Transactions on Networks and Communications – Vol. 9, No. 2
Publication Date: April, 1, 2021
DOI: 10.14738/tnc.92.9760.
Onoja, A., Ejiwale, M. O., & Rewane, A. (2021). Interpretable machine learning approach for predicting COVID-19 risk status of an individual, 9(2), 1-14

Interpretable Machine Learning Approach for Predicting COVID-19 Risk Status of an Individual

[1]A.A. Onoja, [2]M.O. Ejiwale, [3]Z. Yu, [4]Amal Mitra, [5]A. Rewane, [6]B. Amao, [7]G. Bertarelli, [8]C. Giusti, [9]M. Nanni

[1]Department of Data Science, Scuola Normale Superiore, Pisa, Italy; [2,3]Computer Science Department, College of Engineering and Applied Science, University of Wisconsin-Milwaukee, United States; [4]Department of Epidemiology and Biostatistics, School of Public Health, Jackson State University, United States; [5]Rush University Medical Center, Chicago, United States, [6]School of Social Work, Jackson State University, United States; [7,8]Department of Economics & Management, University of Pisa, Italy; [9]KDD Laboratory, ISTI-National Research Council of Italy;

ABSTRACT

This study aimed to identify the best features that predict COVID-19 infection risk status ("Low", "Medium", "High") of an individual, using statistical feature selection methods and interpretable Machine learning models. The research analyzed a publicly available dataset obtained from a web-based COVID-19 risk assessment calculator. Fifty-seven (57) of the 59 features in the dataset were first filtered for multicollinearity using the Pearson Correlation Coefficient, and are further shrunk to 55 features with LASSO GLM. A class imbalance was identified in the dataset, same was addressed with SMOTE resampling technique. Multiple interpretable ML algorithms were explored during the classification phase. The best classifier predictions were saved as a new instance and were perturbed with a single Decision tree classifier. To further build trust and explainability of the best model, the XGBoost classifier was utilized as a global surrogate model to train predictions of the best model. The XGBoost individual's explanation was done using the SHAP explainable AI framework. Random Forest classifier with a validation accuracy score of 96.35 % from 55 features reduced by feature selection emerged as the best classifier model. The decision tree classifier approximated the best classifier correctly with a prediction accuracy score of 92.23 % and Matthew's correlation coefficient of 0.8960. The XGBoost classifier approximated the best classifier model with a prediction score of 99.7 %. This study identified COVID-19 positive, COVID-19 contacts, COVID-19 symptoms, health workers, and public transport count as the five major predictors of COVID-19 risk exposure status of an individual.

Keywords: Imbalanced Class Distribution, Feature Selection, Risk Status, COVID-19, Machine Learning, Model Explainability

INTRODUCTION

Interpretable machine learning has been widely utilized as a methodology to investigate many real-life problems, including those in the healthcare domain [14]. Its acceptability has been linked to a thorough understanding of the inner working metrics of the models in each given problem domain. Since the outbreak of the COVID-19 pandemic, humanity is faced with numerous new challenges. A thorough understanding of how the virus evolves and spreads

(human-to-human transmission) is paramount [14]. Many risk assessment tools powered by Artificial intelligence (AI) and Machine learning (ML) models have been developed to monitor, assess, and predict the risk of contracting the COVID-19 viral infection [29]. Most of the risk assessment predictive models/tools have been very useful in fighting infection transmissibility. However, model complexity plagues some of such tools as a result of noisy variables, imbalanced class distribution, lack of trust, and interpretability of the models [2]. Irrelevant features have little, or no importance and they hamper model performance. Among other impacts of redundant features are increased data frame dimensionality, computational burden, and high false-positive rates which resultantly affect model accuracy and precision. Lack of model explainability/interpretability is linked to users' lack of confidence in the accuracy of AI-based predictive models for COVID-19 risk infection status [16].

The aim of Explainable AI/Machine Learning (XAI/IML) is to understand how black-box models work for different real-world problems. This is vital to AI/ML model utilization in mission-critical applications such as COVID-19 risk assessment tools, not only due to the increased regulatory scrutiny but also to establish trust among the end-users. Roscher et al., [24] identified three core elements of explainability: transparency, interpretability, and explainability. XAI enhances users and parts of the internal system to be more transparent, providing explanations of their decisions in some level of detail [11] [22].

These explanations are important to ensure algorithmic fairness, identify potential bias/problems in the training data, and ensure that the algorithms perform as expected. The main goal of this study is to adopt a combination of statistical feature selection and interpretable ML models to predict, and more importantly, to precisely identify and explain features that contribute more to the risk of contracting COVID-19. Using post-hoc machine learning interpretable analysis methods such as test accuracy, perturbation, surrogate models, and SHAP, this study provides an integrated view of features that best predicts an individual's risk status in COVID-19. The rest of this study is structured as follows: Section 2, existing risk assessment tools; Section 3, materials and methods; Section 4, results and discussion; and in Section 5, the study concluded and made recommendations.

