UNCERTAINTY AND THE MEDICAL INTERVIEW

TOWARDS SELF-ASSESSMENT IN MACHINE LEARNING MODELS

Jakob D. Havtorn

DTU Compute
Department of Applied Mathematics and Computer Science

Outline of Part

- Introduction

Healthcare

Healthcare is the improvement of health via the prevention, diagnosis, treatment, amelioration or cure of disease, illness, injury, and other physical and mental impairments in people.

Medical dialogue

General practitioner

Emergency room

Health insurance

Hospital

Specialist doctor

Psychiatrist

Physiotherapist

Emergency services

Errors in medical dialogue

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.

Errors in medical dialogue

- Communication is everywhere in healthcare.
- It is complex, involving multiple participants, different contexts, and different purposes.
- Failure of communication is a leading cause of medical error contributing to two out of three adverse events [6].
- A considerable fraction of all hospital admissions had preventable adverse outcomes (9% to 16.6% in AU, NZ, UK, DK) [19].

Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.

---

[1]Ambulatory care across four specialties in four states and tertiary care at an academic medical center.
[2]Outpatient visits, Yale-New Haven Hospital.

Documenting medical encounters

- Documentation is a central part of healthcare.
- E.g. patient records, insurance claims, billing, research, training, legal purposes.

- Time: Physicians spend 34-37% of their time on documentation [9, 2, 8][1].
- Quality: Discharge summaries rarely meet all timeline, transmission, and content criteria. [3][2]

---

[1]Ambulatory care across four specialties in four states and tertiary care at an academic medical center.
[2]Outpatient visits, Yale-New Haven Hospital.

How might machine learning help?

- Assist with documentation.
- Augment communication.
- Improve decision-making.

- Reduce errors.
- Save time.

Reliability of machine learning

- Data: Privacy, quality, quantity, diversity.
- Interpretability: Trust, ethics, regulation.
- Explainability: Transparency, accountability.
- Robustness: Adversarial attacks, distribution shift.
- Bias: Fairness, transparency.
- Complexity: Context, domain, language, culture.

Part I

Unsupervised Out-of-Distribution Detection

Outline of Part

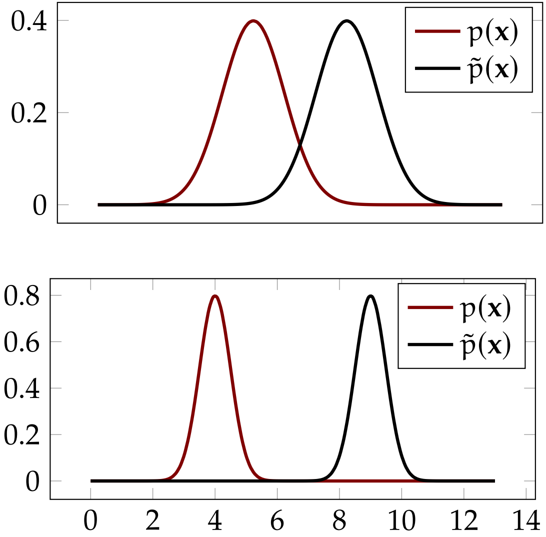
- Out-of-distribution detection
- Latent variable models
- Identifying the issue
- The $\mathcal{L}^{>k}$ likelihood bound
- Likelihood ratio

Defining OOD detection

Out-of-distribution (OOD) detection is about enabling models to distinguish the training data distribution $p(\mathbf{x})$ from any other distribution $\tilde{p}(\mathbf{x})$.

We are concerned with doing this on a per-observation basis, i.e. answering the question:

