**Title**: Linking Genotype to Phenotype: Further Exploration of Mutations in SARS-CoV-2 Associated with Mild or Severe Outcomes

**Authors**: Roshna Agarwal[1], Tyler Leblond[1], Erin M McAuley[1], Ezekiel J Maier[1], Martin Skarzynski[1], Jameson D Voss[2], Shanmuga Sozhamannan[3,4]

[1]Booz Allen Hamilton, Bethesda, MD 20814, USA
[2]US Air Force Medical Readiness Agency, Falls Church, VA 22042, USA
[3]Joint Program Executive Office for Chemical, Biological, Radiological and Nuclear Defense (JPEO-CBRND), Joint Project Lead for CBRND Enabling Biotechnologies (JPL CBRND EB), Frederick, MD 21702, USA
[4]Logistics Management Institute, Tysons, VA 22102, USA

**Summary**:

We previously interrogated the relationship between SARS-CoV-2 genetic mutations and associated patient outcomes using publicly available data downloaded from GISAID in October 2020 [1]. Using high-level patient data included in some GISAID submissions, we were able to aggregate patient status values and differentiate between severe and mild COVID-19 outcomes. In our previous publication, we utilized a logistic regression model with an L1 penalty (Lasso regularization) and found several statistically significant associations between genetic mutations and COVID-19 severity. In this work, we explore the applicability of our October 2020 findings to a more current phase of the COVID-19 pandemic.

Here we first test our previous models on newer GISAID data downloaded in October 2021 to evaluate the classification ability of each model on expanded datasets. The October 2021 dataset (n=53,787 samples) is approximately 15 times larger than our October 2020 dataset (n=3,637 samples). We show limitations in using a supervised learning approach and a need for expansion of the feature sets based on progression of the COVID-19 pandemic, such as vaccination status. We then re-train on the newer GISAID data and compare the performance of our two logistic regression models. Based on accuracy and Area Under the Curve (AUC) metrics, we find that the AUC of the re-trained October 2021 model is modestly decreased as compared to the October 2020 model. These results are consistent with the increased emergence of multiple mutations, each with a potentially smaller impact on COVID-19 patient outcomes. Bioinformatics scripts used in this study are available at https://github.com/JPEO-CBRND/opendata-variant-analysis. As described in Voss et al. 2021, machine learning scripts are available at https://github.com/Digital-Biobank/covid_variant_severity.

**Introduction**:

The Global Initiative on Sharing Avian Influenza Data (GISAID) is a popular and publicly available repository that houses SARS-CoV-2 sequencing data from the global community. GISAID stores SARS-CoV-2 genomic data and sequencing metadata—including sequencing technology and assembly method—as well as some patient metadata, including high-level patient outcomes, region of origin, age, gender and date of collection.

We have previously used the high-level patient metadata submitted by GISAID users to differentiate between severe and mild patient outcomes. Briefly, we aggregated patient outcomes into “Mild” outcomes (e.g., Outpatient, Asymptomatic, Mild, Home/Isolated/Quarantined, Not Hospitalized) or “Severe” outcomes (e.g., Hospitalized (including severe, moderate, and stable) and Deceased (Death)). We excluded entries whereby the severity of the condition could not readily be discerned, (e.g., retesting, screened for travel, not vaccinated, moderate covid, and live). Following this approach, in Voss et al., 2021, of 3,637 SARS-CoV-2 samples from GISAID with patient outcome metadata, 2,870 were classified as severe patient outcomes and 767 as mild.

Using this patient outcome classification described previously (Voss et al., 2021), we showed logistic regression models that included viral genomic mutations outperformed other models that used only patient age, gender, sample region, and viral clade as features. Among individual mutations, we found 16 SARS-CoV-2 mutations that have ≥ 2-fold odds of being associated with more severe outcomes, and 68 mutations associated with mild outcomes (odds ratio ≤ 0.5). While most assessed SARS-CoV-2 mutations are rare, 85% of genomes had at least one mutation associated with patient outcomes.

**Methods**:

Metadata preprocessing, cohort building, mutation and metadata modeling, data visualization, and statistical analyses were all done according to procedures in Voss et al., 2021 and are diagramed in Figure 1. Briefly, an export of raw GISAID SARS-CoV-2 data was curated using Nexstrain’s ncov-ingest shell scripts [2] and FASTA sequences were parsed from the data export using Python (version 3.8.10). Of the 4,646,285 samples available in GISAID through ncov-ingest on October 26$^{th}$ 2021, we used a subset of 53,787 samples with patient outcome metadata for our analyses. We utilized a total of 29,359 severe and 24,428 mild samples for our analyses. FASTA sequences were aligned to the reference sequence, Wuhan-Hu-1 (NCBI: NC_045512.2; GISAID: EPI_ISL_402125) using Minimap2. Resulting VCF (Variant Call Format) files were merged using bcftools and annotated using SnpEff and filtered using SnpSift. We applied a similar bioinformatics analysis pipeline to Rayko et al [7] and is available under the scripts directory at https://github.com/JPEO-CBRND/opendata-variant-analysis.

