Feature Analysis and Machine Learning Techniques to determine severity of COVID-19 infections

Hanna Gloyna [1], Mario Fernando Jojoa Acosta [2], Cristian Castillo [2], and Begonya Garcia-Zapirain [2]

[1]Universidad de Deusto
[2]Affiliation not available

October 30, 2023

Abstract

In 2019 appeared a new Coronavirus Disease (COVID-19) in China, spreading rapidly globally and causing a pandemic with high infection and death numbers. To prevent a collapse of the health institutions, accurate decision making about assignments of intense care units (ICU) is required, depending on the probable outcome. The usage of machine learning (ML) for other medical fields had been successful before. So we applied ML techniques to a dataset of COVID-19 and influenza patients from Mexico to predict the severity of an individual's infection regarding risk factors including, but not limited to, chronic obstructive pulmonary disease (COPD), cardiovascular disease, diabetes, asthma, immunosupression, and obesity. We conducted two experiments, one on hospitalised patients and the other one on a balanced dataset. The resulting applications should not be used as a diagnostic tool yet, due to a relatively short time period of data collection and 74.64% accuracy for the first experiment and 82.61% accuracy for the second one. Nonetheless it is a good starting point to continue research about predicting COVID-19 infection's outcome based on risk factors.

Feature Analysis and Machine Learning Techniques to determine severity of COVID-19 infections

Hanna Helene Gloyna, Mario Fernando Jojoa Acosta, Cristian Castillo, and Begonya Garcia-Zapirain

***Abstract*— In 2019 appeared a new Coronavirus Disease (COVID-19) in China, spreading rapidly globally and causing a pandemic with high infection and death numbers. To prevent a collapse of the health institutions, accurate decision making about assignments of intense care units (ICU) is required, depending on the probable outcome. The usage of machine learning (ML) for other medical fields had been successful before. So we applied ML techniques to a dataset of COVID-19 and influenza patients from Mexico to predict the severity of an individual's infection regarding risk factors including, but not limited to, chronic obstructive pulmonary disease (COPD), cardiovascular disease, diabetes, asthma, immunosupression, and obesity. We conducted two experiments, one on hospitalised patients and the other one on a balanced dataset. The resulting applications should not be used as a diagnostic tool yet, due to a relatively short time period of data collection and 74.64% accuracy for the first experiment and 82.61% accuracy for the second one. Nonetheless it is a good starting point to continue research about predicting COVID-19 infection's outcome based on risk factors.**

***Index Terms*— COVID-19, Feature Analysis, Machine Learning, Neural Network, Random Forest, SARS-CoV-2, Support Vector Machine**

I. Introduction

IN 2019 a new Coronavirus Disease (COVID-19) appeared in the Chinese city Wuhan, which spread from there fast over the whole world, causing a global pandemic [1], [2]. COVID-19 is an infection with Serve Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) [3]. The clinical manifestations are mainly respiratory, with 5% who develop severe pneumonia with acute respiratory distress syndrome (ARDS) and were admitted to intense care units (ICU) [4].

Typical symptoms for a mild COVID-19 infection include fever, a dry cough, fatigue and pneumonia, whereas serve and critical cases also show dyspnea, ARDS and often multi-organ failure [5], [6]. The main risk factors for ARDS or death from SARS-CoV-2 include high age, gender and underlying comorbidities such as hypertension, diabetes, immunosuppression, obesity, tumors as well as cardiovascular disease [6–10].

A precise estimation of the curse of infection in patients influences the applied, and necessary treatment. Especially in a pandemic when a shortage of ICUs occurs an accurate estimator for decisions of ICU assignments is needed. Bordbar *et al.* investigated in [11] the utility of HScore for predicting the risk of having reactive hemophagocytic syndrome, which often accompanies SARS-CoV-2 disease and results in serve cases [12]. They concluded that a higher HScore is associated with a higher probability of ICU admission and increased risk of mortality [11].

Having such estimators, it seems logical to combine risk factors such as comorbidities or age and symptoms of an individual's infection to obtain more precise predictions of COVID-19 outcomes before certain control values are exceeded. Estimating the outcome of one's infection does not only lead to the ability of applying better treatment, but also a more thoughtful and efficient organisation of hospitals [13].

The usage of machine learning (hereinafter: ML) for decision making and diagnosing had become popular over the last decade with a wide variety of data driven applications [14], such as image-based medical diagnosis [15–17], analysis of sentiments [18] or predicting complications in ostomy patients [19].

Trained on a dataset from Mexico we designed a ML application to predict the outcome of a COVID-19 or influenza infection based on the idea, that persons with chronicle diseases, obesity, diabetes or hypertension tend to have more serve courses of infection. Therefore we take these comorbidities, treatment and symptoms into account for training. Moreover we tried to rank the used features by their importance from all models to give an idea which attributes should be especially considered for manual decision making.

This paper presents available input data, general methods to clean up the data and clarification of the used models. Furthermore an explanation about the generation of a feature ranking by importance for decision making and evaluation of models is given. It describes 2 experiments which were performed to determine important features and comparison of the quality of models trained on subsets of features. Finally limitations and possible extensions of this research are discussed.

This research was submitted for review on May 4, 2022

Hanna Helene Gloyna was with eVIDA research group, Avda. de las Universidades, 24, 48007 Bilbao. She is now with University of Potsdam (e-mail: gloyna@uni-potsdam.de).

Mario Fernando Jojoa Acosta is with eVIDA research group, Avda. de las Universidades, 24, 48007 Bilbao (e-mail: mariojojoa@deusto.es)

Cristian Castillo is with eVIDA research group, Avda. de las Universidades, 24, 48007 Bilbao (e-mail: cristian.castillo@deusto.es)

Begonya Garcia-Zapirain is with eVIDA research group, Avda. de las Universidades, 24, 48007 Bilbao (e-mail: mbgarciazapi@deusto.es)

II. Materials

The used dataset was provided by several public hospitals from Baja California, including communities like Tijuana, Mexicali, Ensenada and Rosarito. It consisted initially of 92 features and 6550 entries, where each entry represents a patient.

The population is made up of 49,7% women and 50,3% men, respectively. The average age is 43 years and the average hospital stay lasted 5 days. The data contained 1969 confirmed COVID-19 cases, 4331 cases of atypical pneumonia, 46 entries of influenza with other respiratory manifestations and 250 patients who were affected by identified influenza variants, influenza with pneumonia and unidentified viruses. The distribution of prevalent diseases showed that a chronic disease was most common among patients with 43%, followed by diabetes and obesity, both affecting 20%.

From this data set 193 erroneous entries needed to be deleted. The remaining 6357 patients contained 1968 confirmed COVID-19 cases of those were 650 serve cases. The data were recorded from January 2020 until August 2020 and include 21 categorical, 59 binary or continuous as well as 12 time features.

III. Methods

Two experiments were executed on different subsets of the dataset. This section explains the set up of experiments.

A. Implementation

All implementations were done with Python 3.8.12 using Pandas version 1.3.3 for data analysis and manipulation, Keras version 2.4.3 and eli5 version 0.11.0 for neural networks as well as Sklearn version 1.0.1 [20] for classification algorithms and evaluation. The code and evaluating plots are available in this github repository (https://github.com/hGl0/Covid-19).

B. Input Data

We could remove 9 features because they represented different medical categorisations and their descriptions, respectively. The latter was kept as it provided better understanding. Additionally ID was dropped, because it contained an unique ID for each patient, but did not gave any relevant information for statistic models.

Because the data were collected in the first months of the pandemic, seasonal behaviour of COVID-19 could not be observed [21–24] and is not represented in the data. To avoid overfitting due to coincidental correlations between time features and the resulting outcome, all time features were removed.