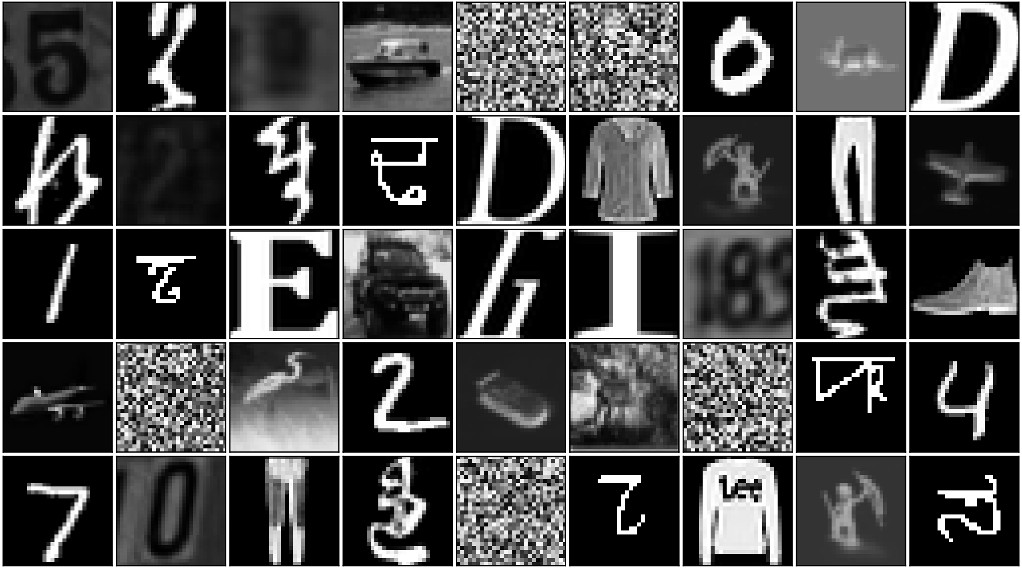
> "Was $\mathbf{x}$ sampled from $p(\mathbf{x})$ or not?"



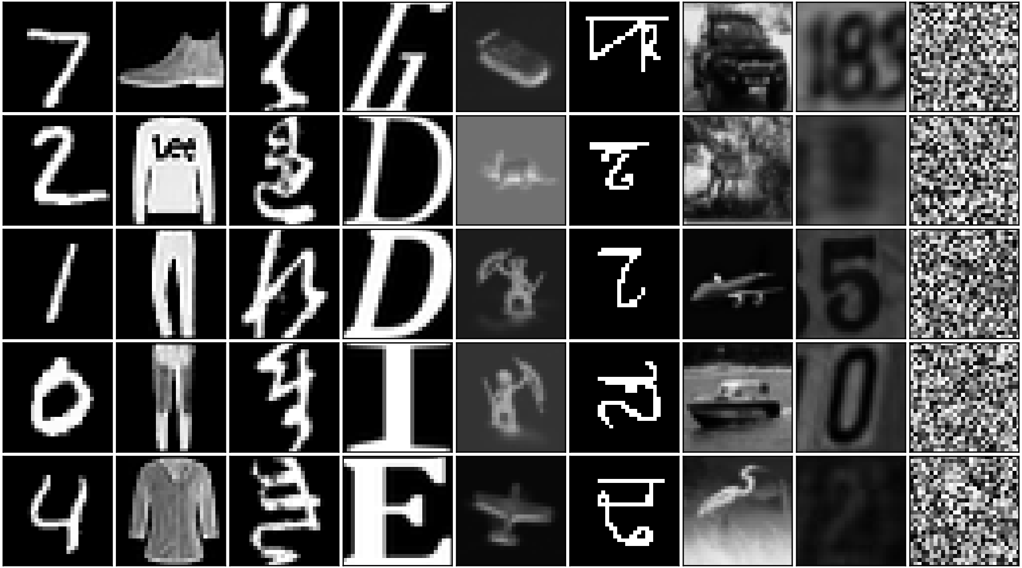
Problem and Contributions

- Deep generative models often fail at OOD detection task when using their likelihood estimate as the score function [13] by, perhaps surprisingly, assigning **higher likelihoods** to the OOD data.
- Contributions:
  - We provide evidence that out-of-distribution detection fails due to learned low-level features that generalize across datasets.
  - We present a fast and fully unsupervised method for OOD detection competitive with the state-of-the-art

In distribution?



Out of distribution?



Hierarchical VAE

We choose the hierarchical VAE as our model [4, 5].

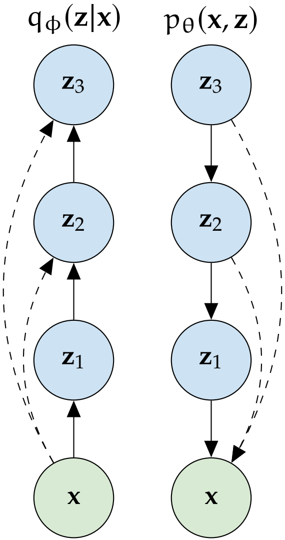
$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z}$$

Specifically we use

1. a three-layered hierarchical VAE with bottom-up inference and deterministic skip-connections for both inference and generation.

$$\begin{aligned} \text{Generative model:} \quad & p_\theta(\mathbf{x}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_1|\mathbf{z}_2) p(\mathbf{z}_3), \\ \text{Inference model:} \quad & q_\phi(\mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}_1|\mathbf{x}) q_\phi(\mathbf{z}_2|\mathbf{z}_1) q_\phi(\mathbf{z}_3|\mathbf{z}_2). \end{aligned}$$

2. a ten-layered layered Bidirectional-Inference Variational Autoencoder (BIVA) [12].



