Healthcare, Machine Learning & Security

Leila Karimi

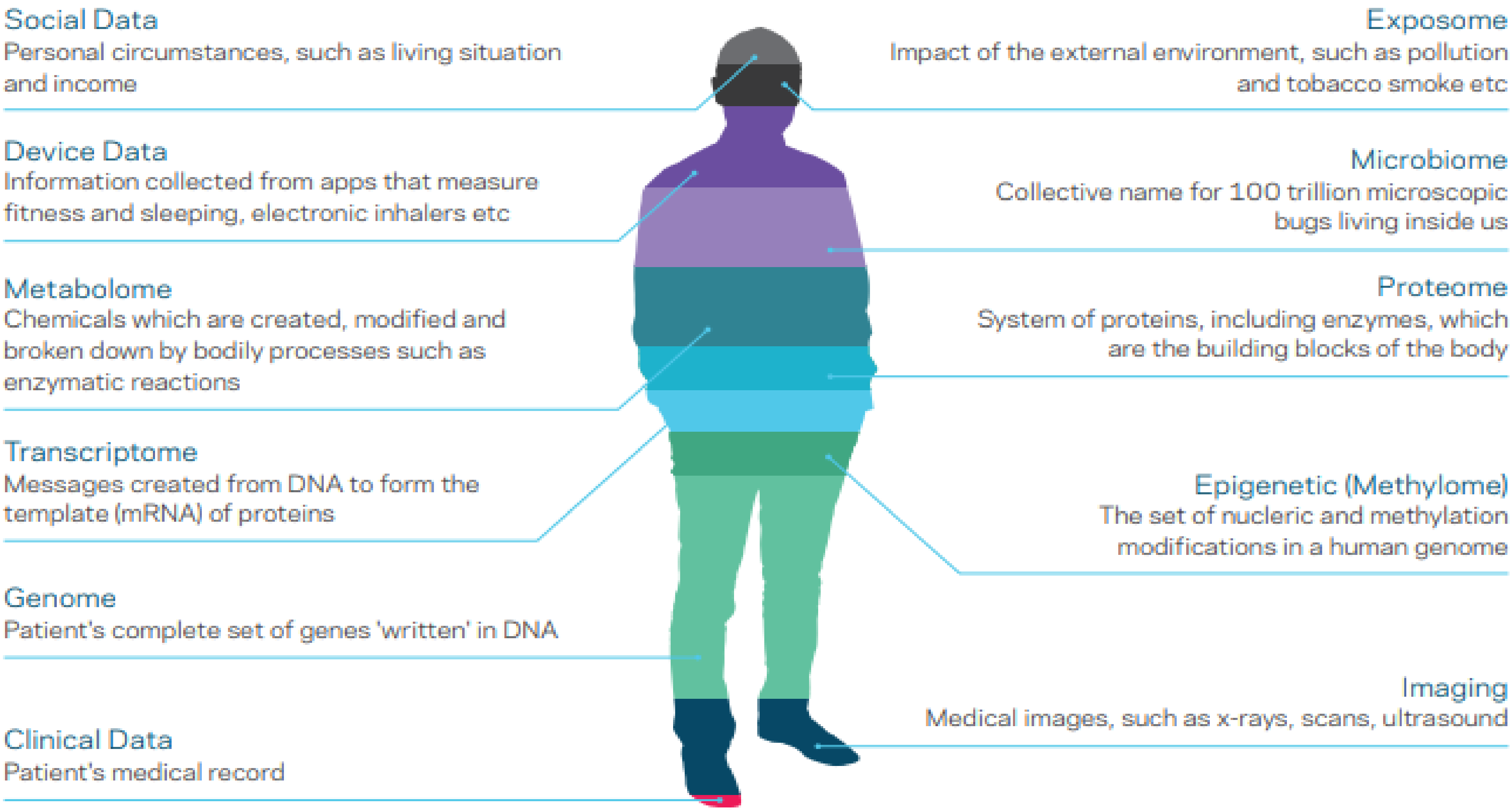
10/10/2018

Machine Learning Applications in Healthcare

Applying Machine Learning to Healthcare

- Healthcare sector is being transformed by the ability to record massive amounts of information
- Machine learning provides a way to automatically find patterns and reason about data
- It enables healthcare professionals to move to personalized care known as precision medicine.

Data useful for the practice of precision medicine



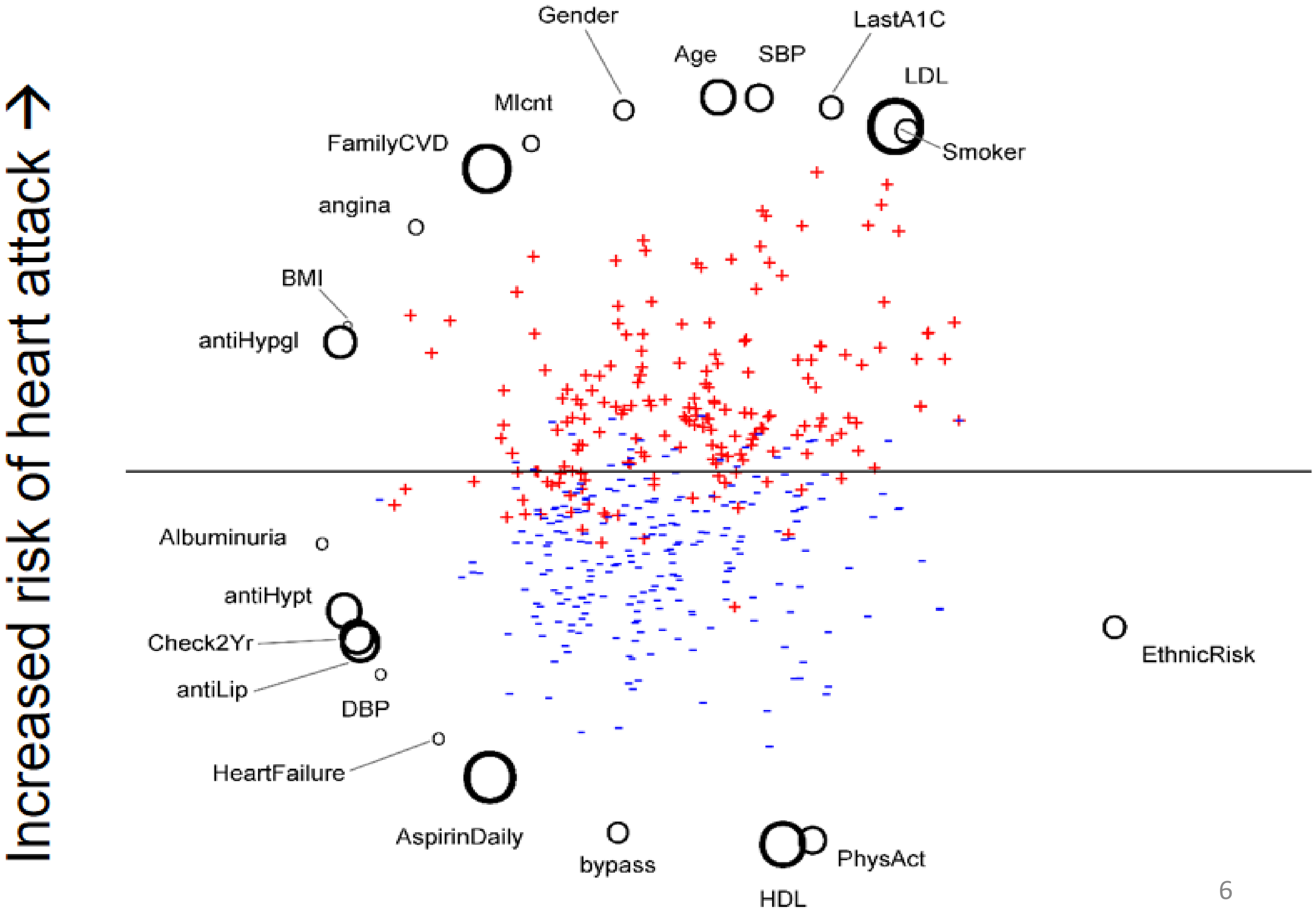
What can machine learning do for the healthcare industry?

