PSI 5889 – Princípios de Neurocomputação

Detecção de anomalias em séries temporais utilizando autoencoders recorrentes

Marina Cardoso de Oliveira

Ricardo Tiosso Panassiol

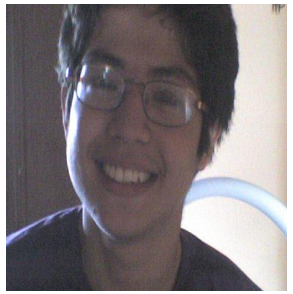
Vitor Augusto Takeuchi

Grupo e Divisão das Tarefas



Ricardo

- Pesquisa e formalização da fundamentação teórica;
- Suporte com a implementação das redes;



Vitor

- Suporte com a implementação das redes;
- Realização de testes com a base de dados da fábrica de papel;

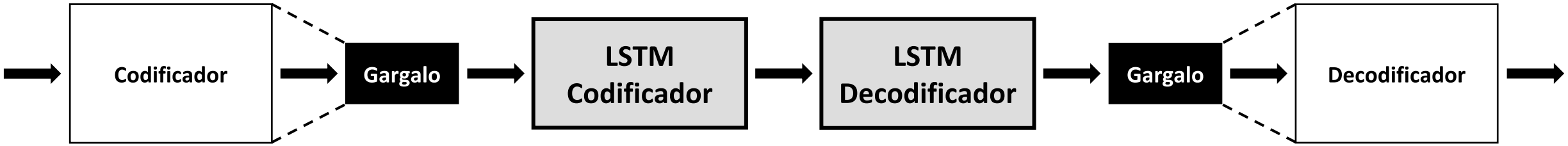


Marina

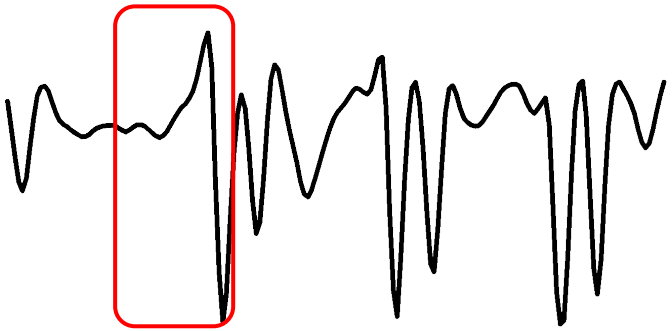
- Implementação da rede neural;
- Realização de testes com a base de dados biológicos.

Autoencoder Recorrente

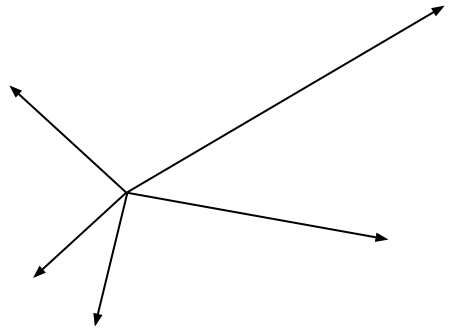
- Um *Autoencoder* Recorrente é uma implementação de um *autoencoder* para dados de sequência usando uma arquitetura LSTM *Encoder-Decoder* que inclui a temporalidade em seu processamento.