What is wrong with the ELBO for OOD detection?

We can split the ELBO into two terms

$$\mathcal{L}(\mathbf{x};\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction likelihood}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))}_{\text{regularization penalty}} . \tag{1}$$

The first term is high if the data is well-explained by $\mathbf{z}$.

The second term we can rewrite as,

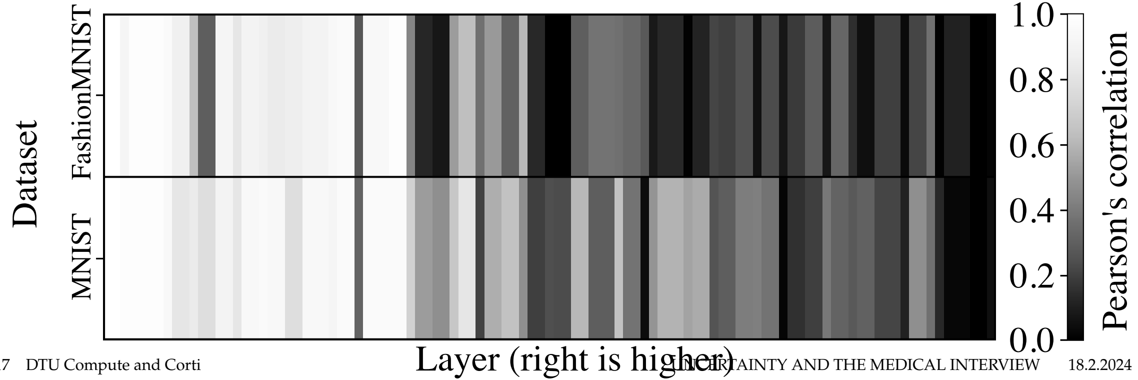
$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{i=1}^{L-1} \log \frac{p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})}{q_\phi(\mathbf{z}_i|\mathbf{z}_{i-1})} + \log \frac{p_\theta(\mathbf{z}_L)}{q_\phi(\mathbf{z}_L|\mathbf{z}_{L-1})}\right] . \tag{2}$$

The absolute log-ratios grow with $\dim(\mathbf{z}_i)$ since the log probability terms are computed by summing over the dimensionality of $\mathbf{z}_i$.

What do the lowest latent variables code for?

Absolute Pearson correlations between data representations in all layers of the inference network of a hierarchical VAE trained on FashionMNIST and of another trained on MNIST.

Correlation computed between the representations of the two different models given the same data, FashionMNIST (top) and MNIST (bottom).



An alternative likelihood bound, $\mathcal{L}^{>k}$

An alternative version of the ELBO that only partially uses the approximate posterior can be written as [12]

$$\mathcal{L}^{>k}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}>k) q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}_{>k})}{q_\phi(\mathbf{z}_{>k}|\mathbf{x})}\right] \tag{3}$$

Here, we have replaced the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ with a different proposal distribution that combines part of the approximate posterior with the conditional prior, namely

$$p_\theta(\mathbf{z}_{\leqslant k}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x})$$

This bound uses the conditional prior for the lowest latent variables in the hierarchy.

Likelihood ratios

We can use our new bound to compute the score used in a standard likelihood ratio test [1].

$$\mathrm{LLR}^{>k}(\mathbf{x}) \equiv \mathcal{L}(\mathbf{x}) - \mathcal{L}^{>k}(\mathbf{x}) \,. \tag{4}$$

We can inspect what this likelihood-ratio measures by considering the exact form of our bounds.

$$\begin{aligned} \mathcal{L} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) , \\ \mathcal{L}^{>k} &= \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned} \tag{5}$$

In the likelihood ratio the reconstruction terms cancel out and only the KL-divergences from the approximate to the true posterior remain.

$$\begin{aligned} \mathrm{LLR}^{>k}(\mathbf{x}) = &-D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &+ D_{\mathrm{KL}}\left(p_\theta(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k}) q_\phi(\mathbf{z}_{>k}|\mathbf{x}) || p_\theta(\mathbf{z}|\mathbf{x})\right) \,. \end{aligned} \tag{6}$$

Importance sampling the ELBO

The well-known importance weighted autoencoder (IWAE) bound is tight with the true likelihood in the limit of infinite samples, $S \to \infty$ [7],

