EDMaL

Ang Jun Ray (RO019)

Enhanced Detection of AI-Generated Text using Machine Learning

The Problem

60% of people do not trust companies to be **ethical** in their use of AI

Inaccurate results

Fake News

Misuse/Dishonest Use

Current Solutions

GPTZero

DetectGPT

Content-At-Scale AI Detector

ZeroGPT

DNA-GPT (Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text)

Our Research: Improving on DNA-GPT

4 Methods

Stemming from the DNA-GPT's original methods, we propose **4 methods** to try to enhance their text-detection:

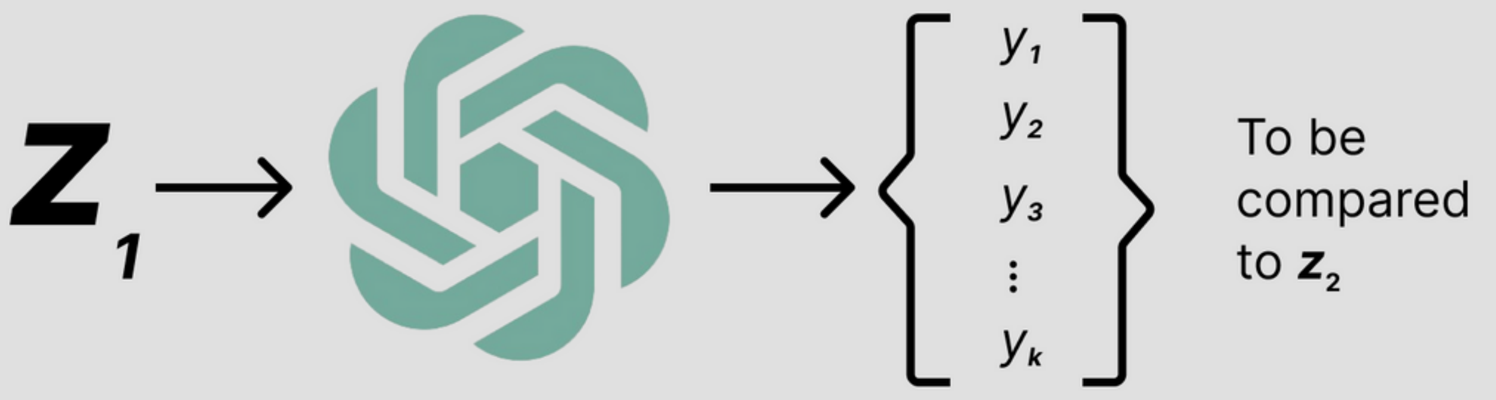
Current DNA-GPT Detection using N-Gram Analysis

1 **Truncate** the text, *z*, using the truncate ratio of γ. (γ = 0.5)

"xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"

z_2

2 **Regenerate** from the **truncated output, z_1,** using an LLM, K times. (K = 10)



3 **Compare** y_i to z_2 to classify *z* as AI-generated or Human-written

We propose and test **4 additional methods** of doing so along with doing analysis of n-grams (like DNA-GPT

4 Using computed scores in step 3, and based on a **threshold**, evaluate if text is AI-generated or human-written

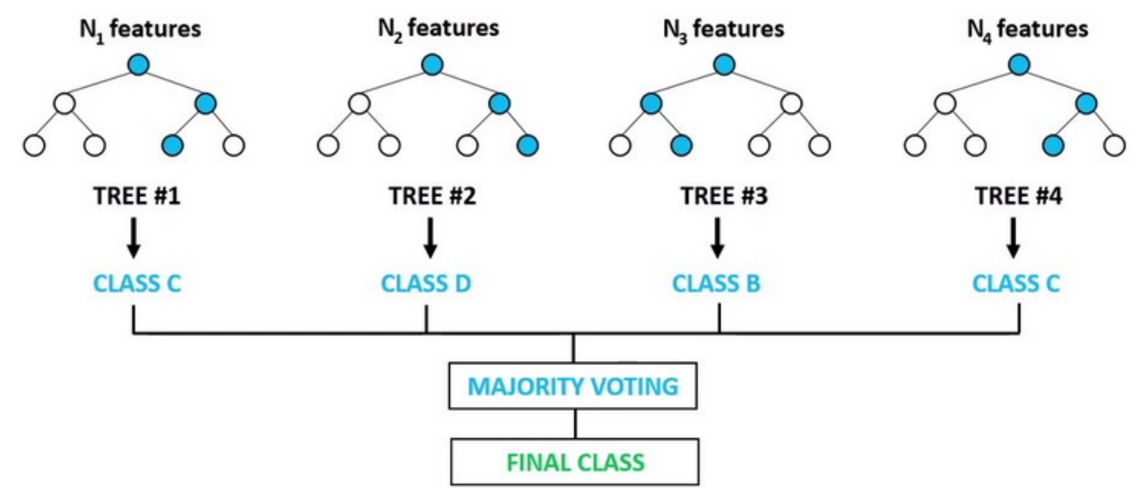
The threshold has to be fine-tuned to maximize scores for metrics

