KNOCKOFF STATISTICS-DRIVEN INTERPRETABLE DEEP LEARNING MODELS FOR UNCOVERING POTENTIAL BIOMARKERS FOR COVID-19 SEVERITY PREDICTION

Qian Liu[1], **Daryl Fung**[2], **Huanjing Liu**[1], **CGEn HostSeq Initiative**[3,*], **Pingzhao Hu**[4,†]
[1] Department of Applied Computer Science, University of Winnipeg, Winnipeg, Manitoba, Canada
[2] Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada
[3] Canada's national platform for genome sequencing & analysis
[4] Department of Biochemistry, Western University, London, Ontario, Canada
phu49@uwo.ca

ABSTRACT

COVID-19 affects individuals differently, with some experiencing severe symptoms while others remain asymptomatic. Identifying genetic determinants behind this variability can improve disease management, resource allocation, and public health decisions. Traditional approaches like genome-wide association studies and polygenic risk scores offer limited interpretability and predictive accuracy. In this study, we developed an computational framework that involves deep generative model and xAI to predict COVID-19 severity based on whole-genome data. Our framework identified 72 significant genetic markers and achieved an improved prediction performance (ROC-AUC = 0.64) using whole-genome data from 6752 samples in Canada's CGEn HostSeq project. Among these markers, 50 are novel, linked to hematopoietic stem cell differentiation, lung fibrosis, and SARS-CoV-2 mitochondrial interactions. This study introduces an interpretable AI tool for personalized COVID-19 severity prediction.

1 INTRODUCTION

The COVID-19 pandemic placed immense strain on healthcare systems. In response, the Canadian government allocated $40 million on April 23, 2020, to support the Canadian COVID-19 Genomics Network (CanCOGeN) led by Genome Canada. As part of this initiative, Canada's national genome sequencing platform (CGEn) launched the HostSeq project, sequencing the genomes of over 10,000 individuals affected by COVID-19 across 14 research studies in Canada (Yoo et al., 2023). This resource provides researchers with extensive genomic data to investigate genetic determinants of COVID-19 severity, aiding diagnostics, treatment strategies, and vaccine development. COVID-19 severity varies widely among individuals, necessitating genetic biomarker discovery for predicting critical outcomes such as Intensive care unit (ICU) admission, ventilation use, and vital use (Pun & et al., 2021). However, large-scale genome-wide association studies (GWAS) focusing on the Canadian population remain scarce. To address this gap, systematic GWAS incorporating advanced models and extensive Canadian COVID-19 data are essential. The HostSeq project presents a unique opportunity for such investigations(Yoo et al., 2023).

This study introduces a Knockoff statistics-driven deep learning (DL) framework for predicting COVID-19 severity using whole-genome sequence data from HostSeq. By identifying key genetic biomarkers, this approach enhances our understanding of COVID-19 severity's genetic architecture. Additionally, it provides an interpretable and automated DL tool for precise severity prediction, offering genetic insights to inform public health decisions and individualized patient care.

[*] https://www.cgen.ca/hostseq-contributing-studies-implementation-committee
[†] Corresponding author: phu49@uwo.ca

2 RELATED WORK

Previous GWAS have identified several risk loci associated with COVID-19 severity. For instance, the Severe COVID-19 GWAS Group discovered significant associations at loci 3p21.31 and 9q34.2, implicating genes such as SLC6A20, LZTFL1, and others, and highlighting the influence of the ABO blood group on disease severity (of Severe Covid-19 with Respiratory Failure, 2020). Similarly, Pairo-Castineira et al. identified variants in genes like TYK2, DPP9, and IFNAR2 linked to critical COVID-19 cases (Pairo-Castineira et al., 2020). Leveraging these genetic insights can enhance patient care through risk stratification. For example, Toh et al. developed a polygenic risk score (PRS) using an XGBoost model to predict severe COVID-19 cases, achieving modest performance (Toh & et al., 2020). More recently, Farooqi et al. utilized a larger cohort from the UK Biobank, achieving improved prediction accuracy (Farooqi et al., 2023). However, traditional PRS-based predictions often rely on simple regression models that may overlook complex interactions among genetic variants, limiting predictive performance. DL models offer improved estimation of SNP effect sizes through nonlinear approaches, though their application in genetics has been limited by interpretability challenges. To address this, integrating DL with knockoff inference—a statistical method providing rigorous false discovery rate control—holds promise for identifying interpretable biomarkers (He & et al., 2021).

3 MATERIALS AND METHODS

The overall workflow of this study could be found in Figure 1.

