RR_Project_Report

June 3, 2024

Team Members

- Dominik Koterwa 454845
- Hüseyin Polat 437969
- Nurdan Beşli 457945
- Maciej Lorens 419763

1 INTRODUCTION

Our project is centered on replicating and expanding upon the groundbreaking machine learning research detailed in the Asian Journal of Computer and Information Systems (AJCIS), specifically the article titled '**Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets**'. This initiative seeks to deepen our understanding of how machine learning can revolutionize the analysis of medical data, thereby enhancing disease detection and improving diagnostic accuracy across various health conditions.

By engaging with this research, we aim to assess the reproducibility of the results and explore potential enhancements in the methodology. Our study will utilize a similar approach to analyze medical datasets, focusing on maximizing the potential of machine learning in healthcare. Through rigorous testing and training, our objective is to contribute meaningful advancements to the medical field, offering insights that could be vital in developing more effective diagnostic tools and treatments.

2 DATASETS

The study we replicated utilized machine learning algorithms to analyze six different healthcare datasets, all sourced from the **UC Irvine Machine Learning Repository**. Each of these datasets varies in the number of attributes and instances, as detailed below.

Dataset	Attributes	Instances
Wisconsin Breast Cancer	32	569
Chronic Kidney Disease	25	400
Immunotherapy	8	90
Cryotherapy	7	90
Hepatitis	19	155
COVID-19	7	14

3   METHODOLOGY

3.1   DATA PREPROCESSING

For each dataset involved in our study, detailed data preprocessing steps are applied to ensure compatibility with machine learning models and to maintain the integrity of the data analysis. Below are the dataset-specific preprocessing methods implemented:

COVID-19 Dataset

- **Binary Remapping:** Each categorical variable representing binary states (e.g., positive '+' or negative '-') is replaced with numerical values where '+' is mapped to 1 and '-' is mapped to 0.
- **Numerical Conversion:** All columns are converted to numeric types to facilitate mathematical operations and model fitting.
- **Column Renaming:** To clearly indicate the binary nature of the data, columns are renamed by adding a 'is_' prefix, enhancing clarity and manageability in the dataset.

Hepatitis Dataset

- **Binary Remapping:** Specific columns that contain categorical data related to the patient's conditions (e.g., 'Steroid', 'Antivirals', 'Fatigue') are converted to binary format, where applicable values are adjusted to 0 and 1 to reflect absence and presence of the condition.
- **Mean and Median Imputation:** Missing values are addressed by calculating and substituting the mean and median for each column, allowing for robustness against outliers and skewed data.
- **Dropping Missing Data:** Rows containing missing data are removed to avoid the introduction of bias or inaccuracies in the model's predictions. Corresponding entries in the target variable are also excluded to maintain dataset alignment.

Chronic Kidney Disease Dataset

- **Binary Mapping:** Categorical descriptions (e.g., 'normal', 'abnormal'; 'present', 'notpresent') in symptoms and diagnostic results are mapped to binary (1, 0) to simplify the input for algorithm processing.
- **Mean and Median Imputation:** Implements both mean and median imputation for handling missing values, providing two sets of data for comparative analysis and model robustness.
- **Dropping Missing Data:** Ensures that only complete data entries are processed, by excluding rows with any missing values from the analysis, thus maintaining the reliability of the dataset.

Breast Cancer Dataset

- **Binary Mapping:** The target variable 'Diagnosis' is converted from its original categorical labels ('M' for malignant, 'B' for benign) into a binary format (1 for malignant, 0 for benign), aligning it with binary classification tasks in machine learning.
- **Data Checks and Cleaning:** Verifies data types, checks for outliers, and ensures that all entries are consistent and appropriate for analysis.

Immunotherapy Dataset

- **Binary Mapping and Renaming:** Converts 'sex' from a categorical to a binary format and updates the column name to 'is_female' to reflect this binary distinction.

- **One-Hot Encoding:** Applies one-hot encoding to the 'Type' column, which involves transforming it into multiple binary columns, each representing a different type of immunotherapy, thereby avoiding ordinal implications in the categorical data.
- **Concatenation:** The one-hot encoded columns are then concatenated back to the main dataset, preserving the original data structure while adding new binary indicators.

