Deep Learning for Predicting In Hospital Mortality

CHRISTINE LEE, MS, UC IRVINE BIOMEDICAL ENGINEERING

IRA HOFER, MD, UCLA ANESTHESIOLOGY

MAXIME CANNESSON, MD PHD, UCLA ANESTHESIOLOGY

PIERRE BALDI, PHD, UC IRVINE COMPUTER SCIENCE

The Need for Identifying Patients At Risk

More than 230 million major surgical procedures are performed annually[1]

Overall mortality rate is less than 2%[2,3]

10% of surgical population at high risk, but 80% of postoperative deaths[2,3]

Less than 15% of high risk surgical patients are admitted to the ICU[3]

Postoperative Risk Scores

ASA Score[4] 1963

- 1 = healthy and 5 = not expected to live 24 hours
- Overall health
- Preoperative

Surgical Apgar Score[5] 2007

- Risk of postoperative complication and poor outcome
- Developed on ~300 patients
- EBL, lowest MAP, and lowest HR

Risk Quantification Index[6] 2011

- Risk of 30 day morbidity and mortality
- Developed on ~635,000 patients
- CPT code of the performed primary procedure, ASA, AGE
- Score + Probability of outcome

POSPOM[7] 2016

- Preoperative score to predict post operative mortality
- Developed on >2 million patients
- 17 preoperative variables (such as age, presence of heart disease, surgery type)

Limitations

ASA is subjective

ASA, POSPOM, and RQI are limited to preoperative information

RQI depends on Procedural Severity Score (PSS)

Surgical Apgar score has been shown to have limited accuracy[8]

Adding Surgical Apgar to RQI to leverage both preoperative and intraoperative information does not significantly improve prediction of mortality[9]

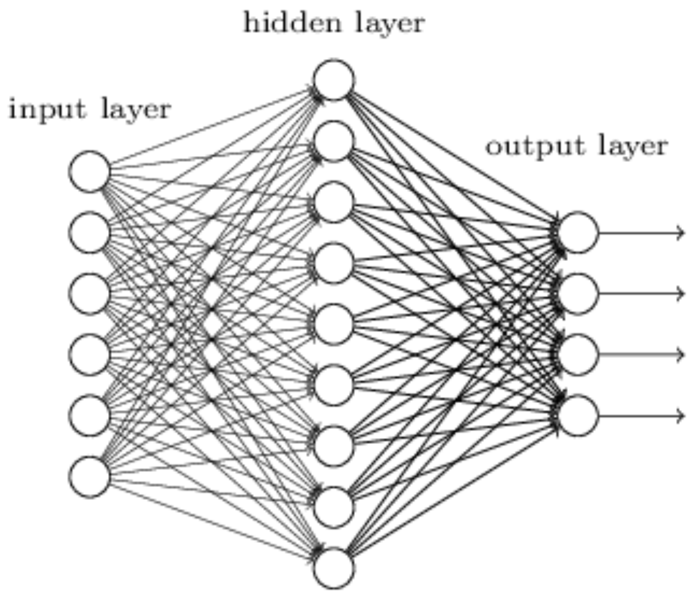
Aim of Study

Predict inhospital mortality in surgical patients by using deep neural network models (DNNs) and intraoperative features

