Deploying Machine Learning Algorithms in Healthcare: Challenges

Machine Learning for Healthcare and Diabetes Workshop

Ilker Demirel

Bilkent University
Department of Electrical and Electronics Engineering
Cognitive Systems, Bandits, and Optimization Research Group

January 19, 2022

CYBORG

Artificial Intelligence (AI)

- computers mimicking human abilities such as perception, decision-making etc.

Artificial Intelligence (AI)

- computers mimicking human abilities such as perception, decision-making etc.
- recent advancements: **deep learning**

Artificial Intelligence (AI)

- computers mimicking human abilities such as perception, decision-making etc.
- recent advancements: **deep learning**
- **input**, **output**, **model**, and a **loss function**

Artificial Intelligence (AI)

- computers mimicking human abilities such as perception, decision-making etc.
- recent advancements: **deep learning**
- **input**, **output**, **model**, and a **loss function**

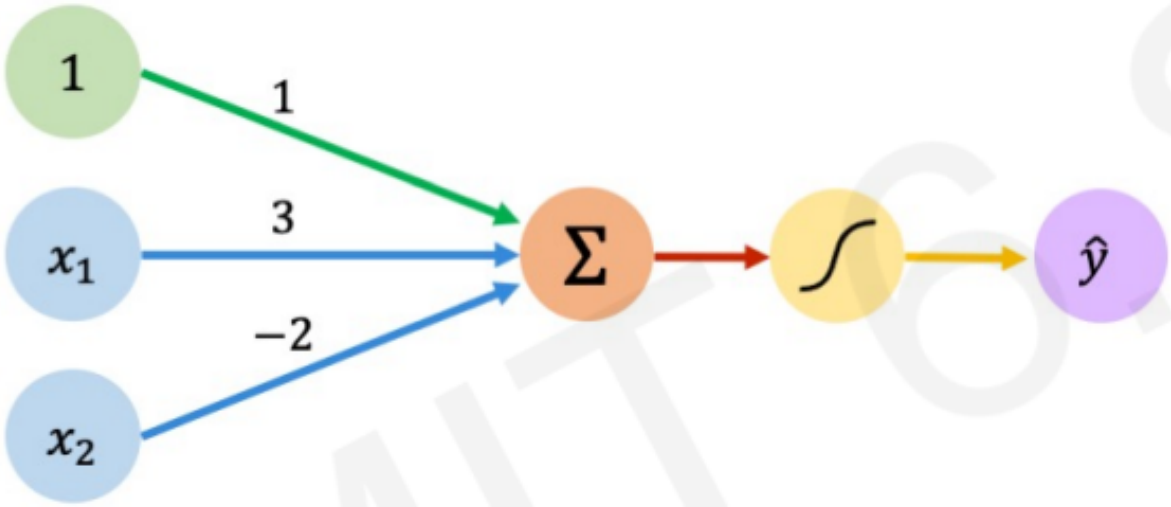
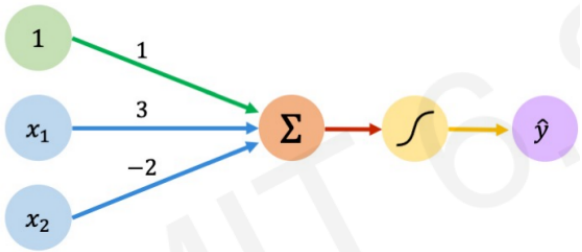
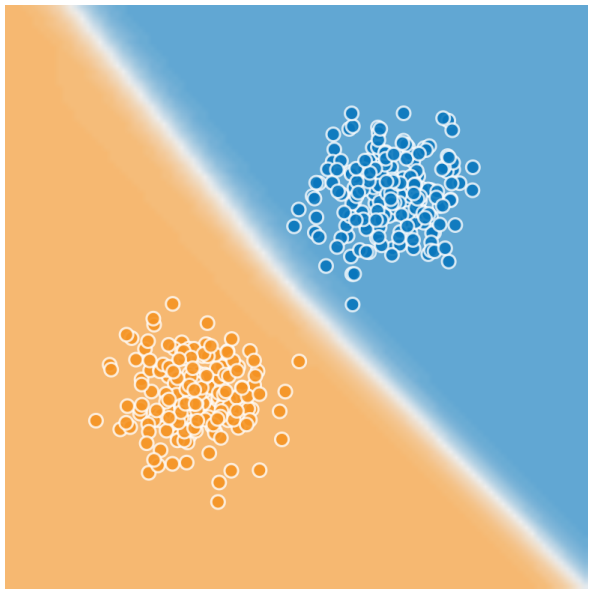


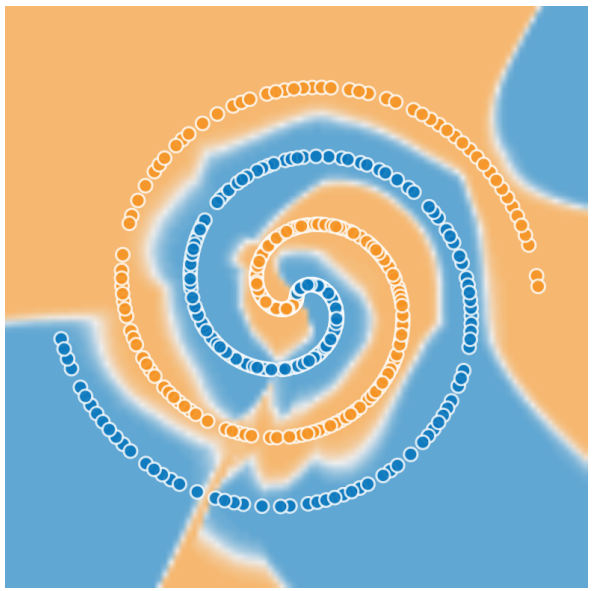
Figure 1: Perceptron [Amini, 2021]

Artificial Intelligence (AI)

- computers mimicking human abilities such as perception, decision-making etc.
- recent advancements: **deep learning**
- **input**, **output**, **model**, and a **loss function**







Learning from the Data

- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**

Learning from the Data

- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**
- does the model work well for **unseen** data?