Cryotherapy Dataset

- **Binary Mapping:** Converts 'sex' to a binary format and adjusts column naming similarly to the Immunotherapy dataset to maintain consistency.
- **One-Hot Encoding:** The 'Type' column is treated with one-hot encoding to convert categorical treatment types into a series of binary columns, which are more suitable for machine learning models.
- **Concatenation of One-Hot Encoded Data:** These binary columns are merged back into the primary dataset, ensuring that all data remains integrated and accessible for subsequent analysis.

3.2 MODELING AND HYPERPARAMETER OPTIMIZATION

In aligning with the methods of the replicated study, we employ the same machine learning models, each optimized through a systematic hyperparameter tuning process. The models and their respective hyperparameters are as follows:

- **Random Forest (RF):** Optimized for **n_estimators** (50, 100, 200), **criterion** ('gini', 'entropy'), **max_depth** (None, 10, 20), **min_samples_split** (2, 5, 10), **min_samples_leaf** (1, 2, 4), **max_features** ('sqrt', 'log2').

- **K-Nearest Neighbors (KNN)**: Optimized for **n_neighbors** (3, 5, 7), **weights** ('uniform', 'distance'), **algorithm** ('auto', 'ball_tree', 'kd_tree', 'brute'), **p** (1, 2).

- **Support Vector Machine (SVM):** Optimized for **C** (0.1, 1, 10), **kernel** ('poly', 'rbf'), **degree** (2, 3).

- **Decision Tree (DT):** Optimized for **criterion** ('gini', 'entropy'), **splitter** ('best', 'random'), **max_depth** (None, 10, 20), **min_samples_split** (2, 5, 10), **min_samples_leaf** (1, 2, 4), **max_features** ('sqrt', 'log2').

Hyperparameter tuning is conducted using **GridSearchCV**, which incorporates a **5-fold cross-validation** to ensure robustness and reliability in model performance. This meticulous optimization process ensures that each model is finely tuned to provide the best possible accuracy and generalizability on diverse medical datasets.

3.3 PERFORMANCE METRICS

In alignment with the replicated study from the Asian Journal of Computer and Information Systems, we utilize the same performance metrics to evaluate the effectiveness of the machine learning algorithms implemented in our research. Each model was applied to the designated healthcare datasets, and their performance was quantitatively assessed using the following metrics:

- **Training Accuracy:** This metric indicates the proportion of correct predictions made by the model on the training dataset. It provides insight into how well the model learns from the data, mirroring the approach taken in the replicated study.

- **Testing Accuracy:** Reflects the proportion of correct predictions on the testing dataset, crucial for understanding how well the model generalizes to new, unseen data.

- **Training Time:** Measures the duration of time the model takes to learn from the training data, highlighting the computational efficiency of the model during the training phase.

- **Testing Time:** The time required by the model to make predictions on the testing dataset, indicative of the model's operational efficiency.

Additionally, to further enhance our evaluation framework, we have incorporated the measurement of the **Training and Testing F1 Score**, metrics not used in the replicated study. The F1 Score is a harmonic mean of precision and recall, providing a more nuanced view of model performance, particularly in the context of imbalanced datasets:

- **Training F1 Score:** This metric helps assess the balance between precision and recall during the model's training phase, offering deeper insights into the model's predictive performance and accuracy.

- **Testing F1 Score:** Evaluates the model's precision and recall on the testing dataset, which is crucial for applications where the cost of false positives and false negatives is significant.

By utilizing these metrics, which were also employed in the original study, we ensure a consistent and comparative approach to evaluating model performance.
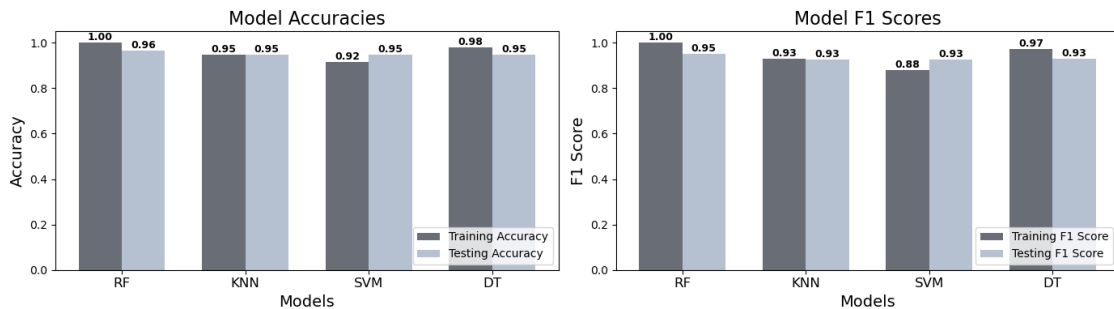
4 RESULTS

