) especially as ML models become increasingly complex [26,
98 27]. Recognition of this issue has led to the development of the subfield of ‘interpretable’ ML
99 and, in particular, the development and application of the SHapley Additive exPlanations
100 (SHAP), an approach based on cooperative game theory [28-33].

101 Our goal, was to train multiple machine learning models, specifically Bernoulli Naïve Bayes
102 (NB), DT, RF, XGB, SVM, logistic regression (LR) and the multi-layer perceptron (MLP, a 3-
103 layer NN) to predict in-hospital mortality for the elderly admitted with HF. We focus on using
104 only those patient features that are readily available in the early phases during a hospital
105 admission, i.e. sociodemographic and comorbidity data. The performances of each model
106 would be compared to identify the most predictive algorithm. Finally, each predictive tool
107 would be analyzed via direct interpretation of model and with calculation of SHAP values.

108 **2. Results**

109 *2.1. Patient cohort characteristics*

110 Of the 3625 patients in the cohort, age was distributed non-normally with median age of 84
111 (interquartile range of 10 years) and 2730 (75.3%) were female; 189 (5.2%) had in-hospital
112 mortality. The most common comorbidity was hypertension (at 2045 [56.4%]). Details are
113 present in Table 1 (with abbreviations defined below).

Table 1. Sociodemographic features, outcomes of HF cohort

Variable	Total Cohort (N=3625)	Female (N=2730, 75.31%)	Male (N=895, 24.69%)	p value ⁽¹⁾
Sociodemographic features				
Age (median [IQR])	84 [10]	85 [10]	82 [12]	<0.001

Aged > 80 years (n,%)	2457, 67.8%	1937, 71.0%	520, 58.1%	<0.001
PRCF resident (n,%)	1208, 33.3%	950, 34.8%	258, 28.8%	0.001
Smoker (n,%)	180, 5.0%	124, 4.5%	56, 6.3%	0.050
Alcohol overuse (n,%)⁽²⁾	144, 4.0%	60, 2.2%	84, 9.4%	<0.001
Walking aids user (n,%)	1300, 35.9%	999, 36.6%	301, 33.6%	0.116
Comorbidities features				
Hypertension, (n,%)	2045, 56.4%	1606, 58.8%	439, 49.1%	<0.001
Anaemia (n,%)	1531, 42.2%	1051, 38.5%	480, 53.6%	<0.001
CKD (n,%)	1444, 39.9%	1106, 40.5%	338, 37.8%	0.152
Dementia (n,%)	1117, 30.8%	858, 31.4%	259, 28.9%	0.172
CAD (n,%)	1073, 29.6%	750, 27.5%	323, 36.1%	<0.001
History of AMI (n,%)	287, 7.9%	191, 7.0%	96, 10.7%	<0.001
AF (n,%)	702, 19.4%	513, 18.8%	189, 21.1%	0.139
COPD (n,%)	561, 15.5%	385, 14.1%	176, 19.7%	<0.001
T2DM (n,%)	482, 13.3%	325, 11.9%	157, 17.5%	<0.001
OP (n,%)	478, 13.2%	410, 15.0%	68, 7.6%	<0.001
CVA (n,%)	431, 11.9%	323, 11.8%	108, 12.1%	0.897
TIA (n,%)	309, 8.5%	227, 8.3%	82, 9.2%	0.474
PD (n,%)	172, 4.7%	97, 3.6%	75, 8.4%	<0.001
Malignancy (n,%)	82, 2.3%	52, 1.9%	30, 3.4%	0.017
PTH>6.8pmol/L	1684, 46.5%	1275, 46.7%	409, 45.7%	0.628
25(OH)vitamin D≤25nmol/L	610, 16.8%	467, 17.1%	143, 16.0%	0.464
25(OH)vitamin D≤50nmol/L	1659, 45.8%	1235, 45.2%	424, 47.4%	0.283
Outcome				
Died (n,%)	189, 5.2%	130, 4.8%	59, 6.6%	0.040

114 ¹Pearson's Chi-squaerr test (Yates corrected).

115 ²Use>3 times a week.

116 **Abbreviations:** PRCF, permanent residential care facility; CKD, chronic kidney disease;
 117 CAD, coronary artery disease; AMI, acute myocardial infarction; AF, atrial fibrillation; COPD,
 118 chronic obstructive pulmonary disease; T2DM, type 2 diabetes mellitus; OP, osteoporosis;
 119 CVA, cerebrovascular accident; TIA, transient ischaemic attack; PD, Parkinson's disease;
 120 PTH, parathyroid hormone

121 2.2. Model Performance – Training

122 The model with the highest area under the receiver operating characteristic (AUROC) was
 123 MLP (AUROC 0.828) followed by LR, RF, XGB and NB (0.733, 0.730, 0.726 and 0.725
 124 respectively, all p>0.05), then DT (AUROC of 0.697) and finally SVM (AUROC 0.533).

125 The model with greatest area under the precision-recall curve (AUPRC) was MLP (AUPRC
 126 0.245), followed by LR, XGB and RF (AUPRCs of 0.134, 0.133 and 0.130 respectively,

127 $p > 0.05$), NB (AUPRC 0.124), DT (AUPRC of 0.094) and finally SVM (AUPRC of 0.058).