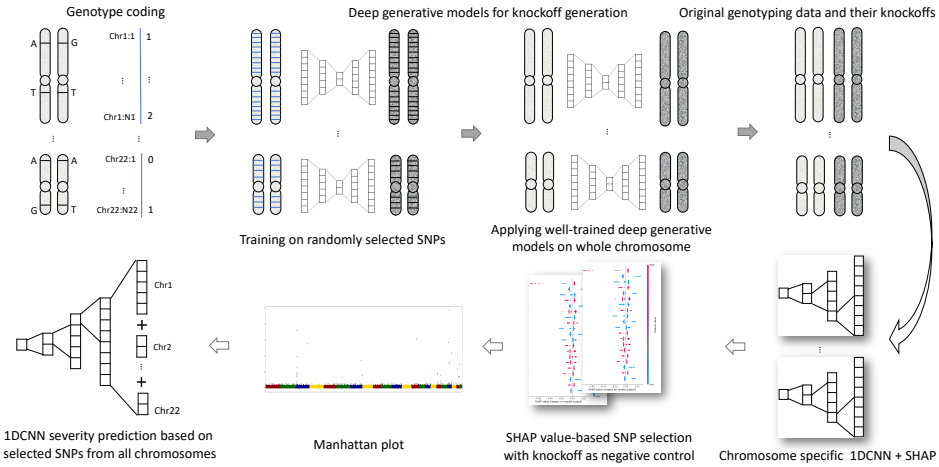

Figure 1: The overall workflow of the study. SNPs were encoded using a numerical scheme (0, 1, 2, 3) to represent missing data (0), no alternative alleles (1), one alternative allele (2), and two alternative alleles (3). One-percent randomly selected SNPs were employed to train a deep generative model, enabling the generation of knockoffs for the entire chromosome. This process was repeated for all 22 chromosomes. The generated knockoff data, alongside the original SNP codes, were then input into chromosome-specific 1DCNN for COVID-19 severity prediction. Subsequently, Shapley values were calculated for each SNP, serving as feature importance scores for knockoff selection. The selected SNPs from each chromosome were combined and utilized as the final predictors in the COVID-19 severity prediction model.

The joint calls of whole genome sequence raw VCF data for 6,752 (4,474 mild cases and 2,278 severe cases) Canadian COVID-19 samples with World Health Organization (WHO) severity classification information were downloaded from the HostSeq project (Yoo et al., 2023). After standard preprocessing and filtering, there were 3,724,619 SNPs left. We coded then code the data as: missing data (0), no alternative alleles (1), one alternative allele (2), and two alternative alleles (3). This study's Research Ethics Board (REB) has been approved by University of Manitoba and Western University.

We first employed a 1-percent random subset of SNPs to train (90%) and test (10%) the deep generative model developed by Romano et al. (Romano et al., 2020). Their model was a moment-matching network that was equipped with a specially designed loss objective which was approved to have the ability to generate robust knockoffs for genetics data. In our application, we tested different hyperparameters until the loss was stabilized and the performance is the best. Then the well-trained model was applied to generate knockoffs for the entire chromosome. This process was repeated for all 22 chromosomes.

The generated knockoffs and the original SNPs for each chromosome were then input into 1DCNNs for feature importance estimation. The primary advantage of a 1DCNN lies in its ability to efficiently capture sequential patterns and dependencies within one-dimensional data. Unlike traditional feedforward neural networks, which treat data as flat input vectors, 1DCNNs are specifically designed to recognize patterns in sequences, making them well-suited for tasks like genomic data analysis. This specialization allows 1DCNNs to effectively extract relevant features from the sequential data, resulting in improved performance and reduced computational complexity compared to more generic neural network architectures. The training/testing split ratio is 80%: 20%. Per node feature importance was estimated using an enhanced version of the DeepLIFT algorithm (Shrikumar et al., 2017), where the conditional expectations of Shapley values were approximated using a selection of background samples (Lundberg & Lee, 2017).

The estimated importance of the original features (Z) and the estimated importance of their knockoffs ($\tilde{Z}$) were then used to calculate the knockoff statistics $W = (Z - \tilde{Z})$ for further FDR estimation. SNPs passed a predefined threshold were identified as significant loci and were kept for the final COVID-19 severity prediction.

SNPs with the knockoff statistics (W) passed the adaptive threshold (T) were selected and input into another 1DCNN model for final COVID-19 severity prediction. Since the SNP size was much smaller after feature selection. The final 1DCNN has six convolutional layers with short kernel lengths (4 and 3), other hyperparameters were set as below: learning rate was 0.0001, batch size was 16, and iteration was 70.