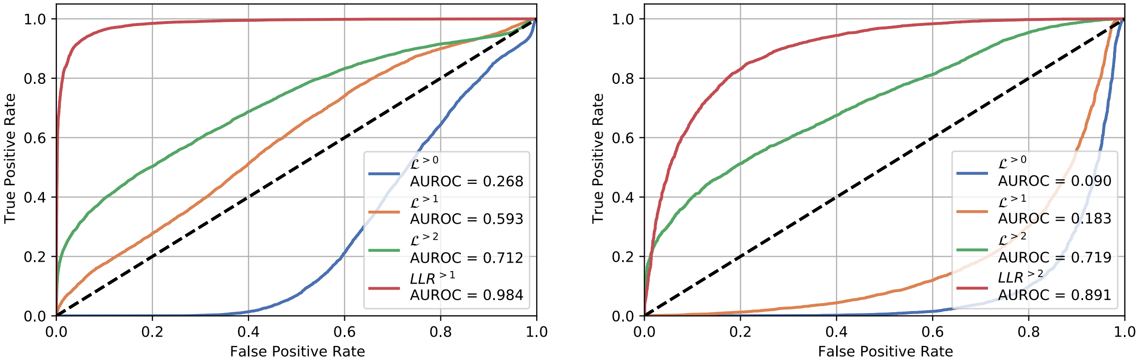
$$\mathcal{L}_S = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N}\sum_{s=1}^{S} \frac{p(\mathbf{x}, \mathbf{z}^{(s)})}{q(\mathbf{z}^{(s)}|\mathbf{x})}\right] \leqslant \log p_\theta(\mathbf{x}) \, , \tag{7}$$

Consequently, by importance sampling the ELBO, the associated KL-divergence associated vanishes and our likelihood ratio reduces to the KL-divergence associated with $\mathcal{L}^{>k}$.

$$\mathrm{LLR}_S^{>k}(\mathbf{x}) \to D_{\mathrm{KL}}(p(\mathbf{z}_{\leqslant}|\mathbf{z}_{>k})q(\mathbf{z}_{>k}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})) \, . \tag{8}$$

We can now see that $\mathrm{LLR}_S^{>k}(\mathbf{x})$ performs OOD detection based on the top-most latent variables.

Results with $\mathrm{LLR}^{>k}$



(a) FashionMNIST HVAE evaluated on MNIST

(b) CIFAR10 BIVA evaluated on SVHN

Results with $\mathrm{LLR}^{>k}$

The score has good performance across many datasets.

OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on CIFAR10				
SVHN	$\mathrm{LLR}^{>2}$	0.811	0.837	0.394
CIFAR10	$\mathrm{LLR}^{>1}$	0.469	0.479	0.835
Trained on SVHN				
CIFAR10	$\mathrm{LLR}^{>1}$	0.939	0.950	0.052
SVHN	$\mathrm{LLR}^{>1}$	0.489	0.484	0.799

OOD dataset	Metric	AUROC↑	AUPRC↑	FPR80↓
Trained on FashionMNIST				
MNIST	$\mathrm{LLR}^{>1}$	0.986	0.987	0.011
notMNIST	$\mathrm{LLR}^{>1}$	0.998	0.998	0.000
KMNIST	$\mathrm{LLR}^{>1}$	0.974	0.977	0.017
Omniglot28x28	$\mathrm{LLR}^{>2}$	1.000	1.000	0.000
Omniglot28x28Inverted	$\mathrm{LLR}^{>1}$	0.954	0.954	0.050
SmallNORB28x28	$\mathrm{LLR}^{>2}$	0.999	0.999	0.002
SmallNORB28x28Inverted	$\mathrm{LLR}^{>2}$	0.941	0.946	0.069
FashionMNIST	$\mathrm{LLR}^{>1}$	0.488	0.496	0.811
Trained on MNIST				
FashionMNIST	$\mathrm{LLR}^{>1}$	0.999	0.999	0.000
notMNIST	$\mathrm{LLR}^{>1}$	1.000	0.999	0.000
KMNIST	$\mathrm{LLR}^{>1}$	0.999	0.999	0.000
Omniglot28x28	$\mathrm{LLR}^{>1}$	1.000	1.000	0.000
Omniglot28x28Inverted	$\mathrm{LLR}^{>1}$	0.944	0.953	0.057
SmallNORB28x28	$\mathrm{LLR}^{>1}$	1.000	1.000	0.000
SmallNORB28x28Inverted	$\mathrm{LLR}^{>1}$	0.985	0.987	0.000
MNIST	$\mathrm{LLR}^{>2}$	0.515	0.507	0.792