- Improve accuracy of diagnosis, prognosis, and risk prediction.
- Reduce med
- Model and s.
- Optimize h and patient flow.
- Identify pa n medicine.
- Discover ne st practices).
- Automate detection of relevant findings in pathology, radiology, etc.

Improve quality of care and population health outcomes, while reducing healthcare costs.

Improve accuracy of diagnosis and risk prediction

- New methods are developed for chronic disease **risk prediction** and **visualization**.
- These methods give clinicians a comprehensive view of their patient population, risk levels, and risk factors, along with the estimated effects of potential interventions.

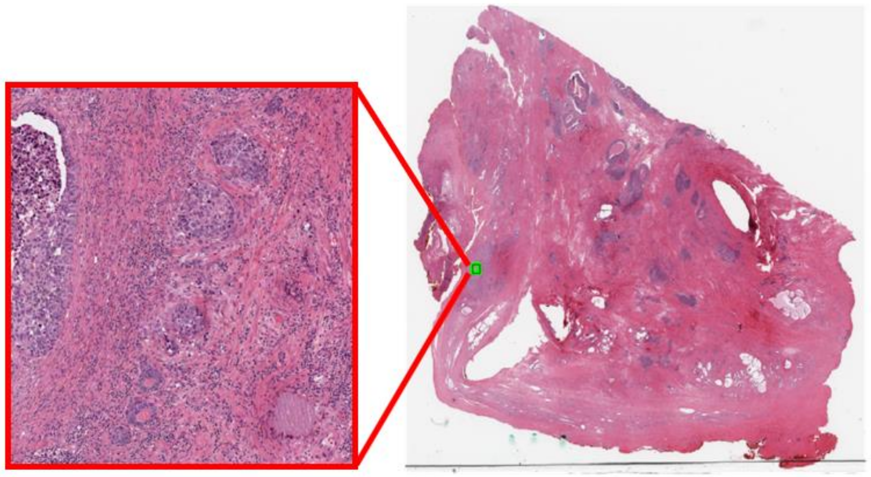


Optimize hospital processes

- By early and accurate prediction of each patient's **Diagnosis Related Group** (DRG), demand for scarce hospital resources such as beds and operating rooms can be better predicted.

Automate detection of relevant findings

- Pattern detection approaches have been successfully applied to detect regions of interest in digital pathology slides, and work surprisingly well to detect cancers.



- Automatic detection of anomalies and patterns is especially valuable when the key to diagnosis is a tiny piece of the patient's health data.

Security of Machine Learning in Healthcare

Machine Learning Security

- Although Machine Learning models are very beneficial in healthcare domain, there are several types of attacks against these models:
  - Model Inversion Attack
  - Membership Inference Attack
  - Poisoning Attack
  - Machine Learning Models that Remember Too Much

Model Inversion Attack

Fredrikson, Matthew, et al. “***Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing***.” *USENIX Security Symposium*. 2014.

The slides are from the author’s presentation in USENIX Security 2014.

Overview

- Model Inversion Attack:
  - Extracting patients’ genetics from *pharmacogenetic dosing models*
- With an end-to-end study, it shows that Differential Privacy prevents the attack
  - However, risk of adverse outcomes is too high with DP

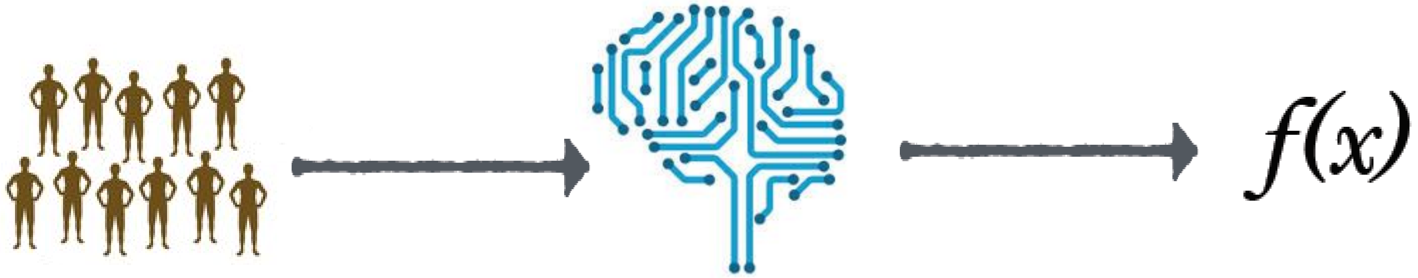
Conclusion

Current methods fail to balance privacy and utility

This really matters when inaccuracy is expensive

Pharmacogenetic

- Using machine learning models to guide medical treatments based on patient’s genotype and background



Population Dataset Learning Algorithm Trained Model

- Genotype: The actual set of genes an individual has, or is made up of is a ***genotype***

Pharmacogenetic

Genotype

Clinical Variable

$f(x)$

Trained Model

Medical Guidance

Warfarin Dosing

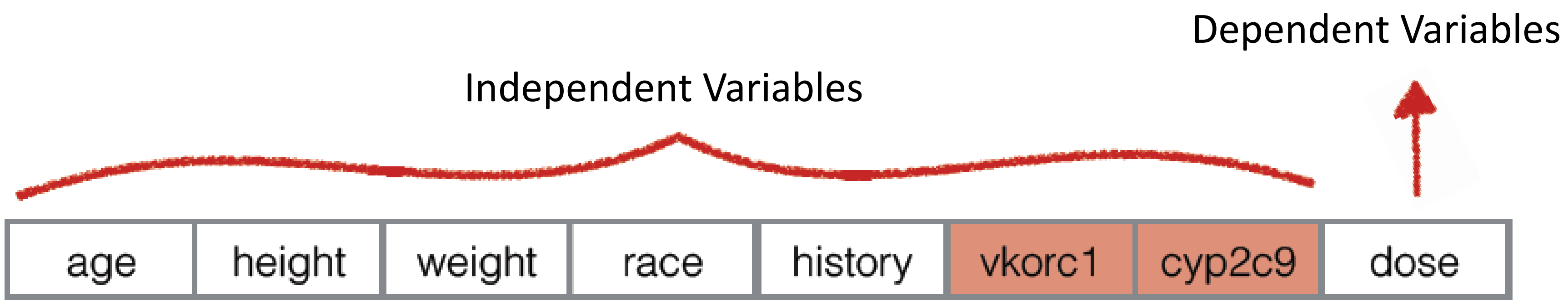
- Warfarin is a drug widely used to help prevent heart attacks, strokes, and blood clots
- Warfarin is one of the most well-studied targets in pharmacogenetics
- Warfarin is notoriously difficult to dose correctly