The machine learning classification workflow, available at https://github.com/Digital-Biobank/covid_variant_severity, first runs 00_long.py to join and converts annotated VCF files to a single parquet file. Next, 00_red.py is run against GISAID metadata and aggregates patient outcomes into positive (‘Mild’), or negative (‘Severe’) outcomes and stores the classification into a recode dictionary. 01_wide.py, 02_var-freq.py, 02_join.py, and 03_clean.py are consecutively run for pivoting to wide format, deriving mutation frequencies per sample, joining the VCF parquet file with GISAID patient data, and further parsing. 04_logit.py runs training and testing for the logistic regression models. Finally, 05_plot-variants.py is run to derive a table of highest and lowest odds of mutations being classified into mild or severe outcomes.

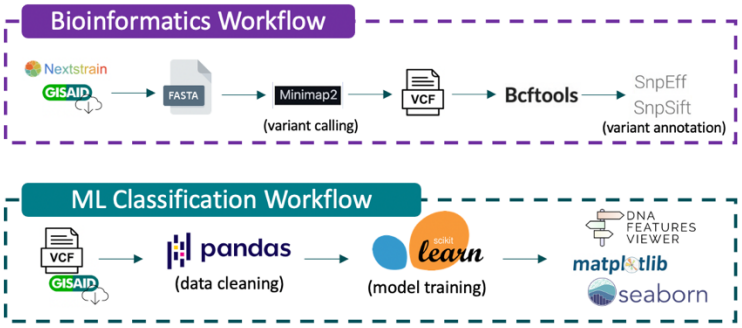



**Figure 1.** Bioinformatics and Machine Learning Classification workflows used for SARS-CoV-2 genetic variant calling and annotation, and mutation and metadata modeling based on procedures used in Voss et al., 2021.

| Analysis Tools | Version |
|---|---|
| Python | 3.8.10 |
| Minimap2 | 2.22 |
| Bcftools | 1.13 |
| SnpEff | 5.0 |
| Pandas | 1.3.3 |
| Matplotlib | 3.4.3 |
| Seaborn | 0.11.2 |
| Scikit-learn | 1.0 |
| NumPy | 1.21.2 |
| Statsmodels | 0.12.2 |
| SciPy | 1.7.1 |
| Joblib | 0.14.1 |

**Table 1.** Versioning for Python utilities and third-party software tools used in shell scripts for machine learning classification and bioinformatics workflows.

Scikit-learn [4] was used to fit logistic regression models with the L1 penalty (Lasso regularization) and the default regularization strength (C = 1) to the patient (rows) and mutation (columns). A train/test split was created on the data (67% train, 33% test). The training data were split into five cross-validation folds using the Scikit-learn stratified K-fold cross-validation generator. Test data was only used for evaluating the performance of trained models. Models were persisted as pickle files using joblib. Scatter and bar plots were created using Pandas [3], Matplotlib [5] and Seaborn [6]. ROC curves were plotted using Scikit-learn [4], and Matplotlib [5]. The versions of all tools used for bioinformatics and machine learning analysis are shown in Table 1.

Similar to previous work (Voss et al., 2021), we trained a total of five logistic regression models using different input features. For each model, Scikit-learn [4] was used to calculate the area under the curve (AUC), a measure of goodness of fit of a binary classification model. AUC confidence intervals, P-values, and diagnostic odds ratios (OR) were calculated using NumPy [9] for each of the five logistic regression models. The Scikit-learn implementation of logistic regression does not provide ORs or P-values for individual variables. ORs and Chi-square test P-values for the association of mutations with Severe and Mild outcomes (Figure 5) were calculated from mutation count data using Statsmodels and SciPy respectively [8]. Mutation frequency was calculated using Pandas [3].

**Results**:

## I. Validation of Previous Methods

Before testing and retraining the logistic regression models on newer data downloaded in October 2021, we first reproduced the previous results using the same dataset (Voss et al., 2021). Reassessing the previous results, the model using age, gender, region, and variants as features had the highest AUC (0.91) and accuracy (91%), followed by models that use fewer features (age/gender/region/clade, age/gender/region, age/gender, and age). The SARS-CoV-2 mutations significantly associated with disease severity identified previously, were also replicated in this reanalysis (Voss et al., 2021).

## II. Evaluation of Model Performance on the Expanded GISAID Dataset

Next, we evaluated the classification performance of the previous logistic regression models (Voss et al., 2021) on the newer, expanded October 26th, 2021 GISAID dataset. For this evaluation, the genetic mutations included in the expanded October 2021 dataset was limited to match the feature space of the trained previous logistic regression models (Voss et al., 2021). Therefore, mutations observed in the October 2021 dataset, but not the October 2020 dataset, were not included in this test dataset. The performance of the trained Voss et al., 2021 logistic regression models was evaluated on the test split (67% train, 33% test) of the expanded October 2021 dataset.