Application of Artificial Intelligence to Curbing COVID-19 Pandemic

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first reported in Wuhan China, is a very contagious virus that causes the COVID-19 infection [5]. Shereen *et al*., [26] state that Genomic analysis has revealed that SARS-CoV-2 is phylogenetically related to severe acute respiratory syndrome-like (SARS-like) bat viruses. It is transmittable from humans to humans through close contact. Tian et al, [29] estimated the ratio of COVID-19 infection in the severe cases to mild cases, asymptomatic and non-pneumonia cases in Beijing. Many AI surveillance tools have been deployed to help mitigate infection [30].

COVID-19 Risk Factors

A study conducted on COVID-19 shows that patients with cardiac diseases, hypertension, or diabetes, who are treated with angiotensin-converting enzyme 2 (ACE2)-increasing drugs, are at higher risk for severe COVID-19 infection. It indicates such patients should, therefore, be monitored for ACE2-modulating medications, such as ACE inhibitors or angiotensin II

URL: https://doi.org/10.14738/tnc.92.9760.

type-I receptor blockers (ARBs) [2]. Currently, there is no established evidence of mother-to-child transmission of the virus. However, a significant proportion of COVID-19 infection cases were related to occupational exposure. Occupational group risk that was first observed was that of persons working in seafood and wet animal wholesale markets in Wuhan [16]. Others include health care workers (HCWs), staff in the tourism, retail, and hospitality industry, transport and security workers, and construction workers. The author further stressed that those who carry this occupational risk are liable to face stigmatization from their families and the public [25].

Another recent study by Sun et al., [28] explored the multivariable statistical model to characterize risk factors in 34,503cases of laboratory-confirmed positive or negative COVID-19 infection. The results of their analysis showed that a higher risk of COVID-19 infection was associated with older age, Asian race, Black/African American race, Latino ethnicity, non-English language, residing in a neighborhood with financial insecurity, low air quality, housing insecurity, transportation insecurity and living in senior living communities. These findings imply that COVID-19 infection is higher among the groups of people who are already affected by health disparities across age, race, ethnicity, language, income, and living conditions. Blood group has been linked to the occurrence of COVID-19. A related observational study [37] involving 1559 individuals who tested for SARS-CoV-2 revealed that a higher proportion (682) of blood group A, and a lower proportion of blood group O (*size not mentioned*) were among SARS-CoV-2 positive patients compared to the negative cases. In both cases, however, the result is significant only in Rh-positive blood types.

EXISTING RISK ASSESSMENT TOOLS

According to Hao *et al.,* [13] the risk assessment application systems are web or mobile-based assessment tools used in identifying patients at risk of certain diseases, exposure, or readmission. The risk assessment tools may be fundamental to design interventions and provide targeted care. The risk assessment tools have been used for validation in an integrated healthcare delivery system to identify the population at risk of disease under study [19]. A useful computational tool was developed to assess the risks of novel coronavirus outbreaks outside of China [3]. They estimate the dependence of the risk of a major outbreak in a country from imported cases. “Symptoma”, an online application, was developed to help curb the spread of COVID-19 [20]. This is a symptom-to-disease digital health assistant that can differentiate more than 20,000 diseases with an accuracy of more than 90 %. The authors tested the accuracy of Symptoma to identify COVID-19 using a set of diverse clinical cases combined with case reports of COVID-19. They further demonstrated that Symptoma can accurately distinguish COVID-19 in 96.32 % of clinical cases. When considering only COVID-19 symptoms and risk factors, Symptoma identified 100 % of those infected when presented with only three signs (shortness of breath, fever, and dry cough). Their findings demonstrated that AI-web-based application system accuracy in predicting COVID-19 far exceeds that of simple “yes-no” questionnaires that are commonly available online [20]. Data obtained from publicly available sources were geo-coded to generate detailed, real-time, and robust data for emerging disease outbreaks stressing the importance of AI to support and inform public health decision making [33] [34]. The COVID-19 survival risk calculator powered by AI was developed by Grantham, Jonathon [25] to unpack the factors that predicted the outcome of COVID-19 among individuals. Their model is being

continuously updated to provide the most accurate predictions possible and make available raw datasets to researchers for download. This study, therefore, utilized the survival risk calculator dataset and interpretable ML models to predict an individual's risk assessment status to COVID-19.

MATERIALS AND METHODS

The target populace consists of individuals from diverse races, gender, countries, and ages 18 –100 years, who were able to access and evaluate their COVID-19 risk via the COVID-19 survival calculator between March 25, 2020, and August 17, 2020. The COVID-19 survival risk calculator dataset was collected by the Nexoid software company, located in London, United Kingdom. More than *1 million* individuals have utilized the COVID-19 survival calculator to obtain an insight into their risk status. The dataset is made of two distinct parts: the likelihood of infection and the mortality rate once someone is infected. These contain key information about the individual filling the survival calculator such as (race, gender, heart disease, kidney disease, liver disease, lungs diseases, public transport usage, nursing home, COVID-19 contact count, rate reducing mask, rate reducing house, opinion infection, risk infection, etc.).