Low dose	**High dose**
Embolism	Intracranial Bleeding
Stroke	Extracranial Bleeding

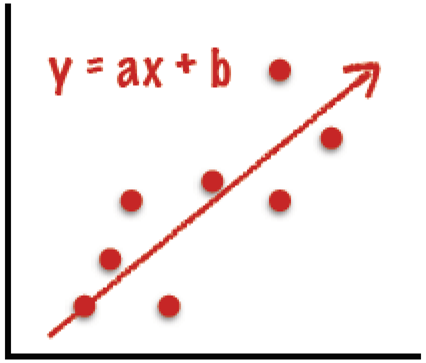
The IWPC Warfarin Model

- Population Dataset: 5700 patients from 21 hospitals in 6 countries, 4 continents

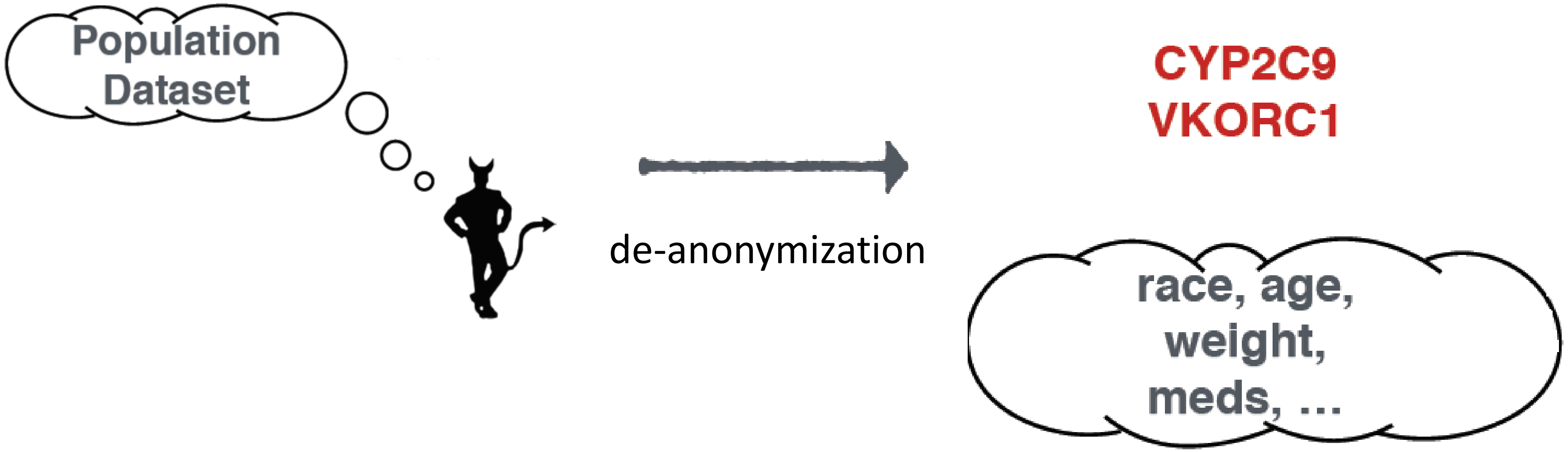
The IWPC Warfarin Model



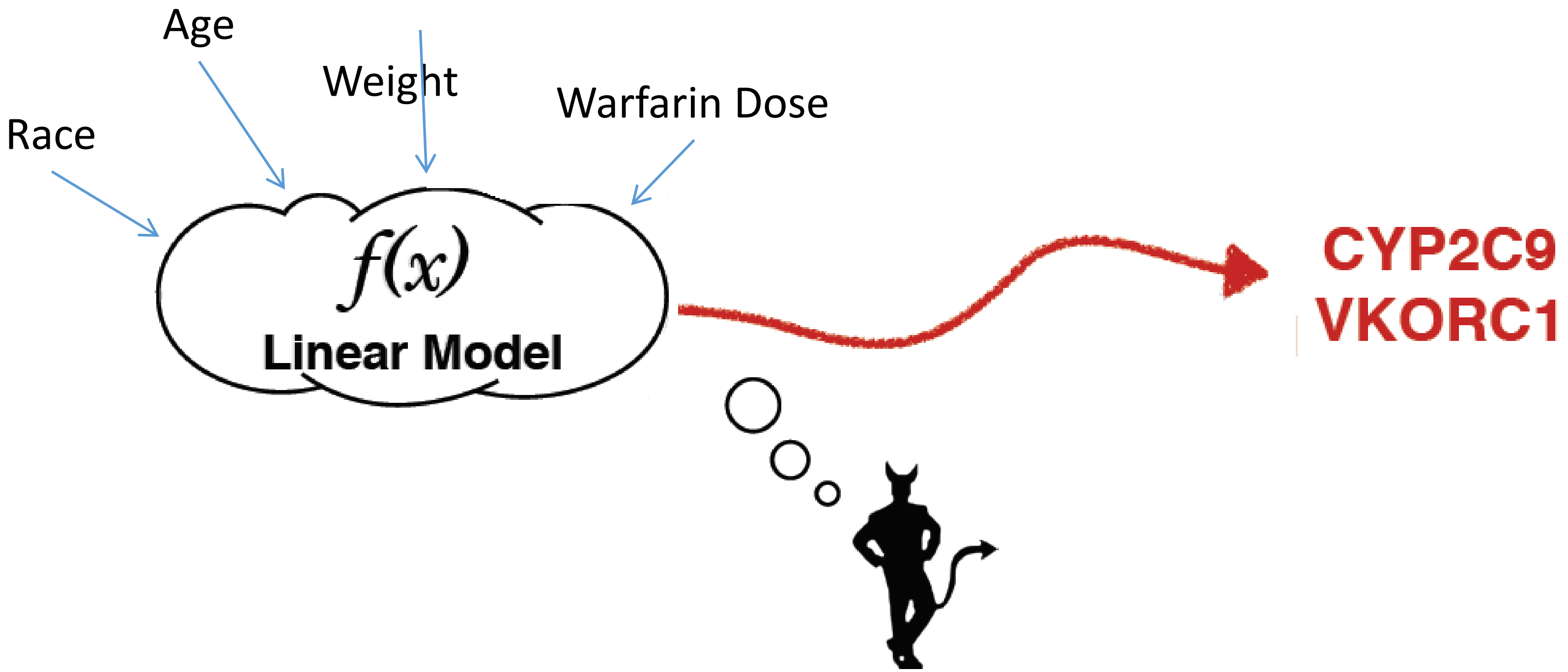
- **The IWPC found ordinary linear regression to be the best learning algorithm**



