*Avrupa Bilim ve Teknoloji Dergisi*
*Özel Sayı 37, S. 52-58, Haziran 2022*

***Araştırma Makalesi***

www.ejosat.com ISSN:2148-2683

*European Journal of Science and Technology*
*Special Issue 37, pp. 52-58, June 2022*

***Research Article***

# Predicting COVID-19 Infection Using Machine Learning Methods Combined with Feature Selection

Umut Ahmet Çetin[1], Fatih Abut[2*]

[1] Çukurova University, Faculty of Engineering, Department of Computer Engineering, Adana, Turkey, (ORCID: 0000-0001-8755-4417), uacetin@student.cu.edu.tr
[2*] Çukurova University, Faculty of Engineering, Department of Computer Engineering, Adana, Turkey, (ORCID: 0000-0001-5876-4116), fabut@cu.edu.tr

(5th International Symposium on Innovative Approaches in Smart Technologies– 28-29 May 2022)

(**DOI:** 10.31590/ejosat.1132337)

**ATIF/REFERENCE:** Çetin, U. A. & Abut, F. (2022). Predicting COVID-19 Infection Using Machine Learning Methods Combined with Feature Selection. *European Journal of Science and Technology*, (37), 52-58.

**Abstract**

COVID-19 is an infection that has affected the world since December 31, 2019, and was declared a pandemic by WHO in March 2020. In this study, Multi-Layer Perceptron (MLP), Tree Boost (TB), Radial Basis Function Network (RBF), Support Vector Machine (SVM), and K-Means Clustering (kMC) individually combined with minimum redundancy maximum relevance (mRMR) and Relief-F have been used to construct new feature selection-based COVID-19 prediction models and discern the influential variables for prediction of COVID-19 infection. The dataset has information related to 20.000 patients (i.e., 10.000 positives, 10.000 negatives) and includes several personal, symptomatic, and non-symptomatic variables. The accuracy, recall, and F1-score metrics have been used to assess the models' performance, whereas the generalization errors of the models were evaluated using 10-fold cross-validation. The results show that the average performance of mRMR is slightly better than Relief-F in predicting the COVID-19 infection of a patient. In addition, mRMR is more successful than the Relief-F algorithm in finding the relative relevance order of the COVID-19 predictors. The mRMR algorithm emphasizes symptomatic variables such as fever and cough, whereas the Relief-F algorithm highlights non-symptomatic variables such as age and race. It has also been observed that, in general, MLP outperforms all other classifiers for predicting the COVID-19 infection.

**Keywords:** Relief-F, mRMR, machine learning, prediction, COVID-19, coronavirus.

# COVID-19 Enfeksiyonunun Nitelik Seçme ile Birleştirilmiş Makine Öğrenmesi Yöntemleriyle Tahmin Edilmesi

**Öz**

COVID-19, 31 Aralık 2019'dan itibaren dünyayı etkisi altına alan ve Mart 2020'de DSÖ tarafından pandemi ilan edilen bir enfeksiyondur. Bu çalışmada, yeni nitelik seçme tabanlı COVID-19 tahmin modelleri oluşturmak ve COVID-19 enfeksiyonunun tahmini için etkili değişkenleri ayırt etmek için minimum fazlalık maksimum önem (mRMR) ve Relief-F nitelik seçiciler ile ayrı ayrı birleştirilmiş Çok Katmanlı Algılayıcı (MLP), Tree Boost (TB), Radyal Temelli Fonksiyon Ağı (RBF), Destek Vektör Makinesi (SVM) ve K-Means Kümeleme (kMC) yöntemleri kullanılmıştır. Veri seti, 20.000 hasta (10.000 pozitif, 10.000 negatif) ile ilgili bilgileri içermektedir ve çeşitli kişisel, semptomatik ve asemptomatik değişkenlerden oluşmaktadır. Modellerin performansını değerlendirmek için doğruluk, duyarlılık ve F1-Skor metrikleri kullanılmıştır ve modellerin genelleme hataları 10 katlı çapraz doğrulama ile değerlendirilmiştir. Sonuçlar, bir hastanın COVID-19 enfeksiyonunu tahmin etmede mRMR'ın ortalama performansının Relief-F'den biraz daha iyi olduğunu göstermektedir. Ek olarak, mRMR'ın, COVID-19 tahmin değişkenlerinin göreceli alaka sırasını bulmada Relief-F algoritmasından daha başarılı olduğu gözlemlenmiştir. mRMR algoritması ateş ve öksürük gibi semptomatik değişkenleri vurgularken, Relief-F algoritması yaş ve ırk gibi asemptomatik değişkenleri öne çıkarmaktadır. Ayrıca, genel olarak MLP'nin COVID-19 enfeksiyonunu tahmin etmede diğer tüm sınıflandırıcılarından daha iyi performans gösterdiği de gözlemlenmiştir.

