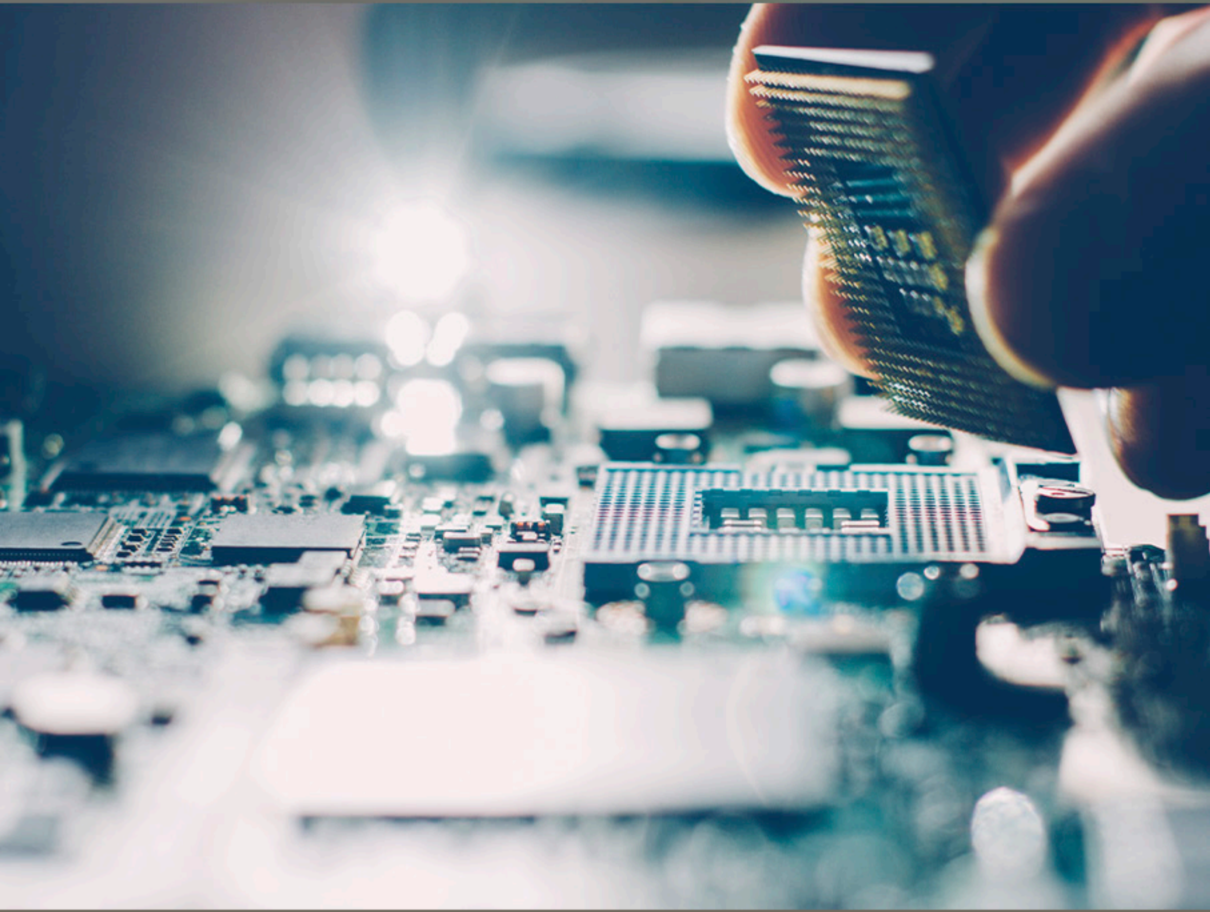
COLEÇÃO

DESAFIOS DAS ENGENHARIAS:

ENGENHARIA DE COMPUTAÇÃO 3



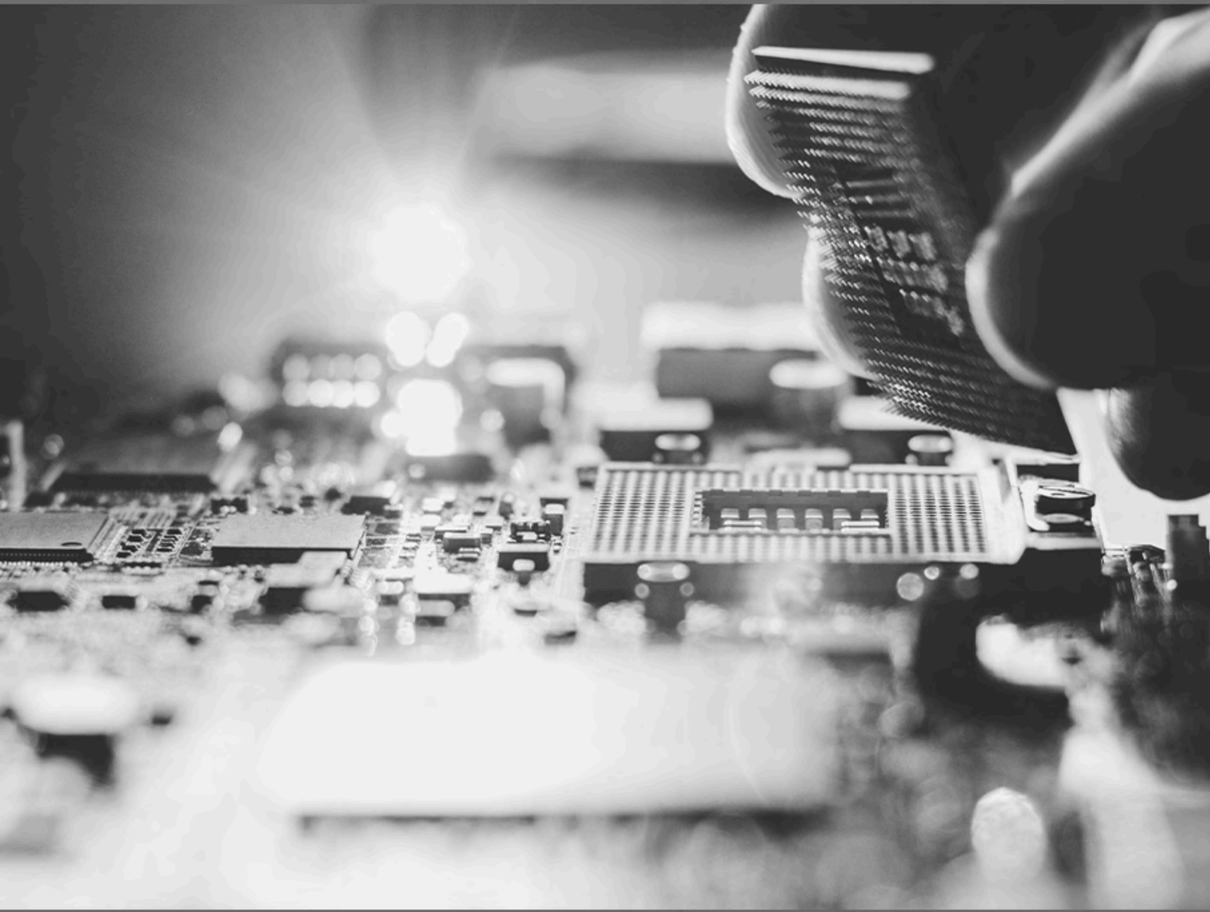
LILIAN COELHO DE FREITAS
(ORGANIZADORA)

Atena
Editora
Ano 2021

COLEÇÃO

DESAFIOS DAS ENGENHARIAS:

ENGENHARIA DE COMPUTAÇÃO 3



LILIAN COELHO DE FREITAS
(ORGANIZADORA)

Atena
Editora
Ano 2021

Editora chefe
Profª Drª Antonella Carvalho de Oliveira
Editora executiva
Natalia Oliveira
Assistente editorial
Flávia Roberta Barão
Bibliotecária
Janaina Ramos
Projeto gráfico
Camila Alves de Cremo
Daphynny Pamplona
Gabriel Motomu Teshima
Luiza Alves Batista
Natália Sandrini de Azevedo
Imagens da capa
iStock
Edição de arte
Luiza Alves Batista

2021 *by Atena Editora*
Copyright © Atena Editora
Copyright do texto © 2021 Os autores
Copyright da edição © 2021 Atena Editora
Direitos para esta edição cedidos à Atena Editora pelos autores.
Open access publication by Atena Editora

Todo o conteúdo deste livro está licenciado sob uma Licença de Atribuição *Creative Commons*. Atribuição-Não-Comercial-NãoDerivativos 4.0 Internacional (CC BY-NC-ND 4.0).

O conteúdo dos artigos e seus dados em sua forma, correção e confiabilidade são de responsabilidade exclusiva dos autores, inclusive não representam necessariamente a posição oficial da Atena Editora. Permitido o *download* da obra e o compartilhamento desde que sejam atribuídos créditos aos autores, mas sem a possibilidade de alterá-la de nenhuma forma ou utilizá-la para fins comerciais.

Todos os manuscritos foram previamente submetidos à avaliação cega pelos pares, membros do Conselho Editorial desta Editora, tendo sido aprovados para a publicação com base em critérios de neutralidade e imparcialidade acadêmica.

A Atena Editora é comprometida em garantir a integridade editorial em todas as etapas do processo de publicação, evitando plágio, dados ou resultados fraudulentos e impedindo que interesses financeiros comprometam os padrões éticos da publicação. Situações suspeitas de má conduta científica serão investigadas sob o mais alto padrão de rigor acadêmico e ético.

Conselho Editorial
Ciências Exatas e da Terra e Engenharias
Prof. Dr. Adélio Alcino Sampaio Castro Machado – Universidade do Porto
ProFª Drª Ana Grasielle Dionísio Corrêa – Universidade Presbiteriana Mackenzie
Prof. Dr. Carlos Eduardo Sanches de Andrade – Universidade Federal de Goiás
Profª Drª Carmen Lúcia Voigt – Universidade Norte do Paraná
Prof. Dr. Cleiseano Emanuel da Silva Paniagua – Instituto Federal de Educação, Ciência e Tecnologia de Goiás

Atena
Editora
Ano 2021

Atena
Editora
Ano 2021

Coleção desafios das engenharias: engenharia de computação 3

Diagramação: Daphynny Pamplona
Correção: Gabriel Motomu Teshima
Indexação: Amanda Kelly da Costa Veiga
Revisão: Os autores
Organizadora: Lilian Coelho de Freitas

Dados Internacionais de Catalogação na Publicação (CIP)

C691 Coleção desafios das engenharias: engenharia de computação 3 / Organizadora Lilian Coelho de Freitas. – Ponta Grossa - PR: Atena, 2021.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-5983-619-2
DOI: https://doi.org/10.22533/at.ed.192212911