4.1 Wisconsin Breast Cancer

Training And Testing Time

Model	Training Time	Testing Time
RF	0.183281	0.027984
KNN	0.002406	0.018223
SVM	0.029788	0.012569
DT	0.017596	0.005135

Accuracy and F1 Score

4.2 Chronic Kidney Disease

Training And Testing Time

Dataset	Model	Training Time	Testing Time
Mean Imputted	RF	0.269601	0.044630
Mean Imputted	KNN	0.005406	0.024386
Mean Imputted	SVM	0.010864	0.006508
Mean Imputted	DT	0.004596	0.001432
Median Imputted	RF	0.191878	0.040303
Median Imputted	KNN	0.005036	0.017453
Median Imputted	SVM	0.010226	0.003573
Median Imputted	DT	0.003194	0.001441
Missing Dropped	RF	0.273432	0.035553
Missing Dropped	KNN	0.004863	0.015425
Missing Dropped	SVM	0.009367	0.004073
Missing Dropped	DT	0.008460	0.003340

Accuracy and F1 Score

Mean Imputted
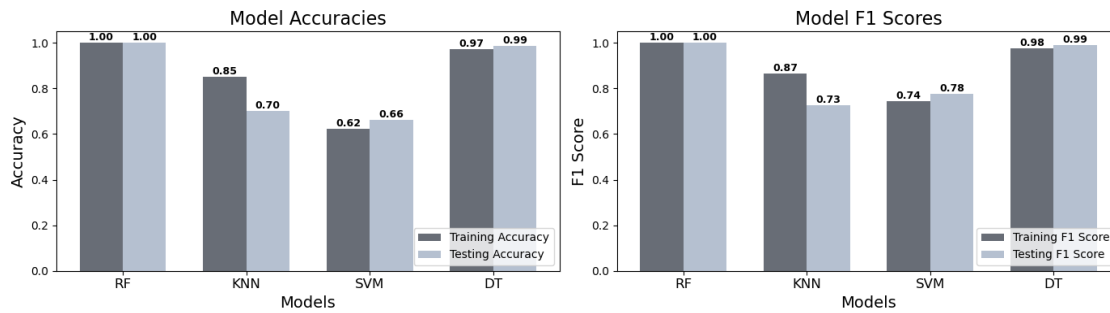
Median Imputted
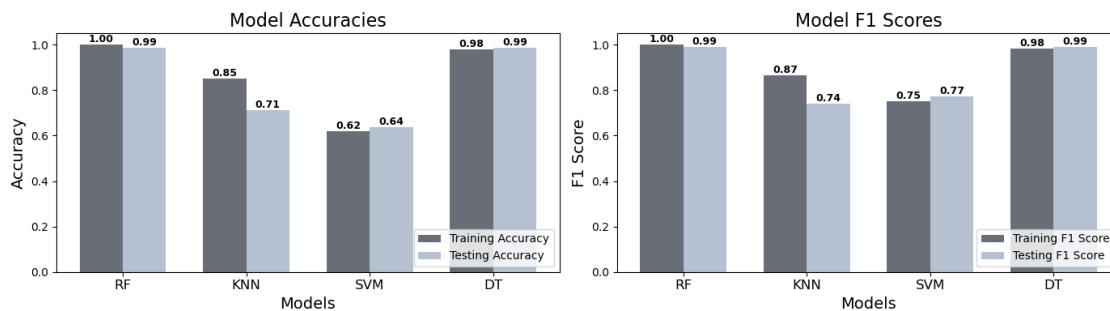
Missing Dropped
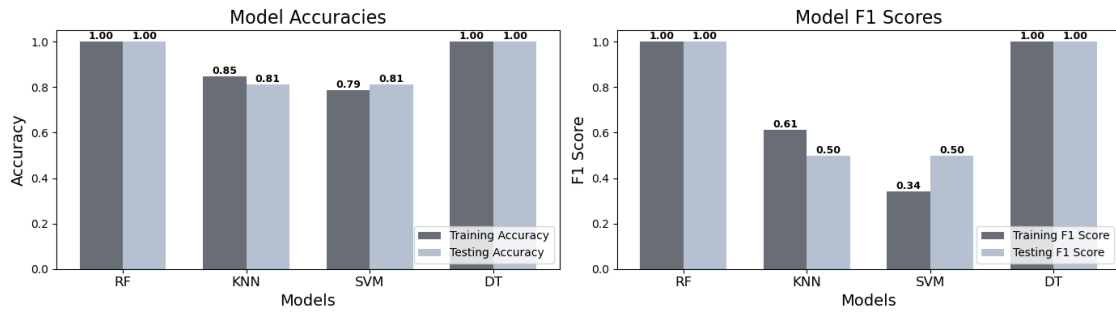
4.3   Covid 19