For the first experiment 1732 entries of hospitalised patients were used. Of those entries were 1206 serve cases and 526 mild cases. The seconded experiment regarded a balance between serve and mild cases in DMET and did use 2412 entries, with 680 entries randomly drawn from ambulatory patients, who were supposed to be mild cases. We encoded serve cases as 1 and mild cases as 0.

The data set was divided in two disjoint and randomly chosen subsets. The larger subset with 80% of the data was used for training and cross validation, and respectively 20% remained for testing and evaluation of the final models on subsets of features.

All categorical features were encoded with a One Hot Encoding (OHE). This means that for each attribute of a feature a new feature was generated which remarked whether an entry had this attribute or not. Although OHE increases the dimensions of the input data, it is a promising encoding of categorical feature, which leads to good results [25], [26]. Furthermore all continuous features were normalised, when necessary. Equation (1) was used for normalisation with standard scaling to train the support vector machine (SVM) and neural network (NN).

$$\hat{x} = \frac{x - \mu}{\sigma} \tag{1}$$

Here x is the original input value of an attribute for an entry and $\hat{x}$ the value which an entry has for an attribute after the normalisation. Moreover μ describes the mean of each feature and σ the corresponding standard deviation. All numerical features are normalised so that after the normalisation the mean of a feature is 0 and the standard deviation 1, disregarding rounding errors due to computational inaccuracy. The normalisation is needed so that features with different scales become comparable for non-linear classifiers, i.e. neural networks and support vector machines.

C. Feature elimination

By the following criteria features were deleted in each data set:

- Too low variance or entropy
- High Pearson correlation coefficient to another feature
- Equality to another feature
- Information of the feature are contained by another feature

An attribute was removed due to too low variance if it contained at least 95% of the same value. Whereas for categorical features a threshold of 80% for one value was used. Although several investigations listed pregnancy as a risk factor [6], [10] for serve cases, we dismissed it as its variance was very low and an accidental correlation could not be excluded. Moreover Gao *et al.* mention in [10] that black patients and south asiens were found to have a higher mortality, which is also not represented in the dataset as all patients are from Mexico.

According to Akoglu in [27] a Pearson correlation coefficient p with $|p| > 0.75$ can be considered as a very strong correlation between two attributes in medicine. Hence $|p| > 0.75$ was used as threshold to obtain highly correlated feature pairs. One feature of each pair was removed.

All deleted features, including the reason for dropping, are listed for both experiments in Table III. Some features have only a high enough variance when just hospitalised patients were considered, which indicates that this attribute is a factor for hospitalisation and therefore for more serve cases

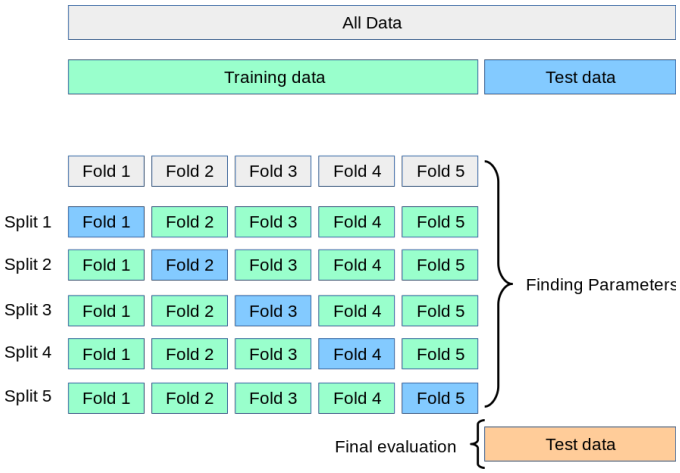

Fig. 1: Process of cross validated grid search

as well. This was not considered when executing the second experiment.

D. Model Selection

For both experiments a cross validated grid search was applied for hyper-parameter tuning of the following basic models. The models with the best found hyper-parameters were used to determine a feature ranking of all features.

Performing a grid search means that a set of parameters is chosen and for all possible permutations of values for each parameter a model is trained on a training set. Afterwards a test score is calculated on a held out testing set. Although this procedure is computationally expensive, it is easy to implement and shows good results [28], [29]. A cross validated grid search is an extension in which more than one testing score is considered [30]. In this application, the process of training and calculating a test score is repeated 5 times with different training and testing sets as seen in Figure 1 from [31] on 80% of the original data. We use more than one hold-out testing set to break accidental correlations between the train and test set leading to better performances. Split 1 to 5 correspond to one model, i.e. one possible set of hyper-parameters, trained on all green marked subsets, while the blue marked subset was used for calculating accuracy as testing score of the model. The final test score is an average of all 5 calculated test scores.

The following part explains the used model types and choice of searched parameters.

Random Forest

Random Forest (hereinafter RF) is a classification algorithm based on ensemble voting. Therefore a set S of decision trees is trained and the prediction done by a majority vote of all trees in S, with $|S| \in \{10 * i^2\}$ for $i \in [1, 7]$. To avoid overfitting a maximum depth $d \in \{3, 4, 5, 7, 9, 15, 20, 30\}$ is estimated for all trees in S. Additionally the splitting criteria of nodes in trees can be varied. Typical criteria are *entropy*, which implements *information gain*, and *gini coefficient* for splitting. As both criteria might select different attributes, both were used for hyper-parameter tuning [32].

Support Vector Machine

A support vector machine (SVM) algorithm aims to find a hyperplane in R^n, with n equal to the number of features, so that data points are distinctly classified and the margin from the hyperplane to both classes is maximised. When no linear hyperplane in R^n can be found, a SVM is extended with kernels to higher dimensions. Kernels can be regarded as a measure for similarity of two different data points x_i and x_j with $i \neq j, 1 \leq i, j, \leq m$, where m is the number of entries. Because the assumption of linear separable data seemed unrealistic, only *radial basis function* (hereinafter *rbf*) and $sigmoid$ kernel were used. They are calculated by the following Equations 2.

$$k_{rbf}(x_i, x_j) = e^{-\gamma||x_i - x_j||^2} \tag{2a}$$

$$k_{sigmoid}(x_i, x_j) = tanh(\gamma\langle x_i, x_j\rangle) \tag{2b}$$

This introduces some more hyper-parameters which need to be tuned. The parameters used for the grid search are C and γ. Here γ defines the influence of one training sample. Tested γ are given by Equation 3, where n again describes the amount of features and $var(X)$ calculates the variance of the input matrix X.

$$auto = \frac{1}{n * var(X)} \tag{3a}$$

$$scale = \frac{1}{n} \tag{3b}$$

Additionally different values for C are tried with an exponential increasing scale, i.e. $C \in \{2^i, i \in [-2, 4]\}$. C is intuitively a parameter for error regularisation. A large C value does encourage a higher accuracy, i.e. a lower margin, while a lower C value supports lower accuracy but a larger margin, respectively.

Neural Network

A neural network (NN) is a collection of connected nodes. They are ordered in layers and a NN consists of an input layer, an output layer and at least one hidden layer in between the two former. Every layer contains a certain amount of units, also called neurons, which are activated by an activation function to transmit information to the units of the next layer. Common activation functions are shown by Equation 4.

$$relu(x) = max(0, x) \tag{4a}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{4b}$$

$$softmax(z)_i = \frac{e_i^z}{\sum_{j=i}^{n} z_j} \tag{4c}$$

$softmax$ and $sigmoid$ are working similar, but $sigmoid$ as well as $relu$ are taking a tensor x as input while $softmax$ is calculated for every vector z separately. $softmax$ is especially desirable when you want to predict probabilities.

