Use of Ensemble Machine Learning in Screening Electronic Health Records: A Scoping Review

Christophe A.T. Stevens, [a] *Alireza Mahani,* [b] *Kausik K. Ray,* [a] *Antonio J. Vallejo-Vaz,* [a,c,d,*] *Mansour T.A. Sharabiani,* [a,*]

[a] *Imperial College London, London, United Kingdom;* [b] *Davidson Kempner Capital Management, New York, United States;*
[c] *Instituto de Biomedicina de Sevilla, Sevilla, Spain;* [d] *University of Seville, Sevilla, Spain, * Joint first PhD supervisors.*

Imperial College London

Electronic Health Record (EHR) Screening

Every individual is born with a genetically predetermined risk of developing a disease over their lifetime.

These initial risks increase or decrease with time depending on exposure to risk factors that are frequently stored within EHRs such as smoking, cholesterol, alcohol, obesity, high blood pressure and/or, high glucose levels.

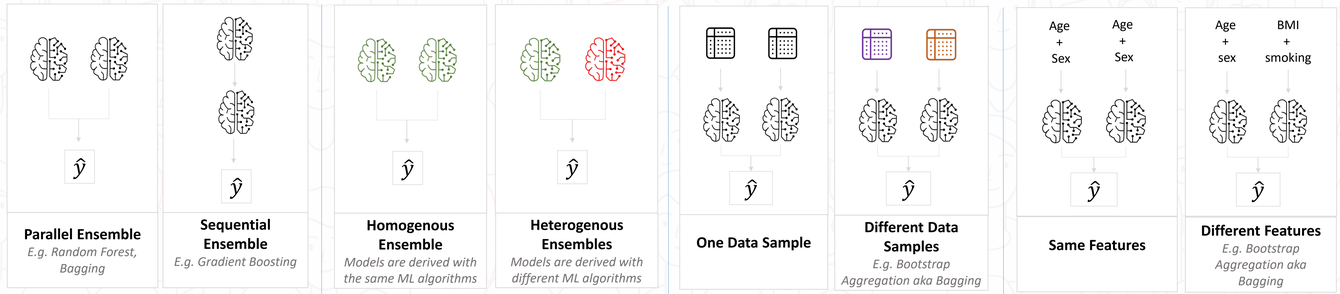
Computer algorithms can easily access these information and estimate risk across entire population, ultimately informing the case-finding process.

Ensemble Machine Learning Models (EML)

EMLs are a type of supervised ML algorithms that combine the predictions from multiple models into one. Their structure and composition can greatly vary and can be organised in 4 dimensions (see Figure 1).

As for panel of medical expert, we would expect decisions originating from an ensemble to be less erroneous than if a single expert (model) were to be consulted.

Figure 1: Four Dimensions of Ensemble Models



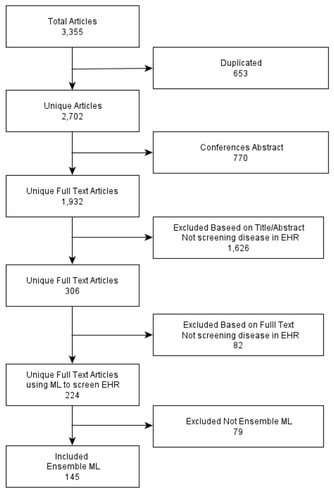
Aims & Objective

To assess the extent, nature and performances of EMLs for screening of EHRs in order to inform future ML studies using ensemble methods.

Methods

A scoping review of the literature reporting the derivation of EMLs for screening of EHRs. EMBASE and MEDLINE databases were searched across all years applying a formal search strategy using terms related to medical screening, EHR and ML.

Figure 2: Inclusion/Exclusion Diagram



Results

EMLS are increasingly being used for screening disease in EHRs across major medical specialties (Figure 3). EMLs where used in 65% of articles reporting the use of ML methods and this rate seems relatively constant over time (Figure 3). EMLs were found best models in 62% of articles comparing EMLs to non-EMLs.

In Table 1, we observe that sequential EMLs were used to a lesser extent than parallel EMLs but had a higher chance to be selected as best models by studies' authors. Within parallel EMLs majority vote was the most used fusion strategy but these EMLS had lower chance to be selected as best model compared to the other fusion strategies. Parallel EMLs with heterogenous base learners were rarer and more likely to be selected as best model than EMLs made of homogenous base learners. EMLs with bagging and random feature selection were not more likely to be selected as best models.

A hypergeometric test showed that parallel EMLs with weighted fusion strategies, XGBOOST, parallel EMLs with average probability fusion, gradient boosting and, deep learning were the less likely to be selected as best models as many time as observed by chance alone (p-value ‹ 0.05; Table 4).