Part II

Medical Applications

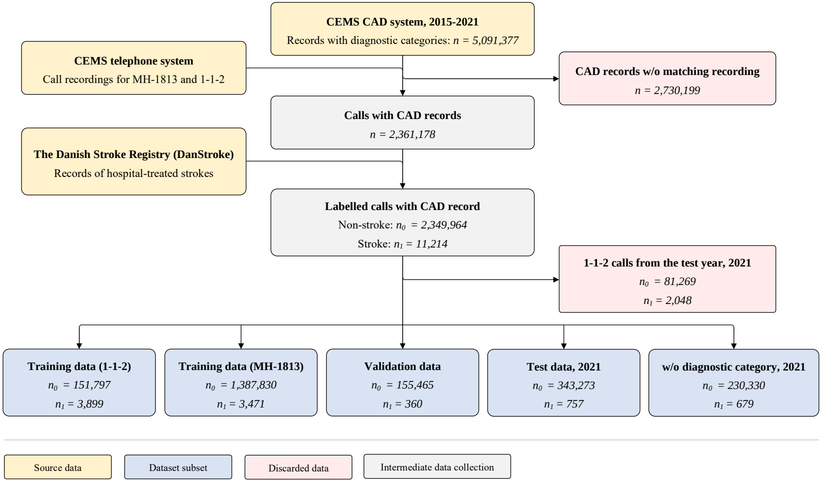
Outline of Part

- A Retrospective Study on Machine Learning-Assisted Stroke Recognition for Medical Helpline Calls

Stroke

- Stroke is a leading cause of disability and death worldwide [16, 11, 10].
- Effective treatment is very time-sensitive. [15, 14].
- The gateway to ambulance transport and hospital admittance is through prehospital telehealth services.
- In the pre-hospital setting, the use of mobile stroke units has made it possible to deliver advanced treatment faster [17, 18].
- As the mobile stroke unit is only dispatched to patients with a suspected stroke, the impact of mobile stroke unit is directly influenced by accurate call-taker recognition of stroke [17, 18].
- Call-taker ability to rapidly and accurately recognize stroke is crucial in facilitating prompt care in both pre-hospital and in-hospital settings.

Population selection and datasets

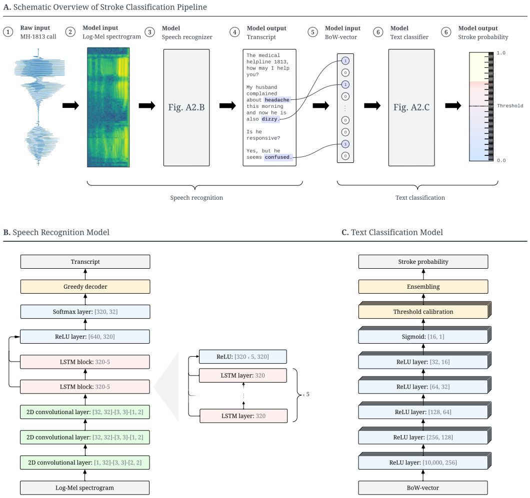


Population characteristics

Table 1. Population characteristics

	Training (1-1-2)	Training (MH-1813)	Validation	Test	2021 w/o category
			All calls		
Num. calls	155,696	1,391,301	155,825	344,030	231,009
Female	74,640 (47.94%)	792,783 (56.98%)	86,959 (55.81%)	190,974 (55.51%)	134,324 (58.14%)
Male	79,564 (51.10%)	596,760 (42.89%)	68,866 (44.19%)	153,050 (44.49%)	96,258 (41.67%)
65+ years	72,930 (46.84%)	335,146 (24.09%)	30,313 (19.45%)	65,652 (19.08%)	81,488 (35.27%)
Age (mean ± std.)	59.47 ± 21.24	47.12 ± 21.38	44.63 ± 20.08	44.31 ± 20.10	50.36 ± 22.77
			Stroke calls		
Num. calls	3,899	3,471	360	757	679
Female	1,784 (45.76%)	1,654 (47.65%)	161 (44.72%)	349 (46.10%)	366 (53.90%)
Male	2,115 (54.24%)	1,815 (52.29%)	199 (55.28%)	408 (53.90%)	313 (46.10%)
65+ years	2,968 (76.12%)	2,421 (69.75%)	250 (69.44%)	555 (73.32%)	567 (83.51%)
Age (mean ± std.)	72.91 ± 12.77	70.68 ± 13.85	70.93 ± 13.83	71.51 ± 13.41	73.41 ± 14.11
			Non-stroke calls		
Num. calls	151,797	1,387,830	155,465	343,273	230,330
Female	72,856 (48.00%)	791,129 (57.00%)	86,798 (55.83%)	190,625 (55.53%)	133,958 (58.16%)
Male	77,449 (51.02%)	594,945 (42.87%)	68,667 (44.17%)	152,642 (44.47%)	95,945 (41.66%)
65+ years	69,962 (46.09%)	332,725 (23.97%)	30,063 (19.34%)	65,097 (18.96%)	80,921 (35.13%)
Age (mean ± std.)	59.12 ± 21.30	47.06 ± 21.36	44.57 ± 20.05	44.25 ± 20.08	50.29 ± 22.76