We grouped the 4 methods to improve step 3 into **2 groups:**

Group A: Extensions of N-gram Analysis

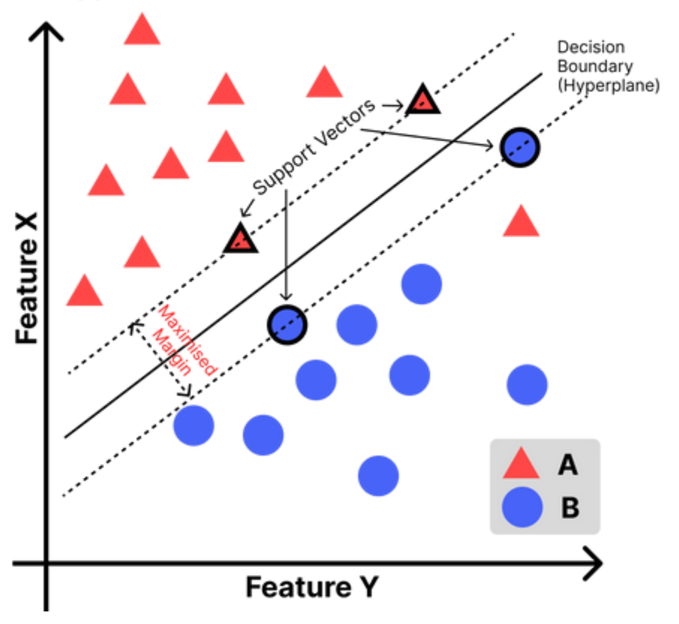
Experimenting with common Machine Learning tools

A.1 Random Forest Classifier



A.2 SVM

(Support Vector Machine)



We trained both by taking **feature vector *x*** as input and returning a label of **1 (AI-generated) or 0 (Human-written)** as *score_z*.

Group B: Alternative Approaches

Other Ad-hoc methods

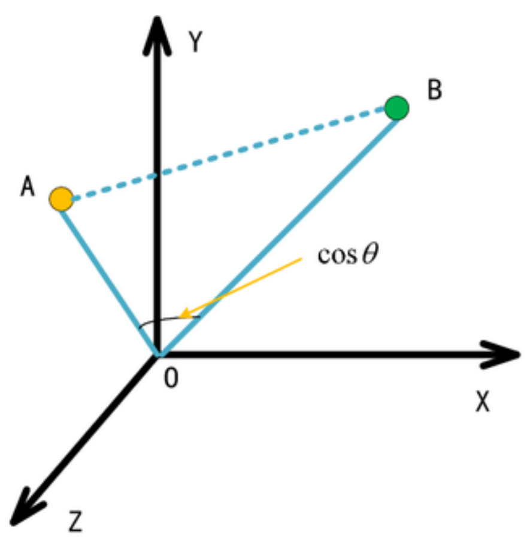
B.1 Levenshtein Edit Distance

(Does not use Machine Learning)

By taking the number of transformations (with **replace, insert and delete** operations) to change *y_i* to *z_2*, we aim to analyze the lexical structure of the text.

This method was added since its similar to the original N-gram analysis method DNA-GPT used.