To improve the training process, initial weights can be set for each unit at each layer. This is done by so called kernel initialisation, which can be completely random, all zero or by drawing randomly from a certain distribution. The used initialisations are *glorot uniform*, *uniform* and *normal*. For *glorot uniform* initial weights are randomly drawn from a uniform distribution within $\left[-\sqrt{\frac{6}{n_{in}+n_{out}}}, \sqrt{\frac{6}{n_{in}+n_{out}}}\right]$, with n_{in} corresponding to the amount of input units in the weight tensor and n_{out} to the number of output units, i.e. the number of units of the current layer and, respectively, next layer [33]. This was proposed by Glorot *et al.* in [33] as this initialisation resulted in a desirable less varying variance between layers and preservation of signal when back propagating through the network. Furthermore *uniform* means that the weights are drawn randomly from a uniform distribution $\mathcal{U}(-0.05, 0.05)$ and for *normal* from a normal distribution $\mathcal{N}(0, 0.05)$.

Kingma and Ba present *adam* in [34] as an efficient algorithm for optimisation of stochastic objective functions, wherefore *adam* is included in the grid search for NN. Additionally *rmsprop* was tried as optimisation function, which was proposed by Hinton [35].

Beside kernel initialisation, activation and optimisation function also the batch size and epoch was tuned for NN. A batch is a set of samples, which are processed through to the network independently at one time and afterwards an update step is performed with an average stochastic gradient from all samples in a batch. The batch size is consequently the number of samples a batch contains. Batches lead to more robustness with respect to hyper-parameters. An epoch is one entire pass of all training data through the NN. Logically the number of epochs is how many times the data pass the NN during its training process.

As neural networks contain a huge amount of hyper-parameters for tuning, a parameter hunt for the activation function was performed before applying grid searching. All hunted NN consisted of 3 layers, one input layer, one hidden layer with 256 units and an output layer. In total 4 NN were tested, two with $relu$ as activation function and once $sigmoid$, once $softmax$ as output. The other two had $sigmoid$ as activation and $softmax$ as output, and $softmax$ as activation and $sigmoid$ as output function.

E. Feature Ranking

A feature ranking was obtained from rankings of 5 different model. Additionally to rankings by feature importance from the 3 models described in the previous part, rankings with ANOVAs f-value and chi-square (χ^2) were used.

The ranking from the RF based on the feature importance, i.e. the mean decrease in impurity of each feature. This is not simply applicable for NN and SVM, except when using a SVM with a linear kernel, because more dimensional models are constructed. Consequently a permutation importance was used to determine the importance of a feature. This means that first a baseline metric with all features was calculated and afterwards each feature was removed or permuted and the metric calculated again. The difference between baseline metric and newly calculated metric can be interpreted as feature importance, i.e. a large difference means a feature is important, and a small difference corresponds to an unimportant feature, respectively. As this is computationally expensive this process was only performed for NN and SVM and not for RF.

Because the resulting rankings are only indicators for how important a feature is for a certain model, they were merged into one ranking based on a score to obtain a more general applicable ranking. The score of a feature was calculated by the sum of ranks from all 5 rankings. This means the minimal score is 5, and the maximal score 375. Afterwards all features got ordered increasingly by their score and the feature with the lowest score is regarded as most important. Table I as well as Table II give an overview of the 40 most important features and their scores for both experiments.

F. Model Evaluation on Subsets of Features

Finally subsets of features were chosen to train the models again and evaluate their performance afterwards on a, through the previous process completely unseen, testing set which contained 20% of the data. Starting with the 2 most important features from the final ranking 3 features were added until a maximum of 38 features was reached. The added features were chosen in order of their positions in the ranking from Table I and Table II for the respective experiment. To evaluate a model several metrics were used which are elaborated below.

The predicted class and actual class are taken as input. For each entry one of the following states is possible.

- *true positives (tp):* A patient is predicted as a serve case and actually is a serve case.
- *false positives (fp):* A patient is predicted as a serve case, but actually is not a serve case.
- *true negatives (tn):* A patient is predicted as a mild case and actually is a mild case.
- *false negatives (fn):* A patient is predicted as a mild case, but actually is a serve case.

In the following *tp, fp, tn* and *fn* remark the amount of occurrences from a performed prediction and corresponding actual value.

The used metrics to evaluate the resulting models are accuracy (ACC), recall (R), precision (P) and F1 score ($F1$). They are calculated by Equations 5 with the amount of *tp, fp, tn* and *fn* given through prediction and actual value. Each function has a maximum value of 1, which is best, and a minimum of 0, which is respectively worst. A model is considered good, when most metrics are close to 1.

$$ACC = \frac{tp + tn}{tp + tn + fp + fn} \tag{5a}$$

$$R = \frac{tp}{tp + fn} \tag{5b}$$

$$P = \frac{tp}{tp + fp} \tag{5c}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{5d}$$

Accuracy (ACC) describes the percentage of correct predictions in total. Recall corresponds to the percentage of correct positive classifications from actual positives, whereas precision P is the percentage of correct positive classifications among the predictions. $F1$ is consequently the weighted average of P and R.

Additionally the area under the receiver operating characteristic curve (AUC) was calculated. AUC can be understood as the probability for a correct prediction from a randomly drawn example. Here a value of 0.5 corresponds to a random process, whereas 1 and 0 describe a perfect model.

To visualise the influence of more features on the quality of models all metrics are plotted in Fig. 2 for the first experiment and Fig. 3 for the second experiment, respectively.

IV. Results

A. Experiment 1

The first experiment did train models on 1732 hospitalised patients to determinate a risk for a serve case.

The performed grid search delivered the following results for RF, SVM and NN. In general all RF with a maximum depth greater than 9 and more than 10 trees performed decent. Unsurprisingly the combination of $d = 30$ and $|S| = 10$ performed worst due to overfitting. The best performance was achieved with a maximum depth $d = 15$, a set size of $|S| = 90$ and $entropy$ as a splitting criterion as parameters. For the SVM all combinations of γ and kernel showed a trend for $C = 1$, which indicates that a low C did have a too large margin whereas the larger C over fitted the training set. The results of the SVM were best with parameters as follows $C = 1$, $\gamma = scale$ and *kernel=sigmoid*. During hunting the network with 2 *relu* layers and *softmax* output achieved the highest *AUC* and was therefore chosen for grid searching. A batch size of 50 and an epoch size of 10 achieved the best results in every setting of optimiser and initialisation. But it worked best with a *glorot uniform* initialisation, and *rmsprop* as an optimiser.

A feature ranking of all features was generated as mentioned in section III. It is interesting to notice that INDICADOR_SOSP_COVID (engl. indicator for suspected COVID-19) was the most important feature as it encodes whether COVID-19 was suspected or not. This indicates that an early diagnosis whether an infection is really SARS-CoV-2 or not might be very important to predict the severity of a case. Furthermore it is remarkable that RINORREA (engl. rhinorrhoea) seems to be an important feature and symptom instead of TOS (engl. cough) or FIEBRE (engl. fever). This might be, because cough and fever are typical symptoms and occur a lot in mild cases as well [5]. Features like EDAD (engl. age) or DISNEA (engl. dyspnea) are important, but not as much as could have been expected [6]. On the other side ANTECED_OBESIDAD (engl. obesity) and ENFERMEDAD_CRONICA (engl. chronicle disease) are of similar importance as dyspnea, which supports our idea to predict the severity of cases based on comorbidities.

The 40 most important features from totally 75 features are listed in Table I including their rank and score. It needs to be noted, that features like DIAGNOSTICO_FINAL (engl. final diagnosis) or ESTANCIA_HOSP (engl. hospitalisation) should be regarded carefully as they can change or are decided during the infection.