Learning from the Data

- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**
- does the model work well for **unseen** data?
- are the model's predictions **interpretable?**

Learning from the Data

- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**
- does the model work well for **unseen** data?
- are the model's predictions **interpretable?**
- is the model **fair?**

Learning from the Data

- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**
- does the model work well for **unseen** data?
- are the model's predictions **interpretable?**
- is the model **fair?**
- could there be **ethical** concerns during the collection and usage of the patient data?

Learning from the Data

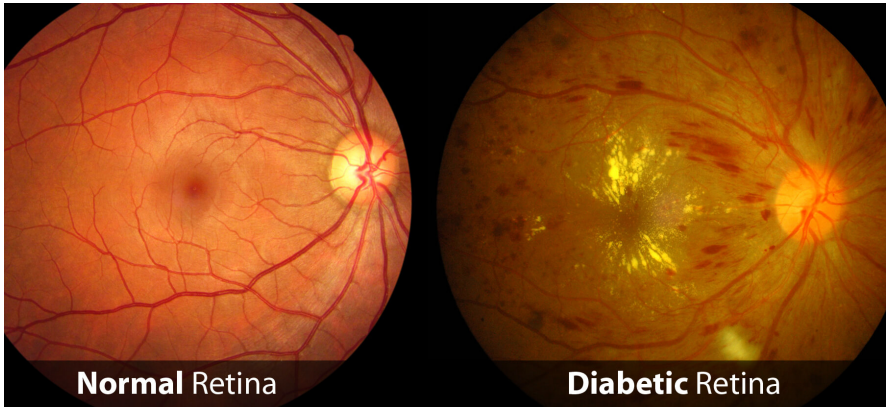
- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**
- does the model work well for **unseen** data?
- are the model's predictions **interpretable?**
- is the model **fair?**
- could there be **ethical** concerns during the collection and usage of the patient data?
- could the model lead clinicians to make **false inferences?**

Learning from the Data

- statistical/traditional/modern machine learning (ML) algorithms: **learning from the data**
- does the model work well for **unseen** data?
- are the model's predictions **interpretable?**
- is the model **fair?**
- could there be **ethical** concerns during the collection and usage of the patient data?
- could the model lead clinicians to make **false inferences?**
- deploying AI in a sensitive area such as healthcare necessitates a careful revisiting of every minor detail, leading to foundational innovations in the field itself

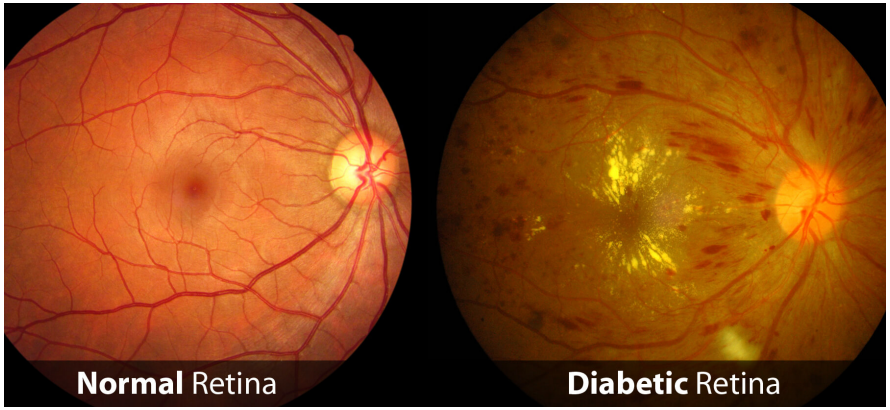
A Success Story: Diabetic Retinopathy and Macular Edema [Gulshan et al., 2016]

- regular screening by an ophthalmologist



A Success Story: Diabetic Retinopathy and Macular Edema [Gulshan et al., 2016]

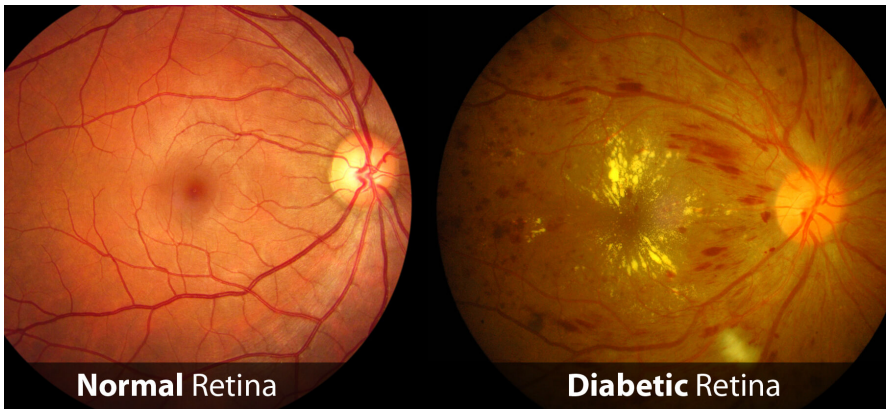
- regular screening by an ophthalmologist



- **sensitivity**: how well can you tell when the complication is **present?**
- **specificity**: how well can you tell when the complication is **NOT** present?