B.2 Word Embeddings With Cosine Similarity

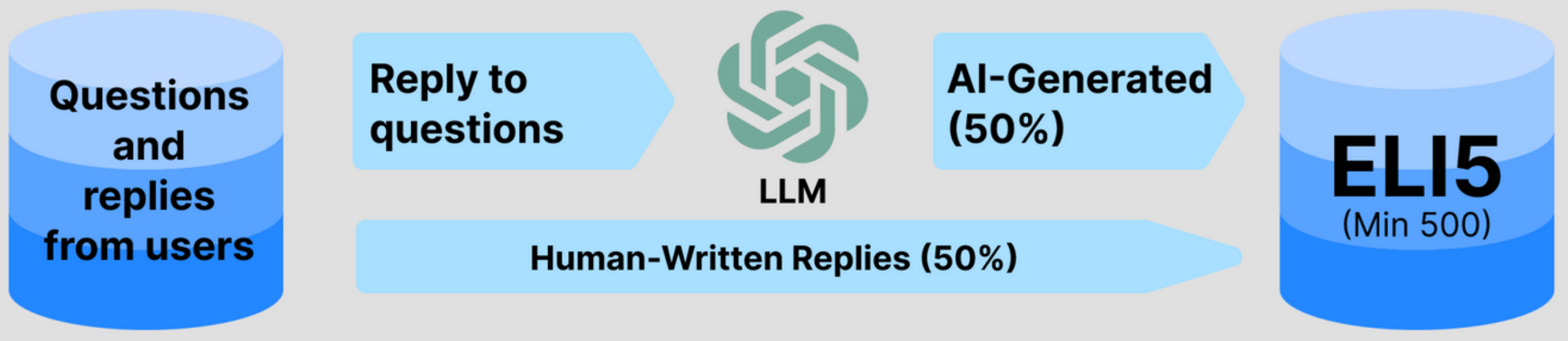


This method aims to capture the **semantic meaning** of text and **determine** texts' semantic distance.

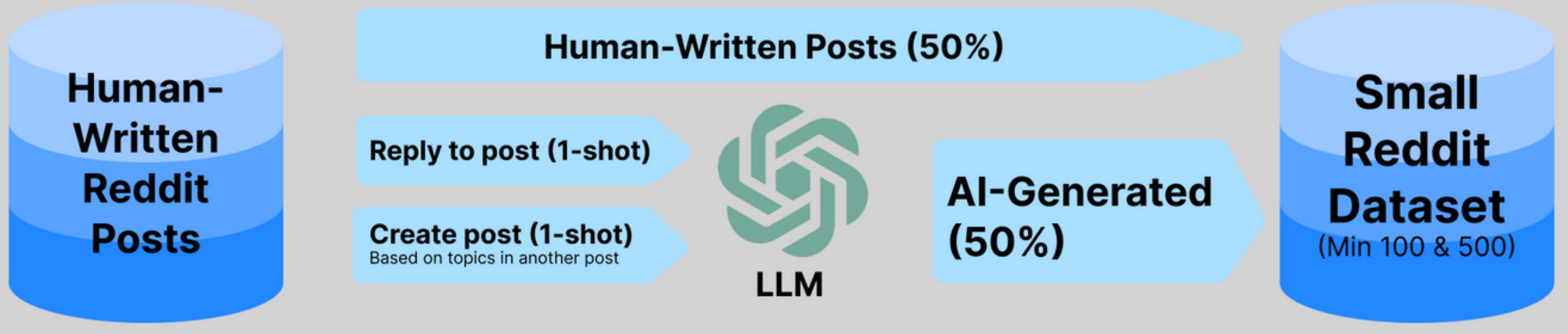
3 Datasets

ELI5 - Explain Like I'm Five (Min 500)

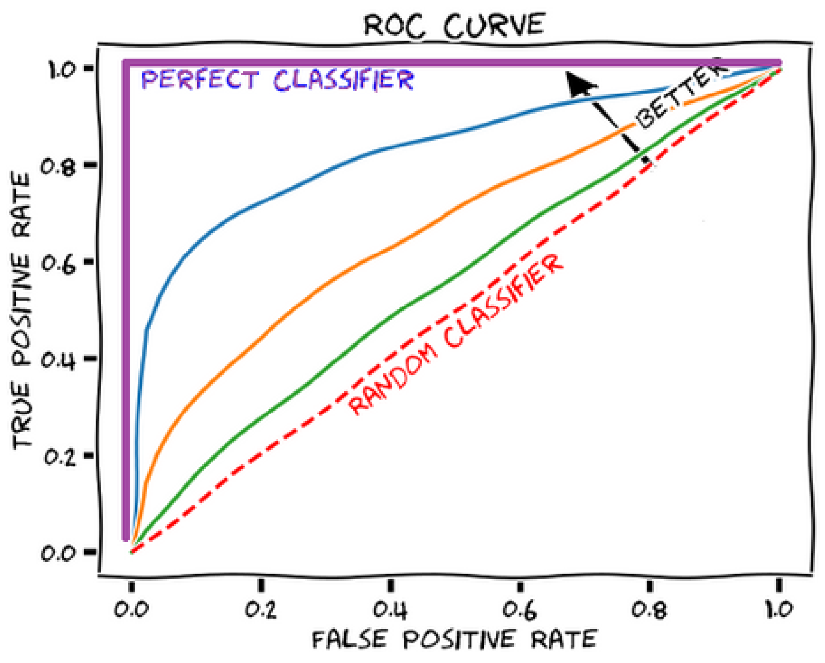
This was also used by DNA-GPT, allowing us to do a ground truth comparison



Small Reddit Dataset (Min 100 & 500)



2 Metrics



AUROC (Area Under Receiver Operating Characteristic Curve)

TPR (True Positive Rate) **at 1% FPR** (False Positive Rate)

Both metrics were used by DNA-GPT, allowing for effective comparison of results.