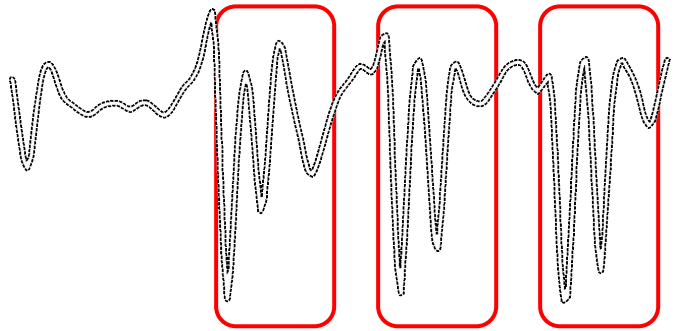
Entrada Temporal Original



Representação Comprimida



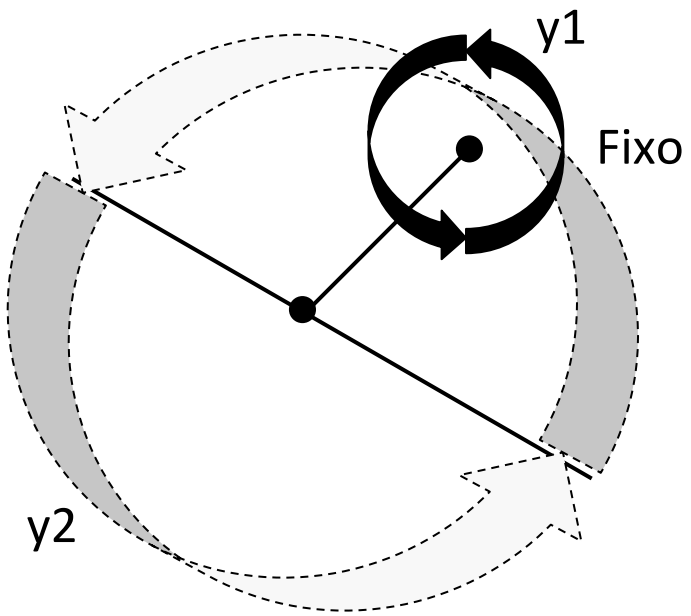
Saída Temporal Prevista

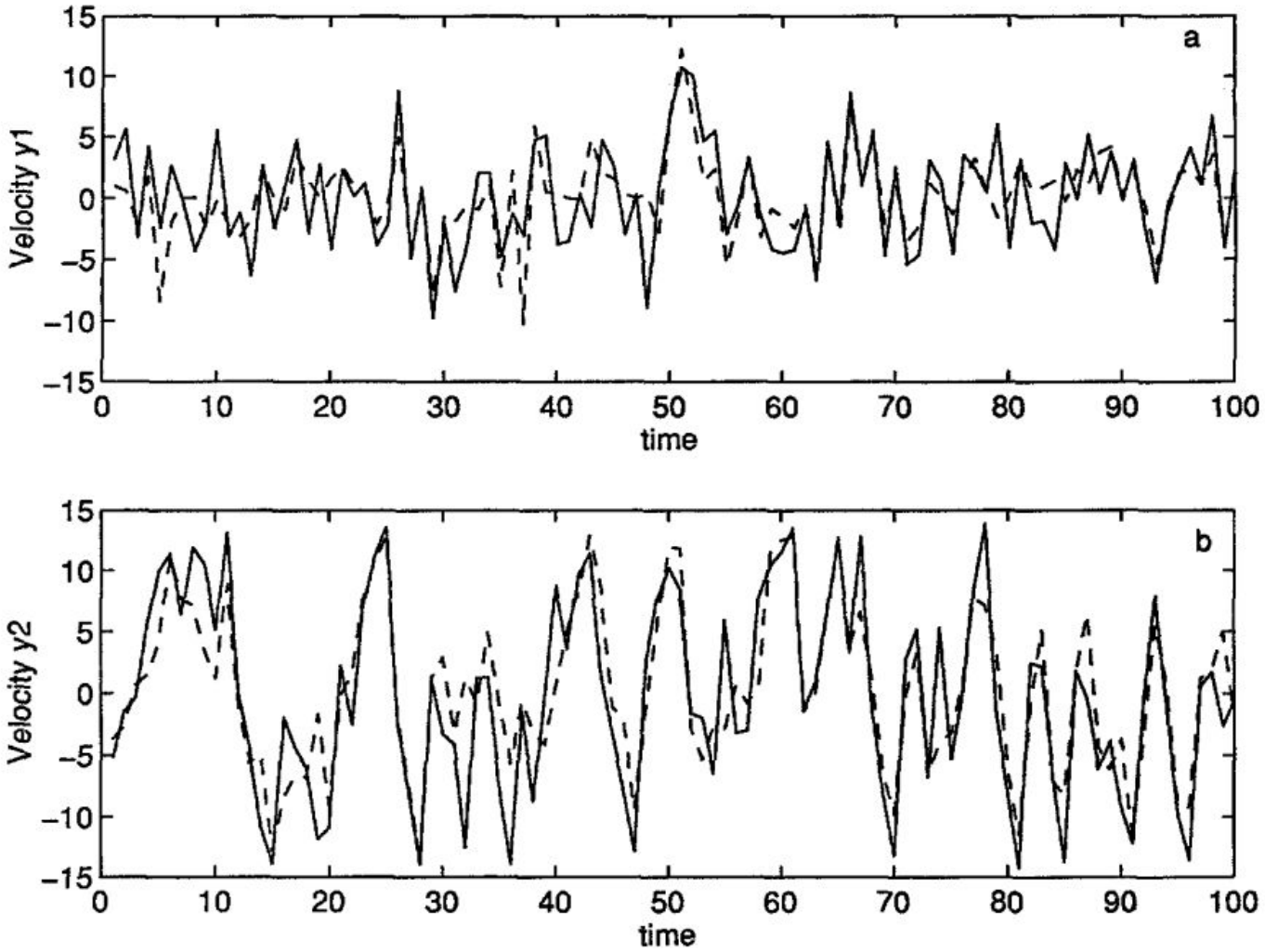


Aplicação: Previsão e Estimativa

Série temporal escalar de velocidade

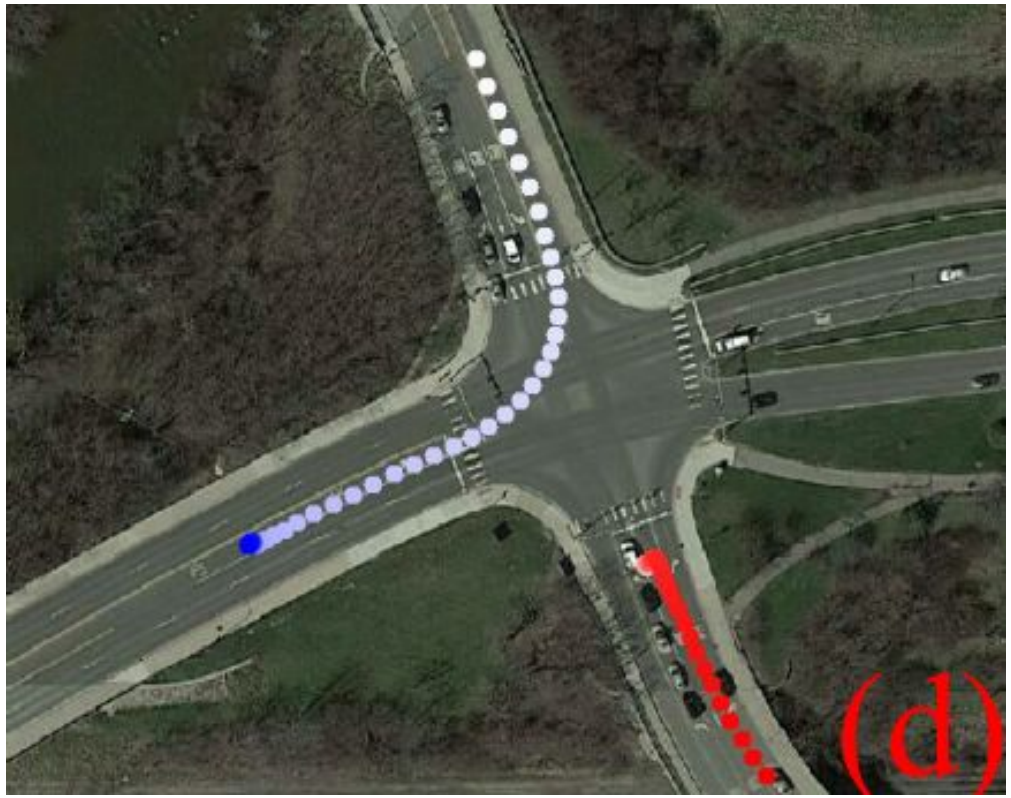
Duplo Rotor

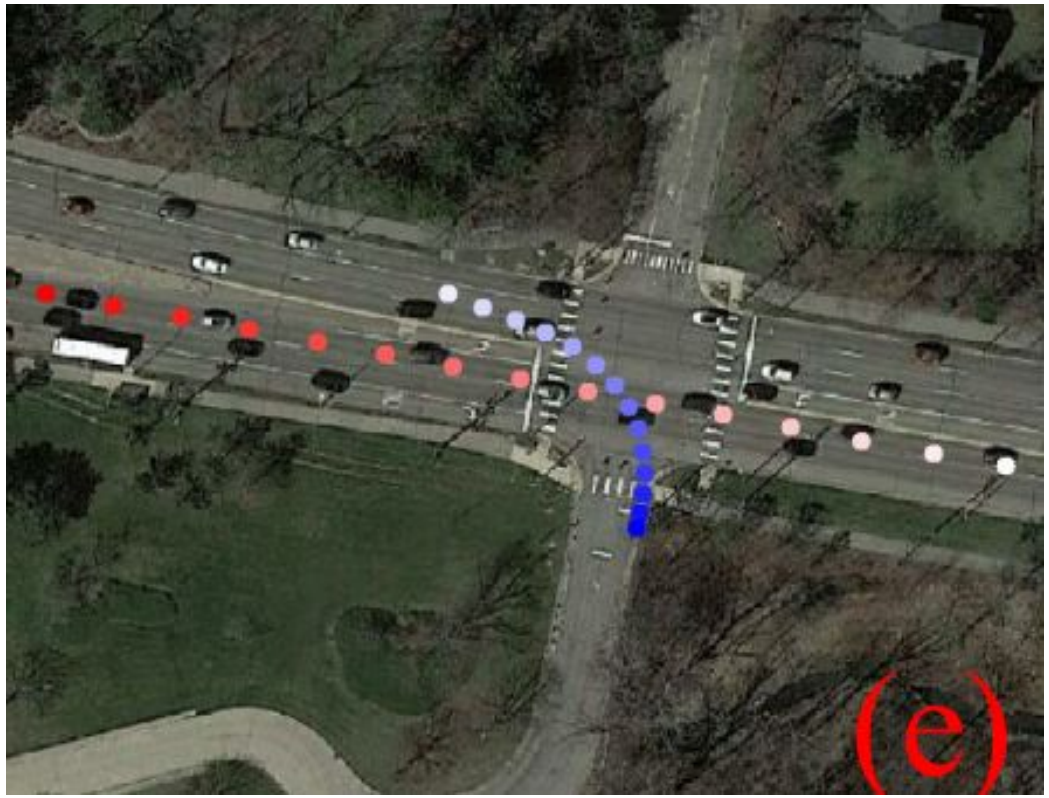




Cao; Mees; Judd (1998). *Physica D*, 121, 75-88

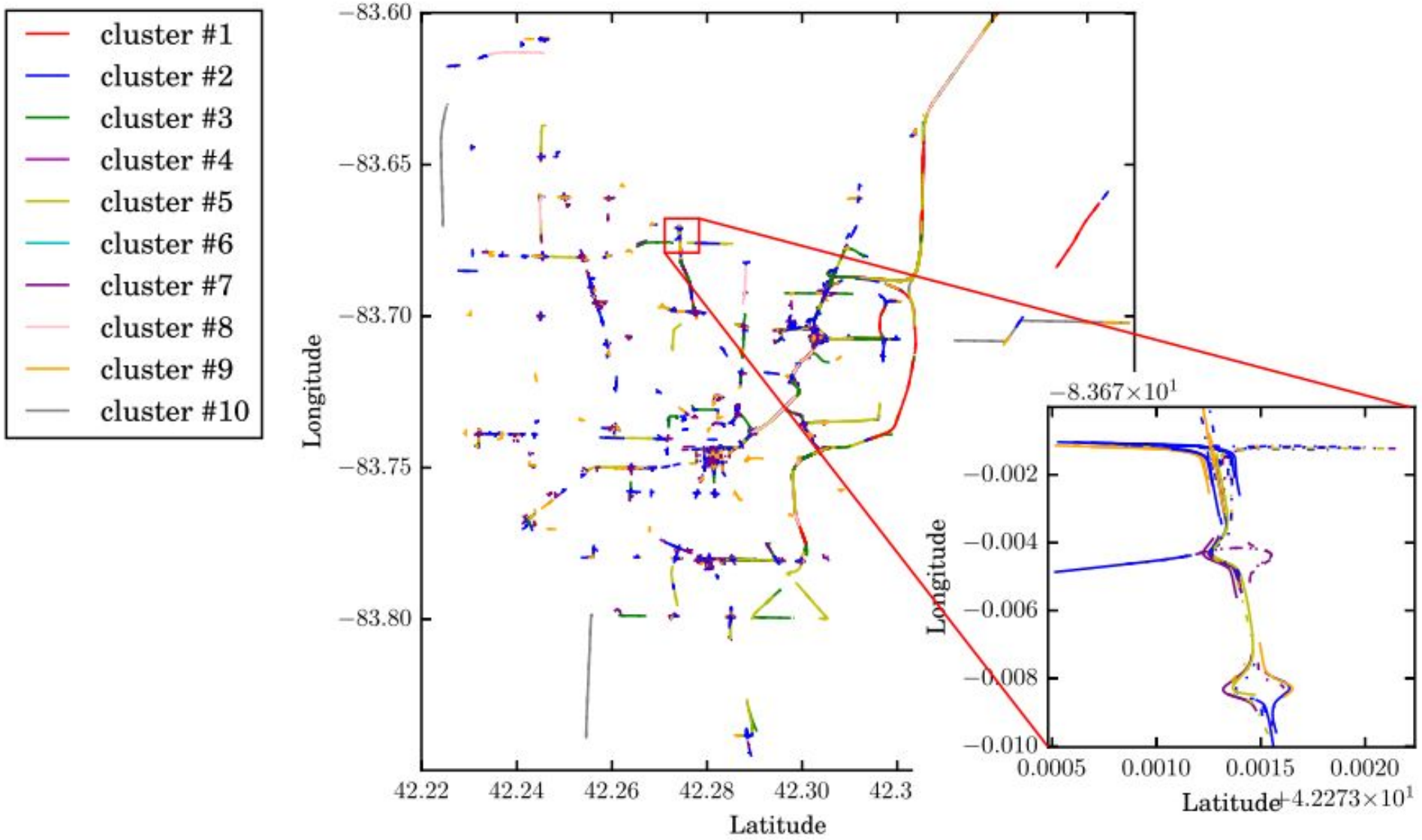
Aplicação: Classificação e Clusterização





Wang; Ramesh; Zhao (2018). *arXiv:1807.08415*.

Aplicação: Classificação e Clusterização



Wang; Ramesh; Zhao (2018). *arXiv:1807.08415*.

Aplicação: Detecção de sinal e anomalias

Anomalias ou outliers são valores atípicos dentro do conjunto de dados, apresentando inconsistência em relação aos demais.

Detecção de anomalias

- Base de dados muito desbalanceada;
- Diferentes tipos de anomalias;
- Os eventos não são rotulados;
- O treinamento é realizado apenas com os dados “normais”.