Rank	Score	Feature
1	27	INDICADOR_SOSP_COVID
2	34	RINORREA
3	36	POSTRACION
4	37	DIAG_CLIN_NEUMONIA
5	41	DESC_RESULTADO_CONF2_NEGATIVO
6	42	DESC_TIPO_MUESTRA_1_EXUDADO FARINGEO
7	43	DESC_TIPO_MUESTRA_1_EXUDADO FARINGEO/NASOFARINGEO
8	58	ESTANCIA_HOSP_MEDICINA INTERNA
9	71	NEUMONIA_RADIOGRAFIA
10	74	EDAD
11	76	DESC_ESTATUS_CONF1_VALIDADA
12	81	ESTANCIA_HOSP_URGENCIAS ADULTOS
13	98	ESTANCIA_HOSP_NEUMOLOGIA
14	108	INICIO_SUBITO
15	118	ANTECED_OBESIDAD
16	131	OCUPACION_Médicos
17	132	DOLOR_TORACICO
18	132	GENERO
19	134	ESCALOFRIO
20	136	DISNEA
21	136	DESC_RESULTADO_CONF2_POSITIVO
22	137	ENFERMEDAD_CRONICA
23	140	CEFALEA
24	141	DESC_TIPO_MUESTRA_1_ EXUDADO NASOFARINGEO
25	141	DESC_ESTATUS_CONF1_POR RECIBIR
26	143	OCUPACION_Enfermeras
27	143	ODINOFAGIA
28	146	DIARREA
29	150	DOLOR_ABDOMINAL
30	156	OCUPACION_Otras Ocupaciones
31	163	CIANOSIS
32	178	DIAGNOSTICO_FINAL_COVID-19
33	181	ANTECED_HIPERTENSION
34	185	DIAGNOSTICO_FINAL_Neumonia atipica
35	189	OCUPACION_Choferes
36	190	FIEBRE
37	191	RESULTADO_DE MUESTRA1
38	196	DIAGNOSTICO_FINAL_Influenza con otras manifestaciones
39	199	ATAQUE_AL_ESTADO_GENERAL
40	199	OCUPACION_Ama de casa

TABLE I: Top 40 features of final ranking for experiment 1

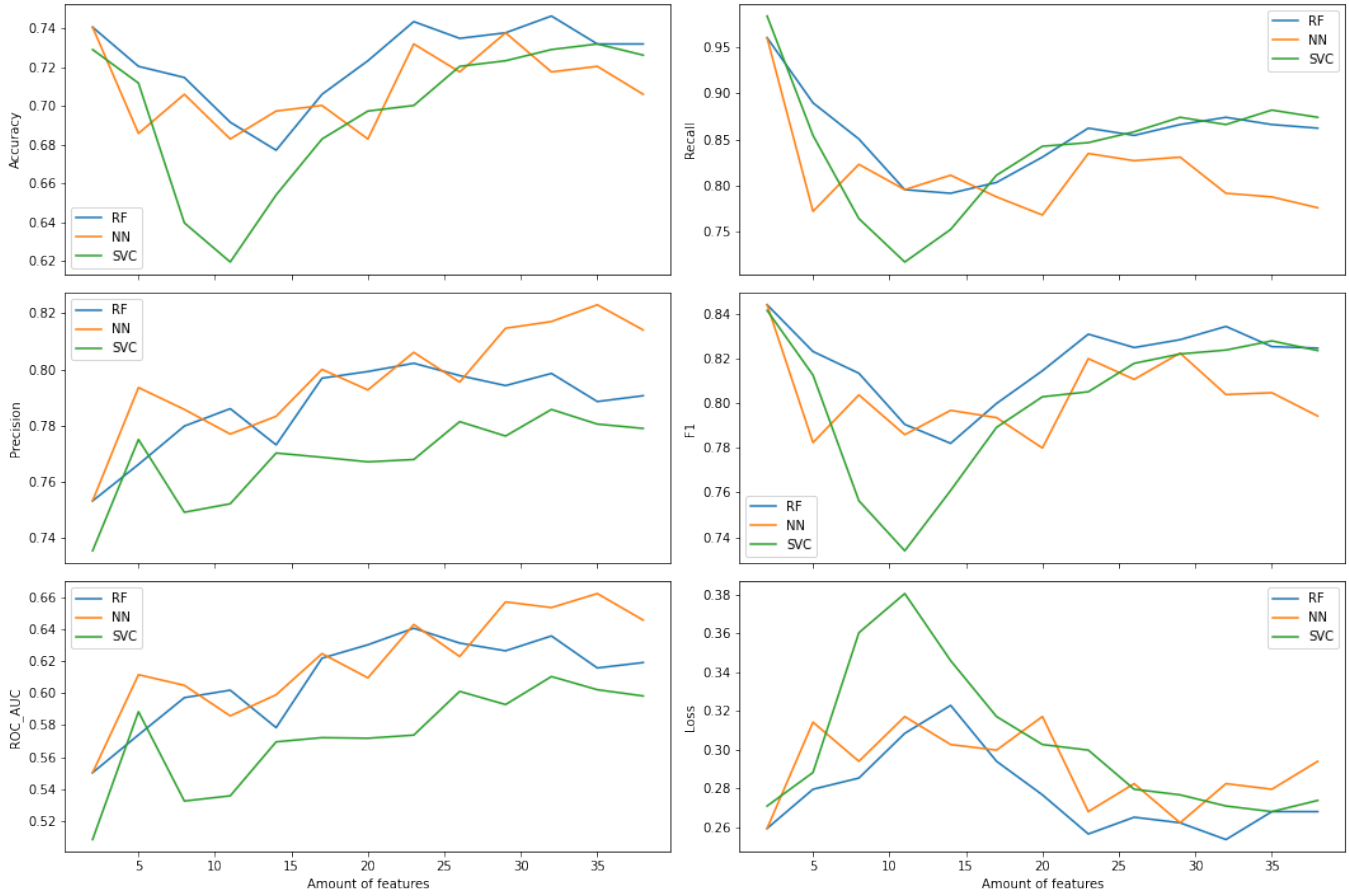

Fig. 2: Different metrics for each model trained on a subset of features for experiment 1

Finally RF, SVM and NN were trained on feature subsets regarding the ranking given by Table I. The results are plotted in Fig. 2 for each metric and model in dependence of the amount of used features.

Surprisingly some features introduce noise in all models when more than 8 and less than 15 features are used as can be seen in the huge peak downwards in Fig. 2. This might be explained by a non-linear correlation between used features and other features which are added later. Especially the SVM seems to be affected as can be observed in Fig. 2. This setting has an impact on the results of RF and NN as well, but not as strong as for the SVM. Furthermore R (5b) is very close to 1 when just INDICADOR_SOSP_COVID and RINORREA are used. This means, that only few *false negatives* were predicted. On the contrary P (5c) is not very high, which leads to the conclusion that the model predict a lot *false positives*. This behaviour does not have to be undesirable as will be discussed later in section V.

The SVM performs poorly regarding P (5c) and AUC. Additionally it does not perform remarkably good in any of the other metrics and has the highest loss. Therefore SVM can be ruled out as a suitable model. The NN on the other hand clearly out performs SVM and RF in P (5c) and AUC when more than 25 features are used. In other metrics it is often a bit inferior to the RF and often superior to the SVM. Finally the RF has the best metrics regarding ACC (5a), $F1$ (5d) and the lowest loss. Furthermore RF is often superior or at least equal to the NN. This observation leads to the statement that the RF is best to predict the severity of a case for a hospitalised patient whereas the SVM is the least appropriate model.

B. Experiment 2

The second experiment trained models on hospitalised and ambulatory patients to estimate a risk for a serve case on a balanced dataset with 2412 entries, so that the amount of serve cases equals the amount of mild cases.