Pharmacogenetic Privacy



Model Inversion Attack



age	height	weight	race	history	vkorc1	cyp2c9	dose
50-60	176.2	185.7	asian	cancer	A/G	*1/*3	42.0

Model Inversion

basic demographics
stable warfarin dose
black-box access to model
marginal priors on patient distribution

- Goal: infer the patient's genetic markers from this information

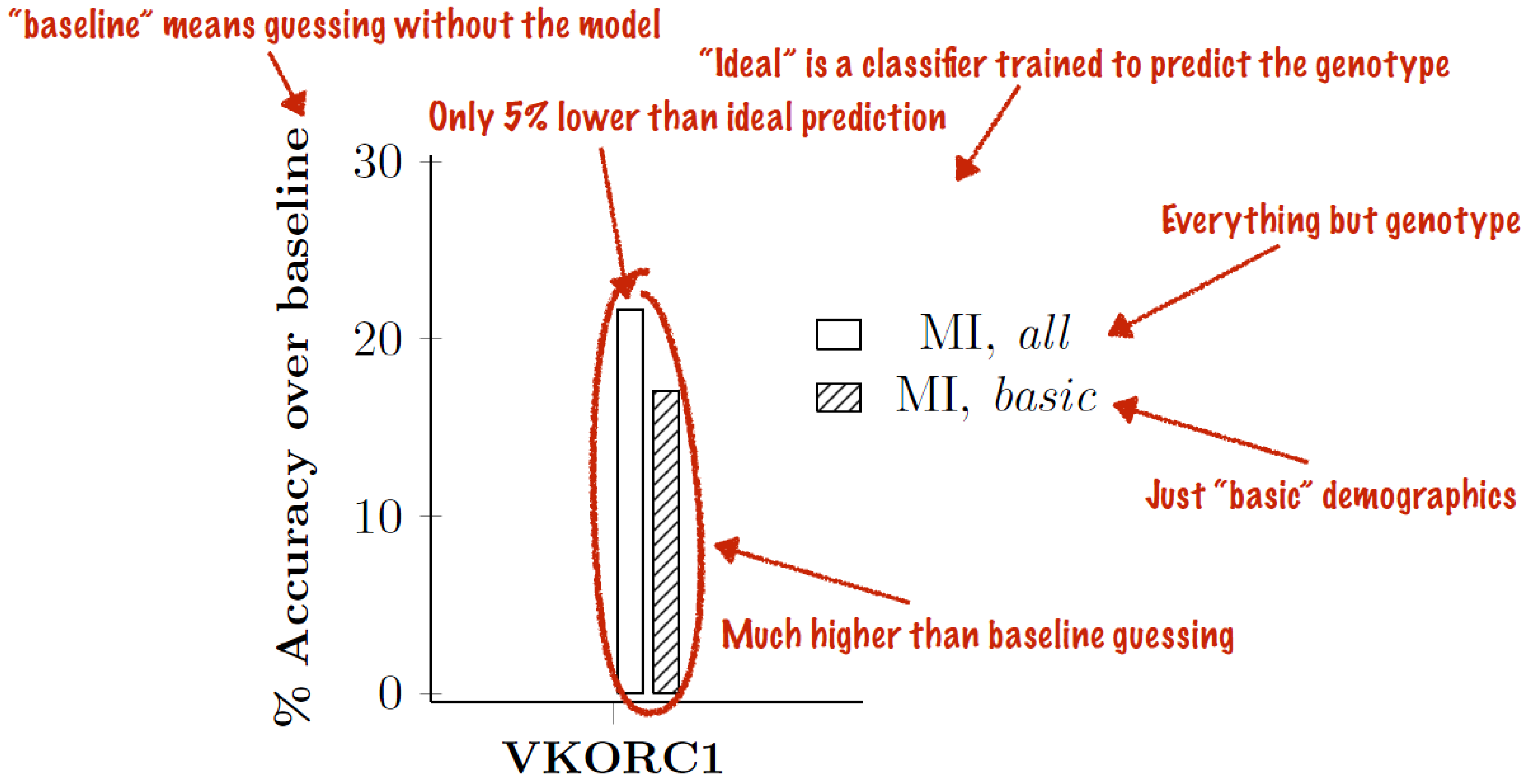
Their Model Inversion

- Compute all values that agree with given information

	age	height	weight	race	history	vkorc1	cyp2c9	dose		
$f(x)$	50-59	176.53	144.2	white	cancer			42.0	49.7	*p=0.23*
	50-59	176.53	144.2	white	heart			42.0	42.0	*p=0.75*
	50-59	176.53	144.2	white	diabetes			42.0	39.2	*p=0.01*

- Find the most likely values among those

Results



- Model Inversion does nearly as well as a linear model trained from the original data

Differential Privacy

- Model Inversion is a problem, so how can it be prevented?
- The paper examines the use of Differential Privacy for preventing Model Inversion
- Most Differential Privacy approaches add noise according to privacy budget.

Differential Privacy (Cont.)

- Any output should be about as likely regardless of whether or not a specific row is in the dataset

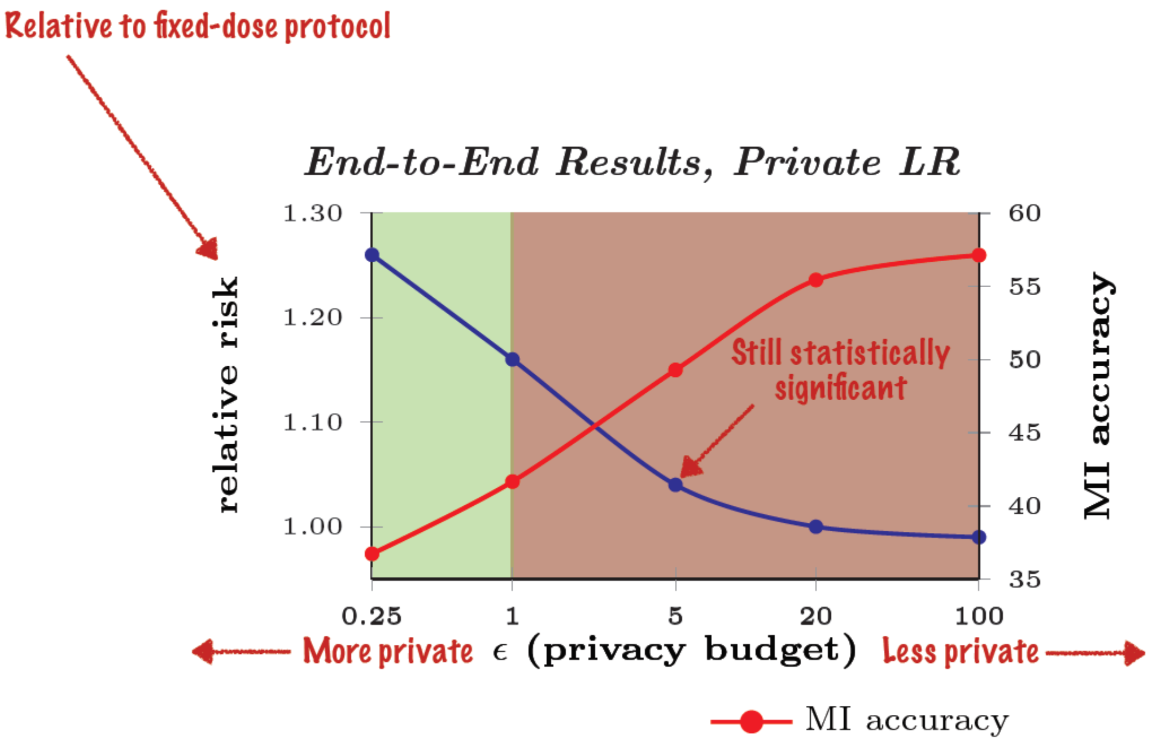
A mechanism K achieves ε-differential privacy if for all databases D_1, D_2 differing in at most one row, and all $S \subseteq \mathrm{Range}(\mathsf{K})$,