Figures 2 and 3 below show comparisons of ROC curves and model performance statistics for the Voss et al., 2021 logistic regression models on the original October 2020 dataset and the expanded October 2021 dataset. An overall decline in performance is observed for the previous models when applied to the expanded October 2021 dataset. Notably, testing the original models on the October 2021 dataset reveals a decrease in performance for models that include the region feature (AGRV AUC: 0.911 vs. 0.580; AGRC AUC: 0.818 vs. 0.571; AGR AUC: 0.817 vs. 0.564).

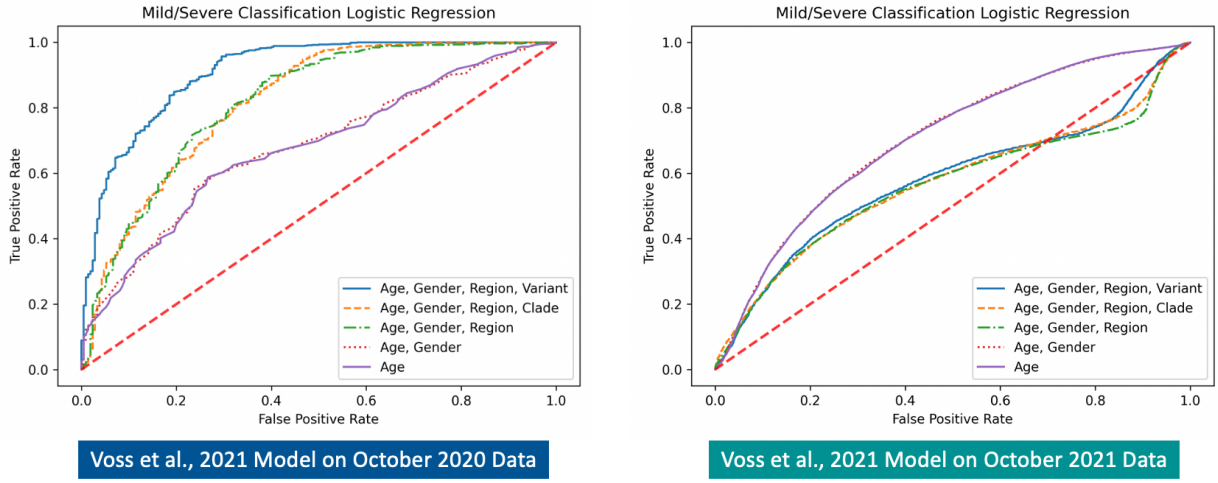



**Figure 2.** Comparison of ROC curves for the previous (Voss et al.,2021) logistic regression models tested on the October 2020 dataset (left) and the expanded October 2021 dataset (right). A decrease in model performance is observed on the expanded October 2021 dataset.

| Model | Accuracy | AUC | Odds Ratio | P-value |
|---|---|---|---|---|
| AGRV | 0.913 | 0.911 | 67.7 | 0.000 |
| AGRC | 0.879 | 0.818 | 28.8 | 0.000 |
| AGR | 0.862 | 0.817 | 17.6 | 0.000 |
| AG | 0.812 | 0.679 | 4.4 | 0.072 |
| A | 0.812 | 0.677 | 4.4 | 0.072 |

Voss et al., 2021 Model on October 2020 Data

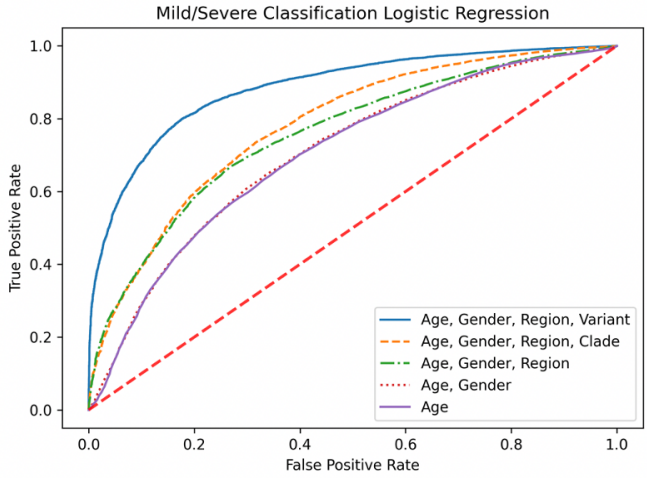
| Model | Accuracy | AUC | Odds Ratio | P-value |
|---|---|---|---|---|
| AGRV | 0.505 | 0.580 | 0.7 | 0.000000 |
| AGRC | 0.494 | 0.571 | 0.5 | 0.000000 |
| AGR | 0.471 | 0.564 | 0.4 | 0.000000 |
| AG | 0.552 | 0.705 | 2.4 | 0.000000 |
| A | 0.550 | 0.703 | 1.8 | 0.000369 |