X

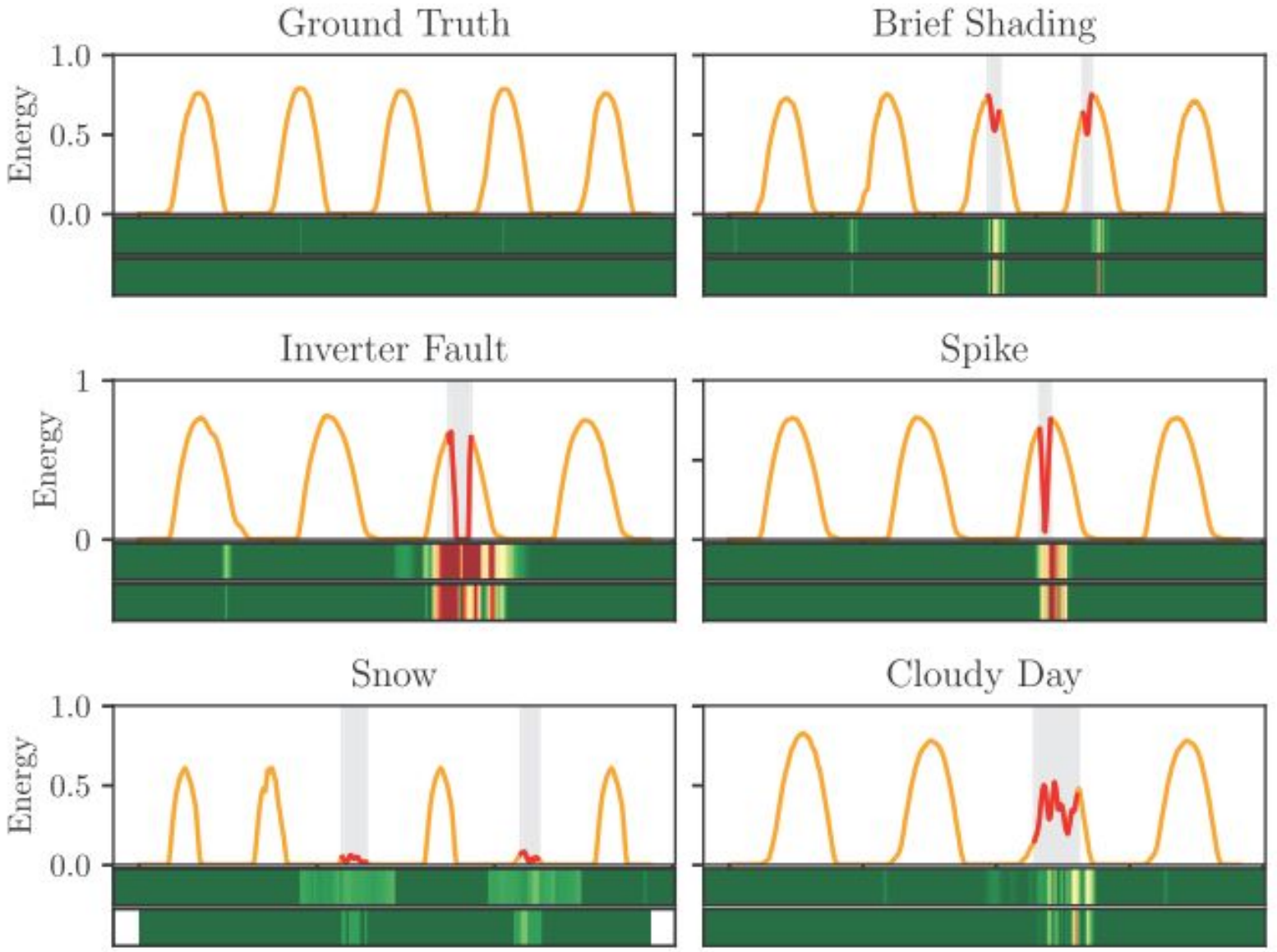
Aprendizado Supervisionado

- Grande número de eventos positivos e negativos;
- Eventos costumam apresentar algum tipo de similaridade;
- Presença de rótulos;
- Treinamento realizado com todas as classes.

Aplicação: Detecção de sinal e anomalias

Série temporal escalar de energia





Pereira; Silveira (2018). *17th IEEE International Conference on Machine Learning and Applications*

Objetivos

Gerais

- Introduzir a arquitetura do autoencoder recorrente, enfatizando a sua aplicação para a detecção de anomalias em séries temporais multivariadas.

Específicos

- Implementar a arquitetura do autoencoder recorrente*;
- Realizar testes com datasets (biológicos ou não) que apresentam diferentes quantidades de anomalias;
- Avaliar a rede por métricas claras.

*https://towardsdatascience.com/lstm-autoencoder-for-extreme-rare-event-classification-in-keras-ce209a224cfb

Base de dados - fábrica de papel

- 18394 elementos – divisão em conjunto de treino, validação e teste.
- 99,34% correspondem ao funcionamento adequado da máquina e 0,66% indicam falhas.
- Diversos sensores na máquina:
  - Variáveis relacionadas à matéria prima, por exemplo quantidade de fibra de celulose, químicos, etc.
  - Variáveis relacionadas ao processo, por exemplo tipo de lâmina e velocidade do rotor.



Fonte: Ranjan, Chitta et al. (2018).Dataset: Rare Event Classification in Multivariate Time Series.

Métricas de avaliação:

- Curva Precision - Recall*
  Precisão:

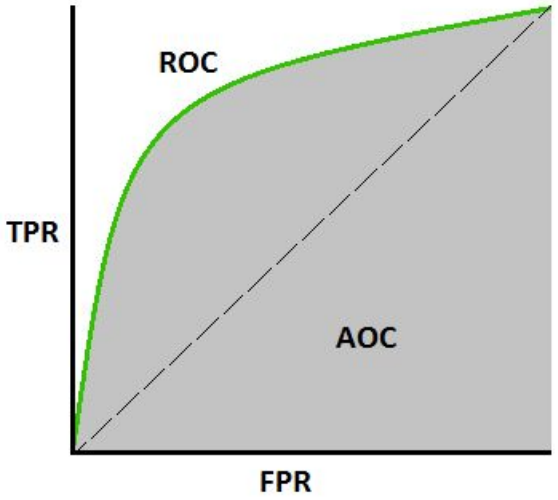
$$P = \frac{T_p}{T_p + F_p}$$

  Sensibilidade (Recall):

$$R = \frac{T_p}{T_p + F_n}$$

A avaliação da precisão e sensibilidade é útil em casos em que existe um desbalanceamento entre as duas classes de observações.

- Curva ROC**:

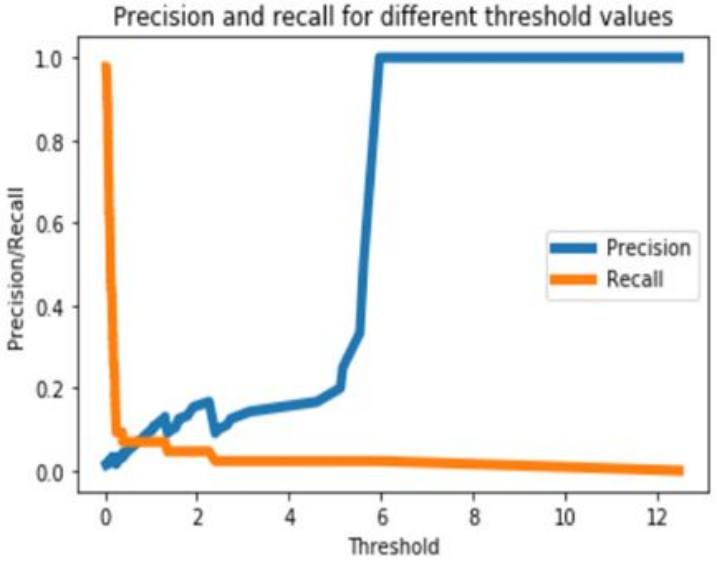


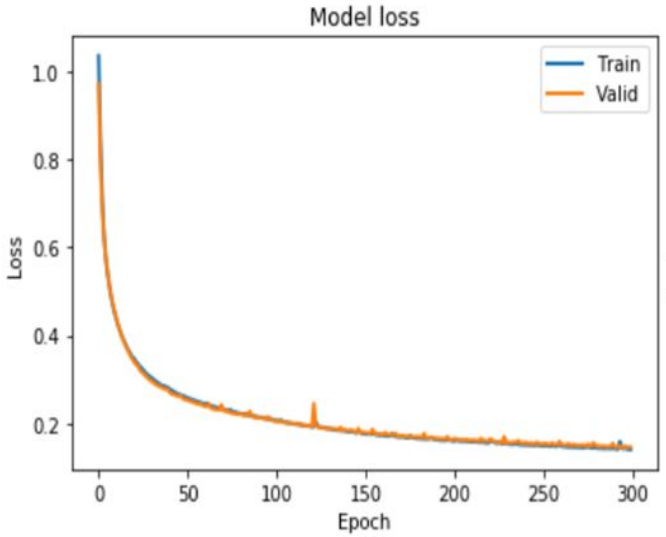
- Matriz de confusão***:

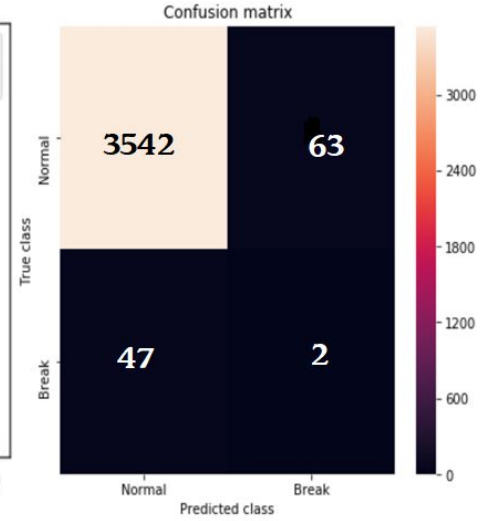
	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