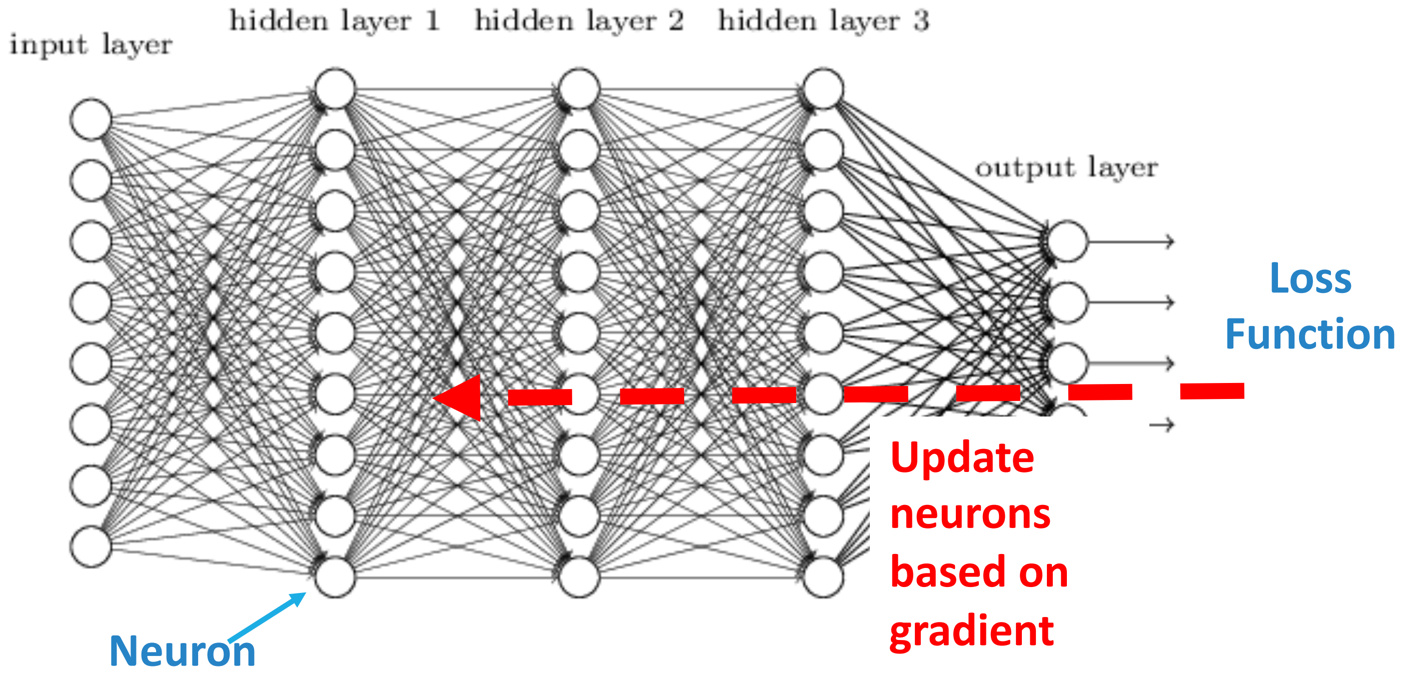
- Compare DNNs to ASA, SAS, RQI, POSPOM, and logistic regression
- Assess DNN with a reduced feature set
- Assess DNN with addition of ASA