The classification label; "Risk Infection", contains five groups, based on a 60 % baseline risk infection rate by the National Health Services (NHS). The categories (0 – 4) illustrate individuals' risk susceptibility to the virus. Assumptions of Risk Infection include 1.) there will be no vaccine and like many other viruses infected people will create antibodies that will dramatically reduce the severity of reinfection; 2.) social distancing by itself is not enough to stop the virus; 3.) the current estimate for herd immunity from the NHS (UK) is 60 %. The risk calculation base infection rate of 60 % (adjusted by geographic area, the number of people user lives with, and the number of people in user's social contact.

$$\text{Infection} = {}^{\text{deaths}}/_{(\text{population} \times 0.0045)} \quad (1)$$

where 0.45 % is the mortality rate estimate in a country.

The dataset used for this study contains 878600 rows and 60 columns (59 predictors, one target variable). The target label (risk infection) originally measured in percentage was binarized to suit our classification task. We combined the categories (Highest + High = "High", Medium = "Medium", and "Lowest + Low" = "Low"). In other words, we reduced the original categories from 5 five to three ("Low", "Medium", "High").

URL: https://doi.org/10.14738/tnc.92.9760.

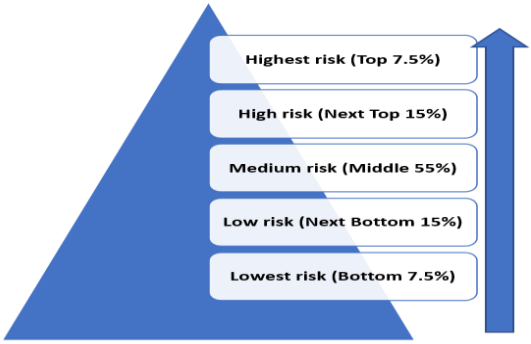

Figure I: Risk Infection pyramid displaying risk status categories of individuals to COVID-19 infection (https://www.covid19survivalcalculator.com/en/research).

While the class distribution of the target variable (Risk infection status) is available in *Figure 2*.

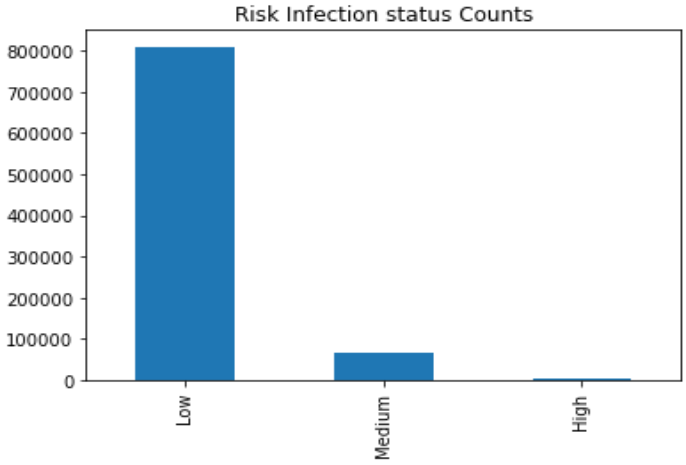

Figure II: Distribution of risk infection status (target variable) showing imbalanced classification problem among the class instances.

The class distribution was: Low risk 92.28 %, Medium risk 7.42 %, and High 0.31 % of the target variable in the dataset.

Analysis and Machine Learning Pipeline

The clean dataset was split into 80-20 % as training and validation set respectively. Five-fold cross-validation with SMOTE resampling approach was applied to handle imbalanced classification problems. Pearson Correlation Coefficient analysis was carried out to filter features with high multicollinearity *(|corr.| < 0.80)*, same was visualized (correlation matrix). We used the LASSO GLM shrinkage method as a wrapper feature selection algorithm. Logistic regression, decision tree, and random forest were the supervised machine learning classification algorithms explored. SHAP was utilized for the interpretability of predictions from the global surrogate approximation of the best performing model *(*see ***Figure VI)***. To assess the performance of the classifiers and any unforeseen biases incurred, Matthew's Correlation Coefficient (MCC), and Balanced Accuracy Score (averages the accuracy score from the multi-class instances) were utilized in this study.

Machine Learning Post-hoc Interpretations of the Best Model Classifier

This research ensured achieving post-hoc machine learning (ML) interpretations (probing, perturbation, and surrogate global model to explain the best ML classifier. see ***Figure III***).