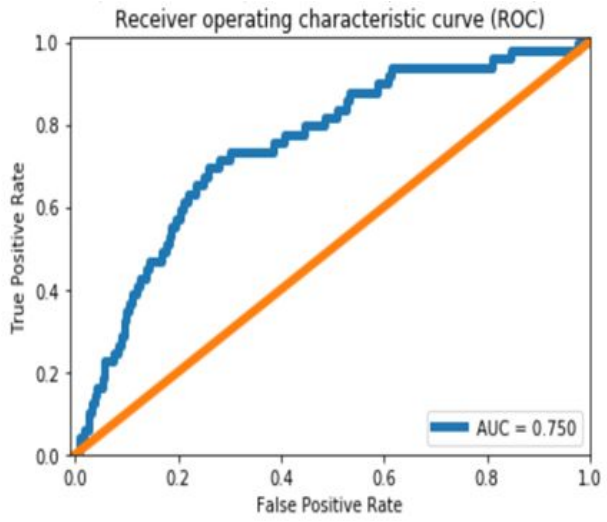
*https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
**https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
***https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

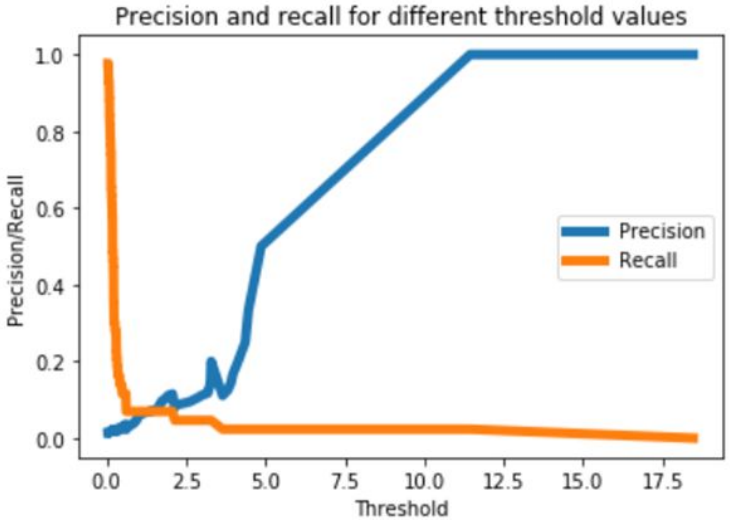
Resultados

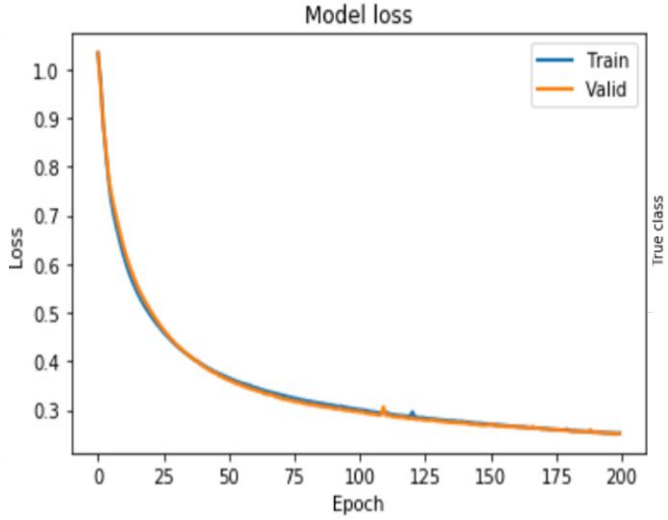

Precision and recall for different threshold values
Precision/Recall
Threshold
Precision
Recall

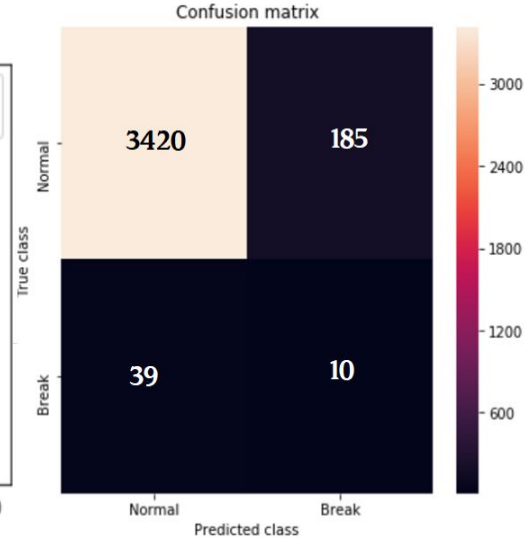

Model loss
Loss
Epoch
Train
Valid

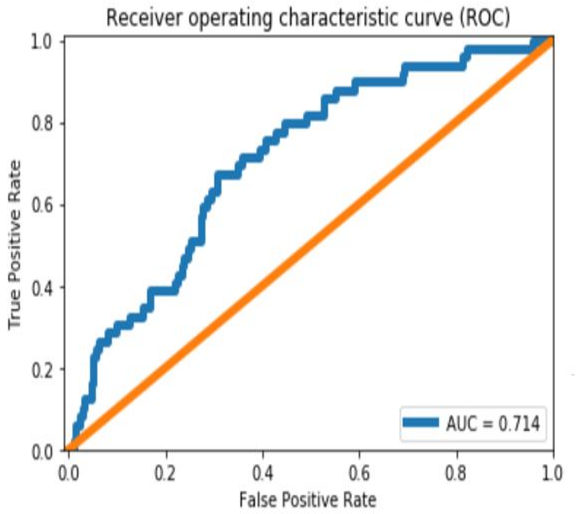

Confusion matrix
3542
63
47
2
True class
Predicted class
Normal
Break


Receiver operating characteristic curve (ROC)
True Positive Rate
False Positive Rate
AUC = 0.750

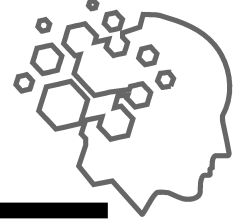

Precision and recall for different threshold values
Precision/Recall
Threshold
Precision
Recall


Model loss
Loss
Epoch
Train
Valid


Confusion matrix
3420
185
39
10
True class
Predicted class
Normal
Break


Receiver operating characteristic curve (ROC)
True Positive Rate
False Positive Rate
AUC = 0.714

Principais Desafios



- Durante a fase de implementação da rede, enfrentamos alguns desafios:
  - Quantidade de neurônios utilizados por camada;
  - Taxa de aprendizado;
  - Número de épocas;
  - Tempo de processamento.