**Anahtar Kelimeler:** Relief-F, mRMR, makine öğrenmesi, tahmin, COVID-19, koronavirüs.

[*] Corresponding Author: fabut@cu.edu.tr

# 1. Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was discovered in Wuhan, Hubei Province, China, in January 2020. Ever since, it has infected more than 465 million people and caused more than 6 million deaths as of March 17, 2022 (Ciotti et al., 2020; COVID Live, 2021). SARS-CoV-2 causes severe pneumonia along with a fatality rate of 2.9%. Patients infected with SARS-CoV-2 may have asymptomatic, mild, or severe symptoms. The most common symptoms of SARS-CoV-2 include cough, shortness of breath, and fever, but also symptoms like vomiting and diarrhea. SARS-CoV-2's transmission is mainly through the respiratory route like other respiratory viruses. Patients suspected of having SARS-CoV-2 are verified mostly through reverse real-time PCR (rRT-PCR) tests (Ciotti et al., 2020).

Using models for predicting the outcome or the trend of an infectious disease is not a new topic in the literature. Several studies have already been conducted to predict diseases like swine fever, H1N1 flu, and influenza. However, the different characteristics and trends of COVID-19 make it spread across the world at an unprecedented scale and forced governments, businesses, and other similar organizations to take harsh measures that were almost never taken before, such as nationwide curfews, temporary closing of all non-emergency government buildings and businesses, vaccine requirement for public spaces or traveling aboard and many more similar measures.

Governments, businesses, and other similar organizations need to develop prediction models to combat the spread of such diseases, prevent any negative consequences caused by such diseases, and plan their decisions to use their budget and health infrastructure effectively. By doing so, they can avoid problems such as using too much budget than needed, overfilled facilities, an insufficient amount of medicine, medical equipment, or medical staff, and many more similar problems.

There are many studies for predicting COVID-19 infection using various machine learning (ML) methods. (Althnian et al., 2020) conducted a study to predict the susceptibility of the individuals based on demographic data using MLP, SVM, Decision Tree (DT), and Random Forest (RF) methods and evaluated their models with accuracy, precision, recall, F1-score, and Area Under Curve (AUC) metrics. (Fayyoumi et al., 2020) conducted a study to forecast potential COVID-19 patients using Logistic Regression (LogReg), SVM, and MLP models. They evaluated their models with accuracy, sensitivity, specificity, Geometric Mean (G_Mean), and precision metrics. (Santana et al., 2021) conducted a study to help the detection of COVID-19 based on the early symptoms of the COVID-19 using RF, SVM, MLP, K-Nearest Neighbors (KNN), DT, Gradient Boosting Machine (GBM), and XGBoost. They evaluated their models with precision, accuracy, recall, and AUC metrics. (de Souza et al., 2020) conducted a study to make a prognosis or early identification of COVID-19 patients using LogReg, LDA (Linear Discriminant Analysis), Naïve Bayes, KNN, DT, XGBoost, and SVM models. They evaluated their models with Receiver Operating Characteristic (ROC), AUC, Precision-Recall (PR) Curve, AUC, precision, recall, and F1-score metrics. (Wollenstein-Betech et al., 2020) conducted a study to forecast hospitalization, mortality, need for Intensive Care Unit, and "need for a ventilator" events using sparse SVM, sparse LogReg, RF, and XGBoost models and evaluated their models with accuracy, weighted F1-score, and AUC metrics. (Prakash et al., 2020) conducted a study to predict which age groups were more affected by COVID-19 using DT, Multi-Linear Regression (MLR), SVM, XGBoost, RF, KNN+NCA, Gaussian Naïve Bayes (GNB), and LogReg models and evaluated their models using R2 and accuracy metrics. To the best of our knowledge, no study has used the ML methods combined with Relief-F (Robnik-Šikonja et al., 2003) or mRMR (Peng et al., 2005) feature selectors to reveal the discriminative predictors of COVID-19 infection.