A Success Story: Diabetic Retinopathy and Macular Edema [Gulshan et al., 2016]

- regular screening by an ophthalmologist



- **sensitivity**: how well can you tell when the complication is **present?**
- **specificity**: how well can you tell when the complication is **NOT** present?
- **sensitivity focused**: 97.5% sensitivity, 93.4% specificity
- **specificity focused**: 90.3% sensitivity, 98.5% specificity

Causality

- **causal** relations between a model's **inputs** and **outputs**

Causality

- **causal** relations between a model's **inputs** and **outputs**
- **counterfactual** questions: "how would the patient respond to this treatment?"

Causality

- **causal** relations between a model's **inputs** and **outputs**
- **counterfactual** questions: "how would the patient respond to this treatment?"
- **IMPORTANT**: is the training data **biased?**

Causality

- **causal** relations between a model's **inputs** and **outputs**
- **counterfactual** questions: "how would the patient respond to this treatment?"
- **IMPORTANT**: is the training data **biased?**

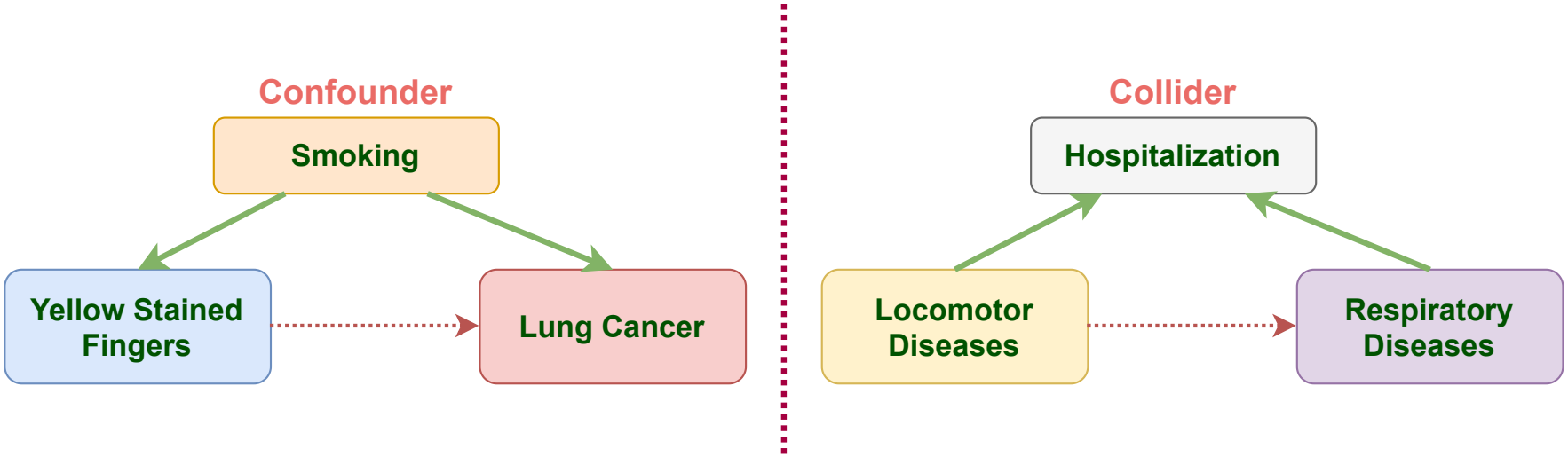


Figure 3: Confounder Bias and Collider Bias [Prosperi et al., 2020]

Causality - Pneumonia

- mortality rate prediction among **pneumonia** patients
- **black-box** deep learning methods

Causality - Pneumonia

- mortality rate prediction among **pneumonia** patients
- **black-box** deep learning methods
- **counter-intuitive** findings: "IF a patient diagnosed with pneumonia has a history of asthma, THEN she is more likely to survive"
- the reason: **bias** in the data

Causality - Pneumonia

- mortality rate prediction among **pneumonia** patients
- **black-box** deep learning methods
- **counter-intuitive** findings: "IF a patient diagnosed with pneumonia has a history of asthma, THEN she is more likely to survive"
- the reason: **bias** in the data
- explanation: patients with asthma are treated more **aggressively** when diagnosed with pneumonia (e.g., in ICU) [Prosperi et al., 2020]

Causality - Pneumonia

- mortality rate prediction among **pneumonia** patients
- **black-box** deep learning methods
- **counter-intuitive** findings: "IF a patient diagnosed with pneumonia has a history of asthma, THEN she is more likely to survive"
- the reason: **bias** in the data
- explanation: patients with asthma are treated more **aggressively** when diagnosed with pneumonia (e.g., in ICU) [Prosperi et al., 2020]
- it may not always be this easy to detect the models' errors, and the kind of bias causing them in the data

Causality - Pneumonia

- mortality rate prediction among **pneumonia** patients
- **black-box** deep learning methods
- **counter-intuitive** findings: "IF a patient diagnosed with pneumonia has a history of asthma, THEN she is more likely to survive"
- the reason: **bias** in the data
- explanation: patients with asthma are treated more **aggressively** when diagnosed with pneumonia (e.g., in ICU) [Prosperi et al., 2020]
- it may not always be this easy to detect the models' errors, and the kind of bias causing them in the data
- especially if we use AI for discovery in the fields where we do not have sufficient domain knowledge

Data, Bias, Holistic Approach [Cabitza et al., 2017]

- data is not **perfect**, and it is biased

Data, Bias, Holistic Approach [Cabitza et al., 2017]

- data is not **perfect**, and it is biased
- some part of data is challenging to digitalize (e.g., psychological state)

Data, Bias, Holistic Approach [Cabitza et al., 2017]

- data is not **perfect**, and it is biased
- some part of data is challenging to digitalize (e.g., psychological state)
- it is hard for AI models to demonstrate a holistic approach similar to clinicians

Data, Bias, Holistic Approach [Cabitza et al., 2017]