Training And Testing Time

Model	Training Time	Testing Time
RF	0.194533	0.030823
KNN	0.003052	0.031629
SVM	0.007811	0.003878
DT	0.007249	0.006057

Accuracy and F1 Score
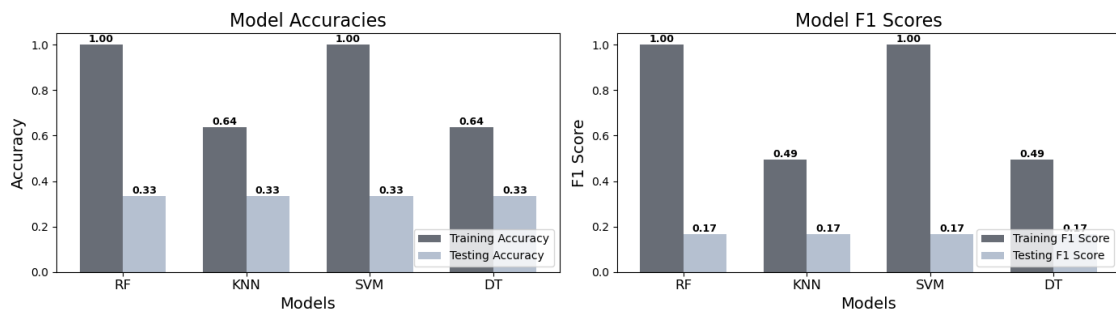
4.4 Cryotherapy

Training And Testing Time

Model	Training Time	Testing Time
RF	0.219233	0.040224
KNN	0.007742	0.034397
SVM	0.007215	0.003353
DT	0.003139	0.002361

Accuracy and F1 Score
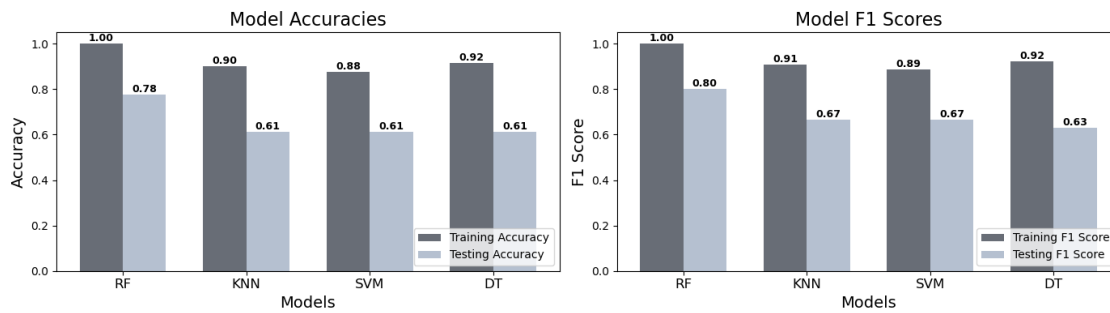
4.5 Hepatitis
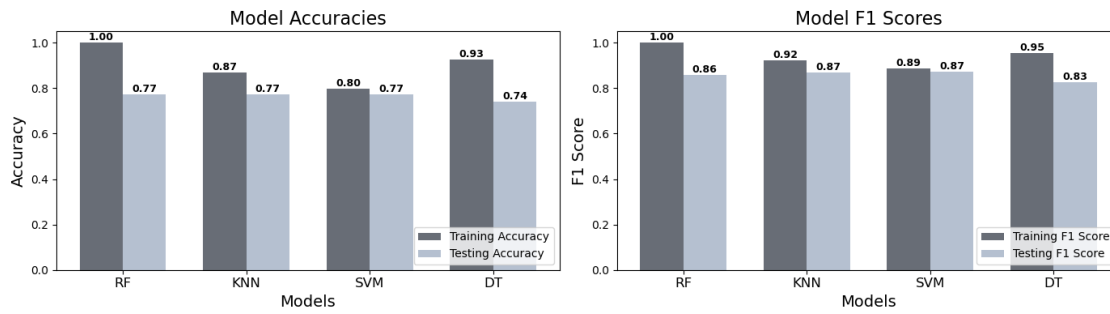
Training And Testing Time