This study aims to create new feature selection-based COVID-19 prediction models using MLP, SVM, RBF, kMC, and TB algorithms individually combined with Relief-F and mRMR. 10-fold cross-validation has been carried out to assess the generalization error, whereas the accuracy, recall, and F1-score metrics are used to evaluate the models' performance.

Table 1. Descriptive statistics of the dataset

| Variables | Mean ± Standard Deviation |
|---|---|
| Age | 39.24±17.43 |
| Sex | 0.56±0.50 |
| Race | 1.92±1.51 |
| Pregnant | 0.006±0.076 |
| Fever | 0.41±0.49 |
| Breathing difficulty | 0.16±0.37 |
| Cough | 0.54±0.50 |
| Runny nose | 0.39±0.49 |
| Throat pain | 0.33±0.47 |
| Diarrhea | 0.16±0.37 |
| Headache | 0.52±0.50 |
| Lung comorbidity | 0.03±0.18 |
| Cardio comorbidity | 0.13±0.34 |
| Renal comorbidity | 0.005±0.07 |
| Diabetes comorbidity | 0.05±0.21 |
| Smoking comorbidity | 0.02±0.14 |
| Obesity comorbidity | 0.02±0.15 |

Table 2. List of ranks of predictor variables assigned by mRMR and Relief-F

| Rank | Predictor Variables Ranked by mRMR | Predictor Variables Ranked by Relief-F |
|---|---|---|
| 1 | Fever | Age |
| 2 | Cough | Race |
| 3 | Headache | Fever |
| 4 | Age | Cough |
| 5 | Breathing difficulty | Headache |
| 6 | Obesity comorbidity | Cardio comorbidity |
| 7 | Throat pain | Runny nose |
| 8 | Diarrhea | Diarrhea |
| 9 | Runny nose | Throat pain |
| 10 | Pregnant | Breathing difficulty |
| 11 | Smoking comorbidity | Sex |
| 12 | Renal comorbidity | Diabetes comorbidity |
| 13 | Race | Obesity comorbidity |
| 14 | Lung comorbidity | Pregnant |
| 15 | Sex | Lung comorbidity |
| 16 | Diabetes comorbidity | Smoking comorbidity |
| 17 | Cardio comorbidity | Renal comorbidity |

Table 3. COVID-19 prediction models along with their predictor variables for mRMR

| Model | Selected Features |
|---|---|
| Model 1 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity, renal comorbidity, race, lung comorbidity, sex, diabetes comorbidity, cardio comorbidity |
| Model 2 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity, renal comorbidity, race, lung comorbidity, sex, diabetes comorbidity |
| Model 3 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity, renal comorbidity, race, lung comorbidity, sex |
| Model 4 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity, renal comorbidity, race, lung comorbidity |
| Model 5 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity, renal comorbidity, race |
| Model 6 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity, renal comorbidity |
| Model 7 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant, smoking comorbidity |
| Model 8 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose, pregnant |
| Model 9 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea, runny nose |
| Model 10 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain, diarrhea |
| Model 11 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity, throat pain |
| Model 12 | Fever, cough, headache, age, breathing difficulty, obesity comorbidity |
| Model 13 | Fever, cough, headache, age, breathing difficulty |
| Model 14 | Fever, cough, headache, age |
| Model 15 | Fever, cough, headache |
| Model 16 | Fever, cough |
| Model 17 | Fever |

Table 4. COVID-19 prediction models along with their predictor variables for Relief-F

| Model | Selected Features |
|---|---|
| Model 1 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex, diabetes comorbidity, obesity comorbidity, pregnant, lung comorbidity, smoking comorbidity, renal comorbidity |
| Model 2 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex, diabetes comorbidity, obesity comorbidity, pregnant, lung comorbidity, smoking comorbidity |
| Model 3 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex, diabetes comorbidity, obesity comorbidity, pregnant, lung comorbidity |
| Model 4 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex, diabetes comorbidity, obesity comorbidity, pregnant |
| Model 5 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex, diabetes comorbidity, obesity comorbidity |
| Model 6 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex, diabetes comorbidity |
| Model 7 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty, sex |
| Model 8 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain, breathing difficulty |
| Model 9 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea, throat pain |
| Model 10 | Age, race, fever, cough, headache, cardio comorbidity, runny nose, diarrhea |
| Model 11 | Age, race, fever, cough, headache, cardio comorbidity, runny nose |
| Model 12 | Age, race, fever, cough, headache, cardio comorbidity |
| Model 13 | Age, race, fever, cough, headache |
| Model 14 | Age, race, fever, cough |
| Model 15 | Age, race, fever |
| Model 16 | Age, race |
| Model 17 | Age |