Figure 3: Use of EMLs over time

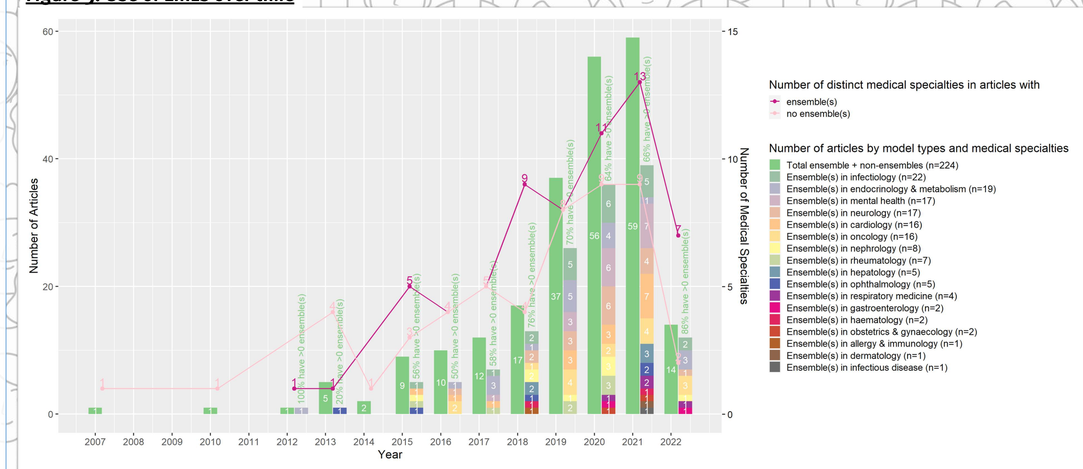


Table 1: Number of articles reporting and finding EMLs "as best model" by dimension.

	Total articles		Articles reporting comparison of EML against other models		Articles reporting that an EML was best	
	N = 145					
Dimensions	N articles (%)	Quality* %	N articles (%)	Quality* %	N articles (%)	Quality* %
Sequential (Boosting)	81 (55.9)	72.2	67 (82.7)	75.4	31 (46.3)	75.8
Parallel	118 (81.4)	63.1	93 (78.8)	68.3	37 (39.8)	64.9
→ Parallel Maj Vote	107 (73.8)	64.5	89 (83.2)	70.2	25 (28.1)	66.0
→ Parallel Average Probability	7 (4.8)	64.3	3 (42.9)	83.3	3 (100.0)	83.3
→ Parallel Weighted Vote	6 (4.1)	58.3	5 (83.3)	60.0	5 (100.0)	60.0
→ Parallel Stacking	6 (4.1)	83.3	4 (66.7)	75.0	3 (75.0)	66.7
Heterogeneous Classifiers (Base learners)	14 (9.7)	82.1	11 (78.6)	81.8	9 (81.8)	77.8
Homogeneous Classifiers (Base learners)	142 (97.9)	64.1	108 (76.1)	69.9	59 (54.6)	68.6
Data Sampling (Bagging)	112 (77.2)	62.9	89 (79.5)	68.5	27 (30.3)	59.3
No Data Sampling (No Bagging)	89 (61.4)	72.5	73 (82.0)	75.3	41 (56.2)	76.8
Random Feature Selection	108 (74.5)	63.9	89 (82.4)	69.1	26 (29.2)	61.5
No Random Feature Selection	93 (64.1)	71.5	74 (79.6)	75.0	42 (56.8)	75.0

Table 4:

ALGORITHMS COMPARED IN THE STUDIES INCLUDED IN THE SCOPING REVIEW	TWO-SIDED HYPERGEOMETRIC TEST : P(X>=X) X 2
FUSION: WEIGHTED VOTE	<0.001
XGBOOST	<0.001
FUSION: AVERAGE PROBABILITY	0.003
GRADIENT BOOSTING	0.02
DEEP LEARNING	0.02
OTHER	0.02
STACKING	0.07
RANDOM FOREST	0.13
ELASTIC NET	0.31
LASSO	0.46
CATBOOST	0.5
RULE BASED	0.79
RANDOM TREE	0.83
ARTIFICIAL NEURAL NETWORK (ANN)	0.97
MULTILAYER PERCEPTRON	1
RIDGE REGRESSION	1
NAÏVE BAYES	1
SUPPORT VECTOR MACHINE	1
LOGISTIC REGRESSION	1

Conclusion

EML methods are increasingly being adopted in medical screening of EHRs, which can have a significant impact on public health due to their ability to identify undiagnosed individuals with a potential disease with more sensitivity and specificity than non-ensemble models.

EMLs with the highest performances, such as heterogeneous EMLs or stacking/weighted average fusion EMLs, are used to a lesser extent than EMLs with more modest performances such as homogeneous and majority voting fusions EMLs.