4 RESULTS

The severity prediction performance of the chromosome-specific 1DCNN models is shown in Table 1. After extracting the SHAP values from these well-trained 1DCNNs, the knockoff statistics (W) were calculated and visualized using the Manhattan plot in Figure 2. There were 72 genetic factors that passed the knockoff adaptive threshold (T) and these 72 genetic factors were input into the final 1DCNN for COVID-19 severity prediction. The performance was listed in the last row of Table 1. These 72 genetic factors were mapped to 49 genes. Among them, DPP9 and SLC6A20 have been previously linked to COVID-19 severity, aligning with findings from GWAS studies that implicated loci 3p21.31 and 9q34.2 in disease progression. Gene set enrichment analysis further highlighted pathways such as Hematopoietic Stem Cell Differentiation, Lung fibrosis, and SARS-CoV-2 mitochondrial interactions, which are the most enriched terms in WikiPathways (Martens et al., 2020) database, providing insights into the genetic basis of COVID-19 severity.

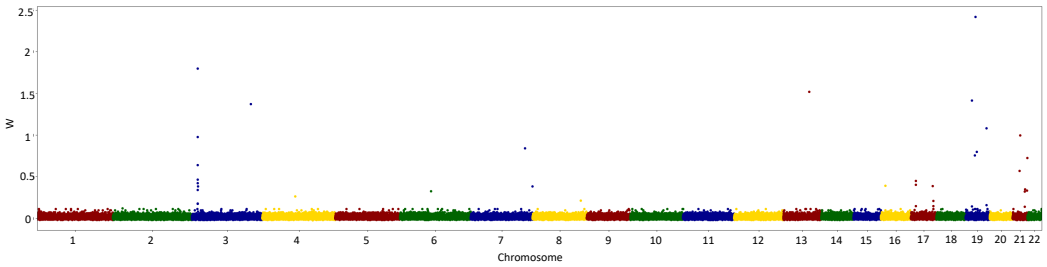

Figure 2: The Manhattan plot of the knockoff statistics. X-axis are the chromosome positions while the y-axis is showing the knockoff statistics.

Table 1: The performance of the chromosome-specific 1DCNNs and the final 1DCNN

Chr[1]	F1	Precision	Recall	ROC-AUC	Chr	F1	Precision	Recall	ROC-AUC
1	0.22	0.14	0.74	0.69	12	0.71	0.63	0.69	0.68
2	0.21	0.13	0.57	0.68	13	0.72	0.66	0.64	0.68
3	0.34	0.48	0.63	0.67	14	0.66	0.27	0.83	0.69
4	0.23	0.13	0.81	0.72	15	0.36	0.50	0.63	0.67
5	0.21	0.45	0.65	0.71	16	0.46	0.46	0.78	0.72
6	0.70	0.64	0.74	0.69	17	0.36	0.51	0.64	0.68
7	0.34	0.48	0.64	0.67	18	0.72	0.64	0.73	0.69
8	0.56	0.34	0.73	0.69	19	0.70	0.67	0.70	0.71
9	0.22	0.14	0.72	0.68	20	0.66	0.54	0.73	0.69
10	0.34	0.46	0.64	0.67	21	0.56	0.50	0.65	0.67
11	0.32	0.47	0.64	0.67	22	0.56	0.46	0.78	0.65
Final	0.42	0.26	0.70	0.64					

5 DISCUSSION

The proposed framework incorporates DL for knockoff generation, optimizing computational costs effectively. The deep knockoff generation model utilized in this research was adopted from (Romano et al., 2020). Its ultimate loss function was intricately structured with three weighted components regulated by corresponding hyperparameters, rendering manual fine-tuning intricate. To streamline our framework and ensure full automation, we set these hyperparameters to a fixed value of 1. This choice led to a scenario where the MMD decreased while the covariance difference increased during training. While the model still produced satisfactory knockoffs, performance could potentially improve with additional tuning efforts. Alternatively, in the future, we could simplify the loss and involve novel generative models such as the probabilistic diffusion model (Ho et al., 2020) for knockoff creation. The proposed framework could also potentially incorporate rare variants with a MAF below 0.05. This addition might enhance the final predictions and contribute additional information to the genetic architecture of the COVID-19 host.

Among the 72 knockoff selected SNPs, 22 had been previously identified in studies by Pairo-Castineira et al.(Pairo-Castineira et al., 2023). These 72 SNPs were subsequently mapped to 49 genes, and within this set of genes, 18 had already been reported in relation to COVID-19 severity (Pairo-Castineira et al., 2023). The enriched WikiPathways by these genes include Hematopoietic Stem Cell Differentiation, Lung fibrosis, and SARS-CoV-2 mitochondrial interactions, suggesting a potential connection between these functional abnormalities and the manifestation of severe COVID-19 symptoms.