$$\Pr[\mathsf{K}(D_1) \in S] \leq \exp(\varepsilon) \times \Pr[\mathsf{K}(D_2) \in S]$$

Seeking a Remedy

- Goal: see if a “reasonable” privacy budget solves the problem
- End-to-End study:
  - Find budget that prevents model inversion
- Two Differential Privacy models
  - **Private Linear Regression**[Zhang et al., VLDB 2012]
    - Differentially private algorithms for learning linear regression models
  - **Private Histograms**[Vinterbo, ECML-PKDD 2012]
    - Differentially private projected histograms for learning binary and multinomial logistic regression models
- Evaluate risk of adverse outputs at these budgets

Results



- For privacy budgets effective at preventing Model Inversion attacks, patients would be exposed to increased risk of mortality

Conclusion

- **The paper did not observe a budget that significantly prevented model inversion, without introducing risk over fixed dosing**

Conclusion

Current methods fail to balance privacy and utility

This really matters when inaccuracy is expensive

Poisoning Attack

Mozaffari-Kermani, Mehran, et al. "***Systematic poisoning attacks on and defenses for machine learning in healthcare***." *IEEE journal of biomedical and health informatics* 19.6 (2015): 1893-1905.

Poisoning Attack

- Is a form of causative attacks in which the adversary feeds carefully crafted poisonous data points into the training set
- Results in degradation of model accuracy

Poisoning Attack in Healthcare Domain

- It can cause two main problems in healthcare domain
  - False Negative:
    - Hindrance of a diagnosis may have life-threatening consequences and could cause distrust
  - False Positive:
    - False diagnosis prompt users to distrust the machine-learning algorithm and even abandon the entire system

Proposed Poisoning Attack

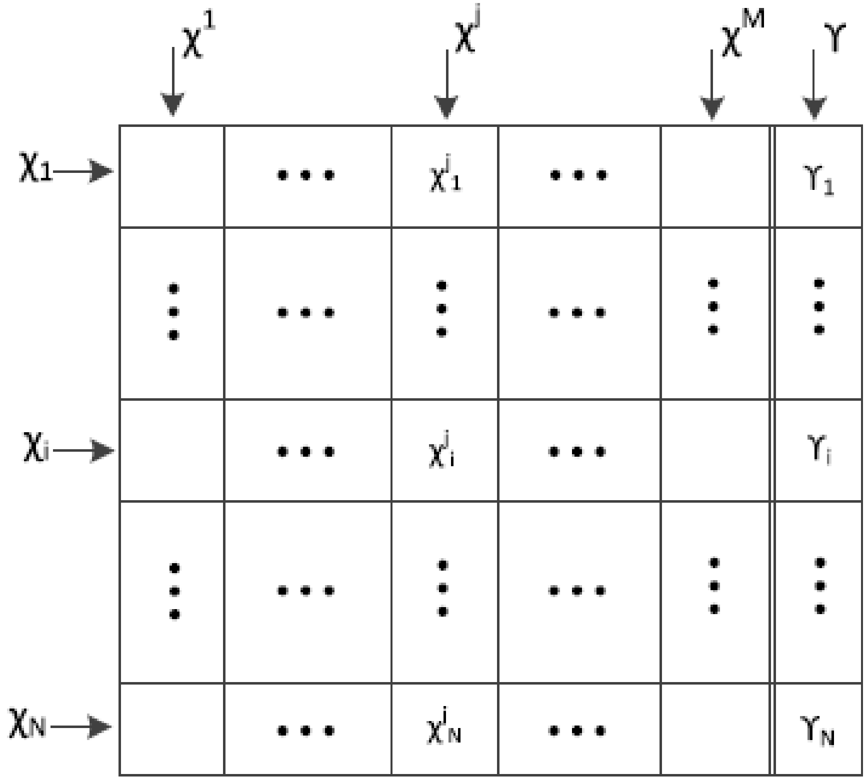
- The proposed attack procedure generates input data, which, when added to the training set, can either
  - Cause the results of machine learning to have targeted errors (e.g., increase the likelihood of classification into a specific class), OR
  - Introduce arbitrary errors (incorrect classification)
- The attacks are algorithm-independent
  - They can be applied to a wide range of machine-learning algorithms
- They can be applied to both <u>fixed</u> and <u>evolving</u> datasets

Attack model

- Assumptions:
  - The attackers have knowledge of the training dataset
    - However, the success of the proposed attacks is only dependent on the knowledge of the statistics of the training dataset
  - The attackers have access to significant computing resources
    - They can repeatedly modify the training dataset and evaluate the effectiveness of the modifications by constructing models and testing them on a validation dataset.
- The attack model considers poisoning attacks in which attackers can only *add* malicious data and they are not capable of arbitrarily manipulating datasets

Notations

Notation	Definition
N	number of instances
M	number of attributes
$\chi_i, 1 \leq i \leq N$	ith instance
$\chi_i^j, 1 \leq i \leq N, 1 \leq j \leq M$	jth attribute value of ith instance
$\chi^j, 1 \leq j \leq M$	jth attribute
$\Upsilon_i, 1 \leq i \leq N$	ith class label

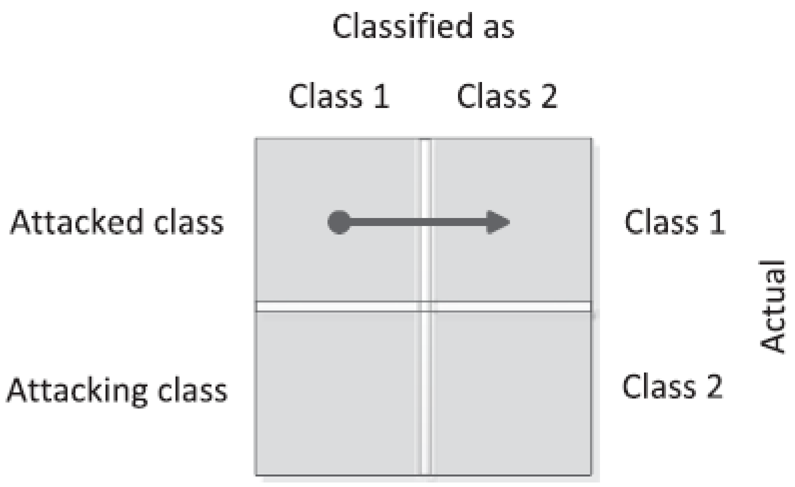


Attack Objectives

- An example of targeted attacks in Thyroid Disease dataset, in which data instances are associated with two classes: normal and hypothyroid
  - Targeted attacks compromise the effectiveness of the machine-learning algorithm either to
    - prevent a hypothyroid diagnosis or
    - to falsely lead to a hypothyroid diagnosis