128 Details are present in Table 2 and Table 3.

129

Table 2. Model performance (training phase)

	AUROC			AUPRC		
	Mean	STD	95%CI	Mean	STD	95%CI
SVM	0.533	0.029	0.475 - 0.591	0.058	0.004	0.050 - 0.067
NB	0.725	0.007	0.711 - 0.739	0.124	0.003	0.117 - 0.131
LR	0.733	0.008	0.717 - 0.750	0.134	0.004	0.127 - 0.141
DT	0.697	0.004	0.690 - 0.704	0.094	0.001	0.093 - 0.095
RF	0.730	0.007	0.716 - 0.745	0.130	0.003	0.125 - 0.136
XGB	0.726	0.007	0.711 - 0.741	0.133	0.005	0.122 - 0.144
MLP	0.828	0.008	0.813 - 0.844	0.245	0.030	0.186 - 0.305

130

Table 3. Comparison of model performance during training. (A) – AUROC (B) – AUPRC

(A) AUROC														
models	t-test statistic							t-test p-value						
	SVM	NB	LR	DT	RF	XGB	MLP	SVM	NB	LR	DT	RF	XGB	MLP
SVM	-	-	-	-	-	-	-	-	0.000	0.000	0.000	0.000	0.000	0.000
NB	-	14.391	14.866	12.527	14.766	14.466	21.927	-	-	0.131	0.000	0.291	0.827	0.000
LR	-	-	-	9.000	0.631	1.472	-	-	-	-	0.000	0.546	0.179	0.000
DT	-	-	-	-	-9.153	-8.043	-	-	-	-	-	0.000	0.000	0.000
RF	-	-	-	-	-	0.904	-	-	-	-	-	-	0.393	0.000
XGB	-	-	-	-	-	-	-	-	-	-	-	-	-	0.000
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

(B) AUPRC														
models	t-test statistic							t-test p-value						
	SVM	NB	LR	DT	RF	XGB	MLP	SVM	NB	LR	DT	RF	XGB	MLP
SVM	-	-	-	-	-	-	-	-	0.000	0.000	0.000	0.000	0.000	0.000
NB	-	29.516	30.042	19.524	40.249	26.191	13.816	-	-	0.002	0.000	0.002	0.009	0.000
LR	-	-	-	21.693	2.236	0.349	-8.201	-	-	-	0.000	0.056	0.736	0.000
DT	-	-	-	-	-	-	-	-	-	-	-	0.000	0.000	0.000
RF	-	-	-	-	80.498	17.103	11.249	-	-	-	-	-	0.217	0.000
XGB	-	-	-	-	-	-	-8.234	-	-	-	-	-	-	0.000
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

131

132 **2.3. Model Performance – Test**

DT	-	-	-	-	-	-	-	-	-	-	-	0.005	0.005	0.128
					3.796	3.887	1.698							
RF	-	-	-	-	-	-	3.130	-	-	-	-	-	0.949	0.014
						0.066								
XGB	-	-	-	-	-	-	3.220	-	-	-	-	-	-	0.012
MLP	-	-	-	-	-	-	-	-	-	-	-	-	-	-

140

141 **2.4. Feature importance – Model interpretation**

142 Feature importance rankings (1 being the most important) according to each model can be
 143 found in Table 6. Corresponding coefficients for NB, LR, XGB and RF can be found in Table
 144 7.

145 For the LR model, the 5 most important patient features in prediction of mortality were
 146 presence of CKD, vitamin D deficiency ($\leq 25\text{nmol/L}$), advanced age (>80 years), COPD, and
 147 AF. In the SVM model the 5 most important patients features in prediction of mortality were
 148 advanced age (>80 years), CKD, vitamin D insufficiency ($\leq 50\text{nmol/L}$), anaemia, and use of
 149 walking aids. For the NB model, the 5 most important features in mortality prediction were
 150 history of MI, AF, CKD and CAD. For the DT model the 5 most important features in mortality
 151 prediction were presence of CKD, hyperparathyroidism ($\text{PTH} > 6.8\text{pmol/L}$), CAD, dementia
 152 and advanced age (>80 years). For the RF model the 5 most important features in mortality
 153 prediction were CKD, hyperparathyroidism ($\text{PTH} > 6.8\text{pmol/L}$), CAD, dementia and advanced
 154 age (>80 years). For the XGB model the 5 most important features in mortality prediction
 155 were CKD, CAD, advanced age (>80 years), $\text{PTH} > 6.8\text{pmol/L}$ and AF. Finally, for the MLP
 156 model the 5 most important features in mortality prediction were AF, CKD, male sex,
 157 dementia, and MI.

Table 6. Feature importance rankings.

Feature	Feature importance rankings*						
	LR	SVM	NB	DT	RF	XGB	MLP
Male	9	6	13	12	12	12	3
Aged > 80	3	1	12	5	5	3	14
Resident of PRCF	10	20	7	7	7	6	13
Smoking	13	17	20	15	15	19	23
Alcohol overuse	22	22	23	16	16	20	21
Walking aids use	16	5	14	17	17	13	6

HT	18	7	19	18	18	22	20
CAD	8	14	4	3	3	2	19
MI	7	8	1	8	8	7	5
AF	5	15	2	13	13	5	1
CVA	20	11	16	19	19	16	22
TIA	15	13	11	20	20	15	10
dementia	12	10	8	4	4	9	4
PD	14	19	18	21	21	17	18
COPD	4	12	6	11	11	11	12
T2DM	17	16	15	10	10	14	11
CKD	1	2	3	1	1	1	2
Anaemia	11	4	10	9	9	10	8
Malignancy	23	23	22	22	22	23	17
Osteoporosis	21	21	21	23	23	18	16
PTH>6.8pmol/L	6	18	9	2	2	4	7
25(OH)vitamin D≤25nmol/L	2	9	5	6	6	8	9
25(OH)vitamin D≤50nmol/L	19	3	17	14	14	21	15

158 *In descending order of importance.

159 2.5. Feature importance – SHAP analysis

160 Features were also ranked by the mean absolute SHAP values as displayed in summary
161 below (Figure 2).

162 For the LR model, the 5 most predictive patient features for mortality in order from highest
163 magnitude to lowest, based on mean SHAP values, were CKD, advanced age (>80 years),
164 hyperparathyroidism (PTH>6.8pmol/L), CAD, and residency from PRCF. Absence of any of
165 these features had a negative SHAP value (i.e. a negative contribution) on the model
166 outcome (in-hospital mortality); the magnitude of this impact was consistent across all
167 patients. Likewise, the presence of any of these features always had a positive SHAP value
168 (i.e. an additive contribution) on in-hospital mortality. The magnitude of this effect was again
169 consistent across all patients.

170 For NB model, the 5 most predictive features were CKD, AF, MI, residency from PRCF,
171 CAD. Again, absence of any of these features most commonly had a negative impact on in-
172 hospital mortality; the magnitude of this effect varied among patients. The presence of any of

173 the above 5 features had a positive contribution to the prediction of in-hospital mortality;
174 similarly, the magnitude of this effect varied significantly among patients.