Deep Feedforward Neural Networks

"Non-deep" feedforward neural network



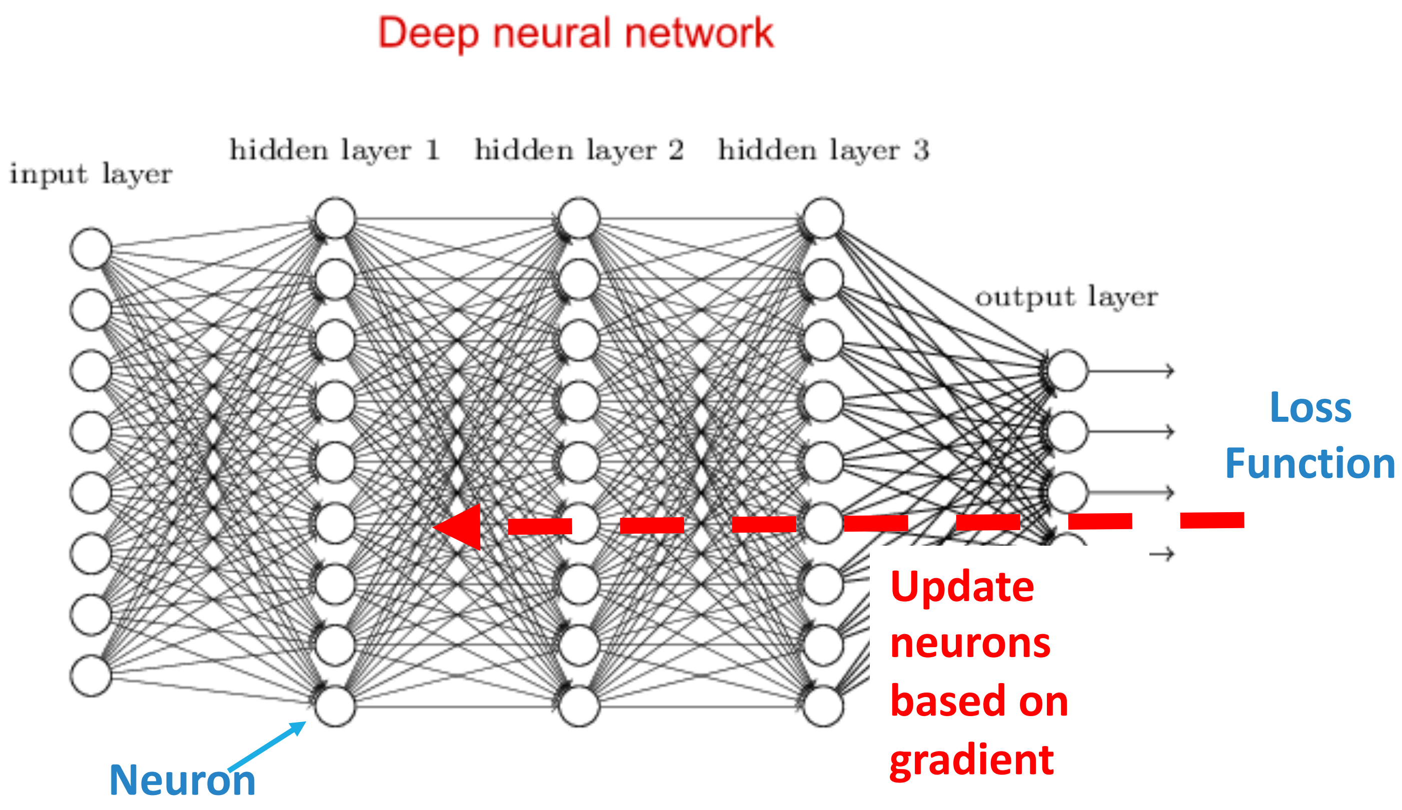
Deep neural network



Deep Feedforward Neural Networks

Hyperparameters

- Number of neurons
- Number of hidden layers
- Batch size
  - Epoch = all training samples have been "seen" and weights updated accordingly
- Activation function
- Learning rate
- Regularization parameters



Data Description

Inclusion Criteria:

- All surgical procedures performed since March 1, 2013 with general anesthesia at UCLA

Exclusion Criteria:

- No general anesthesia
- > 89 years or < 18 years of age
- For patients with >1 procedure, only the first procedure was included

N = 59,985 patients

- 80% Train and 20% Test

87 features calculated/extracted at the end of surgery

- Descriptive intraoperative vital signs
- Summary of drugs and fluids interventions
- Patient anesthesia descriptions

	Train	Test
# of Patients	47,988	11,997
# of Patients with In Hospital Mortality (%)	389 (0.81%)	87 (0.73%)
Age (yrs)	56 ± 17	56 ± 18
EBL (cc)	95 ± 540	94 ± 410
Presence of Arterial Line (%)	17.9	17.8
Presence of PA Line (%)	3.4	0.9
Presence of Central Line (%)	5.1	1.3
ASA (%)		
1	6.3	1.6
2	37.4	9.3
3	49.9	12.5
4	6.1	1.5
5	0.3	0.1
6	0.01	0

Top CPT Codes (# Patients > 100)

1,498 unique CPT codes

167 unique HCUP codes

CPT Code	# Patients	CPT_DESCRIPTION
43239	304	Esophagogastroduodenoscopy
45380	230	Colonoscopy
43259	225	Esophagogastroduodenoscopy
50360	193	Renal allotransplantation, implantation of graft; without recipient nephrectomy
47562	190	Laparoscopy, surgical; cholecystectomy
43242	181	Esophagogastroduodenoscopy
27447	162	Arthroplasty, knee, condyle and plateau
27130	153	Arthroplasty, acetabular and proximal femoral prosthetic replacement (total hip arthroplasty)
59841	149	Induced abortion, by dilation and evacuation
60500	144	Parathyroidectomy or exploration of parathyroid(s)
44970	127	Laparoscopy, surgical, appendectomy
55866	126	Laparoscopy, surgical prostatectomy
61510	106	Craniectomy, trephination, bone flap craniotomy
38724	103	Cervical lymphadenectomy (modified radical neck dissection)