1. Engenharia de computação. I. Freitas, Lilian Coelho de (Organizadora). II. Título.

CDD 621.39

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

Atena Editora
Ano 2021

DECLARAÇÃO DOS AUTORES

Os autores desta obra: 1. Atestam não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao artigo científico publicado; 2. Declaram que participaram ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certificam que os artigos científicos publicados estão completamente isentos de dados e/ou resultados fraudulentos; 4. Confirmam a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhecem terem informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autorizam a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

APRESENTAÇÃO

A Atena Editora tem a honra de presentear o público em geral com a série de *e-books* intitulada “*Coleção desafios das engenharias: Engenharia de computação*”. Em seu terceiro volume, esta obra tem o objetivo de divulgar aplicações tecnológicas da Engenharia de Computação na resolução de problemas atuais, com o intuito de facilitar a difusão do conhecimento científico produzido em várias instituições de ensino e pesquisa do país.

Organizado em 20 capítulos, este volume apresenta temas como utilização de aprendizagem de máquina na avaliação de riscos de infeção por COVID-19; dispositivos automatizados para administração de remédios; comunicação científica apoiada por realidade aumentada; métodos de elementos finitos aplicados na análise de materiais para indústria aeronáutica; aplicações de processamento digital de imagens e de algoritmos genéticos; entre diversas outras aplicações da automação e do desenvolvimento de *software,* combinados para melhorar as atividades do nosso dia-a-dia.

Dessa forma, esta obra contribuirá para aprimoramento do conhecimento de seus leitores e servirá de base referencial para futuras investigações.

Os organizadores da Atena Editora, agradecem especialmente os autores dos diversos capítulos apresentados, parabenizam a dedicação e esforço de cada um, os quais viabilizaram a construção deste trabalho.

Boa leitura.

Lilian Coelho de Freitas

SUMÁRIO

CAPÍTULO 1

EVALUATING THE RISK OF COVID-19 INFECTION BASED ON MACHINE LEARNING OF SYMPTOMS AND CONDITIONS VERSUS LABORATORY METHODS

Data de aceite: 01/11/2021
Data de submissão: 06/08/2021

Daniel Mário de Lima
Laboratorio de Informatica Biomedica, Instituto do Coracao, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo
Sao Paulo – SP
https://orcid.org/0000-0002-7818-6103

João Henrique Gonçalves de Sá
Laboratorio de Informatica Biomedica, Instituto do Coracao, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo
Sao Paulo – SP
http://lattes.cnpq.br/1897087795965443

Ramon Alfredo Moreno
Laboratorio de Informatica Biomedica, Instituto do Coracao, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo
Sao Paulo – SP
https://orcid.org/0000-0002-0548-9365

Marina de Fátima de Sá Rebelo
Laboratorio de Informatica Biomedica, Instituto do Coracao, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo
Sao Paulo – SP
http://lattes.cnpq.br/2603863331633089

José Eduardo Krieger
Laboratorio de Genetica e Cardiologia Molecular, Instituto do Coracao, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo
Sao Paulo – SP
https://orcid.org/0000-0001-5464-1792

Marco Antonio Gutierrez
Laboratorio de Informatica Biomedica, Instituto do Coracao, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo
Sao Paulo – SP
https://orcid.org/0000-0003-0964-6222

ABSTRACT: Objective. Evaluate a platform for daily survey of COVID-19 signs and symptoms in health employees. The platform was developed to indicate the need of additional individual diagnostic procedures, assist institutional planning to prevent the spread of the virus and sustain the hospital operations during the pandemic. Methods. We used information from a recent meta-analysis to simulate datasets of patients with different signs, symptoms and comorbidities to evaluate machine-learning algorithms for each dataset classification. The best performing model identifying COVID-19 from other similar conditions including H1N1 and seasonal influenza was selected as the base model for developing a platform for risk assessment. Results and Conclusion. The platform was deployed for surveillance of 4,200 collaborators from a tertiary hospital on a voluntary basis, but it can be readily adapted for other environments or populational surveillance to assist public authorities devising strategies to prevent the spread of the virus. The evaluation by laboratory data stratification showed that the platform detected a significantly higher risk among infected individuals.

KEYWORDS: Coronavirus Infections, Data Science, Machine Learning.

AVALIAÇÃO DE RISCOS DE INFECÇÃO POR COVID-19 COM BASE EM APRENDIZADO DE MÁQUINA DE SINTOMAS E CONDIÇÕES VERSUS MÉTODOS LABORATORIAIS

RESUMO: Objetivo. Avaliar uma plataforma para levantamento diário dos sinais e sintomas de COVID-19 em profissionais de saúde. A plataforma foi desenvolvida para indicar a necessidade de procedimentos diagnósticos individuais adicionais, auxiliar no planejamento institucional para prevenir a propagação do vírus e sustentar as operações do hospital durante a pandemia. Métodos. Usamos informações de uma meta-análise recente para simular conjuntos de dados de pacientes com diferentes sinais, sintomas e comorbidades para avaliar algoritmos de aprendizado de máquina para cada classificação de conjunto de dados. O modelo de melhor desempenho para identificar COVID-19 de outras condições semelhantes, incluindo H1N1 e influenza sazonal, foi selecionado no desenvolvimento de uma plataforma para avaliação de risco. Resultados e conclusão. A plataforma foi implantada para vigilância voluntária de 4.200 colaboradores de um hospital terciário, mas pode ser prontamente adaptada para outros ambientes ou vigilância populacional para auxiliar o poder público a traçar estratégias de prevenção à disseminação do vírus. A avaliação por estratificação dos resultados laboratoriais demonstrou que a plataforma detectou risco significativamente maior entre indivíduos infectados.