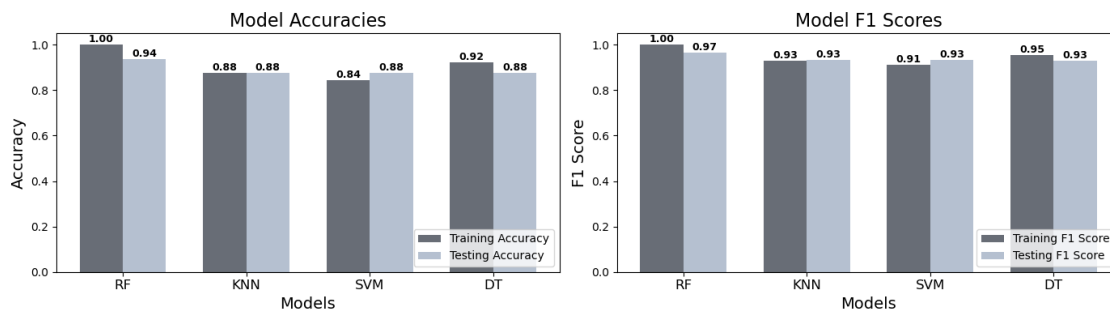| Dataset	Model	Training Time	Testing Time
Mean Imputted	RF	0.237288	0.038766
Mean Imputted	KNN	0.006000	0.018121
Mean Imputted	SVM	0.012891	0.005145
Mean Imputted	DT	0.009227	0.006996
Median Imputted	RF	0.376774	0.041732
Median Imputted	KNN	0.005754	0.022186
Median Imputted	SVM	0.010769	0.006719
Median Imputted	DT	0.007189	0.003993
Missing Dropped	RF	0.208498	0.039103
Missing Dropped	KNN	0.005799	0.018653
Missing Dropped	SVM	0.013068	0.006910
Missing Dropped	DT	0.004419	0.002771

Accuracy and F1 Score

Mean Imputted
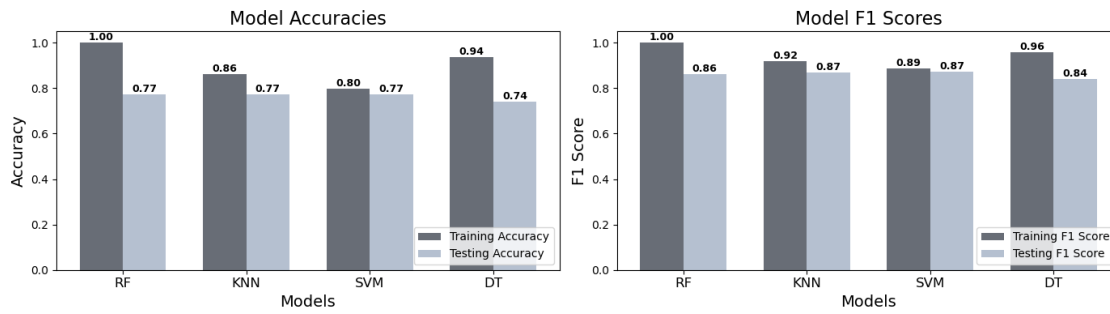
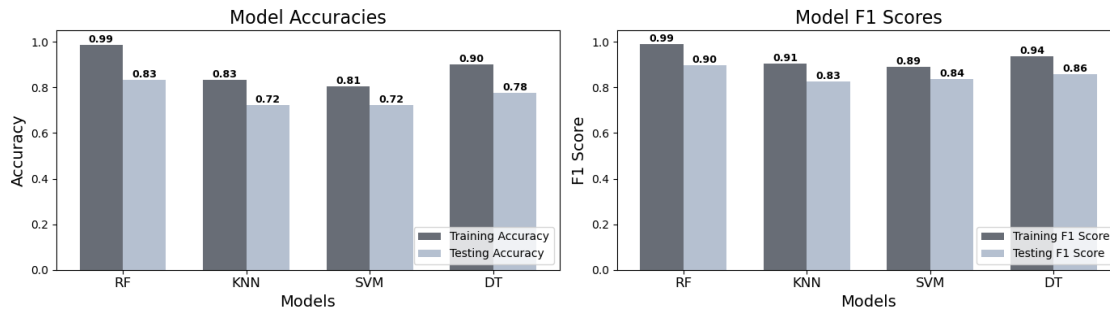## Median Imputted