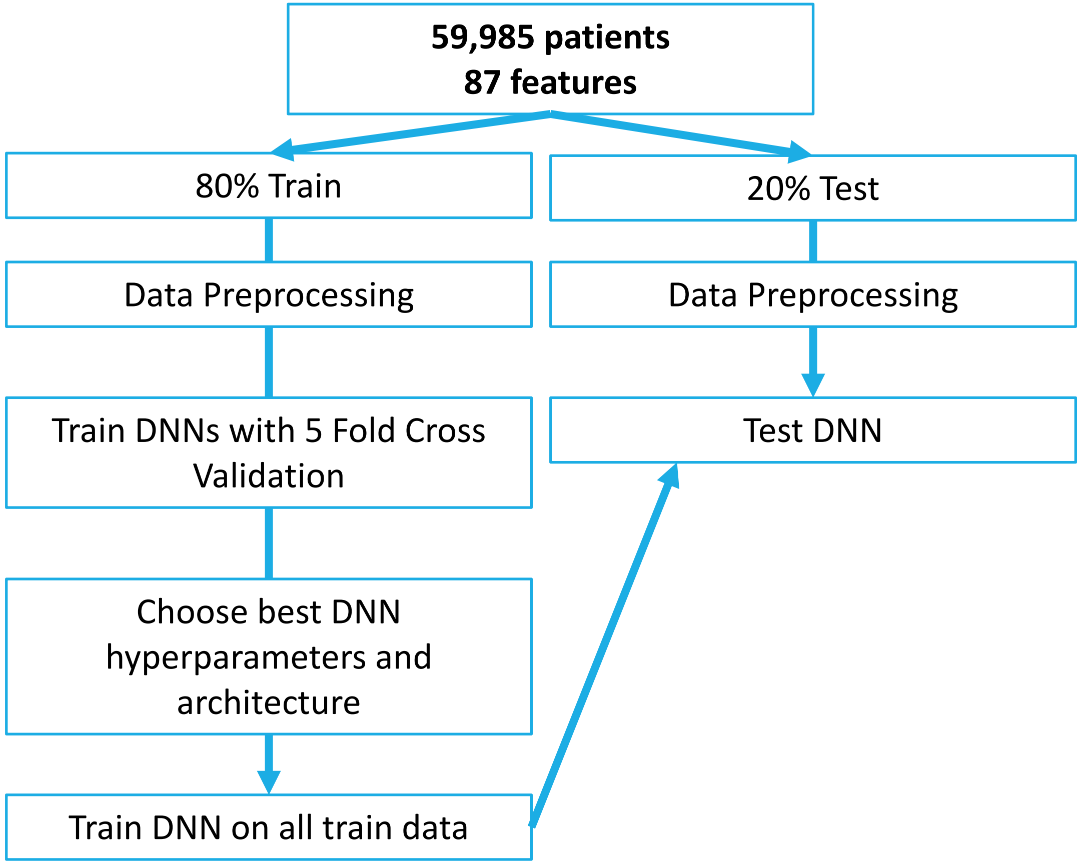
87 Model Features

SBP	min, max, avg, med, std
DBP	min, max, avg, med, std
MAP	min, max, avg, med, std
HR	min, max, avg, med, std
Pulse Ox	min, max, avg, med, std
SBP of the last 10 minutes of the case	min, max, avg, med, std
DBP of the last 10 minutes of the case	min, max, avg, med, std
MAP of the last 10 minutes of the case	min, max, avg, med, std
HR of the last 10 minutes of the case	min, max, avg, med, std
Pulse Ox of the last 10 minutes of the case	min, max, avg, med, std

Current Rate of Phenylephrine
Current Rate of Vasopressin
Current Rate of Epinephrine
Current Rate of Milrinone
Current Rate of Nitroglycerin
Current Rate of Esmolol
Current Rate of Nitroprusside
Current Rate of Nicardipine
Maximum Glucose for the Case
Minimum Glucose for the Case

Nitric Oxide Used for the Case
Presence of invasive central, radial, or pulmonary arterial line
Total Red Blood Cells Transfused
Total Urine Output
Cumulative minutes with MAP<60
Cumulative minutes with MAP<50
Total bolus dose of phenylephrine
Highest infusion rate of phenylephrine during the case
Total bolus dose of ephedrine
Total bolus dose of vasopressin
Highest infusion rate of vasopressin during the case
Total bolus dose of Epinepheine
Highest infusion rate of epinephrine during the case
Highest infusion rate of milrinone during the case
Total bolus does of nitroglycerin during case
Highest infusion rate of nitroglycerin during the case
Total bolus dose of Esmolol during the case
Highest infusion rate of esmolol during the case
Highest infusion rate of nitroprusside during the case
Highest infusion rate of nicardipine during the case
Minimum Hemoglobin during the case
Maximum MAC of isoflurane during the case (note this is not age adjusted)
Maximum MAC of sevoflurane during the case (note this is not age adjusted)
Maximum MAC of desflurane during the case (note this is not age adjusted)

Methods: Summary



Data Preprocessing

- Missing values filled with mean values
- Values clinically out of range filled with clinically normal values
- Train data features rescaled to have a mean of 0 and standard deviation of 1

Methods: Model

Feedforward networks with fully connected layers and a **sigmoid output**

Trained using stochastic gradient descent (SGD) with momentum and a batch size of 200.

Trained 4 DNN models using

1. All 87 features
2. Reduced feature set of 46 features
   - This reduced feature set was created by excluding any average, median, standard deviation, and last 10 minutes of the surgical case features.
3. 87 original features + ASA = 88 features
4. 46 reduced features + ASA = 47 features

Model performance was assessed with AUC

For comparison, the AUCs of logistic regression (87 features), ASA, Surgical Apgar, RQI, and POSPOM were also calculated.

- ASA and POSPOM provided by UCLA
- Surgical Apgar calculated using Gawande et al.[5]
- RQI could not be calculated using published R model from Cleveland Clinic's website [1] due to technical issues with R version
- RQI log probability and score calculated from Sigakis et al.[12]