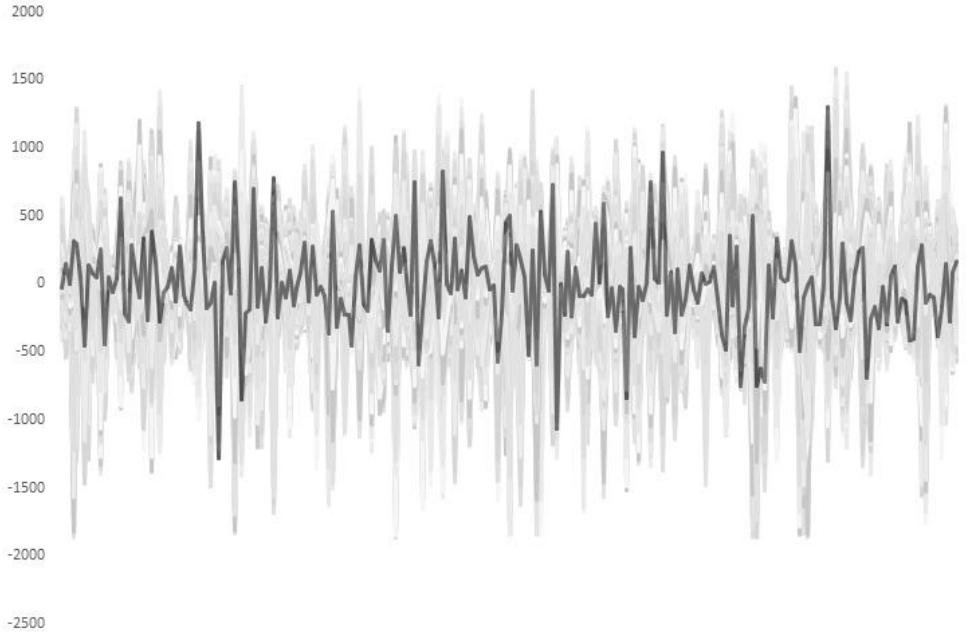
- Localizar datasets para testar a rede.

Base de dados - Reconhecimento de crises epilépticas

Abstract: This dataset is a pre-processed and re-structured/reshaped version of a very commonly used dataset featuring epileptic seizure detection.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	11500	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	179	**Date Donated**	2017-05-24
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	86678

- Dados de 500 indivíduos
- Cada ponto indica o valor do EEG em um instante diferente
- 2300 instâncias (20%) indicam que houve crise epiléptica



Fonte: Autores

Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE (2001) Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E, 64, 061907

https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

Base de dados - Reconhecimento de crises epilépticas

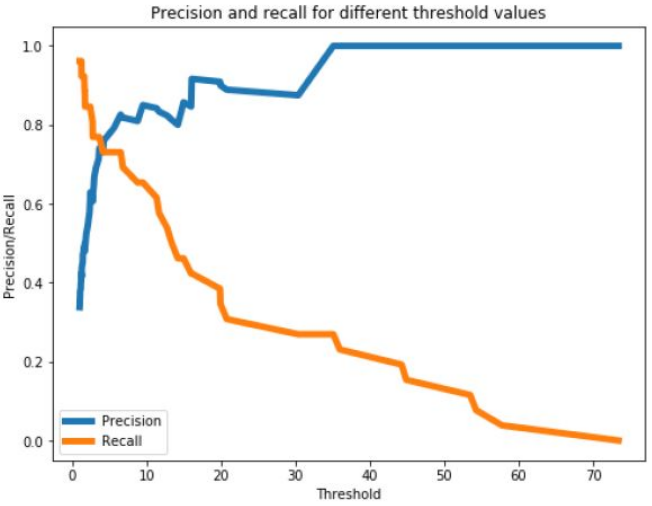
Além dos testes realizados com a base de dados original, contendo 20% de dados positivos, foram realizados testes também com a mesma base de dados contendo apenas 2,44% dos dados positivos. Os dados positivos foram removidos utilizando uma função aleatória.

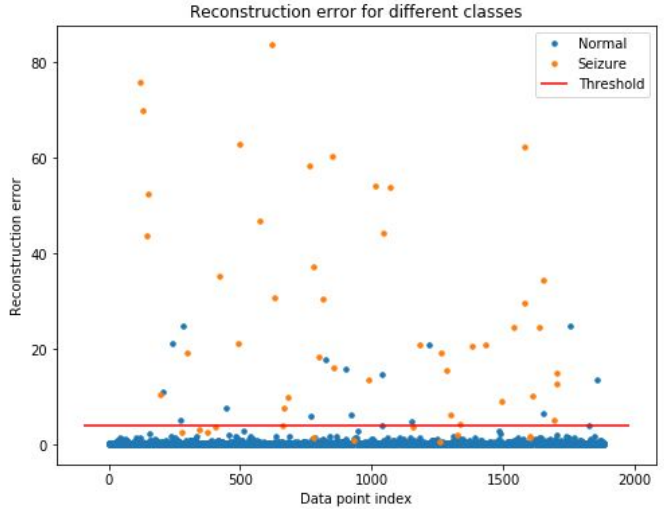
Classe 1 - Crises epilépticas
Classe 0 - Ausência de crises epilépticas

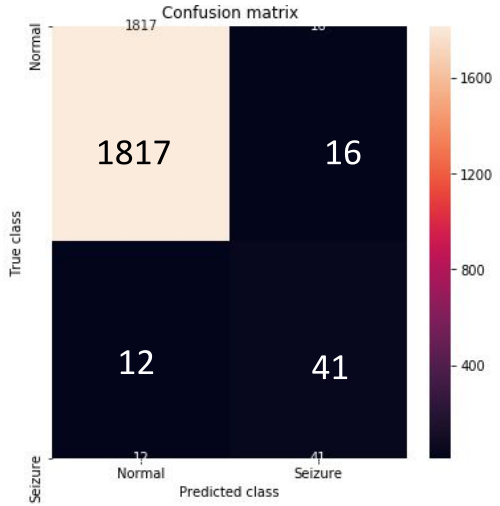
	Quantidade total de instâncias (incluindo colunas categóricas)	Treino		Validação		Teste	
		Classe 1	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0
Base com 20% de dados anormais	11500	0	5893	371	1469	465	1835
Base com 2,44% de dados anormais	9430	0	5881	26	1482	53	1833