175 For the DT model, the 5 most predictive features were CKD, hyperparathyroidism
176 (PTH>6.8pmol/L), advanced age (>80 years), presence of CAD and vitamin D deficiency
177 (≤ 25 nmol/L). Presence of these five comorbidities had a positive contribution to prediction of
178 in-hospital mortality and, conversely their absence had a negative contributory effect on
179 prediction. Interestingly, absence of T2DM had an additive effect and presence of T2DM had
180 a negative effect on mortality prediction. The magnitude of contributions that each of the 5
181 variables had varied among different patients. Finally, it is noteworthy that all other
182 comorbidities had little to no influence on patient outcomes.

183 For the RF model, the 5 most predictive features were CKD, hyperparathyroidism
184 (PTH>6.8pmol/L), CAD, advanced age (>80 years) and residence from PRCF. The presence
185 of these features increased likelihood of mortality and conversely absence decreased the
186 likelihood of mortality; there was only a minor variation of contribution from each feature for
187 each patient.

188 For the XGB model, the 5 most predictive patient features were advanced age (>80
189 years), vitamin D deficiency (≤ 25 nmol/L), CKD, CAD and hyperparathyroidism
190 (PTH>6.8pmol/L). The presence (and absence) of any of these features increased (or
191 decreased) the likelihood of mortality. For each feature, there was only mild variation in the
192 magnitude of contributions among patients.

193 Finally, from the MLP model, the 5 most predictive patient features were male sex,
194 advanced age (>80 years), CVA, HTN and TIA. The presence (or, conversely, the absence)
195 of any of these features except for HTN were associated with an increased (decreased)
196 likelihood of mortality; presence (or absence) of HTN appeared to decrease (increase) the
197 likelihood of mortality.

198 Across all models, the 5 comorbidities most consistently with the greatest influence on
199 mortality prediction were: CKD, advanced age (>80 years), elevated PTH (>6.8pmol/L),
200 cardiovascular disease (CAD, MI, AF or HTN) and PRCF residence.

201 **3. Discussion**

202 We have trained and compared the performances of 7 machine learning models to predict
203 in-hospital mortality for hospitalized elderly minimal trauma HF patients using only
204 categorical data. Overall, the models had reasonable to good performance. We also
205 performed an analysis of each model and applied SHAP analysis to gain insight into feature
206 importance

207 *3.1. Model performance – Training and Test*

208 Notably, but unsurprisingly, classification performance showed some variation among
209 algorithms. The trained models, ordered in decreasing performance (based on both test
210 AUROCs and test AUPRCs), were RF, NB, XGB and LR (all with no statistically significant
211 difference in performance – see Table 4, 5 and Figure 1) followed by MLP and DT (no
212 statistically significant difference in performance) and finally SVM. AUROCs ranged from
213 0.500 (SVM) to almost 0.700 (good performance), while AUPRC values ranged from 0.050
214 (SVM) to 0.115; a reflection of using a simplified model (with binary input data) to perform
215 predictions on a minority class in this imbalanced dataset. There was minimal difference
216 between the training and cross-validation performance for the top 4 models (RF, NB, XGB
217 and LR). A greater variation in training and cross-validation performance scores was noted
218 for DT and MLP, an indicator of overtraining (an infamous tendency in machine learning).
219 That overtraining has occurred despite systematic and meticulous hyperparameter tuning, is
220 strongly suggestive of insufficient data.

221 To our knowledge, most studies have focused on only training and applying one class of
222 machine learning algorithm. Often there is no baseline model trained using traditional
223 statistics (e.g. LR). Indeed, most studies have solely utilized tree-based methods (e.g.

224 applying DT, XGB and RF methods) and this is reflected in a scoping study of ML usage in
225 health economics and research (on 805 studies) which found the most frequent algorithms
226 used were tree-based methods followed by regression-based (linear/logistic) methods, SVM,
227 NN and finally NB [37]. However, it is known that performance on various tasks varies with
228 different ML algorithms [38] and our finding that predictive performance varies among
229 machine learning algorithms (for the same problem, using the same data) is consistent with
230 this. It is thus ideal that in future applications of machine learning, a more comprehensive set
231 of algorithms are trained, or some justification should be provided if possible when certain
232 algorithms are not included.

233 The performance of NB in predicting mortality is on par with RF, XGB and LR which warrants
234 further discussion here as it has received relatively little attention in the literature. Key to its
235 success is the simplifying assumption of conditional independence among all patient input
236 features. The most obvious advantage from this is that, by virtue of such a simplification, it is
237 computationally inexpensive and is fast to train and run. However, with such a large
238 seemingly excessive, simplifying assumption (not strictly satisfied in our current database), it
239 may seem surprising that this model performs so well. Contrary to intuition, its good
240 performance is not a coincidental or even unexpected phenomenon; formal analysis of NBs
241 has established it performs well because the interdependencies, when they do exist, occur in
242 a manner which results in them 'cancel[ling] each other out' [39].

243 *3.2. Feature importance – Model interpretation and SHAP analysis*

244 Rankings of patient comorbidity importance in their role in mortality prediction were
245 determined from all models from direct interpretation of feature coefficients (see Table 6 and
246 Table 7). CKD was most consistently ranked as one of the 5 most important patient
247 comorbidities in predicting mortality. The other most important patient features included
248 markers reflective of bone metabolism (PTH, vitamin D levels) and cardiovascular disease
249 (presence of either one of CAD, MI, AF). Similar trends were found via SHAP value analyses
250 for each model, i.e. CKD, bone metabolism markers and presence of cardiovascular

251 diseases had the strongest influence on prediction of mortality based on mean SHAP values
252 (Figure 2). It is recognized in the literature that cardiovascular comorbidities and renal
253 function are important for prognostication which is reflected in their inclusion as input
254 parameters for non-cardiac surgery risk assessment tools such as the Revised Cardiac Risk
255 Index and the American College of Surgeons - surgical risk calculator [40-44]. However,
256 these features are not explicitly included in HF-specific risk assessment tools (e.g. in O-
257 POSSUM only symptoms and clinical findings suggestive of cardiovascular disease are
258 included and NHFS only the number of comorbidities is included as an input parameter) [7-
259 10]. Moreover, neither PTH or vitamin D levels are included in any of the current tools,
260 despite an increasing number of studies supporting the key role they play in bone
261 metabolism and prevention of fracture [45-57] and, potentially, with increasing recognition of
262 their importance in the immunity [58-60] prevention of post-operative complications such as
263 hospital acquired infections.