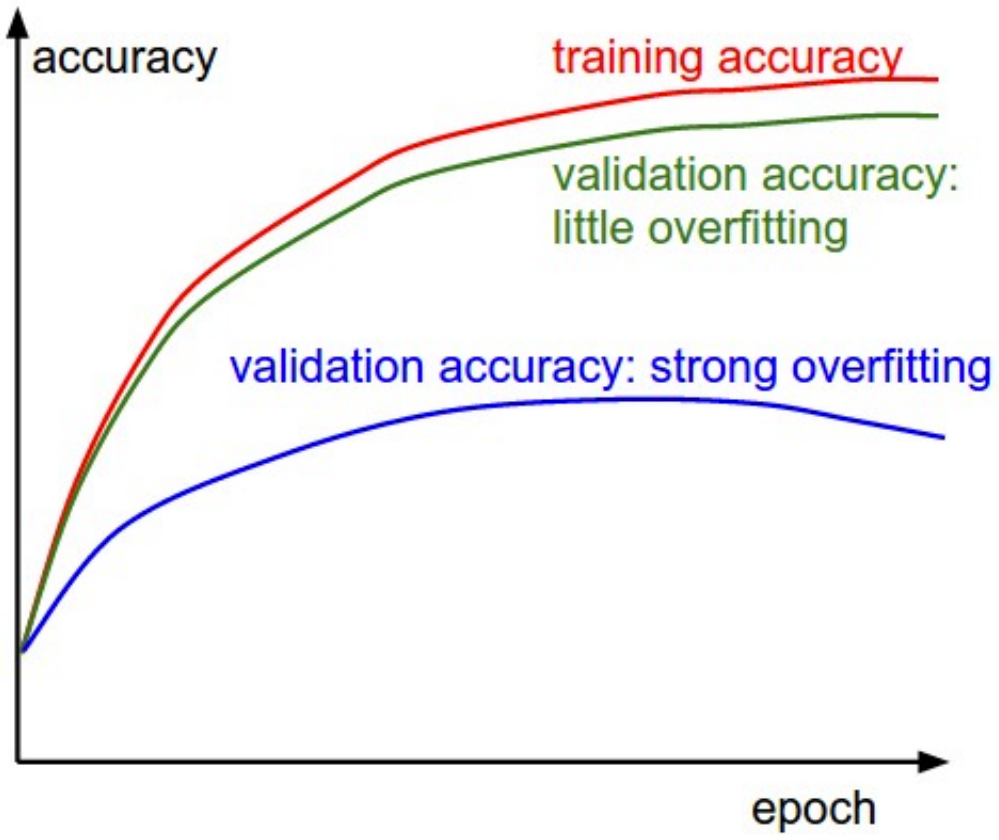
[1] http://my.clevelandclinic.org/departments/anesthesiology/depts/outcomes-research/risk-quantification

Methods: Training Data Augmentation

Prior to training, positive training examples were augmented by adding Gaussian noise with a standard deviation of 0.0001

	% Occurrence	Data Augmentation	Augmented % Occurrence
Inhospital Mortality	0.81	100x	45

Methods: Dealing with Overfitting



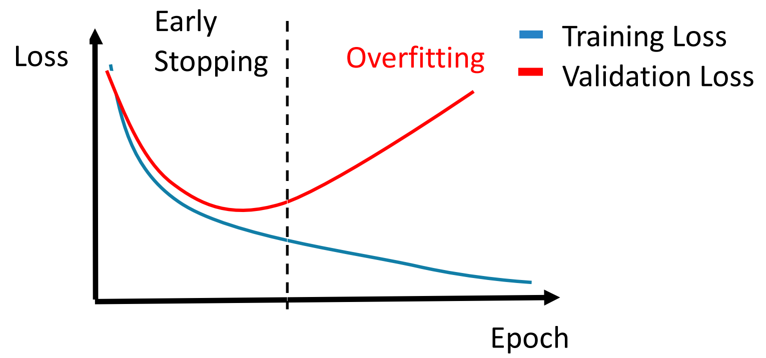
Solutions:

~~Collect more data~~

Methods: Dealing with Overfitting

Early stopping[10] with a patience of 10 epochs

- Stops training when validation loss starts to increase



L2 weight decay

- Penalize squared weights
- Keeps weights small unless error derivative is big

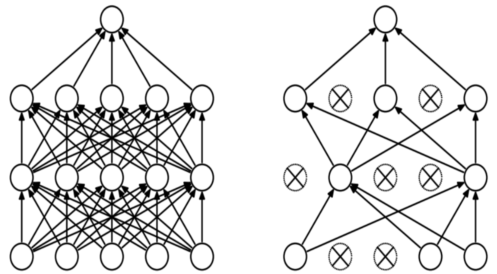
Log Loss Function

w/ L2 Regularization

$$L = -\frac{1}{n}\sum_{x} y \log a + (1-y)\log(1-a) \longrightarrow C = L + \frac{\lambda}{2}\sum_{i} w_i^2$$

Dropout[11] applied at all layers

- Neurons are removed from the network with a specified probability during training.
- This prevents neurons from co-adapting too much.