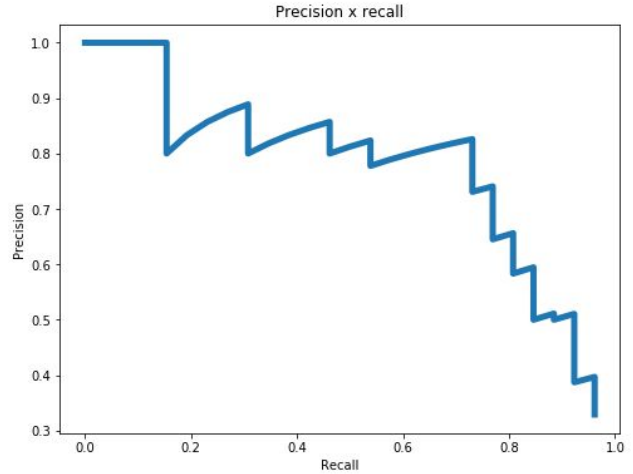
Resultados - 2,44 % de anomalias

Lookback = 1; Camadas: 32, 16, 16, 32; Função de ativação: ReLu

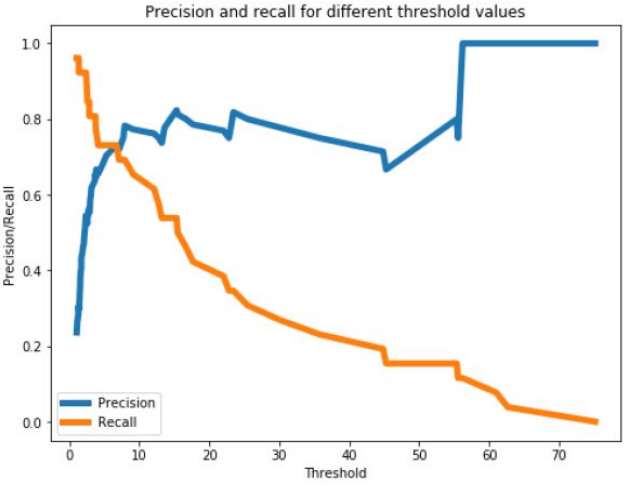


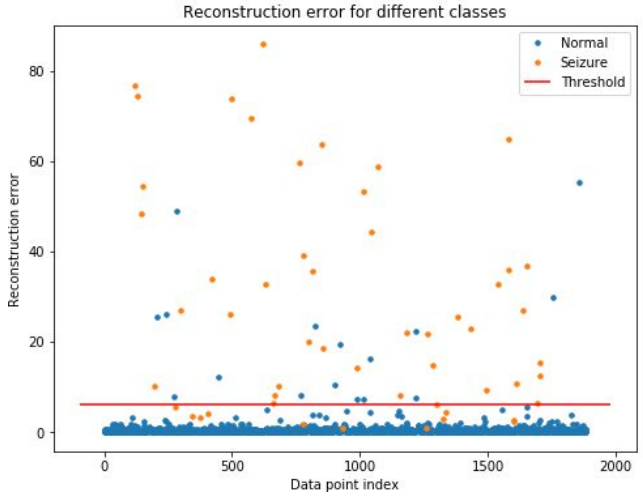


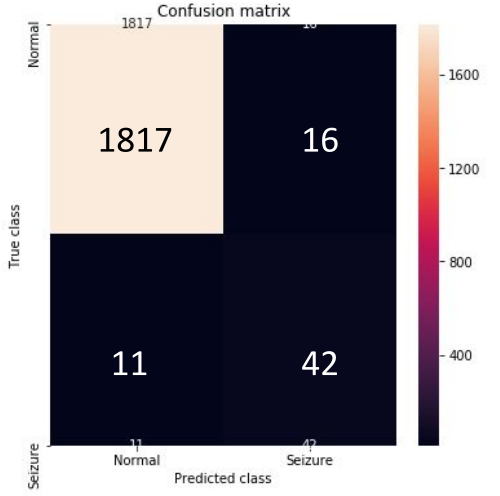


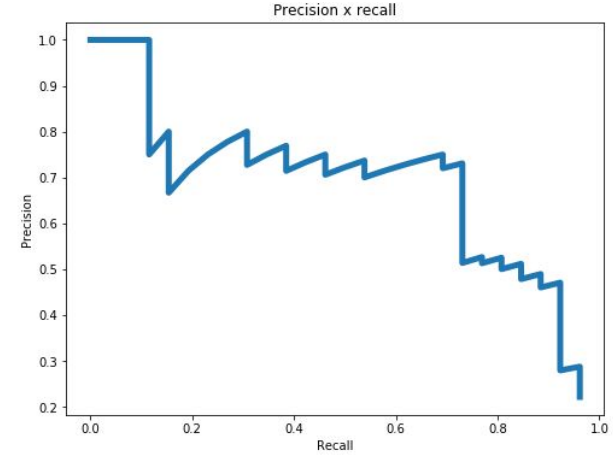


Lookback = 3; Camadas: 32, 16, 16, 32; Função de ativação: ReLu



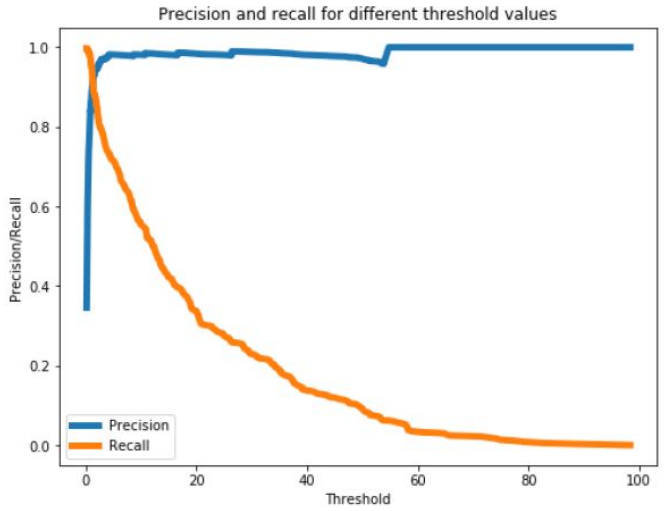


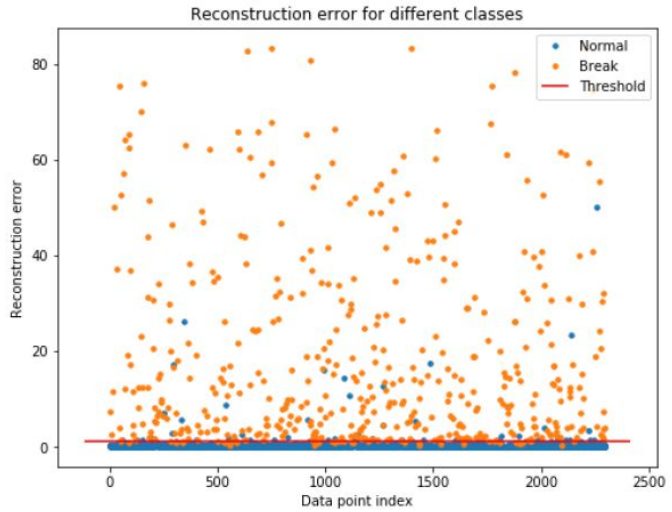


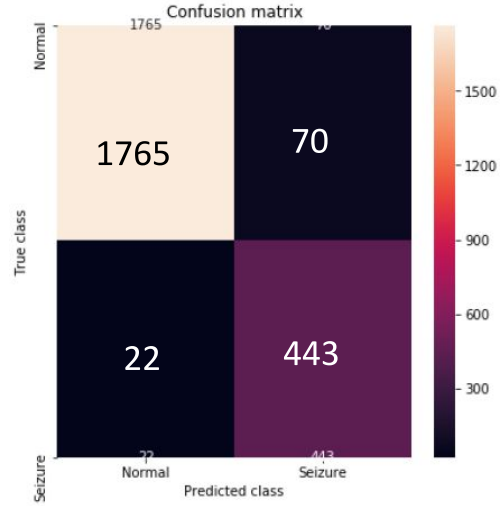


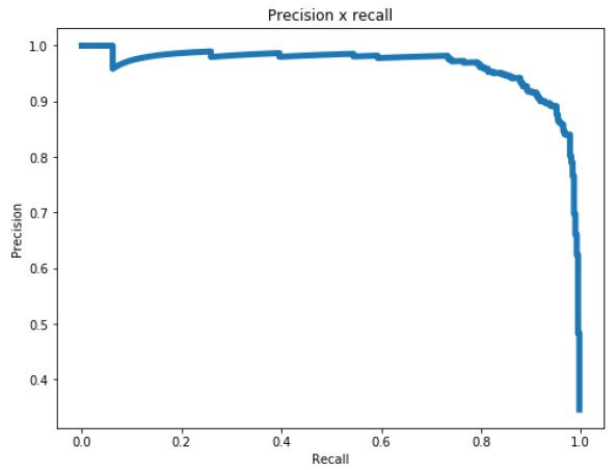
Resultados - 20 % de anomalias

Lookback = 1; Camadas: 32, 16, 16, 32; Função de ativação: ReLu

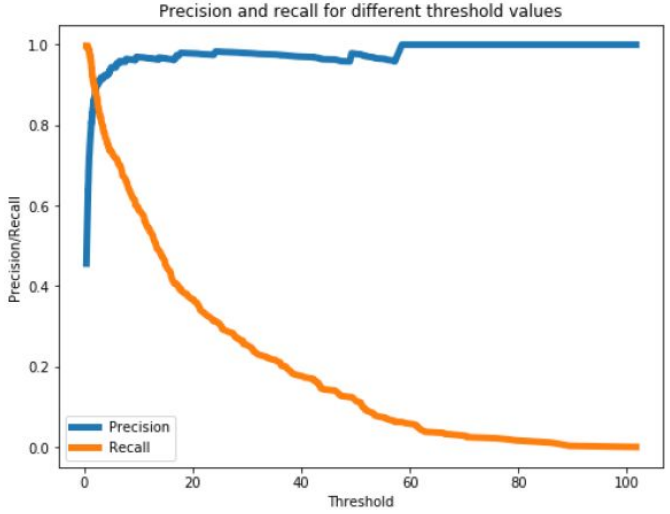


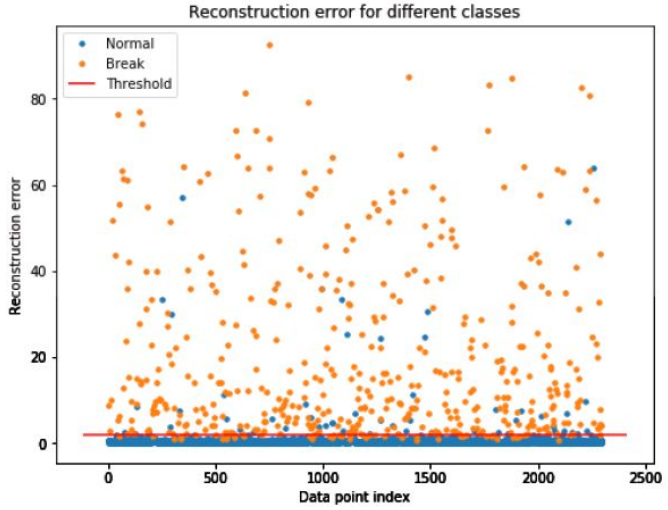


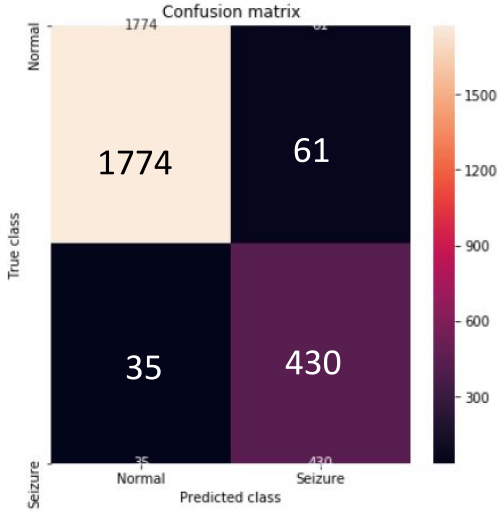


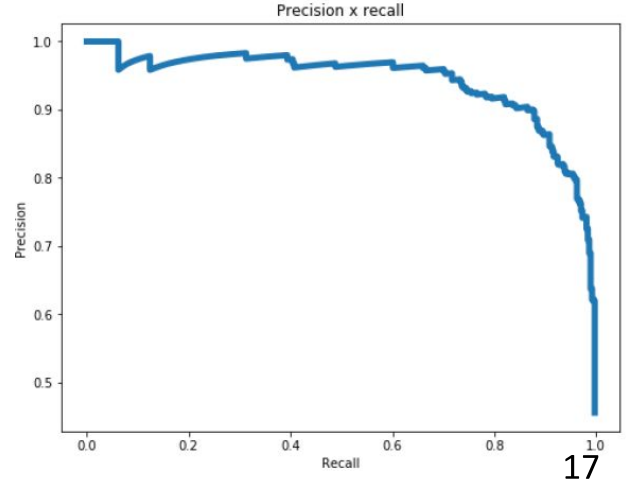


Lookback = 3; Camadas: 32, 16, 16, 32; Função de ativação: ReLu









Outras aplicações

Classification of Normal/Abnormal Heart Sound Recordings: the PhysioNet/Computing in Cardiology Challenge 2016.

- 3126 gravações de sons cardíacos;
- 665 sinais anormais e 2461 sinais normais;
- Duração entre 5 e 120 segundos;
- Condições não ideais – presença de ruídos ambientes, sons intestinais, respiração, movimentação do estetoscópio;

https://www.physionet.org/content/challenge-2016/1.0.0/

MIT-BIH Arrhythmia Database

- 48 trechos de 30 minutos de registros ambulatoriais de ECG de dois canais, obtidos de 47 indivíduos.
- Dados coletados entre 1975 e 1979.

Conclusões

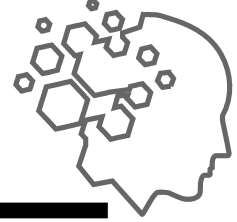
- O dataset de reconhecimento de falhas no maquinário, que contém apenas 0,66% de anomalias, teve o funcionamento esperado.
- É necessário um refinamento nos parâmetros para detectar mais falhas no maquinário.