PALAVRAS-CHAVE: Infecções por Coronavirus, Ciência de Dados**,** Aprendizado de Máquina

1 I INTRODUCTION

In late 2019, a novel type of coronavirus, known as SARS-CoV-2, was discovered, causing several infections and pneumonia cases, first in Wuhan, China, with later spreading worldwide. The World Health Organization named the acute infectious disease caused by the SARS-CoV-2 as COVID-19 (Coronavirus Disease - 2019), a systemic infectious disease characterized by its high transmissibility. A challenge in this type of emergency is the prompt detection of infected people to prioritize health care and prevent the spread of the virus, especially under low virus testing capacity, as was the case in many places including Brazil (depicted in Fig. 1) as of early 2020.

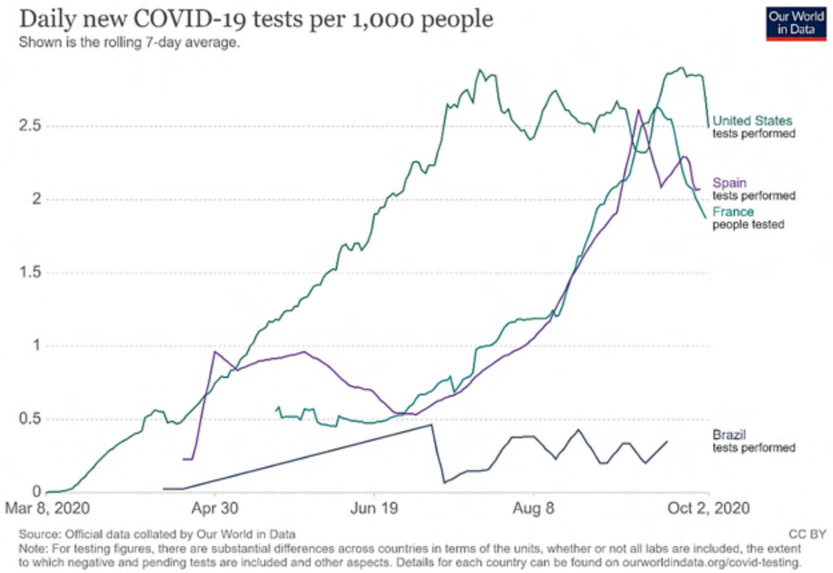


Figure 1 – The number of tests of COVID-19 by 1000 people according to "OurWorldInData.org" for Brazil, EUA, France, and Spain, from March to October 2020. (Source: https://ourworldindata.org/coronavirus#coronavirus-country-profiles)

The lack of capacity for rapid diagnostics is critical for healthcare workers (HCW) who have been overly affected and become an easy target for the infection during the pandemic (GRASSLY et al., 2020), mainly when laboratory tests were unavailable or limited.

According to worldwide publications, these workers represent 19% of all reported COVID-19 cases in Spain (ISCIII, 2020), 9% in Italy (ICN, 2020), 4.6% in Germany (RKI, 2020), and 3.8% in China (WU & MCGOOGAN, 2021). According to PAN et al. (2020) and KLUYTMANS-VAN DEN BERGH et al. (2020), these figures represent a higher incidence of the infection than observed in general population. This outcome is a consequence of the HCW's exposure to infection from patients and daily interaction with fellow staff that may have been already infected and undiagnosed.

The high rate of infection of HCW poses additional risks, as stated by the Imperial College COVID-19 response team, "*Transmission to and potentially among this high-risk group compromises both their health and may contribute to nosocomial spread within hospitals"* (GRASSLY et al., 2020).

Such a spread raises the issue of protecting vulnerable patients from a potentially infectious workforce, as social distancing is not possible while caring for patients. Moreover, MCALONAN et al. (2007) verified that HCW screening programs positively influence the team's self-confidence, decrease absenteeism, and reduce psychological problems. Therefore, establishing strategies for protecting HCW from the infection by the new coronavirus and the screening and surveillance of symptoms should be a priority to assist the management of hospitals and healthcare facilities. Reliable data sources and analytic tools are essential for healthcare decision-makers to establish policies to mitigate the pandemic disruption. Often, surveys rely only on simple models based on the number of symptoms declared by the respondent, e.g. as those assessed by ROSSMAN et al. (2020) and RIVETT et al. (2020), and have limited discrimination power to discriminate from common symptoms associated with other viral diseases, such as influenza and H1N1 – refer to the earlier Influenza study by TANG et al. (2010) and COVID meta-analysis by MA et al. (2020).

Screening of HCW for identifying the most likely employee at risk of infection by the SARS-CoV-2 was paramount considering the availability of other diagnostic tools. Thus, LIMA et al. (2020) proposed a platform to assess COVID-19 risk score based on a machine-learning model. COVIDuc (COVID under control) was then developed as an application based on a set of signs, symptoms and comorbidities that may affect the susceptibility or the intensity of the infection for daily surveys to assess the risk among HCWs. We evaluated the COVIDuc platform after collecting laboratory data from HCWs in a tertiary hospital, and we found that COVIDuc detected a significantly higher risk in infected HCWs.