Three commonly used interpretable supervised machine learning algorithms (Logistic Regression, Decision Tree, and Random Forest) were used for the classification task.

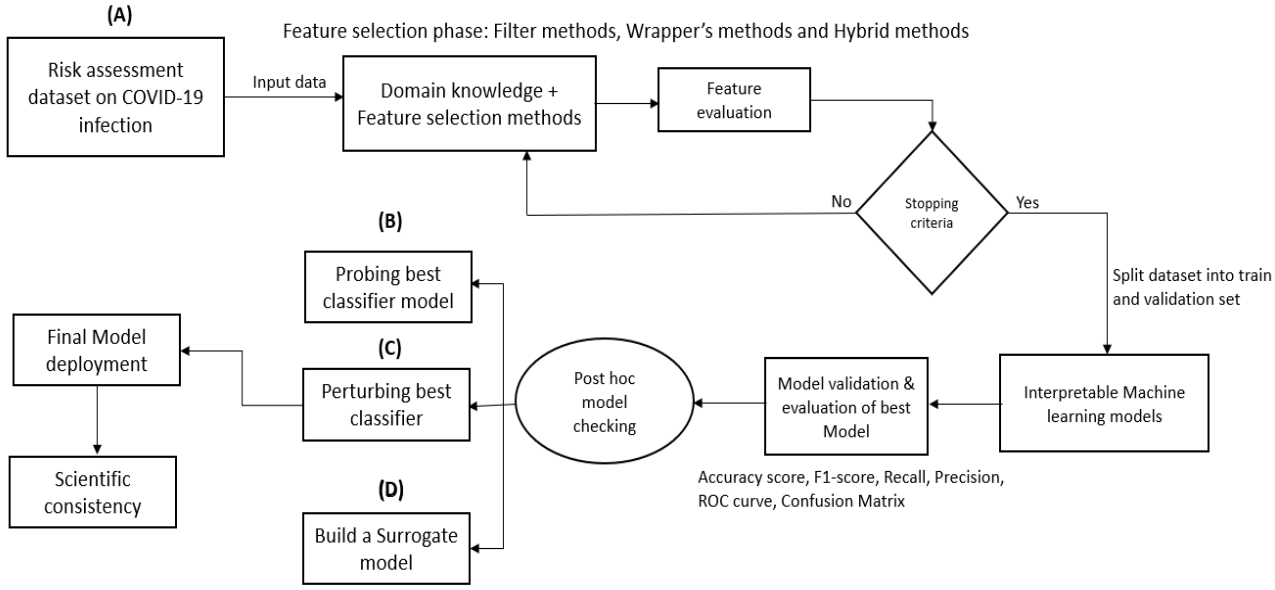

Figure 3.: Interpretable Machine Learning feature selection strategies for the best model classifier

Keynotes:

- **A.** Understanding the problem formulation and modeling process (i.e., clean, apply feature selection, and feed best features to ML algorithms)
- **B.** Probing strategies to inspect the structure and parameters learned by a trained ML model.
- **C.** Perturbing strategies for comparing ML classification results (model performance e.g., test accuracy, MCC, Recall & precision) with 59 features against 55 features selected via filtering (correlation) and LASSO shrinkage. Using the interpretable model (e.g., XGBoost, or decision tree).

RESULTS AND DISCUSSION

This study compared ML classification results with and without feature selection approaches (Correlation and LASSO). The best interpretable ML classifier predictions were saved as new instances, the same perturbed with a single decision tree as a local surrogate. The saved prediction instances were further approximated (trained) with XGBoost as a global surrogate model. Individual interpretations of predictions (risk infection status) from the XGBoost classifier were interpreted using the SHAP interpretable framework.

Feature Selection

The Pearson correlation detected features with a correlation coefficient greater than absolute 0.80. There were 59 features initially. Multicollinearity screening eliminated two of them, leaving us with 57 to analyze *(The correlation heatmap is available in **Figure IV)**.*

URL: https://doi.org/10.14738/tnc.92.9760.