- data is not **perfect**, and it is biased
- some part of data is challenging to digitalize (e.g., psychological state)
- it is hard for AI models to demonstrate a holistic approach similar to clinicians
- poor models can degrade clinicians' performance (e.g., poor MRI segmentation)

Data, Bias, Holistic Approach [Cabitza et al., 2017]

- data is not **perfect**, and it is biased
- some part of data is challenging to digitalize (e.g., psychological state)
- it is hard for AI models to demonstrate a holistic approach similar to clinicians
- poor models can degrade clinicians' performance (e.g., poor MRI segmentation)
- some possible long-term disadvantages of **overreliance** to AI in healthcare:
  - regression in clinicians' practical diagnosis abilities

Data, Bias, Holistic Approach [Cabitza et al., 2017]

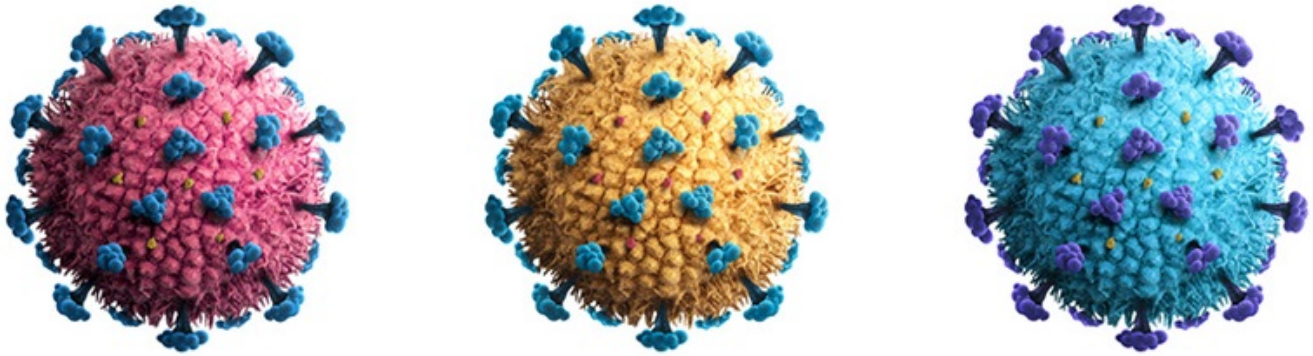
- data is not **perfect**, and it is biased
- some part of data is challenging to digitalize (e.g., psychological state)
- it is hard for AI models to demonstrate a holistic approach similar to clinicians
- poor models can degrade clinicians' performance (e.g., poor MRI segmentation)
- some possible long-term disadvantages of **overreliance** to AI in healthcare:
  - regression in clinicians' practical diagnosis abilities
  - failure to identification of possible model failures

Retrospective Approach

- **static** models trained on **past data**

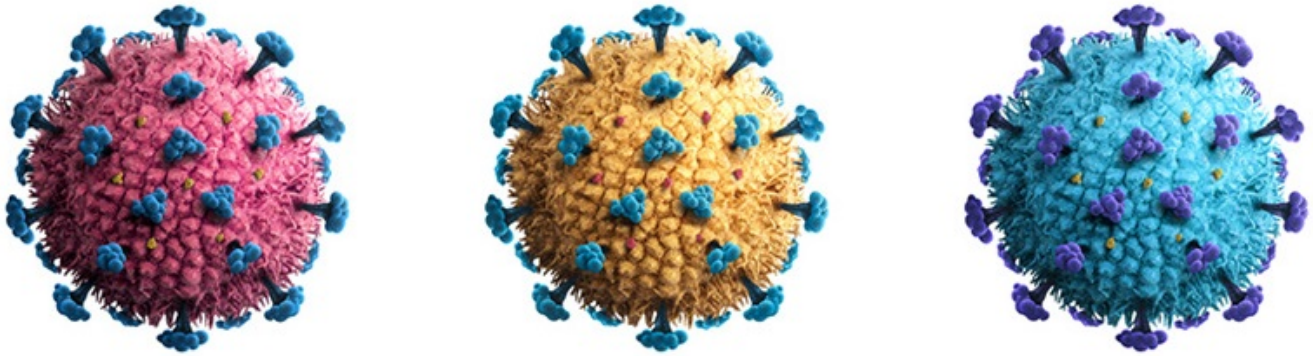
Retrospective Approach

- **static** models trained on **past data**
- need to **adapt** to the changing patient/treatment characteristics



Retrospective Approach

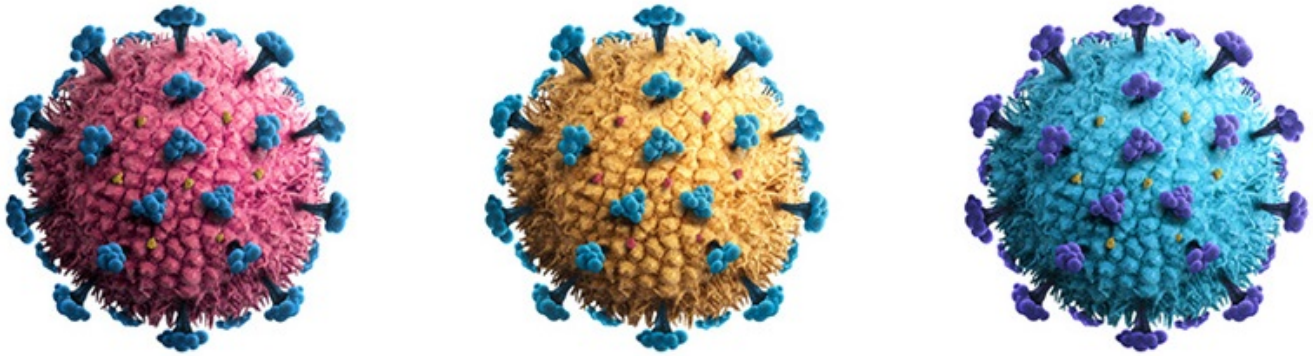
- **static** models trained on **past data**
- need to **adapt** to the changing patient/treatment characteristics



- a clinical decision-making assistance algorithm can alter the clinicians' practice...