- Para base de dados de reconhecimento de crises epilépticas, a rede apresentou desempenho satisfatório.
- Entretanto, tem seu melhor funcionamento para uma base com quantidade maior de anomalias.

Referências

- CHAITANYA, Chakravarty R. Alla et al. Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder. **ACM Transactions on Graphics (TOG)**, v. 36, n. 4, p. 98, 2017.
- D'AVINO, Dario et al. Autoencoder with recurrent neural networks for video forgery detection. **Electronic Imaging**, v. 2017, n. 7, p. 92-99, 2017.
- KIEU, Tung et al. Outlier detection for time series with recurrent autoencoder ensembles. In: **28th international joint conference on artificial intelligence**. 2019.
- KIEU, Tung; YANG, Bin; JENSEN, Christian S. Outlier detection for multidimensional time series using deep neural networks. In: **2018 19th IEEE International Conference on Mobile Data Management (MDM)**. IEEE, 2018. p. 125-134.
- PEREIRA, Joao; SILVEIRA, Margarida. Unsupervised Anomaly Detection in Energy Time Series Data using Variational Recurrent Autoencoders with Attention. In: **2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)**. IEEE, 2018. p. 1275-1282.
- Ranjan, C., Mustonen, M., Paynabar, K., & Pourak, K. (2018). Dataset: Rare Event Classification in Multivariate Time Series. *arXiv preprint arXiv:1809.10717*

Referências

- https://towardsdatascience.com/lstm-autoencoder-for-extreme-rare-event-classification-in-keras-ce209a224cfb
- https://www.tensorflow.org/tutorials/structured_data/imbalanced_data
- https://github.com/cran2367/lstm_autoencoder_classifier/blob/master/lstm_autoencoder_classifier.ipynb
- https://github.com/cran2367/understanding-lstm-autoencoder
- https://machinelearningmastery.com/lstm-autoencoders/



Obrigado!