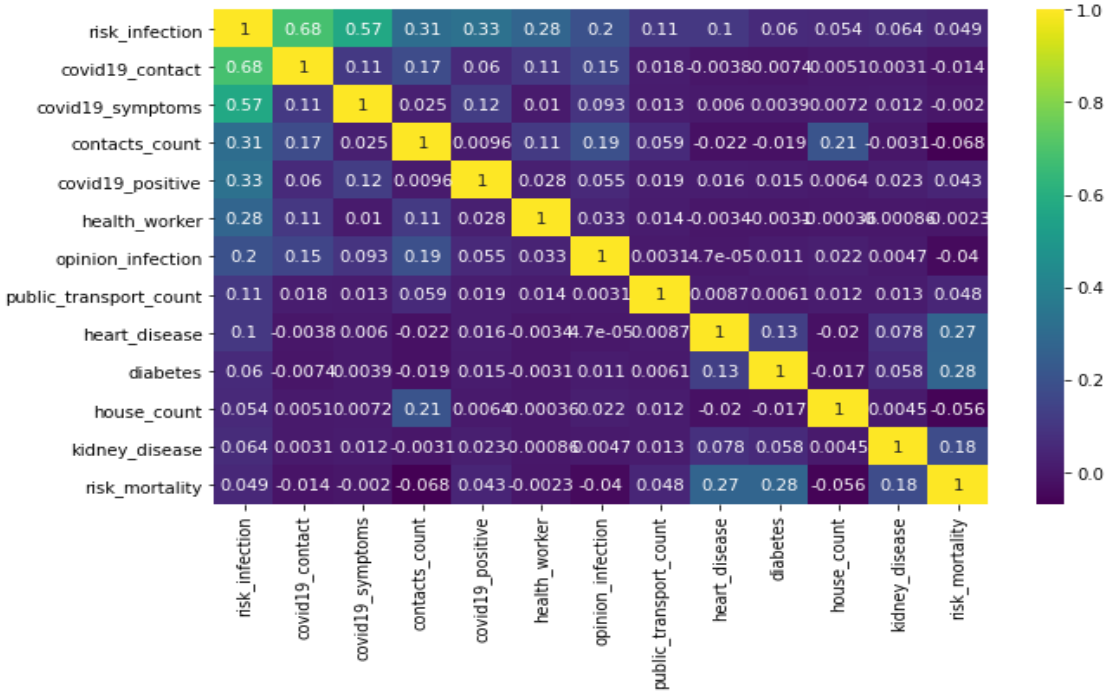

Figure IV: Correlation heatmap after filtering for multicollinearity among features and the target variable.

The LASSO shrinkage method reduced the features from 57 to 55. COVID-19 contact, health workers, COVID-19 symptoms, heart disease, and opinion infection are the five top-notch features that explained individuals' risk status to COVID-19. Table I displays all the top features (predictors) selected via the LASSO GLM shrinkage method.

Table I: Top features selected from the LASSO shrinkage method

S/No.	Features	Importance
1	COVID-19 contact	0.16898
2	Health worker	0.11093
3	COVID-19 symptoms	0.07130
4	Heart disease	0.05853
5	Opinion infection	0.04775
6	Public transport count	0.04452
7	Kidney disease	0.02416
8	Lung disease	0.01874
9	Rate reducing mask	0.01639
10	Race	0.01294
11	Nursing home	0.01238
12	MDMA	0.01219
13	Diabetes	0.01117
14	Rate reducing risk single social distancing	0.01113
15	Contact count	0.01058

Classification Metrics and Results

Table II depicts the classification metrics and results of our interpretable ML algorithms without feature selection methods, while **Table III** shows metrics and results of ML with the use of feature selection methods (Pearson correlation and LASSO).

Table II: Classification Results: Without Feature selection (59 features without feature selection methods)

59 Features & classifiers	Precision (%)	Recall (%)	F1-score (%)	MCC (%)	Balanced train accuracy score (%)	Balanced test accuracy score (%)
Logistic Reg					99.26	95.91
Low	99	95	97	93.68		
Medium	99	100	100			
High	95	93	94			
macro avg.	98	96	97			
weighted avg.	99	99	99			
Decision Tree	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**MCC (%)**	**Balanced train accuracy score (%)**	**Balanced test accuracy score (%)**
Low	99	94	97	80.76	94.33	86.56
Medium	97	100	99			
High	100	66	79			
macro avg.	99	87	92			
weighted avg.	98	97	97			
Random Forest	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**MCC (%)**	**Balanced train accuracy score (%)**	**Balanced test accuracy score (%)**
Low	99	94	97	97.10	99.00	96.1
Medium	100	100	100			
High	100	95	97			
macro avg.	100	97	98			
weighted avg.	100	100	100			

URL: https://doi.org/10.14738/tnc.92.9760.

Table III: Classification Results: 55 features from Pearson Correlation + LASSO shrinkage

55 Features & classifiers	Precision (%)	Recall (%)	F1-score (%)	MCC (%)	Balanced train accuracy score (%)	Balanced test accuracy score (%)
Logistic Reg					99.23	96.16
Low	94	96	95	93.43		
Medium	100	99	99			
High	93	95	94			
macro avg.	97	96	96			
weighted avg.	99	99	99			
Decision Tree	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**MCC (%)**	**Balanced train accuracy score (%)**	**Balanced test accuracy score (%)**
Low	99	94	97	81.20	94.03	87.05
Medium	97	100	99			
High	100	67	80			
macro avg.	99	87	92			
weighted avg.	98	98	97			
Random Forest	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**MCC (%)**	**Balanced train accuracy score (%)**	**Balanced test accuracy score (%)**
Low	99	94	97	97.00	100.00	96.35
Medium	100	100	100			
High	100	95	97			
macro avg.	100	97	98			
weighted avg.	100	100	100			