Attack Objectives

- Hypothyroid diagnosis prevention scenario:
  - hypothyroid class is denoted as the ***attacked*** class and the benign class is denoted as the ***attacking*** class
  - The attacker adds malicious instances to the training dataset such that
    - instances belonging to the attacked class (Class 1) are predicted and classified as belonging to the attacking class (Class 2)



Attack Scheme

- Let the original dataset be denoted as $D \in$ (X, Y) with N instances
- Algorithm 1 adds N' malicious instances to the original dataset to create a manipulated dataset $D' \in$ (X, Y) with $N + N'$ instances
- To add a malicious instance, I pseudorandom candidates are generated
- The candidate that results in the highest degradation in classification accuracy is selected and added to the dataset.

Algorithm 1 Algorithm-independent attacks.

```
Input: Dataset D ∈ (χ, ϒ) with N instances, validation dataset V, number of iterations I.
Output: Maliciously manipulated dataset D' ∈ (χ, ϒ) with N + N' instances, where N' is the number of added malicious instances.
Begin
    Assign D' ← D
    for k = 1 to N' do
        //Select kth malicious instance
        for i = 1 to I do
            Use Algorithm 2 to generate malicious instance candidate i
            Add the candidate to D' to create a temporary training set D_T ∈ (χ, ϒ) with N + k instances
            Build the model using D_T and record its classification accuracy on the validation set V as A_i
        endfor
        Select instance î such that A_î = min(A_i), 1 ≤ i ≤ I
        Add instance î to D'
    endfor
End
Return: D' ∈ (χ, ϒ).
```

Pseudorandom Generator

- Algorithm 2 generates candidates whose attribute values match the statistics of the attacked class
- The labels of these records are set to the attacking class

Algorithm 2 Deriving a malicious instance candidate.

Input: $\chi^j, 1 \leq j \leq M$ and $\Upsilon_i, 1 \leq i \leq N$, g bins (a specified constant).
Output: $\chi_{N+1}, \Upsilon_{N+1}$ (malicious instance candidate).
Denote: $\eta_{k,j}$ and $\eta'_{k,j}$ as the number of entries in $\chi^j(k)$ corresponding to the attacked class and the attacking class, respectively.
Begin
  for $j = 1$ to M do
    for $k = 1$ to g do
      Calculate $\eta_{k,j}$ and $\eta'_{k,j}$
      Assign $W_k \leftarrow \frac{\eta_{k,j}}{\eta'_{k,j}}$
    endfor
  Compute attribute probabilities ($P_k = \frac{W_k}{\sum_{1 \leq i \leq g} W_i}$), $k = 1$ to g
  Weighted function S selects attribute value α_j pseudorandomly based on attribute probabilities
  endfor
End
Return: Malicious instance candidate is $\chi_{N+1} = \{\alpha_j, 1 \leq j \leq M\}$, Υ_{N+1} = Attacking class.

Experimental Evaluation

- The proposed attack procedure is applied to
  - Six machine-learning algorithms
  - Five medical datasets
- The attack is implemented using the Weka 3 machine-learning workbench

MACHINE-LEARNING ALGORITHMS

Name	Details
BFTree (Best-first decision tree)	Tree-based with binary splits on attributes
Ridor (Ripple-down rule learner)	Rule-based through knowledge acquisition
NBTree (Naive Bayes decision tree)	Decision tree with naive Bayes classifiers
IB1 (Nearest-neighbor classifier)	Normalized Euclidean distance-based
MLP (MultilayerPerceptron)	Feedforward artificial neural network-based
SMO (Sequential minimal optimization)	Support-vector machine-based

DETAILS OF DATASETS WITH NUMBER OF ATTRIBUTES IN EACH TYPE IN PARENTHESES

Name	#Inst., #Attr.	Attr. types
Thyroid Disease	7104, 21	Numeric
Breast Cancer	699, 10	Nominal
Acute Inflammations	120, 6	Numeric (1), Nominal (5)
Echocardiogram	132, 12	Numeric (10), Nominal (2)
Molecular Biology	3190, 61	Nominal

Experimental Results

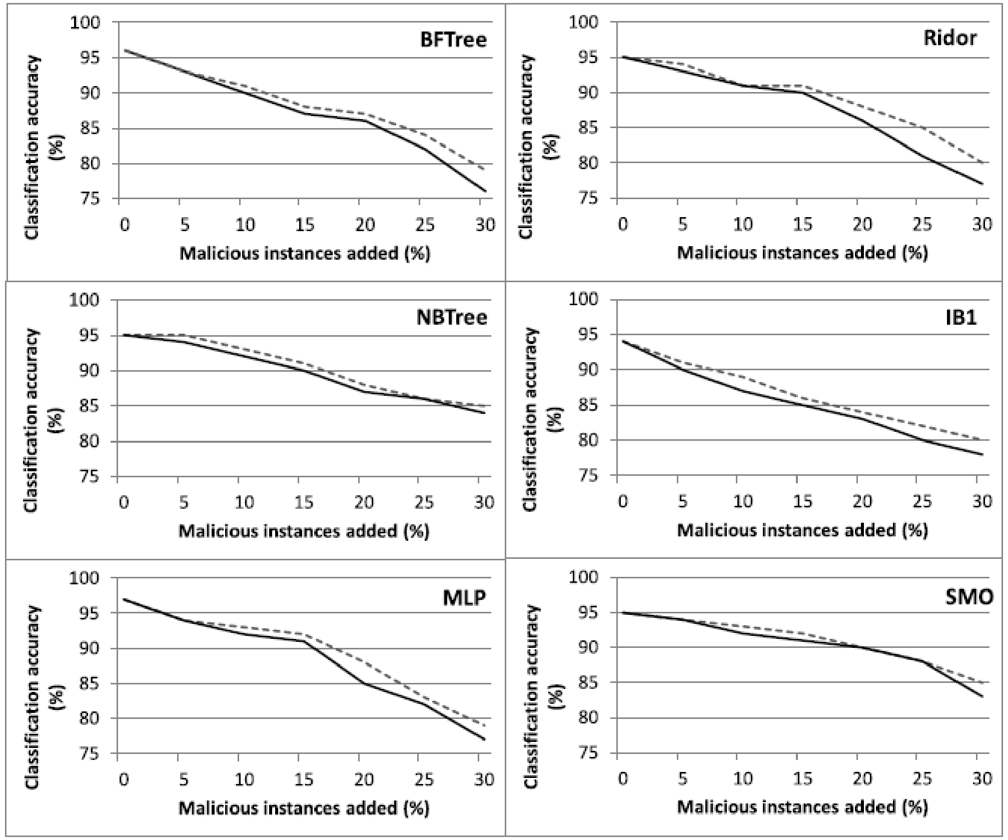


Fig. 4. Results of attacks on the Thyroid Disease dataset for the fixed (solid line) and evolving (dashed line) cases.