Table 5. The intervals for values of the utilized parameters for the COVID-19 prediction models

| Method | Parameter | Range |
|---|---|---|
| MLP | Number of hidden layers | 1 |
| | Hidden layer activation function | Logistic |
| | Output layer activation function | Logistic |
| | Number of neurons in hidden layer | [2-20] |
| TB | Number of trees in series | [10-400] |
| | Depth of individual trees | 5 |
| | Proportion of rows for each tree | 0.5 |
| | Influence trimming factor | 0.01 |
| | Minimum size node to split | 10 |
| RBF | Maximum number of neurons | 100 |
| | Radius | [0.01-400] |
| | Lambda | [0.001-10] |
| | Population size | 200 |
| SVM | Cost (C) | [0.1-2000] |
| | Epsilon (ε) | 0.001 |
| | Gamma (γ) | [0.001-20] |
| kMC | Number of clusters | [2-200] |
| | Search step | 1 |

The rest of the paper is organized as follows. Section II gives information about dataset generation and provides the methodology for developing prediction models. In Section III, results and discussion are presented. Lastly, in Section IV, the paper is concluded.

# 2. Materials and Method

## *2.1. Material and Procedure*

The dataset used in this study was obtained from the Espirito Santo State Portal of Brazil (Data on COVID-19 pandemic, 2021) on May 23, 2021. The dataset contains over 40 columns and has over 1.4 million rows. The dataset contains data that was collected between March 17, 2020, and May 23, 2021, and includes a diverse amount of knowledge such as symptoms of COVID-19 (e.g., fever and cough), biological characteristics of patients (e.g., sex, age, and race), comorbidities of patients (e.g., obesity and diabetes), date-based variables, location-based features (e.g., neighborhood and county), education status, and various test results.

This study aimed to predict whether a patient is infected with COVID-19 by using the patient's symptoms, comorbidities, and biological characteristics. Because of this, we removed columns that we didn't plan to use from the dataset. After this step, 17 predictors remained for developing the models. Next, we removed rows with empty values and converted string-based values to numeral or binary values. Finally, due to the massive amount of data, we created a smaller dataset by randomly choosing 10.000 positive and 10.000 negative cases.

In the dataset, 0 and 1 values for the "sex" variable represent male and female genders, respectively. Similarly, 0, 1, 2, 3, 4, and 5 values for the "race" variable represent "ignored," "white", "indigenous", "brown", "black", and "yellow", respectively. The descriptive statistics of the dataset are shown in Table 1.

## *2.2. Methodology*

By utilizing the mRMR and Relief-F feature selectors, the rank of each variable has been calculated and sorted in descending order by their ranks, as shown in Table 2. In this study, we used the Mutual Information Difference (MID) scheme of the mRMR. Then, we removed each variable with the lowest rank successively. By doing so, we created 17 mRMR-based and 17 Relief-F-based prediction models, as illustrated in Table 3 and Table 4, respectively.

We used five different ML techniques in our study, namely SVM (Hsu et al., 2003), MLP (Popescu et al., 2009), RBF (Orr, 1996), TB (Natekin et al., 2013), and kMC (Kulis et al., 2012). Table 5 lists the intervals for values of the parameters used by SVM-based, MLP-based, RBF-based, kMC-based, and TB-based models.

The performance of each model has been evaluated using 10-fold cross-validation and calculating accuracy, recall, and F1-score values, as given in (1), (2), and (3), respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

In Eqs. (1) through (3), "TP", "TN", "FP", and "FN" refer to true positives, true negatives, false positives, and false negatives.

# 3. Results and Discussion

Table 6 through Table 8 show the accuracy, recall, and F1-score results of all COVID-19 models developed using the different ML classifiers individually combined with Relief-F and mRMR feature selectors.