Voss et al., 2021 Model on October 2021 Data

**Figure 3.** Comparison of previous (Voss et al.,2021) model performance statistics (left) and the previous models run on the expanded October 2021 dataset (right)

## III. Re-Analysis on Expanded GISAID Dataset

To further investigate our previous findings [1], the nested logistic regression models were retrained on the larger October 26th, 2021 GISAID SARS-CoV-2 dataset. Model retraining was done using the train split of the expanded October 2021 dataset, while performance was evaluated using the test split (67% train, 33% test). The performance of the retrained models was then compared to the logistic regression models trained on the October 20th, 2020 GISAID SARS-CoV-2 dataset. ROC curves for the retrained models are shown in Figure 4A (left). The model using age, gender, region, and variants (AGRV) as features continues to show the best performance, as observed previously (Voss et al, 2021). The retrained AGRV model metrics, shown in Figure 4B (right), have an overall decrease in model accuracy (from 0.913 to 0.810) and AUC (from 0.911 to 0.885) as compared to previously published data (Voss et al., 2021) shown in Figure 3A (left). This decrease in retrained AGRV model performance may indicate a modest reduction in power to distinguish between severe and mild outcomes in the expanded October 2021 dataset, or may be explained by inconsistent case severity definitions between the 2020 and 2021 datasets. SARS-CoV-2 mutations most associated with severe and mild outcomes, as measured by odds ratio, from the previous study (Voss et al., 2021) and the expanded October 26th, 2021 GISAID dataset are shown in Figure 5. None of the mutations with the highest (n=20) and lowest (n=20) odds of being associated with severe or mild outcomes from earlier findings (Voss et al., 2021) are also identified in the top 40 mutations of the larger October 2021 dataset analysis. While there is no overlap in the mutations with the highest and lowest mutations between earlier and the expanded October 26th, 2021 GISAID dataset, 10 of the top 20 mutations most associated with severe outcomes from the earlier study (Voss et al., 2021) are also significantly associated with severe outcomes (OR ≥ 2, P-value ≤ 0.05) in the expanded October 2021 dataset. Similarly, 14 of the top 20 mutations most associated with mild outcomes (OR ≤ 0.5, P-value ≤ 0.05) from Voss et al., 2021 are also significantly associated with mild outcomes in the expanded October 2021 dataset.




| Model | Accuracy | AUC | Odds Ratio | P-value |
|---|---|---|---|---|
| AGRV | 0.810 | 0.885 | 17.9 | 0.000 |
| AGRC | 0.715 | 0.779 | 6.0 | 0.000 |
| AGR | 0.695 | 0.757 | 5.0 | 0.000 |
| AG | 0.658 | 0.705 | 3.6 | 0.000 |
| A | 0.656 | 0.703 | 3.5 | 0.000 |

**Figure 4.** ROC curves after retraining on the expanded October 26th, 2021 GISAID dataset (left) and updated model performance metrics (right). Previous (Voss et al., 2021) and retrained model performance is superior when genomic mutations are included as features. A general decrease in model performance was observed for the retrained models on the October 2021 dataset in comparison to the previous models, which used the October 20th, 2020 dataset.