Missing Dropped

4.6 Immunotherapy

Training And Testing Time

Model	Training Time	Testing Time
RF	0.318120	0.035794
KNN	0.006032	0.025356
SVM	0.004922	0.003697
DT	0.005246	0.002195

Accuracy and F1 Score

5 COMPARISON

In our replication of the study, we follow the original approach of categorizing model performance into four levels: **excellent execution, good execution, fair execution, and inadequate execution**. The original study primarily utilizes **Testing Accuracy** to compare the performance of models across various datasets, a method we have also adopted.

In instances where models yield the same Testing Accuracy for the same datasets, we employ a different approach by further comparing results using **Testing F1 Score** and **Testing Time**. This adjustment is necessary because, unlike in the original study where all testing accuracies differed, our replication has encountered scenarios with identical accuracy scores, necessitating additional metrics for a comprehensive evaluation.

Our Results Comparison

Dataset	Excellent Execution	Good Execution	Fair Execution	Inadequate Execution
Wisconsin Breast Cancer	RF	DT	KNN	SVM
Chronic Kidney Disease	RF	DT	KNN	SVM
Immunotherapy	RF	DT	SVM	KNN
Cryotherapy	RF	SVM	KNN	DT
Hepatitis	RF	SVM	KNN	DT
COVID-19	KNN	DT	SVM	RF

Replicated Study Comparison

Dataset	Excellent Execution	Good Execution	Fair Execution	Inadequate Execution
Wisconsin Breast Cancer	SVM	KNN	DT	RF
Chronic Kidney Disease	KNN	SVM	DT	RF
Immunotherapy	RF	DT	KNN	SVM
Cryotherapy	KNN	DT	SVM	RF
Hepatitis	RF	DT	SVM	KNN
COVID-19	RF	SVM	DT	KNN

Our results exhibit significant differences across almost all datasets when compared to the replicated study. These variations can primarily be attributed to the lack of detailed documentation in the replicated study regarding their methodology, data preprocessing processes, and the hyperparameters used in the models. In our replication, we employed a grid search to optimize hyperparameter selection, a method not specified in the replicated study. Without knowledge of how they chose their hyperparameters, our reliance on this systematic approach likely contributed to the observed discrepancies in model performance across the datasets. This highlights the importance of transparency and detailed documentation in research to ensure reproducibility and accurate comparison of results.

6   REFERENCES

Replicated Study

- Mijwil, M. M., Salem, I. E., & Abttan, R. A. (2021). Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets. Asian Journal of Computer and Information Systems, 9(4), 29. Retrieved from https://ajouronline.com/index.php/AJCIS/article/view/6765

Datasets

All datasets are retrieved from UC Irvine Machine Learning Repository:

- Wisconsin Breast Cancer: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- Chronic Kidney Disease: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease
- Immunotherapy: https://archive.ics.uci.edu/dataset/428/immunotherapy+dataset
- Cryotherapy: https://archive.ics.uci.edu/dataset/429/cryotherapy+dataset
- Hepatitis: https://archive.ics.uci.edu/dataset/46/hepatitis
- COVID-19: https://archive.ics.uci.edu/dataset/567/covid+19+surveillance