The final COVID-19 severity prediction performance was slightly worse than that of the chromosome-specific 1DCNNs, because only 72 important genetic factors were involved in the final 1DCNN. Our ultimate goal is to create an easy-to-use prediction tool that takes a reasonable number of genetic factors as predictors. Consequently, a gene signature panel might be designed in the future for quick measurements of these genetic factors for clinical applications.

6 CONCLUSION AND IMPACT

This study explores deep generative model-based knockoff generation for large-scale genetic risk identification and introduces a more efficient genetic filtering framework for whole genome sequencing. The resulting COVID-19 predictive tool and identified SNPs, genes, and pathways enhance understanding of disease severity and inform management strategies. As the first DL-based prediction tool for COVID-19 host genetics in the Canadian population, it contributes to diagnostics, treatment, drug development, and public health. The CGEn HostSeq project, Canada's largest COVID-19 cohort study, advances research on genetic factors in disease severity, supporting efforts to combat COVID-19 and future health threats.

[1]Chromsome-specific 1DCNN is based on all variants in the chromosome while the final 1DCNN is based only on the knockoff selected variants.

ACKNOWLEDGEMENTS

This research has been conducted using CGEn's HostSeq Databank funded by the Government of Canada through Genome Canada under Project DACO-7. Dr. Qian Liu was funded by CGEn HostSeq/CIHR Joint Postdoctoral Fellowship. This research was supported in part by the Canada Research Chairs Tier II Program (CRC-2021-00482).

REFERENCES

Rashid Farooqi, Jaspal S. Kooner, and Weihua Zhang. Associations between polygenic risk score and covid-19 susceptibility and severity across ethnic groups: Uk biobank analysis. *BMC Medical Genomics*, 16:150, 2023.

Yichen He and et al. Identification of putative causal loci in whole-genome sequencing data via knockoff statistics. *Nature Communications*, 12:1–18, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N Slenter, Kristina Hanspers, Ryan A. Miller, Daniela Digles, Elisson N Lopes, Friederike Ehrhart, Lauren J Dupuis, Laurent A Winckers, SusanL Coort, Egon L Willighagen, Chris T Evelo, Alexander R Pico, and Martina Kutmon. WikiPathways: connecting communities. *Nucleic Acids Research*, 49(D1): D613–D621, 11 2020. ISSN 0305-1048.

Genomewide Association Study of Severe Covid-19 with Respiratory Failure. Genomewide association study of severe covid-19 with respiratory failure. *N Engl J Med*, 383:1522–1534, 2020.

E. Pairo-Castineira, S. Clohisey, L. Klaric, AD. Bretherick, K. Rawlik, D. Pasko, and et al. Genetic mechanisms of critical illness in covid-19. *Nature*, 591:92–98, 2020.

E. Pairo-Castineira, K. Rawlik, A.D. Bretherick, and et al. Gwas and meta-analysis identifies 49 genetic variants underlying critical covid-19. *Nature*, 617:764–768, 2023.

Barbara Tsui-Huan Pun and et al. Prevalence and risk factors for delirium in critically ill patients with covid-19 (covid-d): a multicentre cohort study. *The Lancet Respiratory Medicine*, 9:239–250, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Christine Toh and et al. Evaluation of a genetic risk score for severity of covid-19 using human chromosomal-scale length variation. *Human Genomics*, 14:1–5, 2020.

S. Yoo, E. Garg, L. T. Elliott, R. J. Hung, A. R. Halevy, J. D. Brooks, S. B. Bull, F. Gagnon, C. Greenwood, J. F. Lawless, A. D. Paterson, L. Sun, M. H. Zawati, J. Lerner-Ellis, R. Abraham, I. Birol, G. Bourque, J. M. Garant, C. Gosselin, J. Li, J. Whitney, B. Thiruvahindrapuram, J. A. Herbrick, M. Lorenti, M. S. Reuter, O. O. Adeoye, S. Liu, U. Allen, F. P. Bernier, C. M. Biggs, A. M. Cheung, J. Cowan, M. Herridge, D. M. Maslove, B. P. Modi, V. Mooser, S. K. Morris, M. Ostrowski, R. S. Parekh, G. Pfeffer, O. Suchowersky, J. Taher, J. Upton, R. L. Warren, R. Yeung, N. Aziz, S. E. Turvey, B. M. Knoppers, M. Lathrop, S. Jones, S. W. Scherer, and L. J. Strug. Hostseq: a canadian whole genome sequencing and clinical data resource. *BMC Genom Data*, 24 (1):26, May 2023.