1 Model, Baseline and Detection Scenario

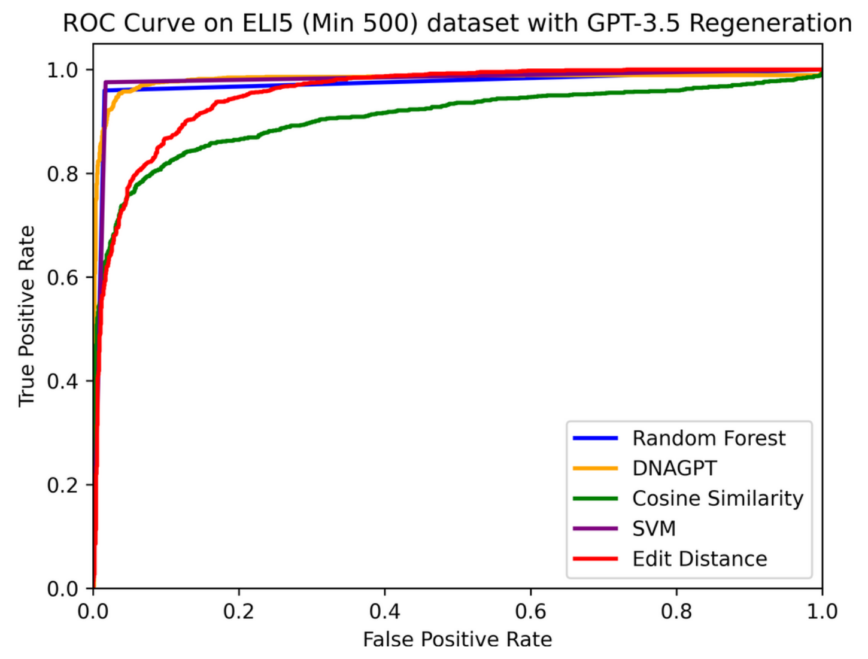
Due to **time constraints**, only one type of each were experimented with.

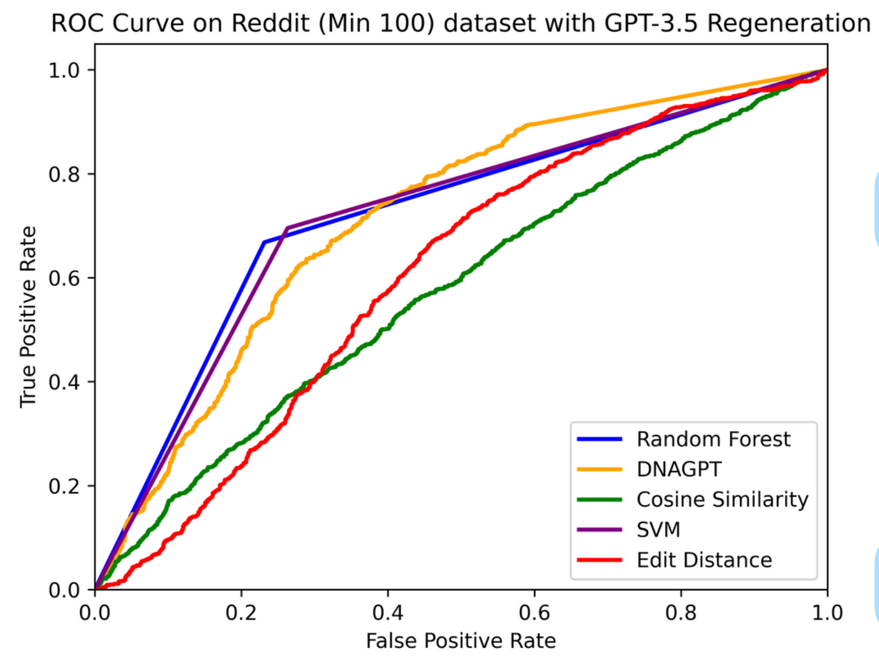
Model: GPT-3.5-Turbo, **Baseline:** DNA-GPT's original method, **Detection Scenario:** Black-Box