The grid search showed that for the balanced dataset the depth of the RF seemed to have less influence on its performance than in the first experiment. Instead the amount of estimators was the main thriving factor to increase the testing score ACC (5a). Additionally no huge difference between the splitting criteria could be noted. The best parameters for RF are $|S| = 490$, a maximum depth of $d = 20$ and $entropy$ as a splitting criteria. Contrary to experiment 1 a clear trend for $C = 1$ could not be observed for SVMs, although $C \geq 8$ resulted in overfitting and respectively worse performance on the hold-out set. SVM performed best on a balanced dataset with $C = 2$, $\gamma = auto$ and a rbf kernel. The hunting search resulted in a network using *sigmoid* as activation function and again *softmax* as output function. The following grid search for NN displayed a clear trend towards *rmsprop* as optimiser, a

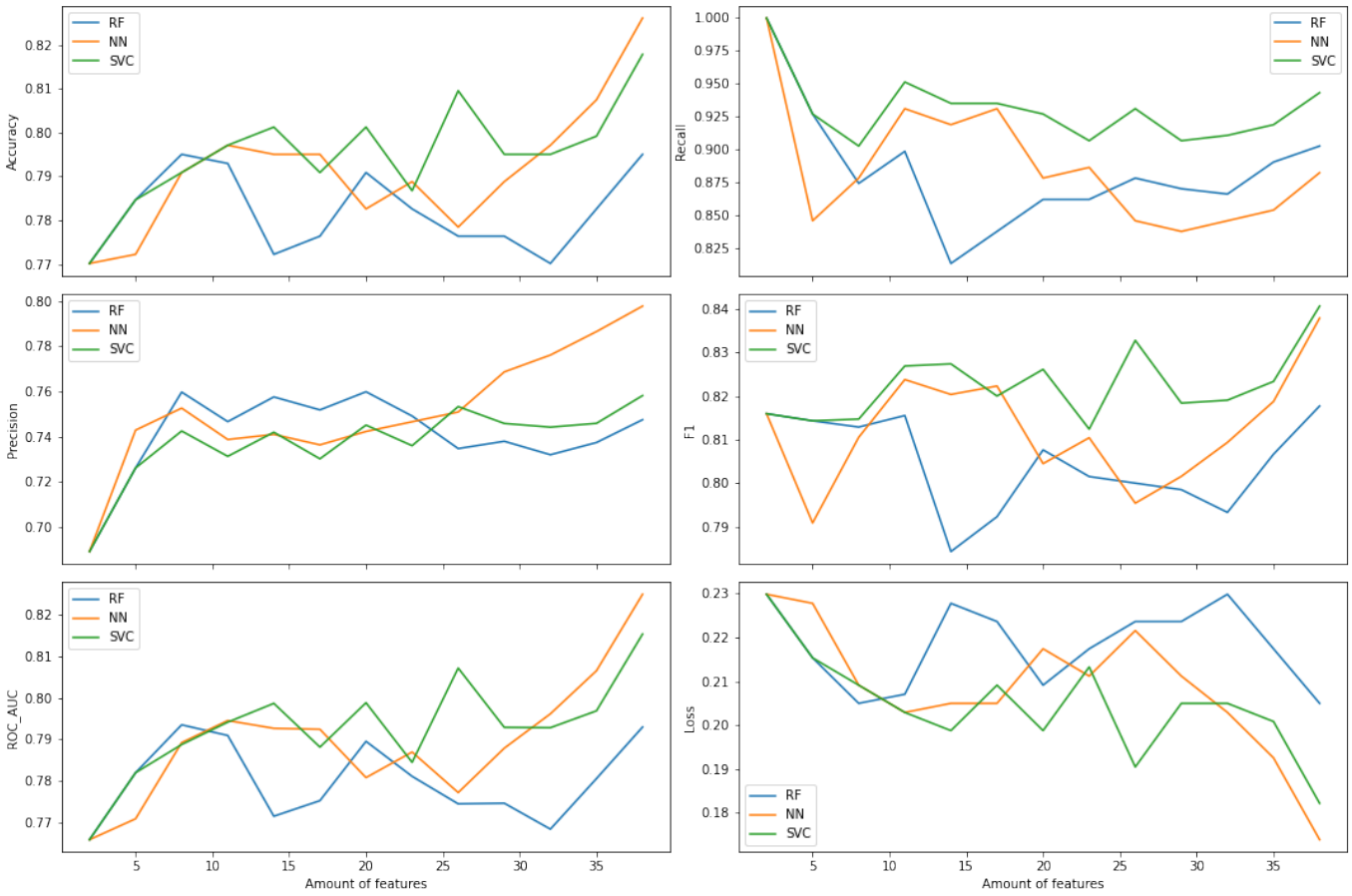

Fig. 3: Different metrics for each model trained on a subset of features for experiment 2

batch size of 20 and an epoch of 100 could be seen. Initialising weights from a *normal* distribution achieved the best results.

The final ranking of features was obtained by the procedure described in section III. The 40 most important features are displayed by Table II.

Aligning with our expectations ESTANICA_HOSP_NO HOSP (engl. no hospitalisation) is very important, because about 28% of the samples are ambulatory patients, which are counted as mild cases. Also the second feature is not really surprising, because a clinical diagnosed pneumonia definitely can indicate a serve infection.

Compared to experiment 1 the age of the patient got less important, which is surprising. But age is ranked as very unimportant for the SVM (rank 74 of 75), which at least explains the bad score. All other models ranked age as important. Moreover dyspnea got more important, which aligns with the characteristics of serve COVID-19 infections [5], [6]. On the other hand INDICADOR_SOSP_COVID and RINORREA, which were the most important feature in the first experiment, got less important.

For the second experiment all metrics are almost perfectly increasing with a higher amount of features. In 5 of 6 metrics the SVM is superior or at least equal to the RF and NN as can be seen in Fig. 3. Only P (5c) of SVM is inferior until more than 26 feature are used. Although a low P does not have to be undesirable when R is adequate.

Interestingly the RF and SVM show the same behavior of overestimating serve cases with very few features as we saw in Fig. 2. On the other hand the peak in between 8 and 15 features, which occurred in the first experiment, disappeared.

Furthermore the NN performs worst with only a few features and later, with more than 30 features better than the 2 other models. The bad performance with just a few features can be explained by overfitting as a lot of parameters need to be trained, but only few features and samples are given. Consequently the given samples can be matched perfectly, but on the unseen testing set new situations are not matched accordingly.

In general this experiment has a better quality and greater significance as it was trained and tested with more samples. Looking at the performance of the SVM, this is clearly the most suitable model to make a decision for this setting.

V. Discussion and Future Work

A. Discussion

In [36], [37], and [38] machine learning approaches were used to predict the severity of cases and assignment of ICU. They used laboratory test results, clinical reports and CT images. Although they achieved very good results, none of them did take comorbidities into account and therefore differ from our work. Additionally all models were trained on smaller data