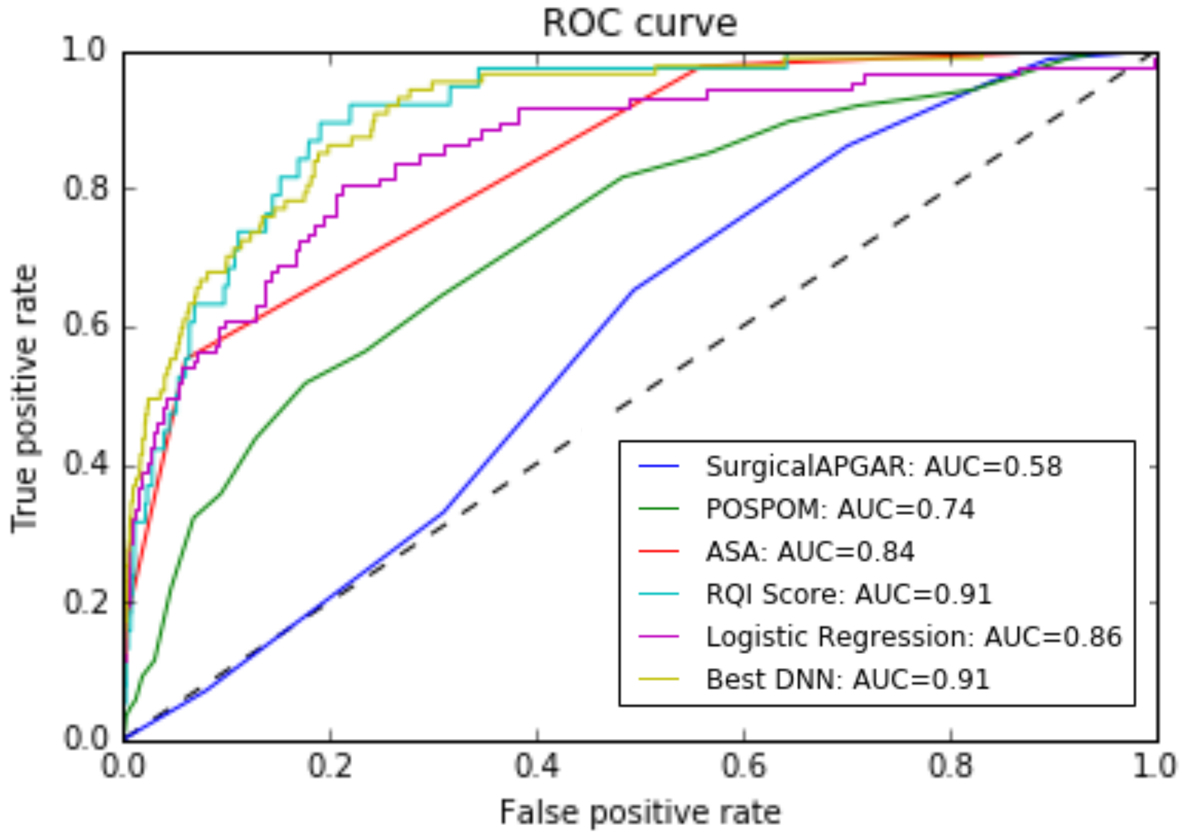
(a) Standard Neural Net (b) After applying dropout.

Figure 1: Dropout Neural Net Model. **Left**: A standard neural net with 2 hidden layers. **Right**: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped.

Results: Best Neural Network Architecture and Hyperparameters

Activation	Output Activiation	Initialization	# Hidden Layers	# Neurons	L2 Weight Decay	Dropout Probability	Learning Rate	Momentum
ReLu	Sigmoid	he_normal	4	[300, 300, 300, 300]	0.0001	0.5	0.01	0.9