The following discussions can be made about the results obtained:

- According to mRMR and Relief-F, fever and age variables have been reported to be the most important variables for predicting COVID-19 infection, respectively. In contrast, the least essential variables for both feature selectors are the non-symptom variables, especially the comorbidities.
- The MLP-based Model 9 created using the mRMR feature selector, and the MLP-based Model 7 created using the Relief-F feature selector performed best in accuracy values with 70.75% and 70.76%, respectively.
- The MLP-based Model 2 created using the mRMR feature selector, and the MLP-based Model 5 created using the Relief-F feature selector achieved the highest recall values with 90.29% and 89.32%, respectively.
- The MLP-based Model 2 created using the mRMR feature selector, and the MLP-based Model 5 created

Table 6. Accuracy results for COVID-19 models using MLP, TB, RBF, SVM, and kMC individually combined with Relief-F and mRMR

| Model | Accuracy (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mRMR | | | | | Relief-F | | | | |
| | MLP | TB | RBF | SVM | kMC | MLP | TB | RBF | SVM | kMC |
| Model 1 | 70.64 | 69.89 | 69.88 | 69.38 | 66.20 | 70.64 | 69.89 | 69.88 | 69.38 | 66.20 |
| Model 2 | 70.56 | 70.04 | 69.72 | 69.38 | 65.81 | 70.62 | 69.78 | 69.88 | 69.38 | 66.06 |
| Model 3 | 70.59 | 69.95 | 69.94 | 69.45 | 65.85 | 70.59 | 70.02 | 69.73 | 69.58 | 66.00 |
| Model 4 | 70.73 | 70.03 | 69.84 | 69.76 | 66.32 | 70.62 | 69.86 | 69.59 | 69.66 | 66.05 |
| Model 5 | 70.61 | 69.80 | 69.88 | 69.72 | 66.50 | 70.45 | 69.84 | 69.49 | 69.64 | 66.24 |
| Model 6 | 70.63 | 69.88 | 69.85 | 70.47 | 66.86 | 70.58 | 69.73 | 69.50 | 69.71 | 65.98 |
| Model 7 | 70.74 | 69.97 | 69.66 | 70.53 | 66.97 | 70.76 | 69.84 | 69.70 | 69.65 | 66.03 |
| Model 8 | 70.67 | 69.95 | 69.67 | 70.56 | 66.54 | 70.61 | 69.70 | 69.48 | 69.69 | 66.34 |
| Model 9 | 70.75 | 69.92 | 69.81 | 70.62 | 66.57 | 70.30 | 69.73 | 69.73 | 69.53 | 66.07 |
| Model 10 | 69.56 | 69.38 | 69.22 | 69.12 | 66.88 | 69.37 | 69.58 | 69.49 | 68.86 | 66.54 |
| Model 11 | 69.56 | 69.35 | 69.27 | 69.13 | 65.75 | 69.73 | 69.55 | 69.55 | 68.96 | 66.40 |
| Model 12 | 69.14 | 68.99 | 69.03 | 68.89 | 65.98 | 68.79 | 68.89 | 69.30 | 68.46 | 65.96 |
| Model 13 | 69.13 | 68.86 | 68.98 | 68.95 | 66.30 | 68.93 | 69.92 | 68.97 | 68.77 | 66.50 |
| Model 14 | 69.03 | 68.88 | 69.12 | 68.99 | 65.75 | 68.52 | 68.38 | 68.62 | 68.49 | 64.94 |
| Model 15 | 67.95 | 67.95 | 67.95 | 67.95 | 53.50 | 65.89 | 65.90 | 65.91 | 65.92 | 62.09 |
| Model 16 | 67.17 | 67.17 | 67.17 | 67.17 | 50.00 | 54.01 | 53.89 | 53.76 | 53.92 | 51.81 |
| Model 17 | 65.05 | 65.05 | 65.05 | 65.05 | 50.00 | 53.35 | 52.65 | 53.09 | 53.34 | 51.80 |