264 3.3. *Further insights from model analysis*

265 Of the four most predictive models, NB and LR models offer intuitive, quantifiable insights
266 into feature contributions to prediction: in LR, the odds ratio can be taken by calculating the
267 exponent of the coefficients, while in NB, from our method of scoring input features (see
268 Appendix A), each coefficient corresponds to the ratio of the rate of the comorbidity in those
269 who experienced in-hospital mortality compared to the comorbidity rate in those who
270 survived. So, for example, in predicting mortality, we can see from the LR model that CAD,
271 with a score of 0.319 (95%CI 0.180 – 0.458) increased mortality risk by 37% (OR 1.37;
272 95%CI 1.20 – 1.58) and CKD, with a score of 0.711 (95%CI 0.505 – 0.918) increased
273 mortality risk by 2.03 (95%CI 1.66 – 2.50). From the NB model, a score of 1.62 (for CAD),
274 and a score of 1.67 (for CKD) indicated that the rate of each comorbidity was greater in
275 mortality than in survival by 62% and 67% respectively.

276 For the other top predictive models, insights gained from direct interpretation of RF and XGB
277 is not so straightforward. Both these methods are based on DTs, which is itself an

278 interpretable and intuitive model. However, a major drawback of DTs is that they are very
279 prone to bias and variance (overfitting). RF and XGB address this issue by constructing
280 multiple DTs and the overall prediction is then made from an ensemble/collection of multiple
281 trees (numbering in the hundreds) and, hence, increased predictive performance is obtained
282 at the expense of interpretability. In our study, the coefficients for each feature correspond to
283 the relatively abstract concept of mean decrease in (Gini) impurity (see Appendix A).

284 *3.4. Further insights from SHAP values*

285 SHAP values revealed that the presence of more ‘severe’ comorbidities in each ML model
286 had a more important additive effect on mortality risk than less severe comorbidities, as one
287 might expect. For instance, patients with a history of acute MI (a higher severity subcohort of
288 CAD patients) typically had the greatest SHAP value indicating that the presence of history
289 of past MI had the greatest additive effect on mortality prediction. Similarly, the presence of
290 vitamin D deficiency ($\leq 25\text{nmol/L}$) was correlated with greater SHAP values compared to
291 vitamin D insufficiency ($\leq 50\text{nmol/L}$) (see Figure 2). In contrast, the absence of both MI and
292 vitamin D deficiency in patients had less of a negative effect on mortality prediction
293 compared to the other comorbidities (and hence was why they had a lower overall
294 importance based on mean SHAP values). This reflects their relatively low prevalence in the
295 cohort (Table 1).

296 For LR, RF, and XGB the SHAP values had low variability and were highly concentrated –
297 an indication that the corresponding input patient features were consistently strong
298 contributors to mortality prediction; a corollary of this was that these models offered good
299 population level insight into mortality risk. Of the top four models, NB was the only model in
300 which the SHAP values themselves varied among individuals. This variability in SHAP
301 values among patients suggested that the influence of each singular comorbidity was not
302 constant, and that each prediction appeared to be tailored toward individual.

303 *3.5. Limitations*

304 We note limitations to our study. Firstly, this was a study on a retrospective cohort, and all
305 members of cohort were from a single-centre study. We recognize internal nested cross-
306 validation, though relatively rigorous, is no substitute for external validation of our findings
307 and that our tools need to be tested on external cohorts. Importantly, though the dataset
308 used here is not of unreasonable size, we acknowledge that it may still be insufficient: firstly,
309 because of the overfitting noted in MLP models and secondly because of the imbalance
310 inherent to the issue class imbalance of mortality in HF – reflected in our dataset with a 5%
311 mortality rate (and with only 191 cases the mortality population may be under-represented
312 from a machine-learning perspective which typically requires cohort sizes numbering in the
313 1000s or greater to be trained effectively). We have restrained ourselves to conducting
314 analysis using only categorical features. Model predictions were not calibrated, and it is
315 known that certain machine learning models, particularly NB are notoriously poor at
316 estimating probabilities despite being good classifiers.

317 *3.6. Conclusion and final comments.*

318 In summary, NB was the most optimal ML model having the optimal virtues of strong
319 predictive performance, model interpretability and potential for making individualized
320 predictions. While RF, XGB and LR had similar performance capabilities, by nature they are
321 not readily interpretable (i.e. RF and XGB) or are not optimal for individualized predictions
322 (i.e. LR).

323 With ongoing development of digital infrastructure in the healthcare industry it is inevitable
324 that machine learning algorithms will only become increasingly powerful and commonplace.
325 As we await this reality, we hope that the findings here will provide physicians and clinicians
326 with a tool that can be used to rapidly identify patients at higher risk of mortality early by
327 knowledge of patient comorbidities; currently most prognostication tools can only be applied
328 later in the admission. Moreover, we hope to provide valuable insights in applying machine
329 learning models in healthcare for clinicians and researchers, in particular the advantages of

330 the computationally inexpensive NB models highlighting its simplicity and interpretability with
331 negligible compromise in performance.

332 **4. Materials and Methods**

333 *4.1. Ethics Statement*

334 The study was conducted in accordance with the Declaration of Helsinki (1964) and the
335 Council for International Organisations of Medical Sciences International Ethic Guidelines
336 and approved by the Australian Capital Territory Human Research Ethics Committee
337 (reference number: 2023.LRE.00063). Because the analysis was based on a digital
338 anonymized database, the patients' written informed consent was waived.