Model design



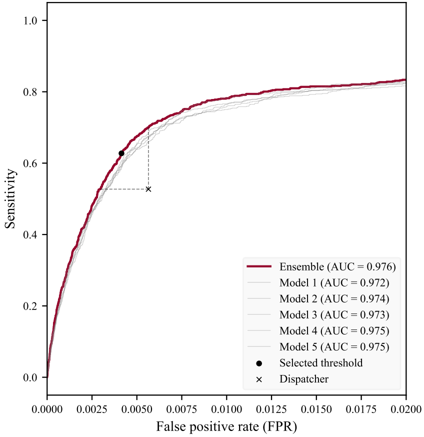
Model performance

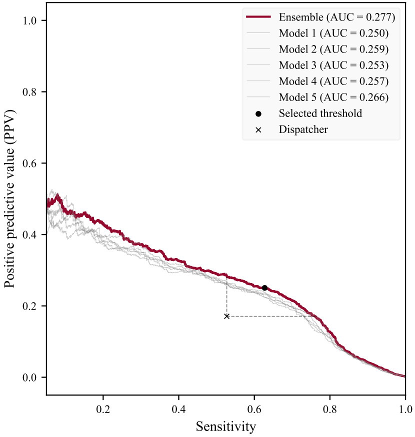
Table 2. Overall performance on MH-1813 test data, performance without 1-1-2 training data, and performance on data from 2021 without diagnostic categories as well as performance on MH-1813 based on demographic subgroups (age/sex) [mean (95% CI)].

	F1-score [%] ↑	Sensitivity [%] ↑	PPV [%] ↑	FOR [%] ↓ (1 - specificity)	FPR [%] ↓ (1 - NPV)
	Overall				
Call takers	25.8 (23.7–27.9)	52.7 (49.2–56.4)	17.1 (15.5–18.6)	0.105 (0.094–0.116)	0.565 (0.539–0.590)
Model	35.7 (35.0–36.4)	63.0 (62.0–64.1)	24.9 (24.3–25.5)	0.082 (0.079–0.085)	0.419 (0.413–0.426)
	Without 1-1-2 training data				
Model	32.4 (31.8–33.1)	60.4 (59.3–61.4)	22.2 (21.6–22.7)	0.088 (0.085–0.091)	0.467 (0.460–0.474)
	On MH-1813 data without diagnostic category				
Model	32.6 (31.9–33.4)	48.3 (47.2–49.4)	24.7 (23.9–25.3)	0.153 (0.148–0.158)	0.435 (0.427–0.443)
	18-64 years				
Call takers	15.9 (13.1–18.5)	50.5 (43.6–57.2)	9.40 (7.61–11.18)	0.036 (0.028–0.043)	0.353 (0.331–0.375)
Model	22.9 (21.8–24.0)	54.1 (52.1–56.3)	14.5 (13.8–15.3)	0.033 (0.031–0.035)	0.231 (0.226–0.236)
	65+ years				
Call takers	32.9 (30.1–35.7)	53.5 (49.4–57.6)	23.7 (21.4–26.0)	0.401 (0.352–0.449)	1.467 (1.373–1.560)
Model	42.8 (41.9–43.7)	66.3 (65.1–67.5)	31.6 (30.8–32.4)	0.290 (0.278–0.303)	1.224 (1.198–1.249)
	Male				
Call takers	30.2 (27.2–33.3)	53.9 (49.1–58.9)	21.0 (18.5–23.5)	0.124 (0.105–0.141)	0.542 (0.506–0.580)
Model	39.0 (38.0–40.1)	63.7 (62.3–65.2)	28.1 (27.3–29.0)	0.097 (0.093–0.102)	0.435 (0.425–0.445)
	Female				
Call takers	21.9 (19.1–24.6)	51.3 (46.0–56.6)	13.9 (12.0–15.8)	0.090 (0.076–0.103)	0.582 (0.547–0.616)
Model	32.4 (31.4–33.4)	62.3 (60.7–63.8)	21.9 (21.1–22.7)	0.069 (0.066–0.073)	0.407 (0.399–0.416)

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; FOR, false omission rate; CI, confidence interval

Model performance





Model performance

Call taker predictions		Ground truth labels: Positives	Ground truth labels: Negatives
	Positives	True positives 399	False positves 1,938
	Negatives	False negatives 358	True negatives 341,335

Model predictions		Ground truth labels: Positives	Ground truth labels: Negatives
	Positives	True positives 477	False positves 1,440
	Negatives	False negatives 280	True negatives 341,833

Which features are important?