2 I 2. METHODS

2.1 Predictive model for COVID-10 risk from symptoms and conditions

We reviewed diagnostic studies of influenza viruses and SARS-CoV-2 for the affected populations' characteristics, the prevalence of symptoms, conditions, and laboratory results in severe and moderate risk groups. TANG et al. (2010) described the similar symptomatology of pandemic, seasonal, and no-influenza cases in a large prospective study with 2,683 patients. MA et al. (2020) performed a systematic review and meta-analysis of 20 studies, involving 19 Health Centers and 11 Multi-Centers, and with around 53,000 patients with COVID-19. The symptoms and conditions found in these studies related to the disease were fever, cough, fatigue, expectoration, headache, diarrhea, myalgia, shortness of breath, sore throat, nausea or vomiting, chill, nasal congestion or rhinorrhea, dyspnea, anorexia, dizziness, hypertension, diabetes, cardiovascular disease (CVD), cerebrovascular disease, chronic obstructive pulmonary disease (COPD).

Assuming symptoms and conditions as independent and normal variables, we simulated groups at different risk levels by random sampling of multivariate normal distributions, with similar prevalence of the reviewed clinical studies' symptoms and conditions. We prepared a data set comprising five groups with the following numbers of case studies: Group 1: 5000 simulating the "non-influenza"; Group 2: 5000 simulating the "seasonal influenza"; Group 3: 5000 simulating the "H1N1-2009"; Group 4: 7500 simulating the "non-severe COVID+"; Group 5: 7500 simulating the "severe COVID+" group. In this data set, 50% of the simulated case studies were COVID-positive, and 50% were COVID-negative (Fig. 2). We then adjusted linear support vector classifiers (SVC) to verify the "severe COVID+ " group's separability from the rest of the data.

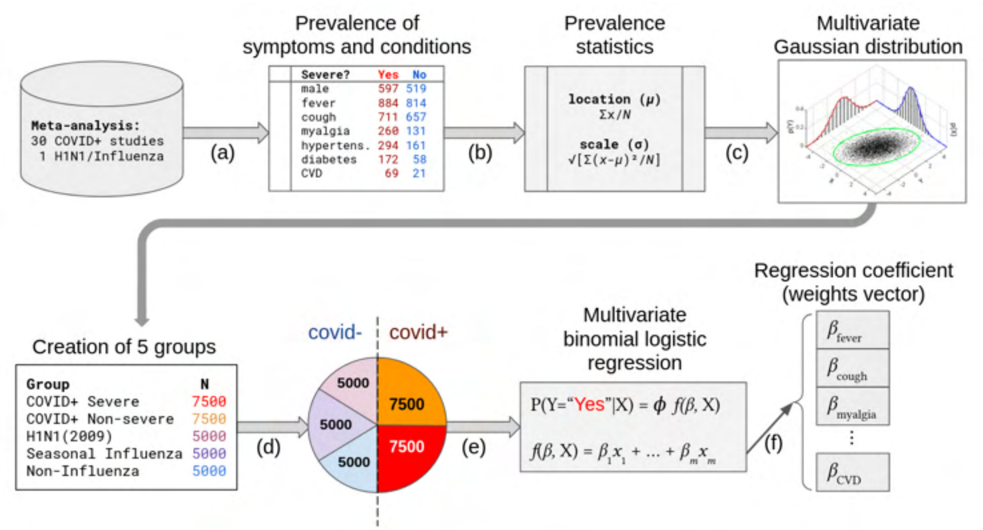


Figure 2 – Principal steps involved in the modeling of the risk score for COVID-19.

Following, we evaluated several machine learning (ML) algorithms for the classification of the groups. (1) GaussianNB and BernoulliNB: Naïve Bayes models for multivariate Gaussian and Bernoulli distributions, that is, with normal and binary-valued variables, respectively (MCCALLUM & NIGAM, 1998). (2) DecisionTree: if-then-else decision trees learned by splitting attributes' values into subsets of minimal (Gini index) classification error (BREIMAN et al., 1984; QUINLAN, 1987). (3) Ridge and LogisticRegression: the regularized and logit regression models for binary-valued output variables, adjusted by liblinear's dual coordinate-descent method (CHOU et al., 2020) – for theory and other linear models, see NELDER & WEDDERBURN (1972). (4) KNneighbors and RadiusNeighbors: indexed search of $k = 5$ neighbors in a kd-tree (BENTLEY, 1975) or ball-tree (OMOHUNDRO, 1989), optionally limited to a radius $r = 1$ in the Minkowski metric space near the query point. (5) Decision Tree Ensembles (Random Forest, Extra Trees, AdaBoost): which are compound versions based on decision trees, such as of randomized and bootstrapped decision trees (BREIMAN, 1996; 2001), with random splitting thresholds (GEURTS, ERNST & WEHENKEL, 2006), and adaptive gradient boosting (FREUND & SCHAPIRE, 1995). (6) Perceptron: the classic linear model with step function output, inspired by how biological neurons work (ROSENBLATT, 1958).

We compared each classifier's predictive performance using receiver operating characteristics (ROC) analysis at each new iteration of evaluation and adjustment of the symptoms and conditions list. Finally, we scaled the weights of the probability estimator of the logistic output to weights of a weighted average with an output between 0 and 10, which defines the range of the risk score for COVID-19. Figure 2 represents the main steps involved in modeling the risk score for COVID-19, where the ML algorithms replace the logistic regression.