Retrospective Approach

- **static** models trained on **past data**
- need to **adapt** to the changing patient/treatment characteristics



- a clinical decision-making assistance algorithm can alter the clinicians' practice...
- ...leading to **previously unattended** patient states for which the **model may not be representative anymore**

Data Sparsity

- deep learning methods are generally trained on **big datasets**

Data Sparsity

- deep learning methods are generally trained on **big datasets**
- challenging to construct big medical datasets
- finding diagnostic MRI images is harder than finding cat/dog images

Data Sparsity

- deep learning methods are generally trained on **big datasets**
- challenging to construct big medical datasets
- finding diagnostic MRI images is harder than finding cat/dog images
- medical datasets are heterogenous with lots of missing entries [Ghassemi et al., 2020]

Data Sparsity

- deep learning methods are generally trained on **big datasets**
- challenging to construct big medical datasets
- finding diagnostic MRI images is harder than finding cat/dog images
- medical datasets are heterogenous with lots of missing entries [Ghassemi et al., 2020]
- sometimes we can resort to: **transfer learning**
  - leveraging another model's expertise on some different task
  - generally used in vision domain

Transfer Learning: Diabetic Foot Ulcer

- a common diabetes complication
- one of the primary causes of hospitalization among diabetes complications
- **amputation risk**

Transfer Learning: Diabetic Foot Ulcer

- a common diabetes complication
- one of the primary causes of hospitalization among diabetes complications
- **amputation risk**
- can **early diagnosis** help avoid amputation?
- evaluation of the patient's medical history, examination by a diabetic foot specialist, supplementary tests such as MRI, X-Ray, CT...

Transfer Learning: Diabetic Foot Ulcer

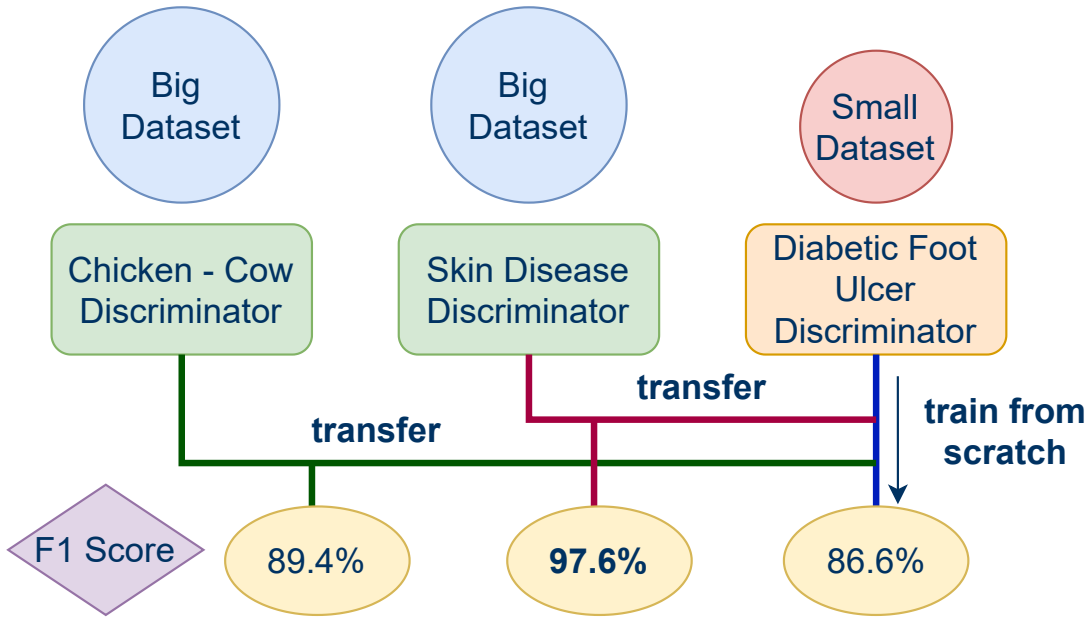
- a common diabetes complication
- one of the primary causes of hospitalization among diabetes complications
- **amputation risk**
- can **early diagnosis** help avoid amputation?
- evaluation of the patient's medical history, examination by a diabetic foot specialist, supplementary tests such as MRI, X-Ray, CT...
- can AI help in early diagnosis?

Transfer Learning: Diabetic Foot Ulcer [Alzubaidi et al., 2020]

- not easy to obtain diagnostic images for diabetic foot ulcer
- images need to be **labeled** by experts
- not too much publicly available data online

Transfer Learning: Diabetic Foot Ulcer [Alzubaidi et al., 2020]

- not easy to obtain diagnostic images for diabetic foot ulcer
- images need to be **labeled** by experts
- not too much publicly available data online



Missing Data
[Ghassemi et al., 2020, Sterne et al., 2009]

- **missing completely at random** (MCAR)
  - no implications about the value of the missing data
  - power outage, forget to record

Missing Data
[Ghassemi et al., 2020, Sterne et al., 2009]

- **missing completely at random** (MCAR)
  - no implications about the value of the missing data
  - power outage, forget to record
- **missing at random** (MAR)
  - there may be a systematic relation between the missing and the available data
  - younger people are more likely to have missing blood pressure measurements

Missing Data
[Ghassemi et al., 2020, Sterne et al., 2009]