339 *4.2. Data Collection*

340 Our cohort comprised 3625 elderly (i.e. aged ≥ 65 years of age) patients consecutively
341 admitted to the Department of Orthopaedic Surgery at the Canberra Hospital between 1999
342 – 2019 with osteoporotic hip fracture. Patients admitted with hip fracture secondary to
343 moderate-high energy trauma, or secondary to minimal trauma but with malignancy
344 associated pathological fracture were excluded. Data on in-hospital mortality,
345 sociodemographics (age, sex, smoking status, active history of overuse of alcohol, use of
346 walking aids, and if the patient was a resident of an permanent residential care facility
347 [PRCF]) and comorbidities (presence of hypertension [HT], coronary artery disease [CAD],
348 previous history of acute myocardial infarction [MI], atrial fibrillation [AF], past history of
349 stroke [cerebrovascular accident, CVA], transient ischaemic attack [TIA], dementia,
350 Parkinson's disease [PD], chronic obstructive pulmonary disease [COPD], type 2 diabetes
351 mellitus [T2DM], chronic kidney disease [CKD], anaemia, history of solid organ malignancy,
352 osteoporosis and hyperparathyroidism [parathyroid hormone/PTH >6.8 pmol/L] and vitamin D
353 insufficiency/deficiency; (25)OH vitamin D $\leq 50/25$ nmol/L) were collected.

354 *4.2. Model Development*

355 Seven machine learning algorithms, LR (as the baseline), SVM, NB, DT, RF, XGB, and the
356 multi-layer perceptron (MLP, a 3-layer NN) were trained to predict mortality. For each
357 algorithm the following steps were taken: identification of key hyperparameters to be trained
358 (using a fractional factorial design of experiments approach), tuning of these key
359 hyperparameters (using an inner 3-fold cross-validation to identify optimal hyperparameter
360 and an outer 5-fold cross-validation to evaluate performance). Computations were performed
361 using the Python packages, sklearn and pandas [34, 35].

362 *4.3. Model Performance (and comparisons)*

363 Performance was measured using the area under the receiver operating curve (AUROC)
364 and the area under the precision-recall curve (AUPRC). Respective scores for each model
365 were evaluated with 5-fold cross-validation; the mean and standard deviation of these scores
366 was taken, and the 95% confidence interval was calculated. The student *t*-test was used to
367 compare the mean performance scores. Computations were performed using the Python
368 package SciPy (in particular 'scipy.stats' routines) [36].

369 *4.4. Feature importance – Model Interpretation*

370 Each trained model was analyzed directly. In general, the training of each model involved
371 optimization of coefficients corresponding to each patient feature (comorbidity). The trained
372 models were analyzed; for each patient comorbidity a corresponding coefficient or score was
373 computed (see Appendix A). Features were ranked by importance based on the values of
374 these scores.

375 *4.5. Feature importance – SHAP analysis*

376 For each patient, the SHAP value allocates a quantifiable credit to each variable (i.e. patient
377 comorbidity) in its contribution to the model output (i.e. the final prediction). Feature
378 importance analysis with SHAP was performed using the Python implementation [30, 33].
379 Features were ranked based on the mean SHAP values for each comorbidity.

380

381 **5. References**

- 382 1. Veronese, N. and S. Maggi, *Epidemiology and social costs of hip fracture*. Injury,
383 2018. **49**(8): p. 1458-1460.
- 384 2. White, S.M. and R. Griffiths, *Projected incidence of proximal femoral fracture in*
385 *England: a report from the NHS Hip Fracture Anaesthesia Network (HIPFAN)*. Injury,
386 2011. **42**(11): p. 1230-3.
- 387 3. Groff, H., et al., *Causes of in-hospital mortality after hip fractures in the elderly*. Hip
388 Int, 2020. **30**(2): p. 204-209.
- 389 4. Sheehan, K.J., et al., *In-hospital mortality after hip fracture by treatment setting*.
390 Cmaj, 2016. **188**(17-18): p. 1219-1225.
- 391 5. Welfare, A.I.o.H.a., *Hip fracture incidence and hospitalisations in Australia 2015-16.*,
392 A.I.o.H.a. Welfare, Editor. 2018: Canberra: AIHW.
- 393 6. Wu, T.Y., et al., *Admission rates and in-hospital mortality for hip fractures in England*
394 *1998 to 2009: time trends study*. J Public Health (Oxf), 2011. **33**(2): p. 284-91.
- 395 7. Sun, L., et al., *Validation of the Nottingham Hip Fracture Score in Predicting*
396 *Postoperative Outcomes Following Hip Fracture Surgery*. Orthop Surg, 2023. **15**(4):
397 p. 1096-1103.
- 398 8. Olsen, F., et al., *Validation of the Nottingham Hip Fracture Score (NHFS) for the*
399 *prediction of 30-day mortality in a Swedish cohort of hip fractures*. Acta Anaesthesiol
400 Scand, 2021. **65**(10): p. 1413-1420.
- 401 9. Jones, H.J. and L. de Cossart, *Risk scoring in surgical patients*. Br J Surg, 1999.
402 **86**(2): p. 149-57.
- 403 10. Mohamed, K., et al., *An assessment of the POSSUM system in orthopaedic surgery*.
404 J Bone Joint Surg Br, 2002. **84**(5): p. 735-9.