The training was carried out with 80% of the data in 5-fold cross-validation, and the remaining 20% of the data were used to analyze the classifiers' receiver operating characteristics curve (ROC curve). These analyses were executed in a Dell XPS8930 workstation with Intel i7-8700 CPU, 16GB RAM, 1TB HDD, and NVIDIA 1050Ti-4GB GPU; using Microsoft Excel [RRID SCR_016137], Python [RRID SCR_008394], Pandas [RRID SCR_018214], and NumPy [RRID SCR_008633] for data preparation; MatPlotLib [RRID SCR_008624] for visualization; SciPy [RRID SCR_008058] and StatsModels [RRID SCR_016074] for statistical analyses; and Scikit-Learn [RRID SCR_002577] for ML.

2.2 Development and Deployment of the COVIDuc web application

We have elaborated an online survey in collaboration with the hospital's clinical specialists to implement a pilot application for COVID-19 risk assessment of the hospital's employees. The questionnaire was based on a simplified subset of the symptoms and conditions used in the predictive model. We regrouped and selected new variables (the subset of symptoms and conditions) and trained the model using this simplified data set, then implemented the online application prototype.

The probability of positive COVID-19 estimated by the logistic regression model was adjusted as weights of a weighted average, with output between 0 and 10. After the revision performed by clinicians, and the re-training of the developed model with the reduced set of symptoms and conditions, we implemented a web application for risk assessment of the employees in a tertiary hospital, called COVID Under Control (COVIDuc). The implemented model takes into account the possible presence of the following pre-conditions and symptoms:

1. **Signals and Symptoms**: dry cough or phlegm; muscle aches; headache; sore throat; shortness of breath; stuffy or runny nose; diarrhea, nausea or vomit; chills or fever (> 38° Celsius); loss of taste or smell.

2. **Comorbidities:** hypertension; diabetes mellitus; heart disease; pulmonary disease; smoking.

The COVIDuc application's front-end was developed using Angular 10, Enterprise Java Beans was used in the back-end, and Oracle 12C as the application database. In order to monitor the use of the questionnaire by hospital employees and the resulting scores, we implemented two dashboards with metrics and statistics of accesses, using the Qlik Sense tool[1].

---

[1] Qlik Sense tool. Link: https://qlik.com/products/qlik-sense. Accessed 04 aug. 2021.

3 I RESULTS

3.1 Predictive model for COVID-10 risk from symptoms and conditions

Figure 3 (left) presents the ROC curves displaying the performance of the classifiers tested for the simulated groups with the complete set of symptoms from TANG et al. (2020) and MA et al. (2020). Figure 3 (right) displays the classification performance considering only the simplified set of variables (Section 2.2), which have been adopted in the online questionnaire that was implemented. It is evident that SVC, Nu-SVC, Bernoulli Naïve Bayes, Ridge, Logistic Regression, Extra Trees, Random Forest and AdaBoost obtained the largest areas under the curve (AUC) of 96% with the complete set, and Ridge, Logistic Regression and Extra Trees obtained the largest AUCs of 91% with the simplified variables.

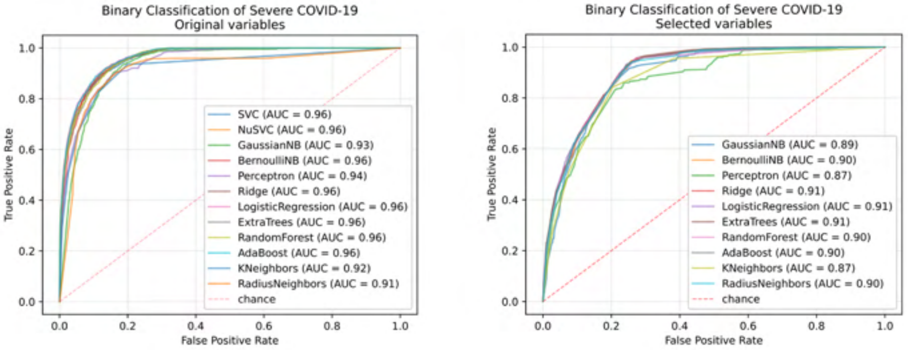


Figure 3 – Comparative ROC analysis of binary classification algorithms between "COVID+ Severe" and "other groups". Left: ROC curves for SVC and other ML algorithms using the original variables from TANG et al. (2020) and MA et al. (2020). Right: ROC curves for ML algorithms without SVC, using only the simplified set of symptoms and conditions.

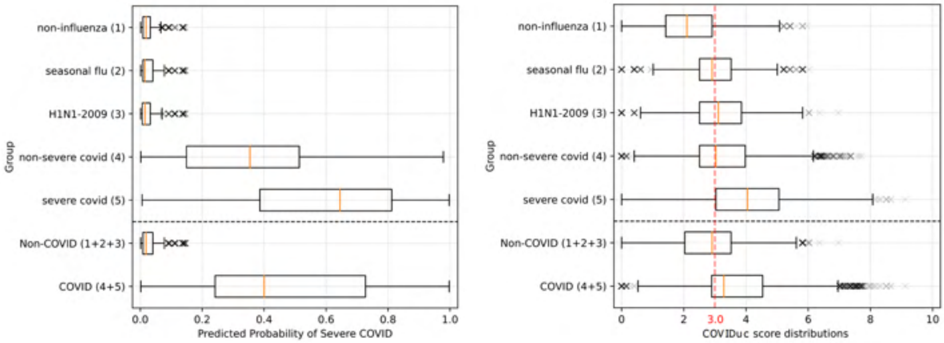


Figure 4 – Left: Distribution of predicted probabilities of "severe COVID" in each patient group (1 to 5), and aggregated COVID (non-severe or severe COVID) and Non-COVID (H1N1-2009, seasonal flu or non-influenza) groups. Right: Distribution of COVIDuc risk scores, after adjustment, in each group (1 to 5), and aggregated COVID and Non-COVID groups.