Table 7. Recall results for COVID-19 models using MLP, TB, RBF, SVM, and kMC individually combined with Relief-F and mRMR

| Model | Recall (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mRMR | | | | | Relief-F | | | | |
| | MLP | TB | RBF | SVM | kMC | MLP | TB | RBF | SVM | kMC |
| Model 1 | 87.96 | 81.18 | 80.36 | 77.19 | 70.44 | 87.96 | 81.18 | 80.36 | 77.19 | 70.44 |
| Model 2 | 90.29 | 81.61 | 80.38 | 77.88 | 68.97 | 87.99 | 80.69 | 80.40 | 77.23 | 70.58 |
| Model 3 | 86.36 | 81.21 | 80.62 | 78.60 | 67.55 | 85.46 | 81.04 | 80.22 | 77.64 | 70.04 |
| Model 4 | 86.57 | 81.24 | 80.31 | 79.67 | 69.37 | 86.77 | 81.23 | 80.38 | 78.15 | 69.24 |
| Model 5 | 86.26 | 80.58 | 80.68 | 79.62 | 71.32 | 89.32 | 80.55 | 80.43 | 78.14 | 69.42 |
| Model 6 | 88.04 | 81.87 | 80.97 | 82.77 | 70.73 | 85.70 | 81.30 | 80.26 | 78.50 | 69.74 |
| Model 7 | 85.97 | 81.49 | 80.72 | 83.03 | 71.08 | 88.29 | 81.13 | 80.17 | 78.87 | 69.45 |
| Model 8 | 86.40 | 81.43 | 80.71 | 83.34 | 71.01 | 84.94 | 81.30 | 80.10 | 88.80 | 72.08 |
| Model 9 | 85.61 | 81.75 | 80.62 | 83.46 | 71.10 | 88.06 | 81.27 | 80.51 | 79.83 | 69.92 |
| Model 10 | 79.87 | 81.50 | 78.85 | 79.39 | 70.17 | 81.15 | 82.34 | 80.00 | 79.22 | 68.61 |
| Model 11 | 81.26 | 81.50 | 78.63 | 81.31 | 67.23 | 83.04 | 82.09 | 79.94 | 79.33 | 69.01 |
| Model 12 | 78.27 | 78.29 | 76.65 | 77.78 | 68.58 | 78.50 | 78.09 | 76.27 | 78.25 | 67.61 |
| Model 13 | 78.72 | 77.67 | 77.03 | 77.72 | 67.74 | 79.26 | 78.70 | 76.84 | 77.52 | 68.13 |
| Model 14 | 77.91 | 77.96 | 75.67 | 76.83 | 68.48 | 79.27 | 79.98 | 80.65 | 80.96 | 65.04 |
| Model 15 | 74.87 | 74.87 | 74.87 | 74.87 | 68.11 | 53.87 | 54.39 | 54.24 | 54.96 | 57.87 |
| Model 16 | 85.29 | 85.29 | 85.29 | 85.29 | 68.45 | 57.86 | 62.26 | 65.13 | 60.43 | 52.15 |
| Model 17 | 55.92 | 55.92 | 55.92 | 55.92 | 68.28 | 61.30 | 56.17 | 55.43 | 54.99 | 52.07 |

using the Relief-F feature selector achieved the highest F1-scores with 75.41% and 75.14%, respectively.

- When the accuracy and F1-score values of the models created by using the Relief-F feature selector are compared, it is seen that the MLP-based models, in general, exhibit the best performance. The performance of TB-based, RBF-based, and SVM-based models are similar/comparable. In other words, there is no superiority of one classifier over the other two classifiers, but their accuracy and F1-score values are lower than the ones of MLP-based models. Finally, the kMC-based models produce the lowest accuracy and F1-score values. In contrast, when the recall values of the models created using the Relief feature selector are compared, the general order of ML classifiers leading from the best to the worst results is MLP, TB, RBF, SVM, and kMC.
- Similarly, when the accuracy and F1-score values of the models created by using the mRMR feature selector are compared, it is seen that again the MLP-based models, in general, produce the highest accuracy and F1-scores. The performance of TB-based, RBF-based, and SVM-based models are similar/comparable, occupying the second rank. Finally, the kMC-based models produce the