Results: AUC

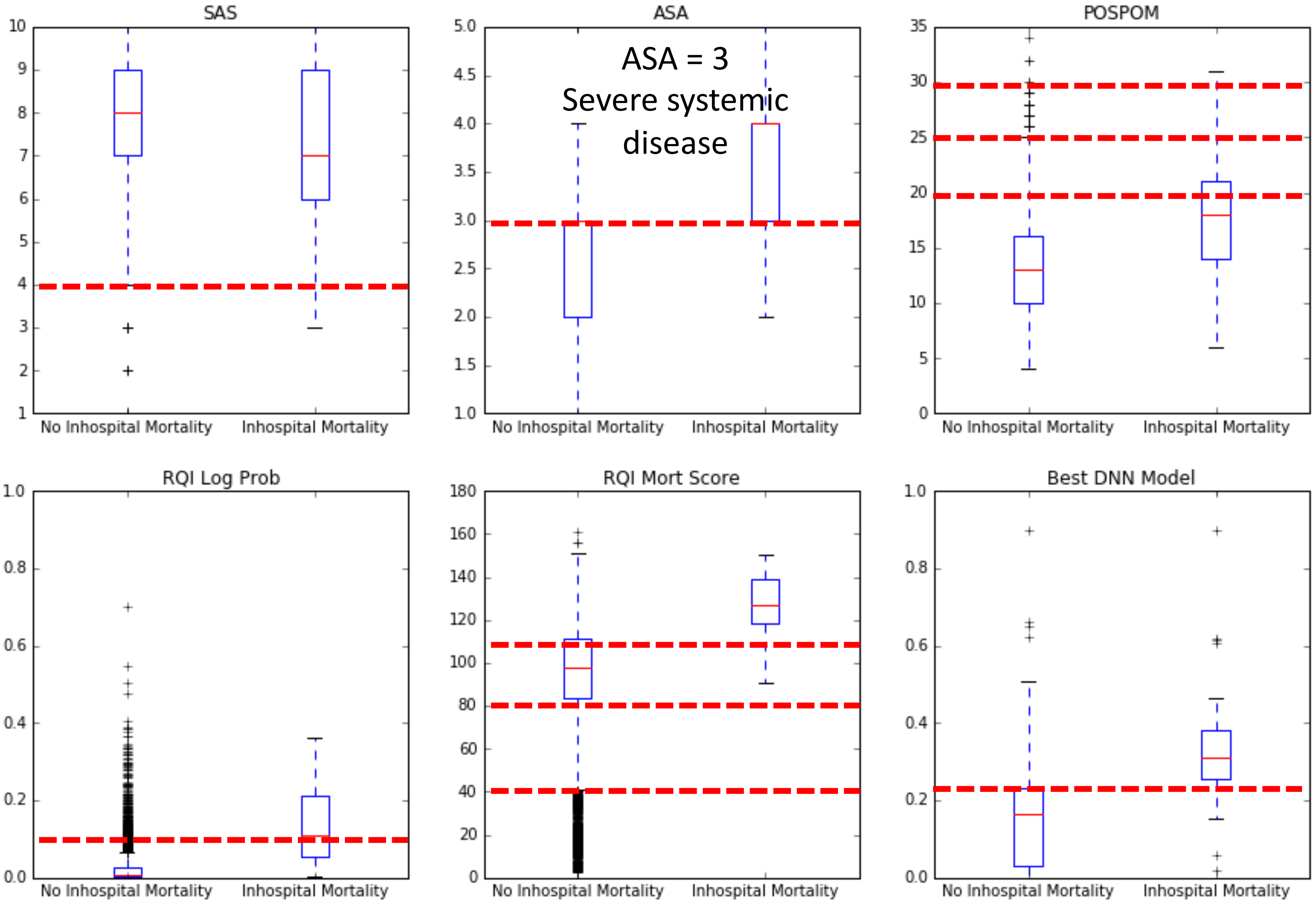


Risk Score	AUC [95% CI]
Surgical Apgar	0.58 [0.52 - 0.64]
POSPOM SCORE	0.74 [0.69 - 0. 78]
ASA	0.84 [0.81 - 0.87]
RQI Log Prob**	0.90 [0.87 - 0.93]
RQI Score**	**0.91 [0.87 - 0.94]**
Model	**AUC [95% CI]**
Logistic Regression	0.86 [0.81 - 0.89]
DNN	0.88 [0.85 - 0.91]
DNN w/ ASA	0.90 [0.87 - 0.93]
DNN w/ Reduced Feature Set	0.89 [0.85 - 0.92]
BEST DNN: DNN w/ Reduced Feature Set & ASA	**0.91 [0.88 - 0.93]**

**It should be noted that RQI could not be calculated for 6,406 of the test patients due to lack of Procedural Severity Scores for their CPT codes.

Results: Boxplots

SAS < 4:
50% risk of major complications, including a 14% mortality rate[5]



POSPOM < 20: probability of in-hospital mortality < 0.32%
POSPOM = 25 : 1.37%
POSPOM = 30 : 5.65%[7]

RQI score = 40: ~0.01% predicted probability of 30 day mortality
RQI score = 80: ~0.05%
RQI score = 110: ~1%[6]

RQI log prob > 10%: 40-50% of 30 day mortality[6]

Results: Choosing Threshold based on F1 Score

Best DNN Threshold	F1 Score	RQI Score	F1 Score	POSPOM	F1 Score
0.1	0.02	100	0.03	10	0.02
0.2	0.03	120	0.08	15	0.03
0.3	0.14	130	0.12	**20**	**0.05**
0.4	**0.22**	**140**	**0.16**	25	0.04
0.5	0.1	145	0.08	30	0.02
0.6	0.1	150	0.05	35	0

	# True Negative	# False Positive	# False Negative	# True Positive
Best DNN Model (n=11,997 all test patients)	11,875	35	72	15
Best DNN Model (n=5,591 RQI Score Calculated Patients)	5,540	13	32	6
RQI Score (n=5,591 RQI Score Calculated Patients)	5,502	51	30	8
POSPOM (n=11,997 all test patients)	10,782	1,128	56	31
POSPOM (n=5,591 RQI Score Calculated Patients)	4,948	605	20	18

Our DNN has the highest F1 Score.

Choosing thresholds based on best F1 score optimizes true negatives.

DNN and RQI model were comparable.