After comparing the performance of the classification algorithms, we adopted the Logistic Regression, whose coefficients define the weights for each symptom and condition. Figure 4 (left) presents the distribution of predicted probabilities for “severe COVID” occurrence in groups 1 to 5, where we see a clear distinction between Non-COVID and COVID symptoms and conditions but a significant overlap between non-severe and severe COVID groups. Figure 4 (right) presents the distribution of post-adjustment risk scores in groups 1 to 5. After evaluating this distribution, we set the risk score “3.0” as the threshold for risk of COVID-19 based on the set of symptoms and conditions from the chosen model’s distribution of risk scores.

3.2 Deployment of the COVIDuc web application

To register in the COVIDuc app, the user has to provide some personal information relevant to the application, such as age, weight, and height, and point out (if present) the symptoms and conditions. The COVIDuc application presents the resulting score with the indication (or not) of the suspicion for COVID-19. After a few minutes, an SMS is sent to the user showing the risk score. Figure 5 shows the sequence of COVIDuc screens of the first access, from the initial page, with the consent form, to the result score and SMS.

Figure 5 (In Portuguese) – The sequence of COVIDuc screens for the first access (in Portuguese). (1) From the QR code, the user can access the application. The QR code has been sent previously by email to all employees and is still available at multiple points at the hospital. (2) Disclaimer page. (3) Personal information page. (4) Screen with symptoms and pre-conditions. In the example screen, the person reported having diabetes as a pre-condition and the following symptoms: body pain, chills, or fever. (4) COVIDuc score page. For this example, there is a suspect of infection. Thus, the person is advised to go to the medical department for tests. (5) SMS is sent to the mobile phone number provided on the personal information page (6).

#Acessos
25.186
#Pessoas
4,29k
#Risco3+
751
Méd. Acessos/Pessoa
5,87
RiscoMéd
0,88
RiscoMáx
9,53

Figure 6 – Dashboard presenting the data related to the respondents of the COVIDuc, daily access for area, how many persons in each risk range, among others.

Since the beginning of May 2020, more than 4,200 employees from a tertiary hospital have been using the web application voluntarily to assess their daily risk for COVID-19. The application has proven a valuable tool for HCW surveillance. The data collected is available for the managers at an Institutional dashboard. Figure 6 shows a dashboard screenshot with daily accesses from all hospital areas and all risk ranges. The scores were available for the occupational medicine department, which contacts the persons that have obtained high scores and did not show up for health evaluation with a physician.

Figure 7 presents the number of people using the application, both daily and cumulative; they also present the number of persons that had any COVIDuc score greater or equal to 3.0, which was the adjusted threshold value of risk for COVID-19 by the developed model. The panels display how many people have been answering COVIDuc, how many had higher risks, the cumulative and the daily percentages of HCW with higher risk scores.

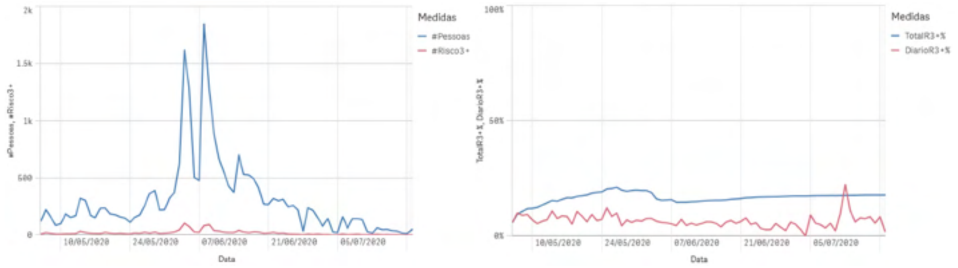


Figure 7 (in Portuguese) – Left: The daily number of people answering the questionnaire (blue line) and the number of persons with COVIDuc scores equal to or higher than 3.0 (red line), who were considered at risk for COVID-19 between May 2020 and July 2020. Right: The accumulated (blue) and daily (red) percentages of individuals with COVIDuc scores equal to or higher than 3.0.

3.3 Model evaluation with Laboratory data

After its deployment, COVIDuc was used by more than 4,200 HCWs of the hospital to assess their risk score for COVID infection. From those HCWs, around 3,200 individuals were tested for SARS-CoV2 presence in blood serum by quantifying anti-SARS-CoV2 immunoglobulin M (IgM) and immunoglobulin G (IgG) antibodies, also called viral load. Furthermore, 311 HCWs had SARS-CoV2 infections confirmed by external reverse transcriptase–polymerase chain reaction (RT-PCR) testing. Figure 8 displays the detection windows for the three methods: PCR, IgM and IgG.