| Mutation | Amino Acid Change | Mutation Type | Variant Frequency | Lower | Odds Ratio | Upper | Odds Ratio P-Value |
|---|---|---|---|---|---|---|---|
| G28321T | T16T | Silent | 0.00443 | 0.0005 | 0.008377 | 0.140377 | 0.000884 |
| G5554T | K1763N | Missense | 0.00443 | 0.0005 | 0.008377 | 0.140377 | 0.000884 |
| C1281T | A339V | Missense | 0.003839 | 0.000574 | 0.009694 | 0.163589 | 0.001302 |
| G25249T | M1229I | Missense | 0.002953 | 0.000738 | 0.012658 | 0.216972 | 0.002579 |
| C22938A | S459Y | Missense | 0.002953 | 0.000738 | 0.012658 | 0.216972 | 0.002579 |
| C25466T | P25L | Missense | 0.002067 | 0.001029 | 0.018162 | 0.320427 | 0.006198 |
| G15652T | D5130Y | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| C15857T | T5198I | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| T13575C | F4437F | Silent | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| C10868T | L3535L | Silent | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| C6445T | D2060D | Silent | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| C27532T | H47Y | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| G23438T | A626S | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| G19891T | D6543Y | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| T772C | V169V | Silent | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| A2869G | V868V | Silent | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| A2550G | D762G | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| C20081T | S6606F | Missense | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| T9880C | N3205N | Silent | 0.001772 | 0.001184 | 0.02122 | 0.380372 | 0.008887 |
| C17518T | L5752F | Missense | 0.001477 | 0.001392 | 0.025501 | 0.467325 | 0.013412 |
| C19524T | L6420L | Silent | 0.054932 | 2.027021 | 3.584261 | 6.337837 | 1.14E-05 |
| T27299C | I33T | Missense | 0.026875 | 1.620729 | 3.724927 | 8.561012 | 0.001953 |
| T29148C | I292T | Missense | 0.027466 | 1.661172 | 3.815497 | 8.763701 | 0.001598 |
| G25979T | G196V | Missense | 0.014471 | 1.237142 | 3.989678 | 12.86637 | 0.020549 |
| C28863T | S197L | Missense | 0.015948 | 1.379027 | 4.431706 | 14.24194 | 0.012436 |
| T9477A | F3071Y | Missense | 0.015948 | 1.379027 | 4.431706 | 14.24194 | 0.012436 |
| C28657T | D128D | Silent | 0.016539 | 1.435939 | 4.608986 | 14.79363 | 0.010227 |
| C6310A | S2015R | Missense | 0.048435 | 2.826946 | 6.055803 | 12.97257 | 3.59E-06 |
| C14724T | F4820F | Silent | 0.011518 | 1.354999 | 9.886295 | 72.13204 | 0.023847 |
| C28854T | S194L | Missense | 0.036621 | 3.424594 | 10.80049 | 34.06262 | 4.9E-05 |
| T16542C | N5426N | Silent | 0.010337 | 1.115817 | 18.21737 | 297.4258 | 0.041663 |
| C2836T | C857C | Silent | 0.022445 | 2.746461 | 19.78809 | 142.572 | 0.003049 |
| C18894T | C6210C | Silent | 0.011223 | 1.21471 | 19.8012 | 322.783 | 0.036034 |
| G29616T | R20I | Missense | 0.011518 | 1.247725 | 20.32994 | 331.2482 | 0.034396 |
| G17427T | V5721V | Silent | 0.011518 | 1.247725 | 20.32994 | 331.2482 | 0.034396 |
| C23185T | F541F | Silent | 0.015062 | 1.645871 | 26.70602 | 433.3336 | 0.020865 |
| C22444T | D294D | Silent | 0.032782 | 4.099582 | 29.41563 | 211.0653 | 0.00077 |
| C12053T | L3930F | Missense | 0.018901 | 2.081322 | 33.67909 | 544.9809 | 0.013285 |
| C26735T | Y71Y | Silent | 0.08417 | 10.05259 | 40.49564 | 163.1318 | 1.93E-07 |
| G25088T | V1176F | Missense | 0.036621 | 4.148285 | 66.77665 | 1074.931 | 0.003042 |