- 405 11. Hill, B.L., et al., *An automated machine learning-based model predicts postoperative*
406 *mortality using readily-extractable preoperative electronic health record data.* Br J
407 Anaesth, 2019. **123**(6): p. 877-886.
- 408 12. Hu, X.Y., et al., *Automated machine learning-based model predicts postoperative*
409 *delirium using readily extractable perioperative collected electronic data.* CNS
410 Neurosci Ther, 2022. **28**(4): p. 608-618.
- 411 13. Bishara, A., et al., *Postoperative delirium prediction using machine learning models*
412 *and preoperative electronic health record data.* BMC Anesthesiol, 2022. **22**(1): p. 8.
- 413 14. Zhang, J., L. Jiang, and X. Zhu, *A Machine Learning-Modified Novel Nomogram to*
414 *Predict Perioperative Blood Transfusion of Total Gastrectomy for Gastric Cancer.*
415 Front Oncol, 2022. **12**: p. 826760.
- 416 15. Peng, X., et al., *Machine learning prediction of postoperative major adverse*
417 *cardiovascular events in geriatric patients: a prospective cohort study.* BMC
418 Anesthesiol, 2022. **22**(1): p. 284.
- 419 16. Neto, P.C.S., et al., *Developing and validating a machine learning ensemble model to*
420 *predict postoperative delirium in a cohort of high-risk surgical patients: A secondary*
421 *cohort analysis.* Eur J Anaesthesiol, 2023. **40**(5): p. 356-364.
- 422 17. Forssten, M.P., et al., *Predicting 1-Year Mortality after Hip Fracture Surgery: An*
423 *Evaluation of Multiple Machine Learning Approaches.* J Pers Med, 2021. **11**(8).
- 424 18. Li, Y.Y., et al., *Implementation of a machine learning application in preoperative risk*
425 *assessment for hip repair surgery.* BMC Anesthesiol, 2022. **22**(1): p. 116.
- 426 19. Lei, M., et al., *A machine learning-based prediction model for in-hospital mortality*
427 *among critically ill patients with hip fracture: An internal and external validated study.*
428 Injury, 2023. **54**(2): p. 636-644.
- 429 20. Zhao, H., et al., *Machine Learning Algorithm Using Electronic Chart-Derived Data to*
430 *Predict Delirium After Elderly Hip Fracture Surgeries: A Retrospective Case-Control*
431 *Study.* Front Surg, 2021. **8**: p. 634629.

- 432 21. Metsis, V., I. Androutsopoulos, and G. Paliouras, *Spam Filtering with Naive Bayes -*
433 *Which Naive Bayes?*, in *Conference on Email and Anti-Spam*. 2006: Mountain View,
434 California USA.
- 435 22. Blanco, R., et al., *Feature selection in Bayesian classifiers for the prognosis of*
436 *survival of cirrhotic patients treated with TIPS*. *Journal of Biomedical Informatics*,
437 2005. **38**(5): p. 376-388.
- 438 23. Galiatsatos, D., et al., *Prediction of 30-Day Mortality after a Hip Fracture Surgery*
439 *Using Neural and Bayesian Networks*. Vol. 436. 2014. 566-575.
- 440 24. Cui, S., et al., *Using Naive Bayes Classifier to predict osteonecrosis of the femoral*
441 *head with cannulated screw fixation*. *Injury*, 2018. **49**(10): p. 1865-1870.
- 442 25. Yun, K., et al., *Prediction of Mortality in Surgical Intensive Care Unit Patients Using*
443 *Machine Learning Algorithms*. *Front Med (Lausanne)*, 2021. **8**: p. 621861.
- 444 26. Vellido, A., *The importance of interpretability and visualization in machine learning for*
445 *applications in medicine and health care*. *Neural Computing and Applications*, 2020.
446 **32**(24): p. 18069-18083.
- 447 27. Lu, S.C., et al., *On the importance of interpretable machine learning predictions to*
448 *inform clinical decision making in oncology*. *Front Oncol*, 2023. **13**: p. 1129380.
- 449 28. Sathyan, A., A.I. Weinberg, and K. Cohen, *Interpretable AI for bio-medical*
450 *applications*. *Complex Eng Syst*, 2022. **2**(4).
- 451 29. Ejiyi, C.J., et al., *A robust predictive diagnosis model for diabetes mellitus using*
452 *Shapley-incorporated machine learning algorithms*. *Healthcare Analytics*, 2023. **3**: p.
453 100166.
- 454 30. Duckworth, C., et al., *Using explainable machine learning to characterise data drift*
455 *and detect emergent health risks for emergency department admissions during*
456 *COVID-19*. *Scientific Reports*, 2021. **11**(1): p. 23017.
- 457 31. Tang, S., et al., *Data valuation for medical imaging using Shapley value and*
458 *application to a large-scale chest X-ray dataset*. *Scientific Reports*, 2021. **11**(1): p.
459 8366.

- 460 32. Lundberg, S.M., et al., *From local explanations to global understanding with*
461 *explainable AI for trees*. *Nature Machine Intelligence*, 2020. **2**(1): p. 56-67.
- 462 33. Lundberg, S.M. and S.-I. Lee. *A Unified Approach to Interpreting Model Predictions*.
463 *in Neural Information Processing Systems*. 2017. Long Beach, CA, USA.
- 464 34. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python*. *Journal of Machine*
465 *Learning Research*, 2011. **12**(85): p. 2825-2830.
- 466 35. McKinney, W.o. *Data Structures for Statistical Computing in Python*. *in Proceedings*
467 *of the 9th Python in Science Conference*. 2010.
- 468 36. Virtanen, P., et al., *SciPy 1.0: fundamental algorithms for scientific computing in*
469 *Python*. *Nat Methods*, 2020. **17**(3): p. 261-272.

470

471 6. Appendix A

472 In the following section we give some mathematical context to the models used and show
473 how the most predictive models (RF, XGB, NB and LR) were used to directly obtain feature
474 importance coefficients/scores for each evaluation in the 5-fold cross validation. The average
475 of these scores was used as the final feature score. The feature scores for NB and LR have
476 the added benefit of offering intuitive insight as noted above in the discussion.

477 6.1. LR (logistic regression).

478 The LR predicts a probability using the function:

$$479 f(x) = \frac{1}{1 + \exp(-x \cdot \beta - \beta)}$$

480 where $x = [x_1, x_2, \dots, x_n]$ is a vector for a patient encoding the presence or absence of a
481 comorbidity (values of '1' and '0' respectively), $\beta = [\beta_1, \beta_2, \beta_3, \dots, \beta_n]$ is a vector containing the
482 coefficient weights (corresponding, in this study, to each patient comorbidity) and β the
483 intercept value.

484

485

486 6.2. (Bernoulli) Naïve Bayes (NB).

487 Naïve Bayes uses the following classification rule:

$$488 \hat{y} = \underset{y}{\operatorname{argmax}} P(y) \prod_{i=1}^n P(x_i|y)$$

489 In Bernoulli Naïve Bayes (i.e. for binary classification, using binarized input):

$$490 P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y))(1 - x_i)$$

491 For each patient comorbidity x_i , the model coefficient (i.e. the feature importance score) was
492 computed as:

$$493 \text{coefficients} = \frac{P(x_i|y = 1)}{P(x_i|y = 0)}$$

494 where $P(x_i|y = 1)$ and $P(x_i|y = 0)$ is the probability of the i^{th} comorbidity occurring given
495 they experienced in-hospital mortality ($y = 1$) and survived to discharge ($y = 0$) respectively.