Rank	Score	Feature
1	7	ESTANCIA_HOSP_NO HOSP
2	26	DIAG_CLIN_NEUMONIA
3	46	DESC_TIPO_MUESTRA_1_EXUDADO FARINGEO/NASOFARINGEO
4	48	NEUMONIA_RADIOGRAFIA
5	61	ESTANICA_HOSP_MEDICINA INTERNA
6	82	DOLOR_TORACICO
7	86	DISNEA
8	94	ANTECED_DIABETES
9	94	DIAGNOSTICO_FINAL_COVID-19
10	104	OCUPACION_Jubilado
11	104	RINORREA
12	105	ESTANICA_HOSP_URGENCIAS ADULTOS
13	109	DIAGNOSTICO_FINAL_Neumonia atipica
14	113	EDAD
15	114	ANTECED_HIPERTENSION
16	126	DESC_ESTATUS_CONF1_VALIDADA
17	135	ENFERMEDAD_CRONICA
18	135	DESC_RESULTADO_CONF2_POSITIVO
19	135	RESULTADO_DE MUESTRA1
20	140	CIANOSIS
21	140	POSTRACION
22	140	ANTECED_OBESIDAD
23	140	ANTECED_RENAL
24	141	DESC_TIPO_MUESTRA_1_EXUDADO NASOFARINGEO
25	143	DESC_TIPO_MUESTRA_1_EXUDADO FARINGEO
26	146	OCUPACION_Ama de casa
27	151	DESC_RESULTADO_CONF2_NEGATIVO
28	154	ODINOFAGIA
29	154	ANTECED_EPOC
30	157	INICIO_SUBITO
31	161	DIARREA
32	166	FIEBRE
33	167	OCUPACION_Otras Ocupaciones
34	169	OCUPACION_Sin ocupación
35	170	INDICADOR_SOSP_COVID
36	174	MIALGIAS
37	183	GENERO
38	184	OCUPACION_Médicos
39	185	ATAQUE_AL_ESTADO_GENERAL
40	185	CEFALEA

TABLE II: Top 40 features of final ranking for experiment 2

sets containing mainly clinical test results, whereas our models rely mainly on common and obvious symptoms.

Jojoa Acosta and Garcia-Zapirain proposed successfully a multilayered perceptron (MLP) and SVM in [39] to predict the number of new daily infections in America. This tool is useful for decision making at public health strategies and organisation of hospitals in advance. Similar to this work a model for epidemic spread of COVID-19 was presented in [40] by Hosseini *et al.*, so governments can take action accordingly. Our work can be regarded as an extension to help organising hospitalisation further and vary not only because of another geographical location, but its ability to predict severity of hospitalised cases.

In [41] and [42] applications to detect COVID-19 infections through cough sounds or chest X-ray images are presented. Their work produces great results with accuracies over 89.2%. [41] differentiates between *healthy* and *COVID-19*, but not between COVID-19 and other respiratory diseases, whereas [42] considers multi classifier with *COVID-19*, *pneumonia* and *healthy*. The final diagnosis of COVID-19 an important feature for both our models. Ideally an early and precise diagnosis of COVID-19 would also improve the quality of our models and the applied health care. If not only the diagnosis is known early, but moreover the severity of infection, further appropriate actions can be taken.

In [43] another ML approach is presented, where patient's basic information and clinical data were used for predicting whether an infection is a COVID-19 case or not. Because all models on the original data set performed insufficient, a Generative Adversarial Network (GAN) was used to generate a balanced dataset. This supports our approach for the second experiment. Training with a balanced dataset achieved much better results in [43], but as other works mentioned before did not take comorbidities and prevalent diseases into account. Furthermore the works differs from ours as it predicts whether an infection is COVID-19 or not.

Although other models perform better regarding prediction of a COVID-19 diagnosis, our models were trained on a larger amount of data from a longer time period and are therefore better for generalisation. Additionally our prediction based more on previous diseases, comorbidities and obvious symptoms of a patient whereas information like CT images or results from a huge variety of laboratory tests were not included in detail. Our models are not computationally expensive and can be trained easily again on a larger amount of data as well as adjusted to deal with data sets with further features to improve the predictions.

B. Future work

Possible extensions of this work could be done by repeating the experiments with more data collected over a longer time period or different locations. This might result in a more representative evaluation of the actual importance of time features as SARS-CoV-2 seems to be correlated to seasonal behaviour, but also regarding environmental influences and accessibility of health institutions [21–24]. Additionally aspects like SARS-CoV-2 mutations, times of vaccination and date of last vaccination could be included and improve the reliability of the model by now.

Another direct extension of our work would be to extended the classifiers to multiclassing and differentiate between mild cases which need no treatment at all, only treatment at home or treatment at the hospital. Especially the differentiation between no treatment and treatment at home cases could be challenging for a statistical model. But this is probably less important to physicians and consequently could be ignored.

We noticed that both experiments resulted in high *recall R* when the models were trained on only a few features. When *precision P* is not equal to *accuracy ACC*, i.e. not everything is predicted as a serve case, this can be desirable as this models is good at predicting serve cases. On the other hand if a shortage of ICU units occurs, this behaviour is unsuitable. For this case a model to predict the probability of death in case of lack of sufficient treatment may be a better fit.

How suspiciously COVID-19 infections, whose SARS-CoV-2 diagnosis is ruled out in the end, influence the organisation of hospitalisation flows and help avoiding an overload of the

health system is discussed by Rogier *et al.* [44]. According to Rogier *et al.* especially the time of being symptomatic until hospitalisation is shorter for COVID-19-negative patients [44] . Applying a machine learning technique and developing a simple application to recognize COVID-19 cases earlier would help to organise hospitalisation as well as necessary treatment of patients as proposed by [41], [43]. Here again a dataset over a longer time period would be helpful to determinate how important time features, like the time period between first symptoms and hospitalisation, are.

Additionally varying treatment and regularisation depending on the mutation would be more clear with an application predicting not only COVID-19 or not COVID-19, but also the possible mutation or other respiratory infections regarding the symptoms and results of laboratory tests.

VI. Conclusion

It needs to be noted that there are certain limitations for this research. First of all only data from the first 6 months of the pandemic were collected, where COVID-19 was still relatively new to treat and no vaccinations available. Additionally the periodical behaviour of the disease was unknown and is not represented in the data [21–24]. Although our data set is larger than for most other research in this topic, it is still relatively small and more data would increase the significance and quality of the models. Furthermore it should be regarded that by now a lot of mutations did occur, which immensely influence the course of infection as well as infection rate [45].

Our model can be helpful for early clinical decision making, because of its good results, usage of common symptoms and prevalent diseases. It is not computationally expensive and can be understood intuitively. Although the results are not as precise as desirable, both models perform well and can provide in combination with diagnosing tools support for health care decisions. Beside this the ranking of feature importance aligns with current research on COVID-19 risk factors, which supports the reliability of our model and provides help for manual decision making. For future work the main focus is retraining and extending our models to perform multiclassing and consider further features from a larger dataset.

References

[1] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, null null, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, and S. V. Scarpino, "The effect of human mobility and control measures on the covid-19 epidemic in china," *Science*, vol. 368, no. 6490, pp. 493–497, 2020.

[2] J. Huang, L. Zhang, X. Liu, Y. Wei, C. Liu, X. Lian, Z. Huang, J. Chou, X. Liu, X. Li, K. Yang, J. Wang, H. Liang, Q. Gu, P. Du, and T. Zhang, "Global prediction system for covid-19 pandemic," *Science Bulletin*, vol. 65, no. 22, pp. 1884–1887, 2020.

[3] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, and W. Tan, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, 2020. PMID: 31978945.

[4] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui, B. Du, L.-j. Li, G. Zeng, K.-Y. Yuen, R.-c. Chen, C.-l. Tang, T. Wang, P.-y. Chen, J. Xiang, S.-y. Li, J.-l. Wang, Z.-j. Liang, Y.-x. Peng, L. Wei, Y. Liu, Y.-h. Hu, P. Peng, J.-m. Wang, J.-y. Liu, Z. Chen, G. Li, Z.-j. Zheng, S.-q. Qiu, J. Luo, C.-j. Ye, S.-y. Zhu, and N.-s. Zhong, "Clinical characteristics of coronavirus disease 2019 in china," *New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[5] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, "Characteristics of sars-cov-2 and covid-19," *Nature Reviews Microbiology*, vol. 19, 03 2021.