Table 3. English translation of words with the largest positive and negative ranking score in calls predicted as stroke and non-stroke, respectively

	Positive ranking score		Negative ranking score	
	Stroke predictions, D = 1,897		Non-stroke predictions, D = 342,133	
	Word, w (translated)	Occurrences, $D^{(w)}$	Word, w (translated)	Occurrences, $D^{(w)}$
1.	Ambulance	1,680	Tetanus	4,378
2.	Blood clot	895	Pregnant	8,749
3.	Left	1,108	Cut	7,592
4.	Right	1,050	Bandage	4,561
5.	Double vision	84	Amager (a location)	23,776
6.	The words	344	O'clock	94,436
7.	Suddenly	783	The emergency room	42,809
8.	Arm	709	The police	2,903
9.	Side	1,139	Swollen	60,559
10.	Stroke	117	Over the counter (OTC)	4,641
11.	Double	113	The neck	30,151
12.	Control	134	Fever	112,586
13.	Call	39	Prescription	5,450
14.	Numb	94	Centimetre	12,026
15.	Minutes	763	The knee	8,875
16.	Difficulties speaking	44	The pharmacy	10,085
17.	Haemorrhagic stroke	133	The stomach	42,105
18.	Hand	297	Psychiatric	3,688
19.	The ambulance	521	Pneumonia	7,597
20.	Slurred speech	58	Stomach pain	10,551
21.	Blood clots	224	Stool	19,155
22.	Fast	663	The ribs	3,928
23.	Express	44	Bleed	10,501
24.	Blood thinner	259	Bleeding	24,313
25.	Incoherent	15	Ribs	2,941
26.	Lopsided	211	Broken	19,415
27.	Reduced	528	Inflammation	10,050
28.	Hangs	628	Common cold	8,127
29.	Transient	48	Morning *or* morrow	78,558
30.	Not making sense	14	Swelling	17,762

Simulated prospective study

I. When is the model prediction presented to the call-taker?

1. Notify the call-taker of potential false positive or negative stroke cases after the call ends.
2. Notify the call-taker of potential false positive or negative stroke cases during the call.

II. How does prediction influence the diagnostic code the call-taker assigns to the call?

A. Call-takers change any stroke prediction from negative to positive if the model predicts a positive (call-takers mirror model positives).
B. Call-takers change any stroke prediction from positive to negative if the model predicts a negative (call-takers mirror model negatives).

Simulated prospective study

Predictor	Call-taker	Model		Call-taker supported by the model (simulated)			
When	-	After call	During call	After call	During call	After call	During call
Method	-	1.C	2.C	1.A	1.B	2.A	2.B
F1-score [%] ↑	25.8 (23.7-27.9)	35.7 (35.0-36.4)	33.1 (32.4-33.7)	28.9 (28.3-29.5)	33.3 (32.5-34.1)	27.6 (27.0-28.1)	32.7 (31.8-33.5)
Sensitivity [%] ↑	52.7 (49.2-56.4)	63.0 (62.0-64.1)	58.7 (57.7-59.8)	72.4 (71.5-73.3)	43.4 (42.3-44.5)	72.3 (71.4-73.3)	39.1 (38.1-40.1)
PPV [%] ↑	17.1 (15.5-18.6)	24.9 (24.3-25.5)	23.0 (22.5-23.6)	18.0 (17.6-18.4)	27.0 (26.3-27.8)	17.0 (16.7-17.4)	28.1 (27.3-28.9)
FOR [%] ↓ (1 - NPV)	0.105 (0.094-0.116)	0.082 (0.079-0.085)	0.091 (0.088-0.094)	0.061 (0.059-0.064)	0.125 (0.121-0.129)	0.061 (0.059-0.064)	0.134 (0.131-0.138)
FPR [%] ↓ (1 - specificity)	0.565 (0.539-0.590)	0.419 (0.413-0.426)	0.432 (0.426-0.439)	0.726 (0.717-0.735)	0.258 (0.253-0.263)	0.776 (0.767-0.786)	0.221 (0.216-0.226)

Thank you for your attention

Bibliography I

[1] Adolf Buse. "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note". In: *The American Statistician* 36 (3a 1982), pp. 153–157.