Supplementary Table 1 from Voss et al., 2021

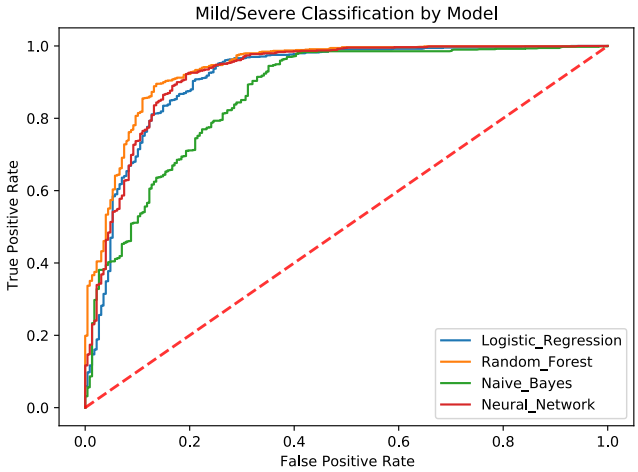
| Mutation | Amino Acid Change | Mutation Type | Variant Frequency | Lower | Odds Ratio | Upper | Odds Ratio P-Value |
|---|---|---|---|---|---|---|---|
| C26885A | N121K | Missense | 0.00303 | 0.000158 | 0.002535 | 0.040707 | 2.44E-05 |
| T14654C | V4797A | Missense | 0.000688 | 0.000689 | 0.011227 | 0.18288 | 0.001615 |
| C12194G | L3977V | Missense | 0.000576 | 0.00082 | 0.013403 | 0.219117 | 0.002487 |
| G5461T | M1732I | Missense | 0.000502 | 0.000938 | 0.015391 | 0.252447 | 0.003451 |
| T4149C | V1295A | Missense | 0.000502 | 0.000938 | 0.015391 | 0.252447 | 0.003451 |
| C10002G | S3246C | Missense | 0.000446 | 0.001052 | 0.017317 | 0.284942 | 0.004531 |
| C8169T | S2635L | Missense | 0.000892 | 0.002438 | 0.01767 | 0.128076 | 6.51E-05 |
| G25699GCCCCCC | P104_F105insPP | | 0.000428 | 0.001097 | 0.018071 | 0.297712 | 0.004992 |
| A5995G | Q1910Q | Silent | 0.00039 | 0.001198 | 0.019794 | 0.327019 | 0.006124 |
| A10333T | T3356T | Silent | 0.000372 | 0.001256 | 0.020784 | 0.343945 | 0.006823 |
| T28245C | L118L | Silent | 0.000335 | 0.00139 | 0.023095 | 0.383647 | 0.008585 |
| A14010G | V4582V | Silent | 0.000335 | 0.00139 | 0.023095 | 0.383647 | 0.008585 |
| C26415T | Y57Y | Silent | 0.000335 | 0.00139 | 0.023095 | 0.383647 | 0.008585 |
| C8802A | T2846N | Missense | 0.000335 | 0.00139 | 0.023095 | 0.383647 | 0.008585 |
| A27477T | T28T | Silent | 0.000651 | 0.003345 | 0.024439 | 0.178541 | 0.000254 |
| C29548T | | | 0.000316 | 0.001469 | 0.024455 | 0.407139 | 0.009704 |
| T14190C | A4642A | Silent | 0.000297 | 0.001557 | 0.025984 | 0.43369 | 0.011032 |
| C29245T | V324V | Silent | 0.000595 | 0.003659 | 0.026807 | 0.196382 | 0.000368 |
| C26509T | | | 0.000576 | 0.003777 | 0.027702 | 0.203147 | 0.000419 |
| G23499C | R646P | Missense | 0.000279 | 0.001656 | 0.027718 | 0.463938 | 0.012625 |
| C11074CT | L3606fs | Frameshift | 0.000688 | 3.784569 | 61.649 | 1004.236 | 0.003794 |
| TC22032T | F157fs | Frameshift | 0.000706 | 3.888891 | 63.31735 | 1030.907 | 0.003568 |
| T21291C | Y7009Y | Silent | 0.001432 | 8.815626 | 63.39692 | 455.9142 | 3.75E-05 |
| G9500A | A3079T | Missense | 0.000725 | 3.993221 | 64.98581 | 1057.581 | 0.003359 |
| G11389T | V3708V | Silent | 0.001469 | 9.051207 | 65.0697 | 467.7903 | 3.34E-05 |
| C6195A | P1977H | Missense | 0.000744 | 4.09756 | 66.65439 | 1084.257 | 0.003167 |
| T15726A | A5154A | Silent | 0.000781 | 4.30626 | 69.99188 | 1137.614 | 0.002824 |
| A9900C | N3212T | Missense | 0.000837 | 4.619372 | 74.99898 | 1217.665 | 0.002397 |
| C14220T | D4652D | Silent | 0.018666 | 42.88242 | 77.70532 | 140.8064 | 1.04E-46 |
| C12733A | A4156A | Silent | 0.000874 | 4.828152 | 78.33761 | 1271.041 | 0.002159 |
| AGGGG28880A | R203fs | Frameshift | 0.000892 | 4.932554 | 80.0071 | 1297.733 | 0.002052 |
| A497G | T78A | Missense | 0.000892 | 4.932554 | 80.0071 | 1297.733 | 0.002052 |
| C2942T | L893L | Silent | 0.002045 | 12.70696 | 91.02711 | 652.0784 | 7.11E-06 |
| A10075G | A3270A | Silent | 0.002231 | 13.88792 | 99.41221 | 711.6101 | 4.65E-06 |
| C18998T | A6245V | Missense | 0.001116 | 6.185969 | 100.0498 | 1618.173 | 0.001182 |
| C23457G | T632S | Missense | 0.001171 | 6.499492 | 105.0631 | 1698.325 | 0.001045 |
| G13903A | D4547N | Missense | 0.002417 | 13.51858 | 217.2938 | 3492.717 | 0.000146 |
| T20391G | R6709R | Silent | 0.002919 | 16.35635 | 262.6667 | 4218.166 | 8.39E-05 |
| G2150A | D629N | Missense | 0.004165 | 23.42096 | 375.6219 | 6024.168 | 2.82E-05 |
| A25336C | E1258D | Missense | 0.007028 | 39.78295 | 637.2302 | 10206.94 | 5.05E-06 |

Supplementary Table 1 on October 2021 dataset

**Figure 5.** The 40 SARS-CoV-2 mutations with the highest (n=20) and lowest (n=20) odds of being associated with severe or mild outcomes from the previous study (Voss et al., 2021) on the left, and this updated analysis using a GISAID dataset from October 2021 (right). Mutations are ordered by odds ratio, and amino acid change, variant frequency, confidence intervals, and P-values are provided.