[6] C. Wu, X. Chen, Y. Cai, J. Xia, X. Zhou, S. Xu, H. Huang, L. Zhang, X. Zhou, C. Du, Y. Zhang, J. Song, S. Wang, Y. Chao, Z. Yang, J. Xu, X. Zhou, D. Chen, W. Xiong, L. Xu, F. Zhou, J. Jiang, C. Bai, J. Zheng, and Y. Song, "Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China," *JAMA Internal Medicine*, vol. 180, pp. 934–943, 07 2020.

[7] H. M. Zawbaa, A. El-Gendy, H. Saeed, H. Osama, A. M. A. Ali, D. Gomaa, M. Abdelrahman, H. S. Harb, Y. M. Madney, and M. E. A. Abdelrahim, "A study of the possible factors affecting covid-19 spread, severity and mortality and the effect of social distancing on these factors: Machine learning forecasting model," *International Journal of Clinical Practice*, vol. 75, no. 6, p. e14116, 2021.

[8] A. K. Upadhyay and S. Shukla, "Correlation study to identify the factors affecting covid-19 case fatality rates in india," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 3, pp. 993–999, 2021.

[9] B. Gallo Marin, G. Aghagoli, K. Lavine, L. Yang, E. J. Siff, S. S. Chiang, T. P. Salazar-Mather, L. Dumenco, M. C. Savaria, S. N. Aung, T. Flanigan, and I. C. Michelow, "Predictors of covid-19 severity: A literature review," *Reviews in Medical Virology*, vol. 31, no. 1, p. e2146, 2021.

[10] Y.-d. Gao, M. Ding, X. Dong, J.-j. Zhang, A. Kursat Azkur, D. Azkur, H. Gan, Y.-l. Sun, W. Fu, W. Li, H.-l. Liang, Y.-y. Cao, Q. Yan, C. Cao, H.-y. Gao, M.-C. Brüggen, W. van de Veen, M. Sokolowska, M. Akdis, and C. A. Akdis, "Risk factors for severe and critically ill covid-19 patients: A review," *Allergy*, vol. 76, no. 2, pp. 428–455, 2021.

[11] M. Bordbar, A. Sanaei Dashti, A. Amanati, E. Shorafa, Y. Mansoori, S. J. Dehghani, and H. Molavi Vardanjani, "Assessment of the hscore as a predictor of disease outcome in patients with covid-19," *BMC Pulmonary Medicine*, 10 2021.

[12] L. Fardet, L. Galicier, O. Lambotte, C. Marzac, C. Aumont, D. Chahwan, P. Coppo, and G. Hejblum, "Development and validation of the hscore, a score for the diagnosis of reactive hemophagocytic syndrome," *Arthritis & Rheumatology*, vol. 66, no. 9, pp. 2613–2620, 2014.

[13] A. Balkhair, M. Al Jufaili, K. Al Wahaibi, D. Al Riyami, F. Al Azri, S. Al Harthi, M. Al Busaidi, S. Al Mubaihsi, Z. Al Muharrmi, N. Al Riyami, Z. Al Belushi, R. Abdawani, A. Al Hashar, A. Al Mahrezi, K. Al Maamari, I. Al Busaidi, Z. Al Hinai, F. B. Alawi, H. B. Taher, and M. Al Jabri, ""virtual interdisciplinary covid-19 team": A hospital pandemic preparedness approach.," *Oman medical journal*, vol. 35, 11 2020.

[14] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[15] M. F. J. Acosta, L. Y. C. Tovar, M. B. Garcia-Zapirain, and W. S. Percybrooks, "Melanoma diagnosis using deep learning techniques on dermatoscopic images," *BMC Medical Imaging*, vol. 21, 2021.

[16] S. Zahia, B. Garcia-Zapirain, I. Saralegui, and B. Fernández-Ruanova, "Dyslexia detection using 3d convolutional neural networks and functional magnetic resonance imaging," *Computer methods and programs in biomedicine*, vol. 197, p. 105726, 2020.

[17] S. I. A. Al-Janabi, B. Al-Khateeb, M. Mahmood, and B. Garcia-Zapirain, "An enhanced convolutional neural network for covid-19 detection," *Intelligent Automation & Soft Computing*, vol. 28, no. 2, pp. 293–303, 2021.

[18] Z. Hameed and B. Zapirain, "Sentiment classification using a single-layered bilstm model," *IEEE Access*, vol. 8, pp. 73992–74001, 04 2020.

[19] O. J. Bastidas, B. Garcia-Zapirain, A. L. Totoricagüena, S. Zahia, and J. U. Carpio, "Feature analysis and prediction of complications in ostomy patients based on laboratory analytical data using a machine learning approach," in *2021 International Conference BIOMDLORE*, pp. 1–8, 2021.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duch-

esnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[21] X. Liu, J. Huang, C. Li, Y. Zhao, D. Wang, Z. Huang, and K. Yang, "The role of seasonality in the spread of covid-19 pandemic," *Environmental Research*, vol. 195, p. 110874, 2021.
[22] A. Mirahmadizadeh, F. Rezaei, K. Jokari, L. Moftakhar, A. Hemmati, S. S. Dehghani, A. H. Hassani, M. Lotfi, A. Jafari, and M. Ghelichi-Ghojogh, "Correlation between environmental factors and covid-19 indices: a global level ecological study," *Environmental Science and Pollution Research*, vol. 29, 01 2022.
[23] A. Kaplin, C. Junker, A. Kumar, M. A. de Amorim Ribeiro, E. Yu, M. Wang, T. Smith, S. N. Rai, and A. Bhatnagar, "Evidence and magnitude of seasonality in sars-cov-2 transmission: Penny wise, pandemic foolish?," *medRxiv*, 2020.
[24] Z. gang Fang, S. qin Yang, C. xia Lv, S. yi An, P. Guan, D. Huang, B. Zhou, and W. Wu, "The correlation between temperature and the incidence of covid-19 in four first-tier cities of china: a time series study," *Environmental Science and Pollution Research International*, pp. 1 – 10, 2022.
[25] Z. Lv, H. Ding, L. Wang, and Q. Zou, "A convolutional neural network using dinucleotide one-hot encoder for identifying dna n6-methyladenine sites in the rice genome," *Neurocomputing*, vol. 422, pp. 214–221, 2021.
[26] S. Bagui, D. Nandi, S. Bagui, and R. White, "Machine learning and deep learning for phishing email classification using one-hot encoding," *Journal of Computer Science*, vol. 17, pp. 610–623, 07 2021.
[27] H. Akoglu, "User's guide to correlation coefficients," *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018.
[28] J.-H. Han, D.-J. Choi, S.-U. Park, and S.-K. Hong, "Hyperparameter optimization using a genetic algorithm considering verification time in a convolutional neural network," *Journal of Electrical Engineering & Technology*, vol. 15, 03 2020.
[29] L. Yao, Z. Fang, Y. Xiao, J. Hou, and Z. Fu, "An intelligent fault diagnosis method for lithium battery systems based on grid search support vector machine," *Energy*, vol. 214, p. 118866, 2021.
[30] A. Panigrahi and M. R. Patra, "Chapter 6 - network intrusion detection model based on fuzzy-rough classifiers," in *Handbook of Neural Computation* (P. Samui, S. Sekhar, and V. E. Balas, eds.), pp. 109–125, Academic Press, 2017.
[31] Serveral, "Scikit-learn documentation: 3.1 cross-validation: evaluating estimator performance." Accessed on 07.02.2022.
[32] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, 2004.
[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 05 2010.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
[35] G. Hinton, "Neural networks for machine learning - lecture," 2012.
[36] A. Salama, "A hybrid two-phase machine learning model for early covid-19 diagnosis prediction," *International Journal of Computer Applications*, vol. 174, pp. 38–49, 04 2021.
[37] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, "Machine learning based approaches for detecting covid-19 using clinical text data," *International Journal of Information Technology*, vol. 12, 09 2020.
[38] M. Chieregato, F. Frangiamore, M. Morassi, C. Baresi, S. Nici, C. Bassetti, C. Bnà, and M. Galelli, "A hybrid machine learning/deep learning covid-19 severity predictive model from ct images and clinical data," 2021.
[39] M. F. Jojoa Acosta and B. García-Zapirain Soto, "Machine learning algorithms for forecasting covid 19 confirmed cases in america," *2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 1–6, 2020.
[40] E. Hosseini, K. Z. Ghafoor, A. S. Sadiq, M. Guizani, and A. Emrouznejad, "Covid-19 optimizer algorithm, modeling and controlling of coronavirus distribution process," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2765–2775, 2020.
[41] R. Islam, E. Abdel-Raheem, and M. Tarique, "A study of using cough sounds and deep neural networks for the early detection of covid-19," *Biomedical Engineering Advances*, vol. 3, p. 100025, 2022.
[42] P. Bhowal, S. Sen, J. H. Yoon, Z. W. Geem, and R. Sarkar, "Choquet integral and coalition game-based ensemble of deep learning models for covid-19 screening from chest x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 12, pp. 4328–4339, 2021.
[43] L. Wang, H. Shen, K. Enfield, and K. Rheuban, "Covid-19 infection detection using machine learning," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4780–4789, 2021.
[44] T. Rogier, I. Eberl, F. Moretto, T. Sixt, F.-X. Catherine, C. Estève, M. Abdallahoui, L. Behague, A. Coussement, L. Mathey, S. Mahy, M. Buisson, A. Salmon-Rousseau, M. Duong, P. Chavanet, Q. Bernard, B. Nicolas, L. Benguella, B. Bonnotte, and L. Piroth, "Covid-19 or not covid-19? compared characteristics of patients hospitalized for suspected covid-19," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 40, 09 2021.
[45] B. Cosar, Z. Y. Karagulleoglu, S. Unal, A. T. Ince, D. B. Uncuoglu, G. Tuncer, B. R. Kilinc, Y. E. Ozkan, H. C. Ozkoc, I. N. Demir, A. Eker, F. Karagoz, S. Y. Simsek, B. Yasar, M. Pala, A. Demir, I. N. Atak, A. H. Mendi, V. U. Bengi, G. Cengiz Seval, E. Gunes Altuntas, P. Kilic, and D. Demir-Dora, "Sars-cov-2 mutations and their viral variants," *Cytokine & Growth Factor Reviews*, 2021.