[2] Matthew D. Tipping et al. "Where Did the Day Go?–A Time-Motion Study of Hospitalists". In: *Journal of Hospital Medicine* 5.6 (2010), pp. 323–328. ISSN: 1553-5606. DOI: 10.1002/jhm.790. pmid: 20803669.

[3] Leora I. Horwitz et al. "Comprehensive Quality of Discharge Summaries at an Academic Medical Center". In: *Journal of hospital medicine : an official publication of the Society of Hospital Medicine* 8.8 (Aug. 2013), pp. 436–443. ISSN: 1553-5592. DOI: 10.1002/jhm.2021. pmid: 23526813. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3695055/ (visited on 02/15/2024).

Bibliography II

[4] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. Banff, AB, Canada, 2014. arXiv: `1312.6114`. URL: `http://arxiv.org/abs/1312.6114`.

[5] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic Backpropagation and Approximate Inference in Deep Generative Models”. In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. International Conference on Machine Learning. Vol. 32. Beijing, China: PMLR, Jan. 16, 2014, pp. 1278–1286. URL: `http://proceedings.mlr.press/v32/rezende14.pdf` (visited on 08/12/2018).

[6] Amy J Starmer et al. “Changes in Medical Errors after Implementation of a Handoff Program”. In: *New England Journal of Medicine* 371.19 (2014), pp. 1803–1812.

Bibliography III

[7] Yuri Burda, Roger Grosse, and Ruslan R. Salakhutdinov. "Importance Weighted Autoencoders". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. San Juan, Puerto Rico, 2016, p. 8. URL: https://arxiv.org/abs/1509.00519 (visited on 10/04/2017).

[8] Christine Sinsky et al. "Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties". In: *Annals of Internal Medicine* 165.11 (Dec. 6, 2016), pp. 753–760. ISSN: 1539-3704. DOI: 10.7326/M16-0961. pmid: 27595430.

Bibliography IV

[9] Erik Joukes et al. "Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record". In: *Applied Clinical Informatics* 09.01 (Jan. 2018), pp. 046–053. ISSN: 1869-0327. DOI: 10.1055/s-0037-1615747. URL: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0037-1615747 (visited on 02/15/2024).

[10] Mira Katan and Andreas Luft. "Global Burden of Stroke". In: *Seminars in Neurology*. Vol. 38. 02. Thieme Medical Publishers, 2018, pp. 208–211.

[11] Hmwe Hmwe Kyu et al. "Global, Regional, and National Disability-Adjusted Life-Years (DALYs) for 359 Diseases and Injuries and Healthy Life Expectancy (HALE) for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017". In: *The Lancet* 392.10159 (2018), pp. 1859–1922.

Bibliography V

[12] Lars Maaløe et al. "BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling". In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*. Conference on Neural Information Processing Systems. Vancouver, Canada, Feb. 6, 2019, pp. 6548–6558. URL: `http://arxiv.org/abs/1902.02102` (visited on 03/19/2019).

[13] Eric Nalisnick et al. "Do Deep Generative Models Know What They Don't Know?" In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations. New Orleans, LA, USA, 2019. arXiv: `1810.09136`. URL: `http://arxiv.org/abs/1810.09136` (visited on 10/02/2019).

Bibliography VI

[14] Guillaume Turc et al. “European Stroke Organisation (ESO)-European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke”. In: *Journal of Neurointerventional Surgery* 11.8 (2019), pp. 535–538.

[15] Eivind Berge et al. “European Stroke Organisation (ESO) Guidelines on Intravenous Thrombolysis for Acute Ischaemic Stroke”. In: *European Stroke Journal* 6.1 (2021), pp. I–LXII.

[16] GBD 2019 Stroke Collaborators et al. “Global, Regional, and National Burden of Stroke and Its Risk Factors, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019”. In: *The Lancet Neurology* 20.10 (2021), pp. 795–820. ISSN: 1474-4422. DOI: 10.1016/S1474-4422(21)00252-0.

[17] Praveen Hariharan et al. “Mobile Stroke Units: Current Evidence and Impact”. In: *Current Neurology and Neuroscience Reports* 22.1 (2022), pp. 71–81.

Bibliography VII

[18] Babak B Navi et al. "Mobile Stroke Units: Evidence, Gaps, and next Steps". In: *Stroke* 53.6 (2022), pp. 2103–2113.

[19] Niki Carver, Vikas Gupta, and John E. Hipskind. "Medical Errors". In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2024. pmid: 28613514. URL: http://www.ncbi.nlm.nih.gov/books/NBK430763/ (visited on 02/13/2024).