Table 8. F1-score results for COVID-19 models using MLP, TB, RBF, SVM, and kMC individually combined with Relief-F and mRMR

| Model | F1-Score (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mRMR | | | | | Relief-F | | | | |
| | MLP | TB | RBF | SVM | kMC | MLP | TB | RBF | SVM | kMC |
| Model 1 | 74.97 | 72.95 | 72.73 | 71.59 | 67.58 | 74.97 | 72.95 | 72.73 | 71.59 | 67.58 |
| Model 2 | 75.41 | 73.15 | 72.63 | 71.78 | 66.85 | 74.97 | 72.75 | 72.74 | 71.61 | 67.53 |
| Model 3 | 74.60 | 72.99 | 72.84 | 72.01 | 66.42 | 74.40 | 72.99 | 72.60 | 71.85 | 67.32 |
| Model 4 | 74.73 | 73.05 | 72.70 | 72.49 | 67.32 | 74.70 | 72.93 | 72.55 | 72.04 | 67.10 |
| Model 5 | 74.59 | 72.74 | 72.82 | 72.44 | 68.04 | 75.14 | 72.76 | 72.50 | 72.02 | 67.28 |
| Model 6 | 74.98 | 73.10 | 72.87 | 73.70 | 68.10 | 74.44 | 72.87 | 72.47 | 72.16 | 67.22 |
| Model 7 | 74.61 | 73.07 | 72.68 | 73.80 | 68.27 | 75.12 | 72.90 | 72.58 | 72.21 | 67.15 |
| Model 8 | 74.65 | 73.04 | 72.69 | 73.90 | 67.97 | 74.29 | 72.85 | 72.41 | 74.55 | 68.17 |
| Model 9 | 74.53 | 73.10 | 72.76 | 73.96 | 68.02 | 74.78 | 72.86 | 72.68 | 72.37 | 67.33 |
| Model 10 | 72.41 | 72.69 | 71.92 | 71.99 | 67.93 | 72.60 | 73.02 | 72.39 | 71.79 | 67.22 |
| Model 11 | 72.75 | 72.67 | 71.90 | 72.48 | 66.25 | 73.28 | 72.94 | 72.41 | 71.88 | 67.25 |
| Model 12 | 71.73 | 71.63 | 71.22 | 71.43 | 66.85 | 71.55 | 71.51 | 71.30 | 71.27 | 66.51 |
| Model 13 | 71.83 | 71.38 | 71.29 | 71.45 | 66.78 | 71.84 | 72.34 | 71.23 | 71.28 | 67.03 |
| Model 14 | 71.56 | 71.47 | 71.01 | 71.24 | 66.66 | 71.57 | 71.67 | 71.99 | 71.99 | 64.97 |
| Model 15 | 70.02 | 70.02 | 70.02 | 70.02 | 34.15 | 61.23 | 61.46 | 61.40 | 61.73 | 60.42 |
| Model 16 | 72.21 | 72.21 | 72.21 | 72.21 | 66.67 | 55.71 | 57.45 | 58.48 | 56.74 | 51.98 |
| Model 17 | 61.54 | 61.54 | 61.54 | 61.54 | 66.67 | 56.79 | 54.26 | 54.16 | 54.10 | 51.93 |

lowest accuracy and F1-score values. In contrast, when the models' recall values created using the mRMR feature selector are compared, the general order of ML classifiers leading from the best to the worst results is MLP, TB, RBF, SVM, and kMC. However, it should be noted that the SVM-based Model 6 through Model 12 exceptionally perform better than TB-based and RBF-based models.

- Models created by using the Relief-F feature selector produced average accuracies of 67.87%, 67.48%, 67.39%, 67.23%, 64.18%; average recall values of 79.93%, 76.69%, 75.96%, 75.29%, 66.55%; and average F1-scores of 71.02%, 70.03%, 69.80%, 69.48%, 64.94% for MLP, TB, RBF, SVM, kMC, respectively.
- Models created by using the mRMR feature selector produced average accuracies of 69.56%, 69.12%, 69.06%, 69.12%, 63.63%; average recall values of 79.88%, 77.67%, 76.21%, 78.41%, 68.55%; and average F1-scores of 70.97%, 70.96%, 70.05%, 71.97%, 66.73% for MLP, TB, RBF, SVM, and kMC, respectively. According to these results, it is seen that mRMR achieves slightly better results than Relief-F on average in terms of accuracy, recall, and F1-score values.
- Average training times of MLP-based, RBF-based, SVM-based, TB-based, and kMC-based prediction models are appr. 16 min, 32 min, 13 h, 2 min, and 4 min, respectively. TB has been observed to be the most time-efficient classifier. It gives the second-best performance in the shortest time.