496

497 6.3. Tree-based methods (DT, RF, XGB).

498 In developing a single decision tree, the features are recursively partitioned to group patients
499 by outcome (mortality). At each step (or 'node') the feature that results in the greatest
500 reduction in 'impurity', or conversely the greatest increase in 'purity' is chosen. The
501 measured used in this paper for all tree based methods (DT, XGB and RF) was the Gini
502 impurity which is given by the formula:

503

504

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

505 Where p_{mk} is the proportion of those who died (k=1) or survived (k=0) at decision step (or
506 node) number m and subsequently simplifies to:

507

508

$$H(Q_m) = p_{m1}(1 - p_{m1}) + p_{m0}(1 - p_{m0})$$

509

510 For RFs and XGBs where features may be used more than once (multiple trees are trained)
511 the mean decrease in impurity across all nodes and trees is computed.

512

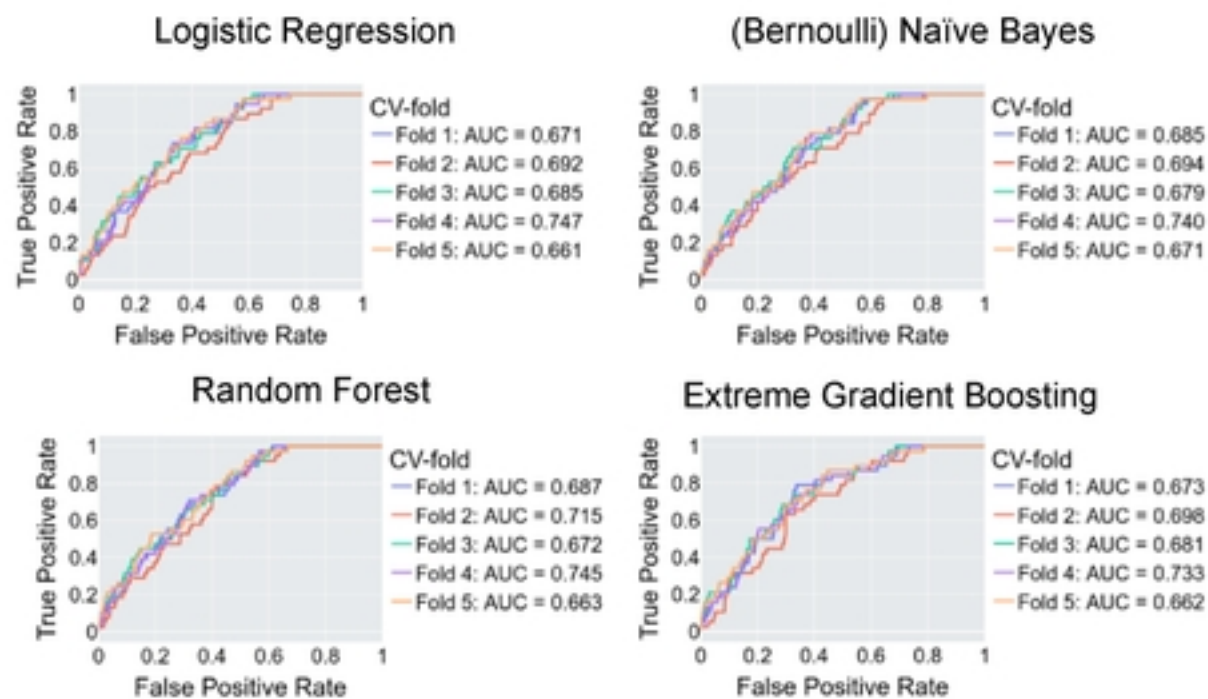
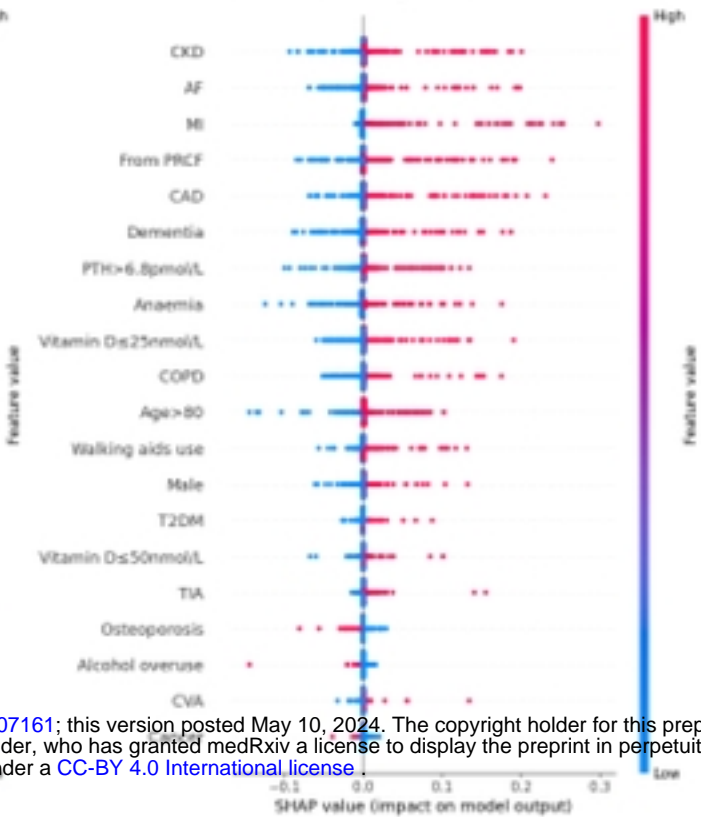
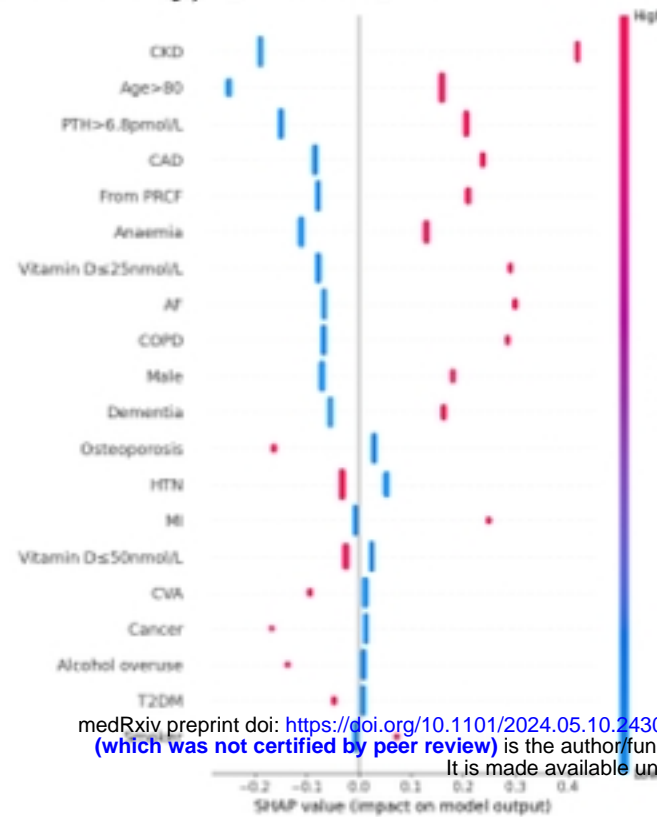