## IV. Expansion of Analytical Models on Original Dataset

We explored additional machine learning binary classifiers, including Random Forest, Naïve Bayes, and Neural Network algorithms, and compared their performance to the logistic regression model. Similar to our previous study (Voss et al., 2021), 3,386 samples were used for this analysis, with 2,694 associated with severe outcomes and 692 with mild patient outcomes. Age, gender, region, and variants (AGRV) were used as features for each model. A stratified 67% train and 33% test data split was created using Sci-kit learn model selection module [4], and a 5-fold cross-validation was performed to select the best parameters for each model. Random Forest, Naïve Bayes, and Neural Network algorithms were run using Sci-kit learn ensemble, naïve_bayes, and neural_network modules respectively. As shown in Figure 6, the random forest model outperformed the other models, including the logistic regression model, with an AUC of 0.936, accuracy of 0.918 and odds ratio of 116.7.




| Model | AUC | Accuracy | Odds ratio |
|---|---|---|---|
| Random Forest | 0.936 | 0.919 | 113.7 |
| Neural Network | 0.921 | 0.909 | 57.8 |
| Logistic Regression | 0.912 | 0.913 | 64.9 |
| Naïve Bayes | 0.865 | 0.895 | 51.1 |

**Figure 6.** Comparison of alternative machine learning algorithms for binary classification (Random Forest, Neural Network, and Naïve Bayes) to Logistic Regression using October 20th, 2020 dataset and AGRV feature set (age/gender/region/variant). The random forest model had the highest performance and the naïve bayes model had the lowest performance.

## **Discussion**:

Herein we further investigated the results published earlier (Voss et al., 2021) by evaluating the performance of the logistic model for classifying COVID-19 severity on a larger SARS-CoV-2 dataset curated from GISAID on October 26th, 2021. Testing the previous (Voss et al. 2021) and retrained logistic regression models on the expanded October 2021 GISAID dataset revealed a general reduction in model performance. Previous results (Voss et al., 2021) indicated that using age, gender, region, and variant (AGRV) features enabled the best performance for COVID-19 patient outcome classification. The performance of the retrained logistic regression models on the October 2021 GISAID dataset continues to demonstrate that inclusion of genomic mutation features improves classification of COVID-19 patient outcomes. A modest 2.7% reduction in AUC was observed for the AGRV logistic regression models re-trained on the October 2021 dataset in comparison to the AGRV logistic regression model from the previous study.

The performance of the previously (Voss et al. 2021) trained AGR (age/gender/region), AGRC (age/gender/region/clade), and AGRV (age/gender/region/variant) models was severely degraded when tested on the expanded October 2021 dataset. This diminished performance could be the result of new individual mutations and SARS-CoV-2 variants that weren't observed in the October 2020 dataset, and vaccination and treatment advances from October 2020 to

October 2021. Notably, the AGR, AGRC, and AGRV models all included region in their feature set, and the differential availability of vaccination and treatment advances would be expected to impact region features. We further investigated the potential deleterious impact of the region feature on the previous (Voss et al. 2021) models when applied to the expanded October 2021 dataset by testing an AGV (age/gender/variant) model and comparing its performance to the AG and AGRV models. We find that the AGV model has an AUC and accuracy that is similar to the AG model, rather than the AGRV model (AUC: AG=0.705, AGV=0.709, AGRV=0.580; Accuracy: AG=0.552, AGV=0.550, AGRV=0.505).

Figure 5 lists the top twenty mutations with highest and lowest odds ratios for association with severe and mild outcomes from the October 2020 and expanded October 2021 datasets. The top three mutations most associated with severe outcomes in the October 2021 dataset are T20391G (ORF1ab: R6709R), G2150A (ORF1ab: D629N), and A25336C (S: E1258D). E1258D, a missense mutation located at the cytoplasmic tail of the spike protein [10] that is observed with a frequency of 0.7%, has the highest odds (OR=637.23) for association with severe outcomes. Interestingly, an independent study using machine learning approaches to model COVID-19 disease outcomes also identified E1258D as a key predictor of disease severity [11]. This linage independent mutation is recurrent and arises independently in samples taken from donors and cell lines, indicating potential selection in host environments [12]. The three mutations most associated with mild outcomes include C26885A (M: N121K; OR=0.0030), T14654C (ORF1ab: V4797A), and C12194G (ORF1ab: L3977V). N121K, a missense mutation in the membrane protein observed with frequency of 0.3%, is most associated (OR = 0.0025) with mild outcomes. The presence of this mutation was identified as a key predictor of asymptomatic outcomes in previous machine learning modeling of COVID-19 disease outcomes [13].

In a comparison of machine learning algorithms, we found Random Forest to be the best performing algorithm for classification. The superior performance of the random forest model over the logistic regression and naïve bayes models may indicate the presence of nonlinear interactions between features (e.g., SARS-CoV-2 mutations). Indeed, Random Forest is a well utilized machine learning method for genomic data analysis because this algorithm applies well to problems with many more features than observations and accounts for interactions between features [14][15]. In addition to exploring machine learning algorithm options, the space of outcome variables can also be explored. For example, a promising direction for future modeling is the investigation of mutations associated with transmissibility.