# 4. Conclusion and Future Work

This study proposed new models based on MLP, RBF, SVM, TB, and kMC individually combined with mRMR and Relief-F feature selectors for predicting COVID-19 infection. The utilized dataset consists of information related to 20.000 patients (i.e., 10.000 positives, 10.000 negatives) and includes several personal, symptomatic, and non-symptomatic variables. Performing 10-fold cross-validation on the dataset, the performance of the models has been assessed by calculating the accuracy, recall, and F1-scores. The results show that the average performance of mRMR is slightly better than Relief-F. Furthermore, mRMR is more successful than the Relief-F algorithm in finding the relative relevance order of the predictor variables to predict COVID-19 infection. Finally, the mRMR feature selector emphasizes symptomatic variables such as fever and cough, whereas the Relief-F algorithm emphasizes non-symptomatic variables such as age and race.

This study compared the performance of models developed by using MLP, TB, SVM, RBF, and kMC with individual combinations of Relief-F and mRMR feature selectors for predicting COVID-19 infection. Even though mRMR performs slightly better than Relief-F, and MLP outperforms other ML models, more studies on more datasets with different characteristics are needed to generalize the advantage of one algorithm over the others in this field.

# 5. Acknowledge

The authors would like to thank Çukurova University Scientific Research Projects Center for supporting this work under grant no FYL-2021-14257.

# References

Althnian, A., Elwafa, A. A., Aloboud, N., Alrasheed, H., & Kurdi, H. (2020). Prediction of COVID-19 Individual Susceptibility using Demographic Data: A Case Study on Saudi Arabia. In Procedia Computer Science (Vol. 177, pp. 379–386). https://doi.org/10.1016/j.procs.2020.10.051

Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.-C., Wang, C.-B., & Bernardini, S. (2020). The COVID-19 pandemic. In

Critical Reviews in Clinical Laboratory Sciences (Vol. 57, Issue 6, pp. 365–388). Informa UK Limited. https://doi.org/-10.1080/10408363.2020.1783198

COVID Live. (2022, May 15). Worldometers. https://www.-worldometers.info/coronavirus/

Data on COVID-19 pandemic. (2021, May 24). Open Data from the State of Espirito Santo. https://dados.es.gov.br/-dataset/dados-sobre-pandemia-covid-19/resource/38cc5066-020d-4c5a-b4c0-e9f690deb6d4

Fayyoumi, E., Idwan, S., & AboShindi, H. (2020). Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan. In International Journal of Advanced Computer Science and Applications (Vol. 11, Issue 5). The Science and Information Organization. https://doi.org/10.14569/ijacsa.2020.0110518

Hanchuan Peng, Fuhui Long, & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. In IEEE Transactions on Pattern Analysis and Machine Intelligence (Vol. 27, Issue 8, pp. 1226–1238). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/-tpami.2005.159

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Kulis, B., & Jordan, M. I. (2011). Revisiting k-means: New Algorithms via Bayesian Nonparametrics (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1111.0352

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. In Frontiers in Neurorobotics (Vol. 7). Frontiers Media SA. https://doi.org/10.3389/fnbot.2013.00021

Orr, M. J. (1996). Introduction to radial basis function networks.

Popescu, M. C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, 8(7), 579-588.

Prakash, K. B. (2020). Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms. In International Journal of Emerging Trends in Engineering Research (Vol. 8, Issue 5, pp. 2199–2204). The World Academy of Research in Science and Engineering. https://doi.org/10.30534/ijeter/2020/117852020

Robnik-Šikonja, M., & Kononenko, I. (2003). In Machine Learning (Vol. 53, Issue 1/2, pp. 23–69). Springer Science and Business Media LLC. https://doi.org/10.1023/a:-1025667309714

Souza, F. S. H., Hojo-Souza, N. S., dos Santos, E. B., da Silva, C. M., & Guidoni, D. L. (2020). Predicting the disease outcome in COVID-19 positive patients through Machine Learning: a retrospective cohort study with Brazilian data. https://doi.org/10.1101/2020.06.26.20140764

Viana dos Santos Santana, Í., CM da Silveira, A., Sobrinho, Á., Chaves e Silva, L., Dias da Silva, L., Santos, D. F. S., Gurjão, E. C., & Perkusich, A. (2021). Classification Models for COVID-19 Test Prioritization in Brazil: Machine Learning Approach (Preprint). JMIR Publications Inc. https://doi.org/10.2196/preprints.27293

Wollenstein-Betech, S., Cassandras, C. G., & Paschalidis, I. Ch. (2020). Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator. https://doi.org/10.1101/2020.05.03.20089813