- **missing completely at random** (MCAR)
  - no implications about the value of the missing data
  - power outage, forget to record
- **missing at random** (MAR)
  - there may be a systematic relation between the missing and the available data
  - younger people are more likely to have missing blood pressure measurements
- **missing not at random** (MNAR)
  - the missingness of the measurement may tell something about its value
  - a patient with high blood pressure may miss an appointment due to headache
  - a patient experiencing depression may skip an appointment when she feels bad

Missing Data
[Ghassemi et al., 2020, Sterne et al., 2009]

- **missing completely at random** (MCAR)
  - no implications about the value of the missing data
  - power outage, forget to record
- **missing at random** (MAR)
  - there may be a systematic relation between the missing and the available data
  - younger people are more likely to have missing blood pressure measurements
- **missing not at random** (MNAR)
  - the missingness of the measurement may tell something about its value
  - a patient with high blood pressure may miss an appointment due to headache
  - a patient experiencing depression may skip an appointment when she feels bad
- **wrong modeling of missing data** may lead to false predictions
  - troponin-T is measured when myocardial infarction is thought to be probable
  - if an imputer assumes troponin-T are MCAR rather than MAR, then the resulting model would overpredict the risk of myocardial infarction

Interpretability [Vellido, 2020]

- "**black-box**" phenomenon

Interpretability [Vellido, 2020]

- "**black-box**" phenomenon
- **interpretability:** *some* kind of explanation as to why or how a model is making its predictions

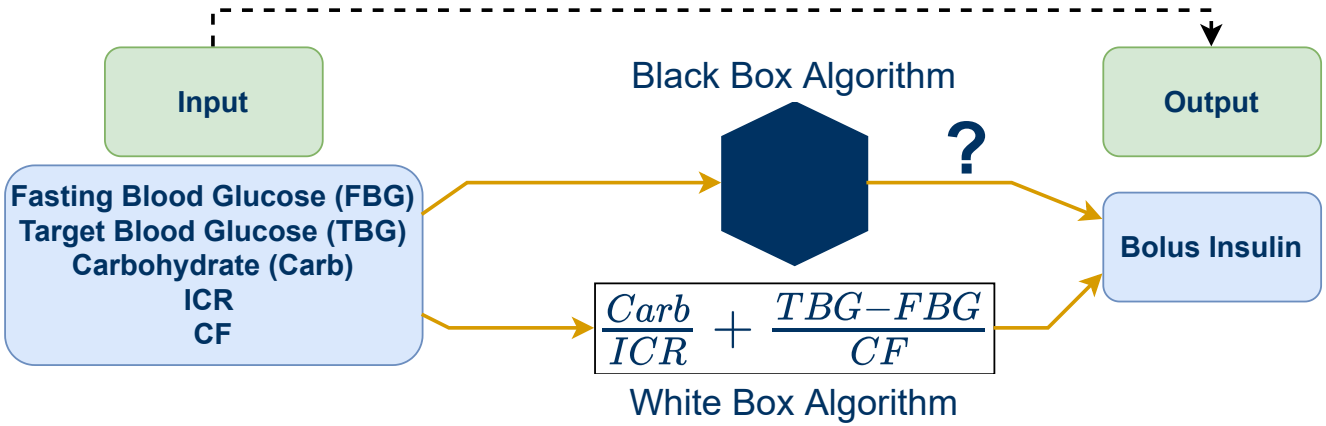


Figure 4: White-Box Algorithm: [Schmidt and Nørgaard, 2014]

Interpretability [Vellido, 2020]

- "**black-box**" phenomenon
- **interpretability:** *some* kind of explanation as to why or how a model is making its predictions

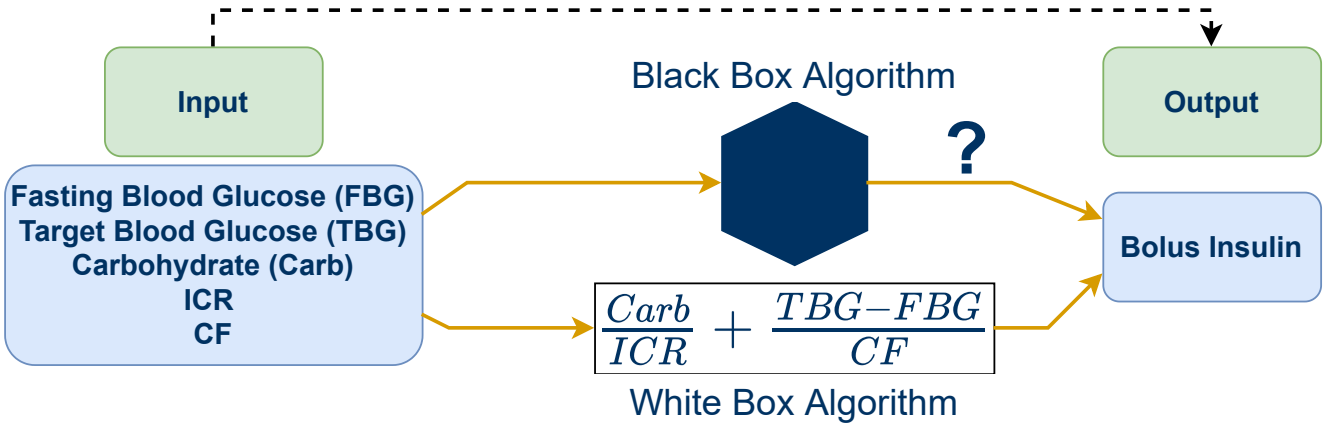


Figure 4: White-Box Algorithm: [Schmidt and Nørgaard, 2014]

- "interpretability" can be ill-posed in the presence of **hidden confounders**

Interpretability [Vellido, 2020]

- "**black-box**" phenomenon
- **interpretability:** *some* kind of explanation as to why or how a model is making its predictions