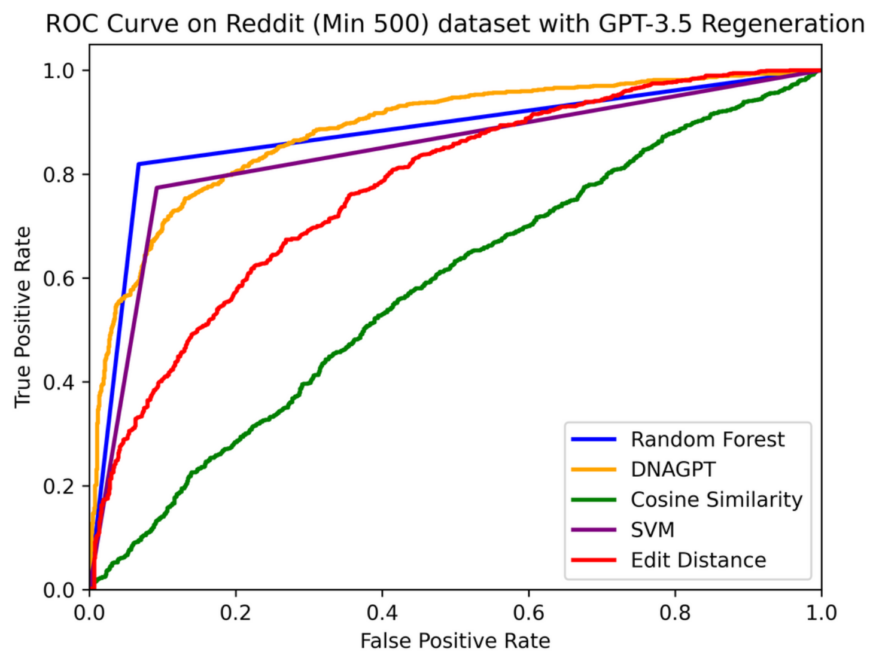
Results & Discussion

Scores for **AUROC** and **TPR** metrics for all **3 datasets** and **5 methods** (including DNA-GPT's original score), with the best-scores (with a margin of 1%) bolded red:

Datasets	ELI5 (Min 500)		Reddit Small (Min 500)		Reddit Small (Min 100)	
Method	**AUROC**	**TPR**	**AUROC**	**TPR**	**AUROC**	**TPR**
DNA-GPT (original)	96.85	63.50	-	-	-	-
DNA-GPT	98.07	59.08	88.32	8.06	71.50	1.36
Random Forest	97.20	61.04	87.60	12.16	71.83	2.89
SVM	97.91	56.78	84.09	8.04	71.62	2.64
Cosine Similarity	90.75	38.41	58.07	1.11	57.58	1.13
Edit Distance	95.45	34.19	77.29	3.96	60.58	0.78







DNA-GPT

- Able to **closely match** their original results
- Original N-Gram Analysis method proved to be **extremely competitive**

Random Forest Classifier and SVM

- Random forest classifier **performs the best**
- SVM has inferior performance, possibly because Random forest classifier relies on **multiple models**
- Both methods require training, **thus with a larger dataset size, improved results can definitely be achieved**

Cosine similarity with Word Embeddings

- Performed **much worse** than other methods
- **Semantic meaning** of regenerated samples likely to match z2 given z1 as context (especially if context is long)

Edit Distance vs N-Gram Analysis

- Lexical analysis of edit distance is **much less effective** than N-Gram analysis
- Edit distance is **unable to differentiate** words with similar spelling could have completely different meanings (eg. "Stationary" vs "Stationery")

Conclusion

- Group A, with a larger training dataset, could **significantly improve** DNA-GPT's original results
- Future Work: More models, detection scenarios and methods can be experimented with

References

[1] Węcel, K., Sawiński, M., Stróżyna, M., Lewoniewski, W., Księżniak, E., Stolarski, P., & Abramowicz, W. (2023). Artificial intelligence—friend or foe in fake news campaigns. The Poznań University of Economics Review, 9(2).

[2] Yang, X. (2023, May 27). DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text.

[3] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346.

[4] The ConvoKit Developers. (n.d.). Reddit Corpus (small) [Dataset]. In ConvoKit.

[5] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305, 2023.

[6] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. arXiv preprint arXiv:2303.13408, 2023

[7] Jwizzed. (2024, January 9). Core algorithms you should know in classification - jWizzed - medium. Medium. https://medium.com/@jwizzed_70966/core-algorithms-you-should-know-in-classification-593bebeae03d

[8] Han, S. (2015). Cosine similarity based fingerprinting algorithm in WLAN indoor positioning against device diversity. https://www.semanticscholar.org/paper/Cosine-similarity-based-fingerprinting-algorithm-in-Han-Zhao/d203f5734f5ee9d49c0adff31805ed93034ca60e

[9] Draelos, V. a. P. B. R., MD PhD. (2020, February 2). Measuring Performance: AUC (AUROC). Glass Box. https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/