Figure 1. ML model test performance (area under the receiver operating characteristic, AUC). Only the test set AUCs evaluated from the 5-fold cross-validation for the four best-performing ML models are shown.

Figure 2. Summary plot SHAP values for patient comorbidities. Each point on the plot represents a SHAP value for an individual patient's comorbidity (SHAP value is on x-axis, corresponding comorbidity on the y-axis). Positive SHAP values corresponds to a positive/additive contribution to the prediction (i.e. in-hospital mortality); conversely a negative SHAP value corresponds to a negative/subtractive contribution. Colours of points represents feature values: magenta/red corresponded to a value of '1' (i.e. presence of the comorbidity) and blue corresponding to value '0' (i.e. absence of comorbidity).
 Logistic Regression Naive Bayes

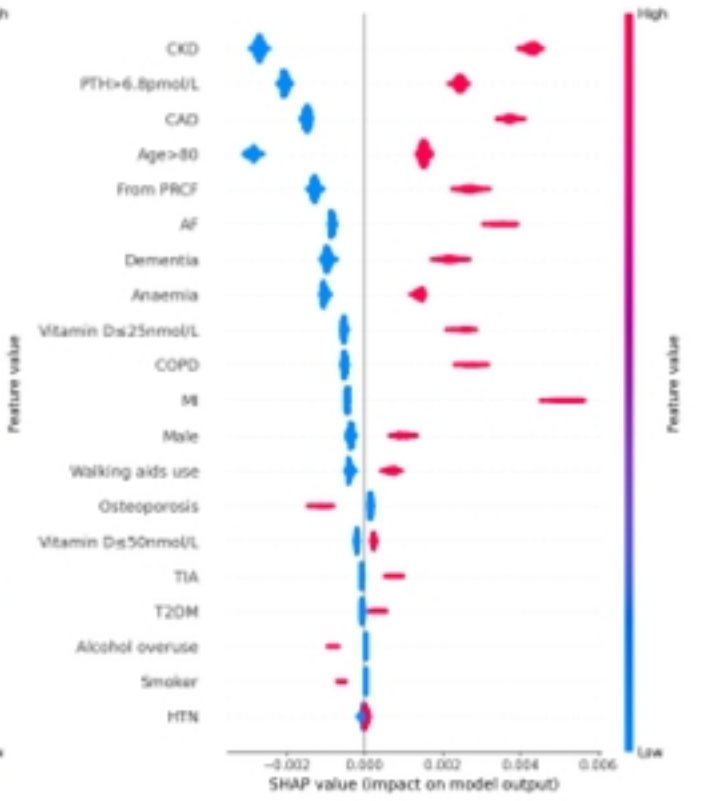


medRxiv preprint doi: <https://doi.org/10.1101/2024.05.10.24307161>; this version posted May 10, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

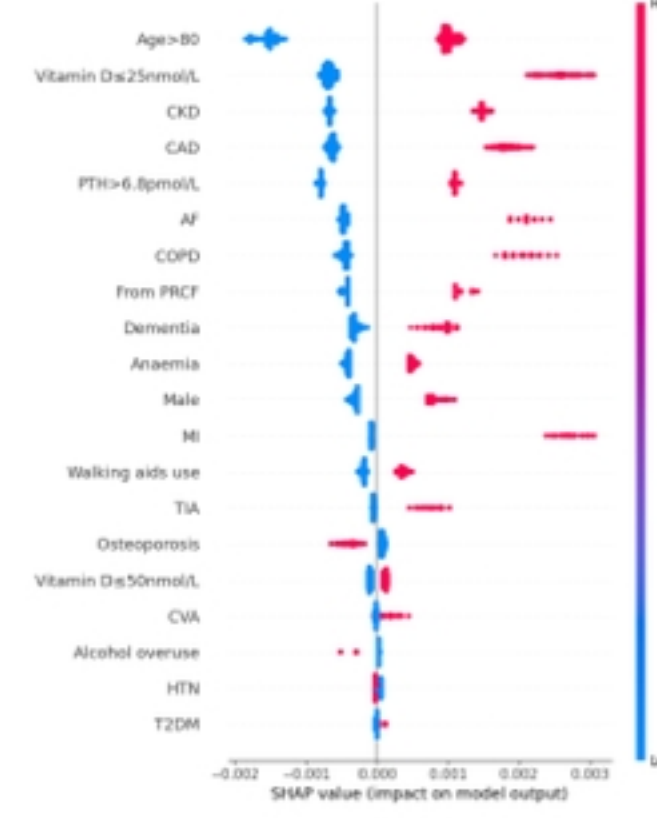
Decision Tree



Random Forest



Extreme Gradient Boosting



Multi-Layer Perceptron



Fig2