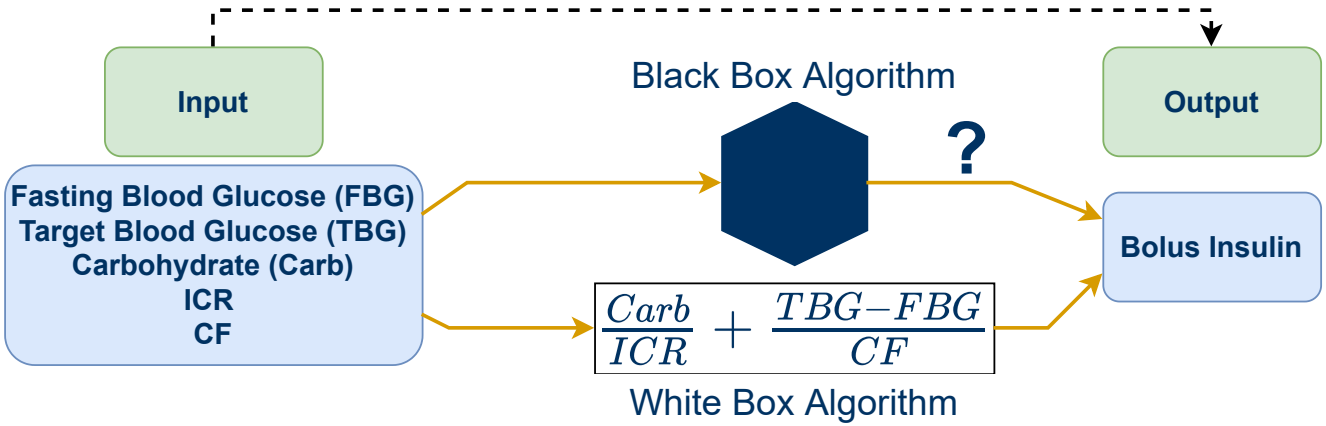


Figure 4: White-Box Algorithm: [Schmidt and Nørgaard, 2014]

- "interpretability" can be ill-posed in the presence of **hidden confounders**
- a (currently) safer reference: **thorough external validation**

Ethical Aspect

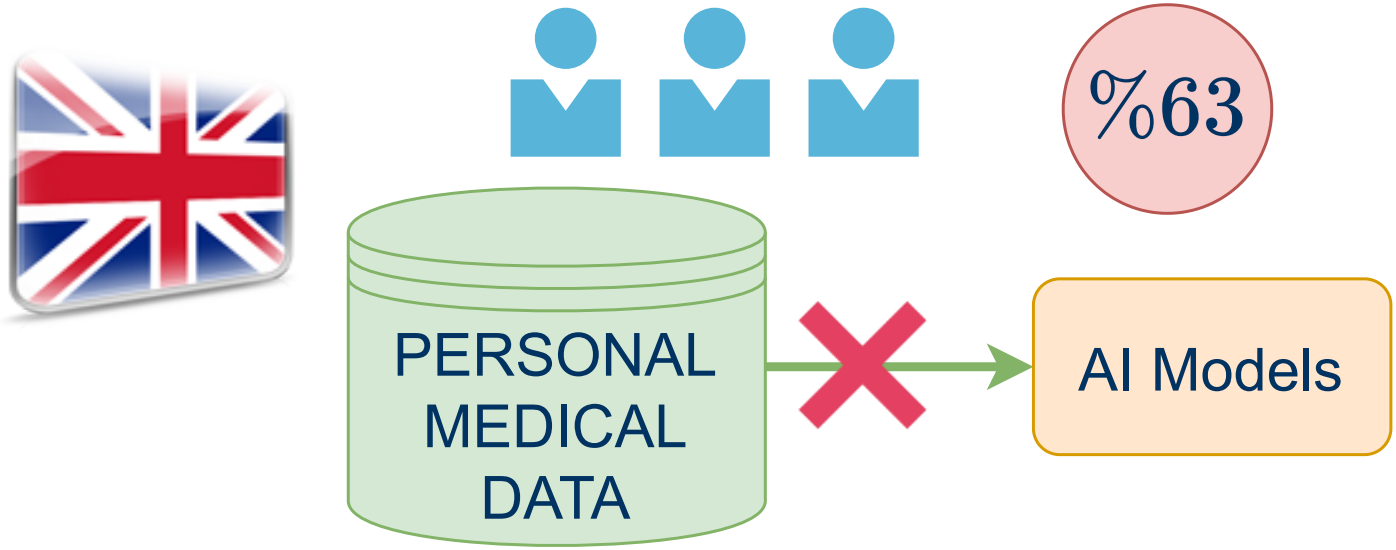


Figure 5: [Vayena et al., 2018]