Matthew's correlation coefficient was employed to measure the quality of the binary classifications (severe versus non-severe) in the dataset. It considers true and false positive. The MCC is in essence a correlation coefficient value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. The Receiver Operating Curve (ROC) and balanced accuracy scores for 55 features were selected via Pearson correlation. LASSO was used as a scoring criterion to select the best ML performing model **(see *Table II & III*)**. Random forest emerged as the best model with a test balanced accuracy score of **96.35 %** ahead of the Logistic regression

classifier with an accuracy score of **96.16 %**. Random forest classifiers in both original features (59) and reduced form (55 features) outperformed logistic regression and decision tree classifiers **(*Figure V*)**.

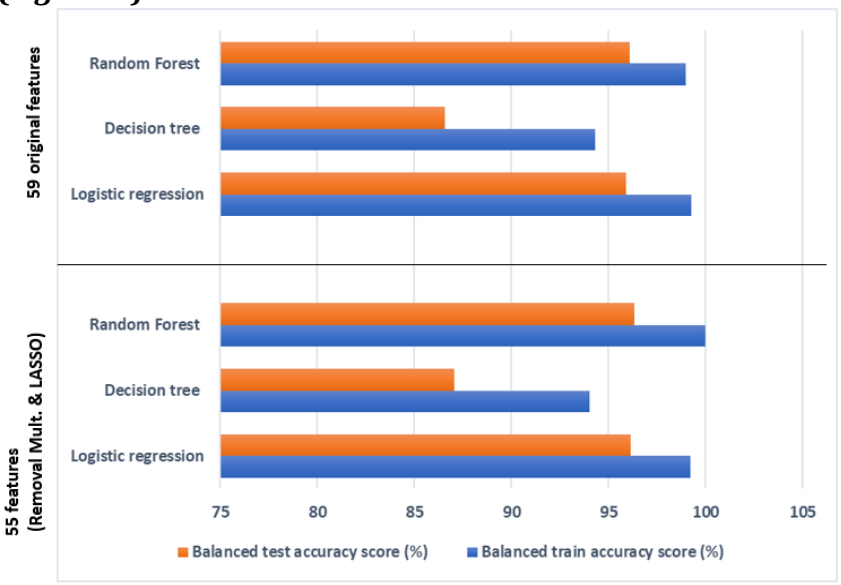

Figure V: Model comparison of 59 features and 55 features selected from Pearson correlation and LASSO.

Therefore, this study adopted the random forest classifier (with 55 features from feature selection) as the best model to predict individuals' risk assessment status to COVID-19. The balanced accuracy score for the best model classifier is 96.35 %.

Local surrogate approximation of the best classifier model with decision tree

A single decision tree classifier was used as a local surrogate model to approximate prediction from the random forest model classifier (from feature selection methods: Pearson correlation and LASSO shrinkage method). The decision tree classifier approximated the random forest classifier correctly with a balanced accuracy prediction score (92.23 %) and Matthew's correlation coefficient of 0.8960. This implies that the decision tree classifier was able to locally approximate the best model (Random Forest) with high fidelity.

Global surrogate approximation of the best classifier model with XGBoost

Predictions from the best classifier model (random forest with 55 features) were saved and trained with the XGBoost classifier model as a global surrogate approximator of the Random forest classifier, ***Table IV*** showed the most relevant features identified by the XGBoost classifier.

URL: https://doi.org/10.14738/tnc.92.9760.

Table III: Feature importance from the global surrogate approx. of the winning model

S/No.	Features Importance: XGBoost	Importance
1	COVID-19 positive	0.498684
2	COVID-19 contact	0.239514
3	COVID-19 symptoms	0.089652
4	Health worker	0.089334
5	Public transport count	0.018367
6	Heart disease	0.01038
7	Rate reducing risk single	0.009176
8	House count	0.005374
9	Amphetamines	0.004739
10	Working	0.003747

The SHAP AI-framework explainer was used to further explain the prediction by the XGBoost model classifier. This was necessary to build the trust and confidence of individuals to understand the inner working metrics of the best model classifier concerning their risk assessment status. More so, using SHAP interpretation further showed the positive and negative relationships of the predictors with the target variable. SHAP feature importance as displayed in **(*Figure VI*)**.