APPENDIX

Experiment 1	Experiment 2	Reason for removal
F_REGISTRO_DEFUNCION	F_REGISTRO_DEFUNCION	contained by other feature
FECHA_DEFUNCION	FECHA_DEFUNCION	contained by other feature
FECHA_MUESTRA_CONF2	FECHA_MUESTRA_CONF2	contained by other feature
DESC_TIPO_MUESTRA_2	DESC_TIPO_MUESTRA_2	contained by other feature
DESC_ESTATUS_CONF2	DESC_ESTATUS_CONF2	contained by other feature
DESC_TIPO_INFLUENZA_CONF1	DESC_TIPO_INFLUENZA_CONF1	contained by other feature
DESC_TIPO_INFLUENZA_CONF2	DESC_TIPO_INFLUENZA_CONF2	contained by other feature
TIPO_INFLUENZA_CONF1	TIPO_INFLUENZA_CONF1	equal to other feature
TIPO_INFLUENZA_CONF2	TIPO_INFLUENZA_CONF2	equal to other feature
TIPO_MUESTRA_CONF_1	TIPO_MUESTRA_CONF_1	equal to other feature
TIPO_MUESTRA_CONF_2	TIPO_MUESTRA_CONF_2	equal to other feature
RESULTADO_DE MUESTRA2	RESULTADO_DE MUESTRA2	equal to other feature
CIE10_DIAGNOSTICO_FINAL	CIE10_DIAGNOSTICO_FINAL	equal to other feature
DIAGNOSTICO_EGRESO_1	DIAGNOSTICO_EGRESO_1	equal to other feature
DIAGNOSTICO_EGRESO_2	DIAGNOSTICO_EGRESO_2	equal to other feature
DIAGNOSTICO_EGRESO_3	DIAGNOSTICO_EGRESO_3	equal to other feature
CONGESTION_NASAL	CONGESTION_NASAL	too low variance
DISFONIA	DISFONIA	too low variance
LUMBALGIA	LUMBALGIA	too low variance
ANTECED_ASMA	ANTECED_ASMA	too low variance
ANTECED_INMUNOSUPRESION	ANTECED_INMUNOSUPRESION	too low variance
ANTECED_VIH_EVIH	ANTECED_VIH_EVIH	too low variance
EMBARAZO	EMBARAZO	too low variance
LACTANCIA	LACTANCIA	too low variance
PUERPERIO	PUERPERIO	too low variance
	TIENE_INTUBACION_ENDOTRAQUEAL	too low variance
	ANTECED_CARDIOVASCULAR	too low variance
ANT_ENF_HEPATICA_CRONICA	ANT_ENF_HEPATICA_CRONICA	too low variance
ANT_ANEMIA_HEMOLITICA	ANT_ANEMIA_HEMOLITICA	too low variance
ANT_ENF_NEUROLOGICA	ANT_ENF_NEUROLOGICA	too low variance
RECIBIO_VAC_NEUMOCOCO	RECIBIO_VAC_NEUMOCOCO	too low variance
DESC_VARIANTE_INFLUENZA	DESC_VARIANTE_INFLUENZA	too low variance
ANTECED_TUBERCULOSIS	ANTECED_TUBERCULOSIS	too low variance
ANTECED_CANCER	ANTECED_CANCER	too low variance
DIAS_PUERP	DIAS_PUERP	too low variance
SEMANAS_DE_GESTACION	SEMANAS_DE_GESTACION	too low entropy
FECHA_VAC_NEUMOCOCO	FECHA_VAC_NEUMOCOCO	too low entropy
DESC_DIAGNOSTICO_EGRESO_1	DESC_DIAGNOSTICO_EGRESO_1	too low entropy
DESC_DIAGNOSTICO_EGRESO_2	DESC_DIAGNOSTICO_EGRESO_2	too low entropy
DESC_DIAGNOSTICO_EGRESO_3	DESC_DIAGNOSTICO_EGRESO_3	too low entropy
FECHA_INICIO_CUADRO_CLINICO	FECHA_INICIO_CUADRO_CLINICO	high correlation
ARTRALGIAS	ARTRALGIAS	high correlation
POLIPNEA	POLIPNEA	high correlation
F_REGISTRO_RESULT_CONF2	F_REGISTRO_RESULT_CONF2	high correlation
DIAGNOSTICO_CONFIRMADO	DIAGNOSTICO_CONFIRMADO	high correlation
DEFUNCION	DEFUNCION	high correlation
MOTIVO_EGRESO	MOTIVO_EGRESO	high correlation
FECHA_INGRESO.1	FECHA_INGRESO.1	high correlation
F_REGISTRO_EGRESO	F_REGISTRO_EGRESO	high correlation
PACIENTE	PACIENTE	high correlation
FECHA_INGRESO	FECHA_INGRESO	not representative
FECHA_MUESTRA_CONF1	FECHA_MUESTRA_CONF1	not representative
F_REGISTRO_RESULT_CONF1	F_REGISTRO_RESULT_CONF1	not representative
FECHA_EGRESO_O_DEFUNCION	FECHA_EGRESO_O_DEFUNCION	not representative
ID	ID	contains no information

TABLE III: Features which were dropped in Experiment 1 and 2, including the reason for dropping.