The utilization of supervised learning machine learning poses a limitation in our analysis. Since labeled outcomes are required to train these models, the number of samples available for training is reduced by 99% (53,787 of 4,646,285). In addition, machine learning models trained on older samples may not be sufficiently exposed to new mutations. For example, while many of the more than 50 mutations present in Omicron were observed previously in other variants of concern, some Omicron mutations were rare or previously unobserved and many previously observed mutations hadn't co-occurred in the same samples [16]. Supervised machine learning models cannot effectively utilize previously unobserved mutations and mutations combinations because parameters have not been fit for these features. A promising approach for addressing these limitations is semi-supervised learning. This machine learning approach uses both labeled data and unlabeled data for model training. Semi-supervised learning may outperform supervised learning approaches when the amount of unlabeled data is much larger than labeled data [17]. Within the field of genomics, recent example uses of semi-supervised learning include

microRNA classification [18], somatic genomic variant classification [19], and identify disease associated genes [20].

**Funding:** The program is partially funded by the DOD's Joint Program Executive Office for Chemical, Biological, Radiological and Nuclear Defense (JPEO-CBRND), in collaboration with the Defense Health Agency (DHA).

**Disclaimer:** *The views and opinions in this article are those of the author and do not necessarily reflect that of the JPEO-CBRND, the U.S. Army, the Department of Defense, or the U.S. Government. Clearance DSS 22055.*

**References**:

1. Voss, Jameson D., et al. "Variants in SARS-CoV-2 associated with mild or severe outcome." *Evolution, medicine, and public health* 9.1 (2021): 267-275.
2. Hadfield, James et al. "Nextstrain: Real-Time Tracking Of Pathogen Evolution". *Bioinformatics*, vol 34, no. 23, 2018, pp. 4121-4123. *Oxford University Press (OUP)*, https://doi.org/10.1093/bioinformatics/bty407. Accessed 21 Dec 2021.
3. McKinney, Wes. "Data structures for statistical computing in python." *Proceedings of the 9th Python in Science Conference*. Vol. 445. No. 1. 2010.
4. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.
5. Hunter, John D. "Matplotlib: A 2D graphics environment." *Computing in science & engineering* 9.03 (2007): 90-95.
6. Waskom, Michael, et al. "Mwaskom/Seaborn: V0. 8.1 (September 2017)." *Zenodo* (2017).
7. Rayko, Mikhail, and Aleksey Komissarov. "Quality Control Of Low-Frequency Variants In SARS-Cov-2 Genomes". 2020. Cold Spring Harbor Laboratory, https://doi.org/10.1101/2020.04.26.062422. Accessed 11 Jan 2022.
8. Virtanen, Pauli, et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python." Nature methods 17.3 (2020): 261-272.
9. Harris, Charles R., et al. "Array programming with NumPy." *Nature* 585.7825 (2020): 357-362.
10. Yang, Xiang-Jiao. "δ subvariants of SARS-COV-2 in Israel, Qatar and Bahrain: Optimal vaccination as an effective strategy to block viral evolution and control the pandemic." *medRxiv* (2021).
11. Sokhansanj, Bahrad A., Zhengqiao Zhao, and Gail L. Rosen. "Interpretable and Predictive Deep Modeling of the SARS-CoV-2 Spike Protein Sequence." *medRxiv* (2021).
12. Rocheleau, Lynda, et al. "Identification of a High-Frequency Intrahost SARS-CoV-2 Spike Variant with Enhanced Cytopathic and Fusogenic Effects." *Mbio* 12.3 (2021): e00788-21.
13. Nagpal, Sunil, et al. "(Machine) Learning the mutation signatures of SARS-CoV-2: a primer for predictive prognosis." *bioRxiv* (2021).
14. Chen, Xi, and Hemant Ishwaran. "Random forests for genomic data analysis." *Genomics* 99.6 (2012): 323-329.
15. Qi, Yanjun. "Random forest for bioinformatics." *Ensemble machine learning*. Springer, Boston, MA, 2012. 307-323.
16. Mallapaty, Smriti. "Where did Omicron come from? Three key theories." *Nature* 602.7895 (2022): 26-28.
17. Libbrecht, Maxwell W., and William Stafford Noble. "Machine learning applications in genetics and genomics." *Nature Reviews Genetics* 16.6 (2015): 321-332.
18. Sheikh Hassani, Mohsen, and James R. Green. "A semi-supervised machine learning framework for microRNA classification." *Human genomics* 13.1 (2019): 1-12.
19. Nicora, Giovanna, et al. "A semi-supervised learning approach for pan-cancer somatic genomic variant classification." *Conference on Artificial Intelligence in Medicine in Europe*. Springer, Cham, 2019.
20. Vitsios, Dimitrios, and Slavé Petrovski. "Stochastic semi-supervised learning to prioritise genes from high-throughput genomic screens." *bioRxiv* (2019): 655449.