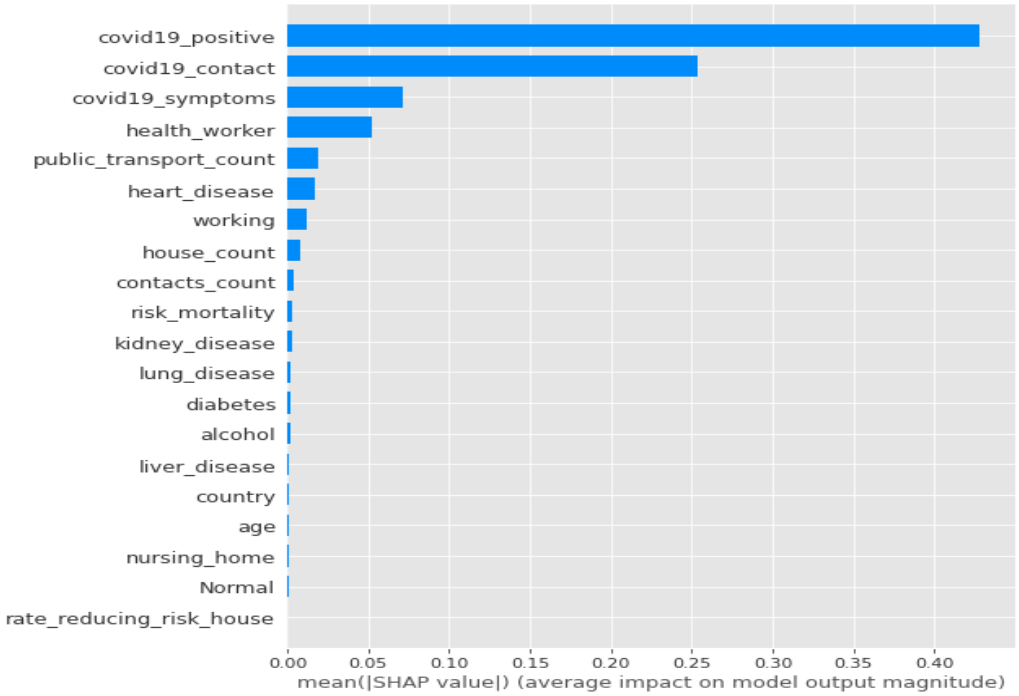

Figure VI: SHAP Feature Importance

The XGBoost classifier was trained with a balanced accuracy score of 99.97 % and a prediction score of 99.7 %. This shows that the model was able to capture significant patterns consistent with the baseline winner's model predictions.

CONCLUSION

The major contribution of this study arises from its application of a novel-based algorithmic interpretable ML approach to identify the best independent variables for predicting an individuals' COVID-19 risk status ("Low", "Medium", "High"). The findings of this study are not only consistent with relevant literature from other researchers on COVID-19 risk infection, but they are also clinically relevant. This study identified **COVID-19 positive, COVID-19 contacts, COVID-19 symptoms, health workers, and public transport count** as the five most consistent features that predict an individual's risk factors for COVID-19. Others include heart disease, working, house count (number of people living in an apartment), contact counts, nursing home, opinion infection (individual's prior beliefs), kidney diseases, lung disease, diabetes, region, and alcohol. These factors exhibited strong consistency when scrutinized with statistical and ML algorithms. This is worthy to note as governments, WHO, health officials, and policymakers plan strategically and resourcefully on how to minimize/stop the virus spread.

Author Contributions

AA, MO, and Z conceived the idea and designed the study. AA, AR, AM, BA, and Z contributed materially to study design and literature. AA, MO, and AR were responsible for data management. AA and MO analyzed the dataset. AA, MO, and AR wrote the article. Z, AM, MN, GB, and CG reviewed the manuscript and contributed to revisions. All authors reviewed, interpreted the results, read, and approved the final version of the manuscript.

Conflict of Interest Statement

No conflict of interest was declared.

References

[1] Alistair Martin, Jama Nateqi., (2020). An artificial intelligence-based first-line defense against COVID-19: digitally screening citizens for risks via a chatbot. bioRxiv 2020.03.25.008805; doi: https://doi.org/10.1101/2020.03.25.008805

[2] Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the Black Box: Interpretable machine learning for geneticists. Trends in Genetics.

[3] Boldog, P., Tekeli, T., Vizi, Z., Dénes, A., Bartha, F. A., & Röst, G. (2020). Risk assessment of novel coronavirus COVID-19 outbreaks outside China. Journal of clinical medicine, 9(2), 571.

[4] Centers for Disease Control and Prevention. 2020. Coronavirus Disease 2019 (COVID-19) In The U.S.. [online] Available at: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html [Accessed 25 April 2020].

[5] Centers for Disease Control and Prevention. 2020. Coronavirus Disease 2019 (COVID-19). [online] Available at: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html[Accessed 25 April 2020].

[6] Centers for Disease Control and Prevention. 2020. Risk Assessment And Management. from https://www.cdc.gov/coronavirus/2019-ncov/php/risk-assessment.html

[7] Chen, X., Chughtai, A. A., & MacIntyre, C. R. (2020). Application of a Risk Analysis Tool to Middle East Respiratory Syndrome Coronavirus (MERS-CoV) Outbreak in Saudi Arabia. Risk Analysis, 40(5), 915-925.

