Project Proposal

CSC499
Zak White
V00899901
January 28, 2022

Motivation

Within the field of biology, machine learning has been able to outperform humans in certain analytical tasks [1], but the use of black box models has been criticized for its absence of explainable output [2]. This project will explore the caveats of machine learning–based algorithms that don't offer a logical interpretation of their findings.

This analysis will focus on the results of an existing study (White, 2021), an investigation into mutations in ACE2 proteins that affect a host's susceptibility to SARS-CoV-2. The approach of the study was naive, comprising an artificial intelligence–based algorithm that compared protein sequences without factoring in the biological significance of the findings.

Objective

Existing at the intersection of bioinformatics and machine learning, this analysis aims to answer questions from the perspective of each field:

- Bioinformatics: what do the findings of the original project mean, ie, why might the identified residues be influential?
- Machine learning: was the original approach too naive without considering the explainability of the results? How effectively can the findings be used as an indicator of a sequence's susceptibility to the virus?

Method

1. Design a binary classification model based on the findings of the original project, and use it to classify some new sequences as susceptible or immune to the virus.
2. Investigate the model's misclassified sequences (especially those that are known to be susceptible but are classified as immune) to recognize the naivety of the original method.
3. Explore the original findings to better understand why some sites may be influential, and isolate those that are explainably influential.
4. Construct a new classification model based on the reduced findings and compare with the original.

References

[1] Kakadiaris, Vrigkas, Yen, Kuznetsova, Budoff, and Naghavi, "Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA," *Journal of the American Heart Association*, vol. 27, no. 22, 2018. DOI: 10.1161/JAHA.118.009476.

[2] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[3] Z. White, "Identifying Mutations in ACE2 That Influence Susceptibility to SARS-CoV-2," Nov. 2021.