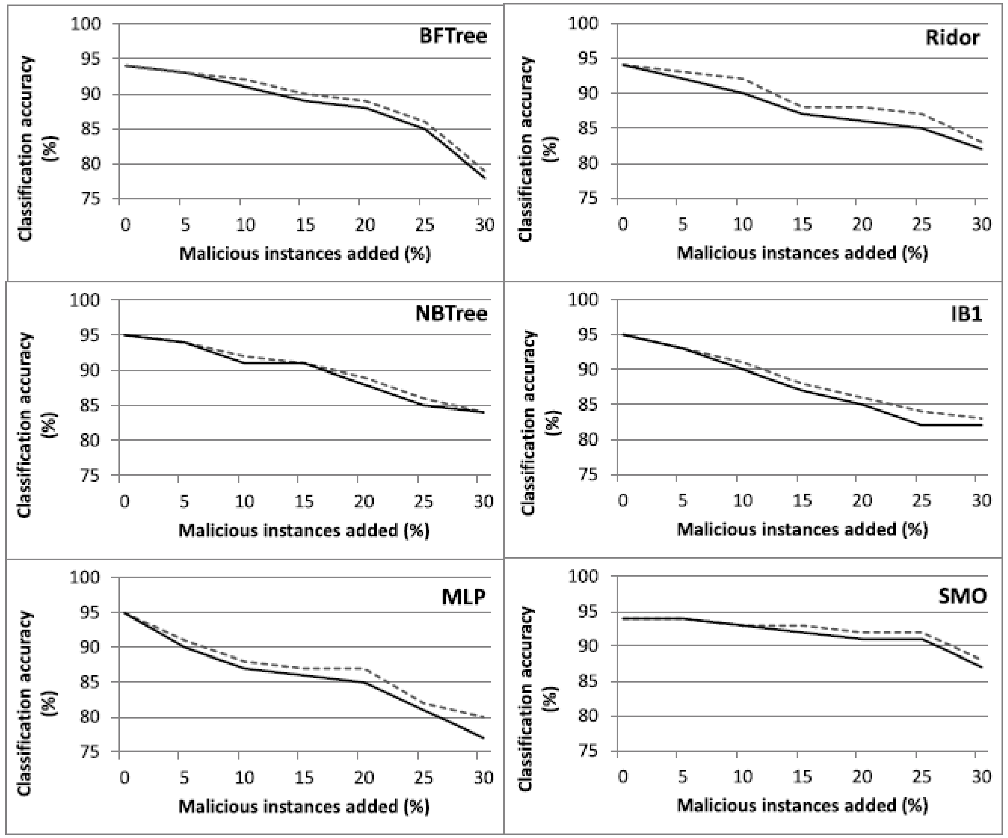


Fig. 5. Results of attacks on the Breast Cancer dataset for the fixed (solid line) and evolving (dashed line) cases.

Experimental Results

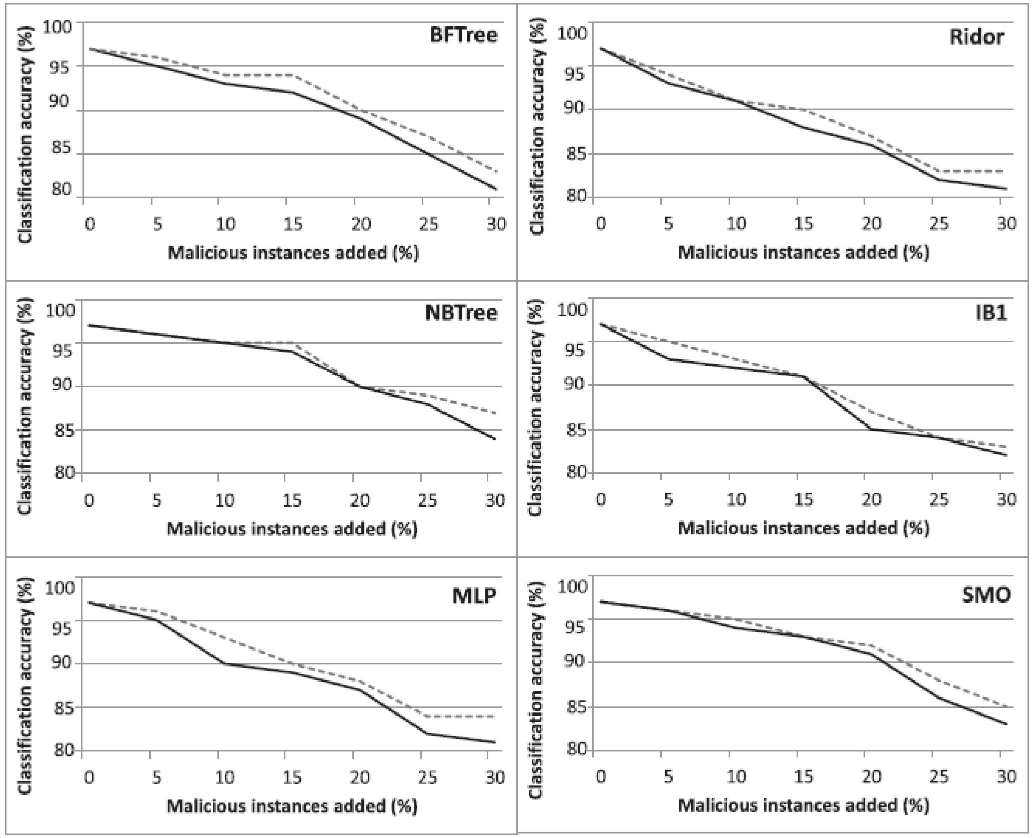


Fig. 6. Results of attacks on the Acute Inflammations dataset for the fixed (solid line) and evolving (dashed line) cases.