Ethical Aspect

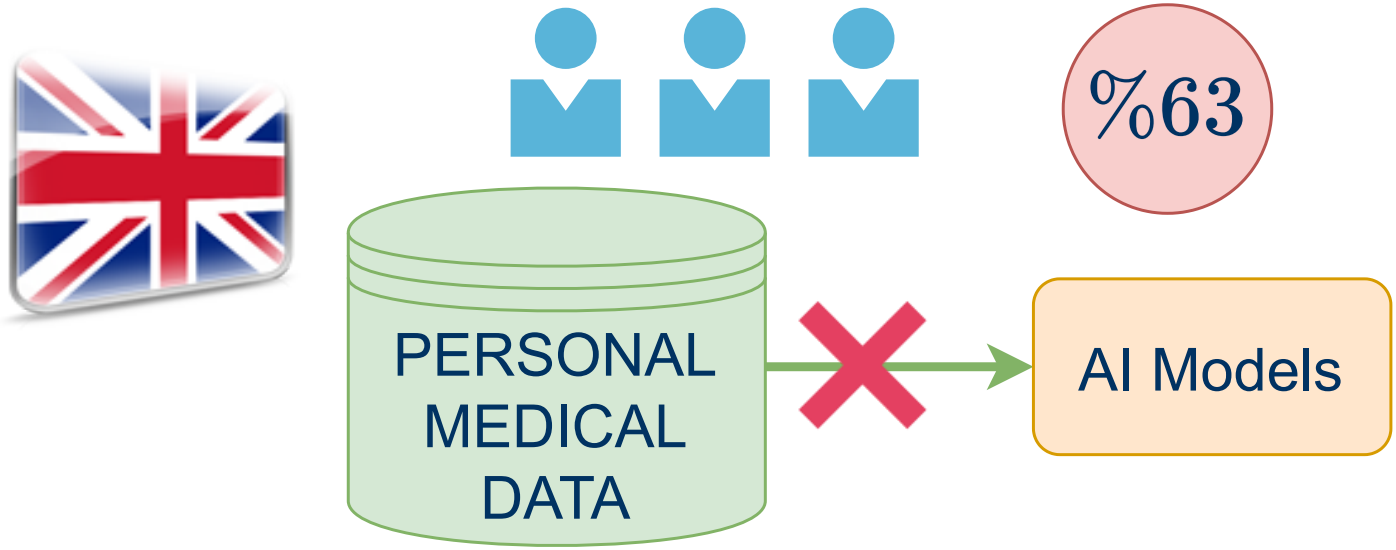
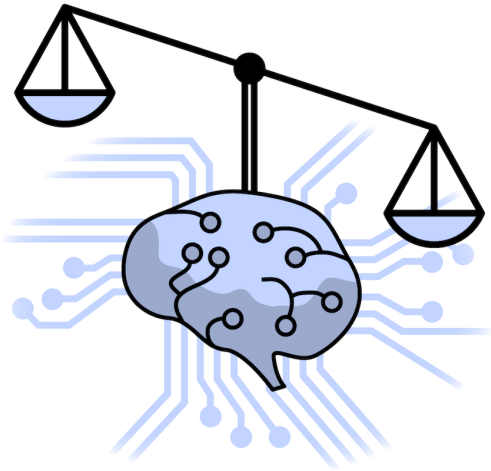


Figure 5: [Vayena et al., 2018]

- if a patient did not share her smoking status or HIV status...
- should we use imputation methods to impute them? [Wiens et al., 2019]

Fairness



- Amazon's job application review algorithm prefers men over women [Dastin, 2018]

Fairness

- Amazon's job application review algorithm prefers men over women [Dastin, 2018]
- COMPAS, an assistive AI model in law, is biased against black people [Angwin, 2016]

Fairness

- Amazon's job application review algorithm prefers men over women [Dastin, 2018]
- COMPAS, an assistive AI model in law, is biased against black people [Angwin, 2016]
- what could "fairness" mean in the context of AI and healthcare?

Fairness

- a model trained on a dataset dominated by people with light-skin may perform poorly for people with dark-skin

Fairness

- a model trained on a dataset dominated by people with light-skin may perform poorly for people with dark-skin
- another interesting bias: **healthy volunteer bias**
- 23andMe genotype dataset, 2399 people: **2,098 (87%) European**, **58 (2%) Asian**, and **50 (2%) African** [Mehrabi et al., 2021]

Fairness

- a model trained on a dataset dominated by people with light-skin may perform poorly for people with dark-skin
- another interesting bias: **healthy volunteer bias**
- 23andMe genotype dataset, 2399 people: **2,098 (87%) European**, **58 (2%) Asian**, and **50 (2%) African** [Mehrabi et al., 2021]
- an exome analysis study to predict hypertrophic cardiomyopathy [Manrai et al., 2016]
  - benign variant classified as pathogenic for people with African ancestry...
  - ...which would not occur should more people with African ancestry were included during training

References I


Alzubaidi, L., Fadhel, M. A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y., and R Oleiwi, S. (2020).
Towards a better understanding of transfer learning for medical imaging: a case study.
Applied Sciences, 10(13):4523.


Amini, A. (2021).
Mit 6.s191 introduction to deep learning online course.


Angwin, J. (2016).
Machine bias—there's software used across the country to predict future criminals. and it's biased against blacks.
ProPublica.

References II


Cabitza, F., Rasoini, R., and Gensini, G. F. (2017).
Unintended consequences of machine learning in medicine.
Jama, 318(6):517–518.


Dastin, J. (2018).
Amazon scraps secret ai recruiting tool that showed bias against women.
Reuters.


Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020).
A review of challenges and opportunities in machine learning for health.
AMIA Summits on Translational Science Proceedings, 2020:191.

References III

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016).
Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.
Jama, 316(22):2402–2410.

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., and Kohane, I. S. (2016).
Genetic misdiagnoses and the potential for health disparities.
New England Journal of Medicine, 375(7):655–665.

References IV


Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021).
A survey on bias and fairness in machine learning.
ACM Computing Surveys (CSUR), 54(6):1–35.


Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. (2020).
Causal inference and counterfactual prediction in machine learning for actionable healthcare.
Nature Machine Intelligence, 2(7):369–375.


Schmidt, S. and Nørgaard, K. (2014).
Bolus calculators.
Journal of diabetes science and technology, 8(5):1035–1041.

References V


Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009).
Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.
Bmj, 338.


Vayena, E., Blasimme, A., and Cohen, I. G. (2018).
Machine learning in medicine: addressing ethical challenges.
PLoS medicine, 15(11):e1002689.


Vellido, A. (2020).
The importance of interpretability and visualization in machine learning for applications in medicine and health care.
Neural computing and applications, 32(24):18069–18083.

References VI


Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al. (2019).
Do no harm: a roadmap for responsible machine learning for health care.
Nature medicine, 25(9):1337–1340.