Results: % Mortality by Model

Best DNN Model	# Mortality	% of Mortality Patients (n=87)	RQI Score	# Mortality	% of Mortality Patients (n=38)	POSPOM	# Mortality	% of Mortality Patients (n=87)
0-0.1	2	2.3	0-100	1	2.63	0-10	6	6.9
0.1-.2	1	1.15	100-120	9	23.68	10-20	50	57.5
0.2-0.3	37	42.53	120-130	11	28.95	20-25	26	29.9
0.3-0.4	32	36.78	130-140	9	23.68	25-30	4	4.6
0.4-0.5	10	11.49	140-150	7	18.42	30-40	1	1.1
>= 0.5	5	5.75	150-160	1	2.63	>40	0	0

Set **0.2** as threshold
97% of Mortality Patients

Set **100** as threshold
97% of Mortality Patients

Set **10** as threshold
93% of Mortality Patients

Results: Choosing Threshold Based on True Positives

At threshold values:

- Best DNN : 0.2
- RQI Score : 100
- POSPOM : 10

The best DNN with a threshold of 0.2 decreases the # of false positives compared to RQI by 352 patients, while comparably labeling true positives.

	# True Negative	# False Positive	# False Negative	# True Positive
Best DNN Model (n=11,997 all test patients)	6,680	5,230	3	84
Best DNN Model (n=5,591 RQI Score Calculated Patients)	3,385	2,168	2	36
RQI Score (n=5,591 RQI Score Calculated Patients)	3,033	2,520	1	37
POSPOM (n=11,997 all test patients)	2,741	9,169	6	81
POSPOM (n=5,591 RQI Score Calculated Patients)	1,312	4,241	1	37

Conclusions

DNN models predict inhospital mortality better or comparably to currently published risk scores

The addition of ASA and reducing the number of features improves the DNN models

RQI is comparable to our models, but can only be calculated on ~50% of patients

Our models can be calculated on all patients and leverages both preoperative and intraoperative information

Future Work:

- Testing on a different hospital’s patient population
- Leveraging time series data during operation
- Patient specific

References

1. Weiser, T., Regenbogen, S., Thompson, K., Haynes, A., Lipsitz, S., Berry, W., et al., An estimation of the global volume of surgery: a modelling strategy based on available data, *Lancet* **372** (2008) 139–144.
2. Pearse, R., Harrison, D., James, P., Watson, D., Hinds, C., Rhodes, A., et al., Identification and characterisation of the high-risk surgical population in the United Kingdom, *Crit Care* **10**(2006) 1–6.
3. Pearse, R., Moreno, R., Bauer, P., Pelosi, P., Metnitz, P., Spies, C., et al., Mortality after surgery in Europe: a 7 day cohort study, *Lancet* **380** (2012) 1059–1065.
4. American Society of Anesthesiologists, A. S. of A. A. S. of A., New classification of physical status, *Anesthesiology* (1963) 11.
5. Gawande, A., Kwaan, M., Regenbogen, S., Lipsitz, S., and Zinner, M., An Apgar Score for Surgery, *J Am Coll Surgeons* **204** (2007) 201–208.
6. Dalton, J.E., Kurz, A., Turan, A., Mascha, E.J., Sessler, D.I. and Saager, L., 2011. Development and validation of a risk quantification index for 30-day postoperative mortality and morbidity in noncardiac surgical patients. *The Journal of the American Society of Anesthesiologists*, *114*(6), pp.1336-1344.
7. Manach, Y., Collins, G., Rodseth, R., Bihan-Benjamin, C., Biccard, B., Riou, B., et al., Preoperative Score to Predict Postoperative Mortality (POSPOM), *Anesthesiology* **124** (2016) 570–579.
8. Haddow, J.B., Adwan, H., Clark, S.E., Tayeh, S., Antonowicz, S.S., Jayia, P., Chicken, D.W., Wiggins, T., Davenport, R., Kaptanis, S. and Fakhry, M., 2014. Use of the surgical Apgar score to guide postoperative care. *The Annals of The Royal College of Surgeons of England*, *96*(5), pp.352-358.
9. Terekhov, Maxim A., Jesse M. Ehrenfeld, and Jonathan P. Wanderer. "Preoperative Surgical Risk Predictions Are Not Meaningfully Improved by Including the Surgical Apgar ScoreAn Analysis of the Risk Quantification Index and Present-On-Admission Risk Models." *The Journal of the American Society of Anesthesiologists* 123, no. 5 (2015): 1059-1066.
10. Yao, Y., Rosasco, L., and Caponnetto, A., On Early Stopping in Gradient Descent Learning, *Constr Approx* **26** (2007) 289–315.
11. Srivastava, N., Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research* **15** (2014) 1929–1958.
12. Sigakis, Matthew JG, Edward A. Bittner, and Jonathan P. Wanderer. "Validation of a risk stratification index and risk quantification index for predicting patient outcomesin-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay." *The Journal of the American Society of Anesthesiologists* 119, no. 3 (2013): 525-540.