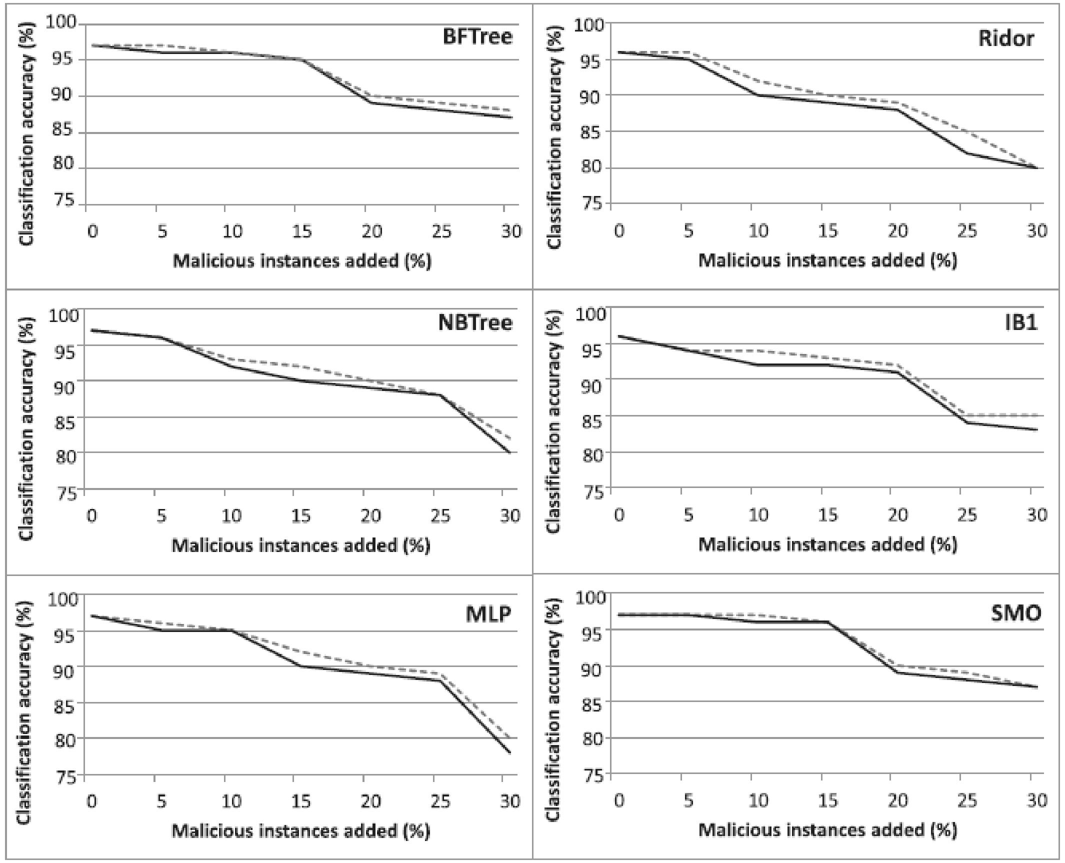


Fig. 7. Results of attacks on the Echocardiogram dataset for the fixed (solid line) and evolving (dashed line) cases.

Comparison of Different ML Algorithms

- The results indicate that SMO is found to be the most robust ML algorithm.

COMPARISON OF THE EFFECTIVENESS OF OUR ATTACKS ON THE MACHINE-LEARNING ALGORITHMS CONSIDERED

Attack	Thyroid Disease (most to least vulnerable)						Breast Cancer (most to least vulnerable)					
15% added	**IB1** (9%)	BFTree (7%)	Ridor (5%)	NBTree (4%)	MLP (3%)	**SMO** (3%)	**MLP** (14%)	IB1 (11%)	BFTree (8%)	Ridor (4%)	NBTree (3%)	**SMO** (3%)
30% added	**MLP** (20%)	Ridor (18%)	BFTree (18%)	IB1 (16%)	NBTree (13%)	**SMO** (12%)	**MLP** (26%)	BFTree (23%)	Ridor (22%)	IB1 (18%)	NBTree (16%)	**SMO** (16%)
	Acute Inflammations (most to least vulnerable)						Echocardiogram (most to least vulnerable)					
15% added	**Ridor** (9%)	BFTree (9%)	IB1 (8%)	NBTree (8%)	MLP (6%)	**SMO** (6%)	**Ridor** (8%)	NBTree (8%)	IB1 (7%)	MLP (6%)	BFTree (3%)	**SMO** (3%)
30% added	**Ridor** (21%)	BFTree (18%)	IB1 (18%)	MLP (14%)	NBTree (12%)	**SMO** (12%)	**NBTree** (20%)	IB1 (18%)	MLP (16%)	Ridor (16%)	BFTree (11%)	**SMO** (11%)
	Molecular Biology (most to least vulnerable)											
15% added	**IB1** (9%)	BFTree (9%)	Ridor (7%)	NBTree (6%)	MLP (6%)	**SMO** (5%)	Note: Changes in the misclassification percentage compared to the original dataset, i.e., the effectiveness of attacks, are shown in parentheses.					
30% added	**BFTree** (18%)	IB1 (17%)	MLP (15%)	NBTree (15%)	Ridor (12%)	**SMO** (12%)						

Countermeasures Against Poisoning Attacks

- The proposed countermeasure is based on
  - Periodically constructing a model using the training dataset
  - Evaluating its accuracy on the validation dataset
  - Raising an alarm in case of any suspicious change in the accuracy metrics
- Metrics for evaluating the accuracy of classification
  - Correctly classified instances (CCI): This statistic indicates the fraction of instances that are classified correctly
  - Kappa statistic: This statistic measures relative improvement over random predictors

Countermeasures Effectiveness

CHANGE IN ACCURACY METRICS UNDER POISONING ATTACKS

Attack				Thyroid Disease							Breast Cancer			
	Metric	SMO	NBTree	BFTree	MLP	Ridor	IB1	*Metric*	SMO	NBTree	BFTree	MLP	Ridor	IB1
15% added	CCI	3%++	4%+	7%	3%++	5%	9%	CCI	3%++	3%++	8%+	14%	4%+	11%
	Kappa	8%++	10%+	18%	8%++	13%	24%	Kappa	8%++	8%++	20%	35%	10%+	26%
30% added	CCI	12%++	13%+	18%	20%	18%	16%	CCI	12%++	13%++	18%	20%	18%	16%
	Kappa	32%++	34%+	47%	52%	47%	42%	Kappa	30%++	32%+	45%	50%	45%	40%
				Acute Inflammations							Echocardiogram			
15% added	CCI	6%++	8%+	9%	6%++	9%	8%+	CCI	3%++	8%	3%++	6%+	8%	7%
	Kappa	15%++	20%+	22%	15%++	22%	20%+	Kappa	8%++	20%	8%++	15%+	20%	17%
30% added	CCI	12%++	12%++	18%	14%+	21%	18%	CCI	11%++	20%	11%++	16%+	16%+	18%
	Kappa	30%++	30%++	45%	35%+	52%	45%	Kappa	26%++	50%	26%++	40%+	40%+	45%
				Molecular Biology										
15% added	CCI	5%++	6%+	9%	6%+	7%	9%							
	Kappa	13%++	15%+	22%	15%+	18%	22%							
30% added	CCI	12%++	15%+	18%	15%+	12%++	17%							
	Kappa	30%++	36%+	45%	36%+	30%++	43%							

Note: The lowest and the second to lowest changes in each of the statistics are depicted by "++" and "+," respectively.

References

- http://web.orionhealth.com/rs/981-HEV-035/images/Introduction_To_Machine_Learning_US.pdf
- Fredrikson, Matthew, et al. “Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing.” *USENIX Security Symposium*. 2014.
- Mozaffari-Kermani, Mehran, et al. "Systematic poisoning attacks on and defenses for machine learning in healthcare." *IEEE journal of biomedical and health informatics* 19.6 (2015): 1893-1905.