[8] Larsen, K., Petersen, J. H., Budtz-Jørgensen, E., & Endahl, L. (2000). Interpreting parameters in the logistic regression model with random effects. *Biometrics*, *56*(3), 909-914.

[9] Coronavirus Live Updates: Virus Keeps Spreading as Governments Clamp Down. (2020). Retrieved 9 March 2020, from https://www.nytimes.com/2020/03/09/world/coronavirus-news.html

[10] Fang, L., Karakiulakis, G., & Roth, M. (2020). Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?. The Lancet. Respiratory Medicine, 8(4), e21.

[11] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

[12] Grantham, Jonathon. "KNOW YOUR RISK". Nexoid's COVID-19 Survival Calculator, 2020, https://www.covid19survivalcalculator.com/calculator . Accessed 18 May 2020.

[13] Hao, S., Wang, Y., Jin, B., Shin, A. Y., Zhu, C., Huang, M., ... & Ling, X. B. (2015). Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine healthcare information exchange. PloS one, 10(10), e0140271.

[14] Hellewell, J., Abbott, S., Gimma, A., Bosse, N., Jarvis, C., & Russell, T. et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health. 2020, doi: 10.1016/s2214109x(20)30074-7

[15] Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A., & Han, T. Y. J. (2019). Reliable and explainable machine-learning methods for accelerated material discovery. npj Computational Materials, 5(1), 1-9.

[16] Koh, D. (2020). Occupational risks for COVID-19 infection. Occupational medicine (Oxford, England), 70(1), 3.

[17] Knight, W. (2020). How AI Is Tracking the Coronavirus Outbreak. Retrieved 9 March 2020, from https://www.wired.com/story/how-ai-tracking-coronavirus-outbreak/

[18] Krawczyk, B., Woźniak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. Applied Soft Computing, 14, 554-562.

[19] Lo, J. C., Pressman, A. R., Chandra, M., & Ettinger, B. (2011). Fracture risk tool validation in an integrated healthcare delivery system. The American journal of managed care, 17(3), 188-194.

[20] Martin, A., Nateqi, J., Gruarin, S., Munsch, N., Abdarahmane, I., Zobel, M., & Knapp, B. (2020). An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. Scientific reports, 10(1), 1-7.

[21] McCall, B. (2020). COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. The Lancet Digital Health, 2(4), e166-e167.

[22] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

[23] Qiao, J. (2020). What are the risks of COVID-19 infection in pregnant women?. The Lancet, 395(10226), 760-762.

[24] Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. IEEE Access, 8, 42200-42216.

[25] Rozenfeld, Y., Beam, J., Maier, H., Haggerson, W., Boudreau, K., Carlson, J., & Medows, R. (2020). A model of disparities: risk factors associated with COVID-19 infection. International journal for equity in health, 19(1), 1-10.

[26] Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research.

[27] Sudjianto, A., Knauth, W., Singh, R., Yang, Z., & Zhang, A. (2020). Unwrapping The Black Box of Deep ReLU Networks: Interpretability, Diagnostics, and Simplification. arXiv preprint arXiv:2011.04041.

[28] Sun, Yijun, Sinisa Todorovic, and Steve Goodison. "Local-learning-based feature selection for high-dimensional data analysis." IEEE transactions on pattern analysis and machine intelligence 32.9 (2009): 1610-1626.

[29] Tian, S., Hu, N., Lou, J., Chen, K., Kang, X., Xiang, Z., ... & Chen, G. (2020). Characteristics of COVID-19 infection in Beijing. Journal of Infection.

[30] Udugama, B., Kadhiresan, P., Kozlowski, H. N., Malekjahani, A., Osborne, M., Li, V. Y., ... & Chan, W. C. (2020). Diagnosing COVID-19: the disease and tools for detection. ACS nano, 14(4), 3822-3835.

[31] World Health Organization. Coronavirus. Retrieved 9 March 2020, from https://www.who.int/health-topics/coronavirus

[32] Writer, R. How effective are travel bans for curbing coronavirus spread?. Retrieved 9 March 2020, from https://www.livescience.com/coronavirus-travel-restrictions-effectiveness.html

[33] Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., ... & Kraemer, M. U. (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. Scientific data, 7(1), 1-6.

[34] Xu, B., Kraemer, M. U., Gutierrez, B., Mekaru, S., Sewalk, K., Loskill, A., ... & Kraemer, M. (2020). Open access epidemiological data from the COVID-19 outbreak. The Lancet Infectious Diseases, 20(5), 534.

[35] Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Huang, X. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 1-6.

[36] Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. ASU feature selection repository, 1-28

[37] Zietz, M., & Tatonetti, N. P. (2020). Testing the association between blood type and COVID-19 infection, intubation, and death. MedRxiv.

[38] Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. Big Data Research, 5, 2-8.

URL: https://doi.org/10.14738/tnc.92.9760.