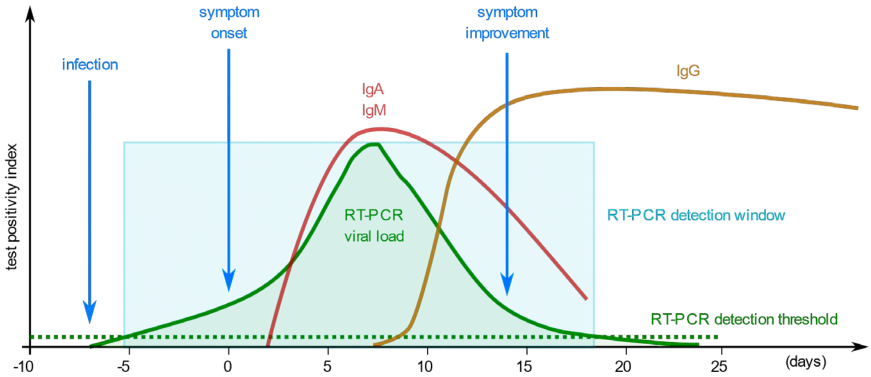


Figure 8 – Diagnostic windows for RT-PCR, IgM and IgG since the onset of symptoms (day 0). The curves represent the relative viral load of anti-SARS-CoV2 for each method. Adapted from NOGUEIRA & SILVA (2020).

According to NOGUEIRA & SILVA (2020), the laboratory methods have different detection windows varying between -7 and +18 days from day 0 of the symptom onset. Therefore, it is not trivial to compare the laboratory results to the onset of symptoms as assessed by COVIDuc risk score. For this reason, we executed separate comparisons for each method. The comparison was between risk scores for positive and non-positive individuals assessed by each method, using independent T-tests, with $p < .05$ indicating significant differences between the mean risk scores. As the reported PCR tests have no negatives, we compare the PCR-positive individuals and PCR-missing individuals directly. In all methods, we considered the maximum and average risk scores from 10 days before to 10 days after the blood sample for each individual.

	n_{max}	$+_{avg}$	**PCR+**	**IgM+**	**IgG+**
Score < 3	2,768	24.0%	8.3%	8.0%	18.6%
Score >= 3	367	45.5%	22.4%	18.2%	36.4%
Score >= 6	33	71.4%	28.6%	42.9%	57.1%

Table 1 – Prevalence of positive tests in low-risk and high-risk HCW (N=3,135), n is the number of individuals in each score stratum, "+" is the percentage of positive individuals (by any test).

Table 1 displays the prevalence of individuals with positive results for each risk stratum. In the low-risk stratum (score < 3), 8% of HCW had active infections when they answered the COVIDuc questionnaire (positive PCR, IgM), while 19% have had past exposure (by IgG). In the high-risk stratum (score >= 3) the SARS-CoV2 infections were 18-22% (2.5 times) by PCR, IgM and 36% (2.0 times) by IgG, showing correspondence between infections and higher risks.

	Max(risk)		**Avg(risk)**	
Method	**T**	**p**	**T**	**p**
PCR+	6.0731	< .0001	6.9537	< .0001
IgM+	5.4454	< .0001	4.7763	< .0001
IgG+	4.9666	< .0001	4.9071	< .0001
IgM ou IgG+	5.6793	< .0001	5.3986	< .0001
IgM, IgG ou PCR+	6.3828	< .0001	6.1771	< .0001

Table 2 – Independent T-tests for asserting that positive individuals have a higher mean risk score than non-positive groups (PCR) or negative groups (others).

Table 2 displays the results of the independent T-tests, asserting that individuals that tested positive for SARS-CoV2 in any laboratory method have had significantly higher risk scores than those individuals with negative, missing or inconclusive results with 95% confidence i.e. $p < .05$ in all hypotheses.

4 I DISCUSSION AND RELATED WORK

Even though SARS-CoV2 infection was common among HCWs (25%), only a few HCWs of those with positive SARS-CoV2 infection (4.2%) answered to have symptoms correlated with severe COVID-19 syndrome. Our evaluation indicated that symptom-based models like COVIDuc cannot detect asymptomatic SARS-CoV2 infections, and ML models for accurate SARS-CoV2 detection must use additional variables considering laboratory and other tests, such as complete blood counts and immunoassays. With this in mind, COVIDuc correctly assigned higher risks to infected individuals, meaning that individuals with SARS-CoV2 may develop worse symptoms than those of concurrent respiratory infections like Influenza, H1N1 and other agents. COVIDuc's main achievement was the ability to indicate the proportion of HCWs with more severe symptoms, and how fast the infection spread among HCWs in early 2020 when the laboratory tests like RT-PCR and immunoassays were unavailable or limited.

Early studies of COVID-19 data were focused on addressing the epidemiological aspects of the disease, the preparedness of hospitals for receiving patients, and the patient conditions in emergency care settings. The more general aspects of infectibility and disease spread are addressed by well-known epidemiological models, and are included in the WHO Global Influenza Surveillance and Response System (GISRS) protocols. Other studies analyze clinical parameters from readily available laboratory methods, e.g., complete blood count analysis (BATISTA et al., 2020). Another category of studies, such as in DREW et al. (2020), focused on active identification of suspects based on symptom assessment, as the COVIDuc objectives but for the general public instead. As the computational approaches require data for model development, many studies also focus on preparing, integrating and publishing the clinical data for future retrospective studies. MELLO et al. (2020) exemplify this approach, describing a collaborative effort in sharing laboratory data related to COVID-19 from several hospitals in São Paulo, Brazil.

5 | CONCLUSIONS

The developed model provides a method to assess the risk of COVID-19 based on the severity of signs and symptoms and conditions for COVID-19. We used data from a meta-analysis to define the weight of the factors contributing to influence the risk, and this measure can be modified as other specific diagnostics verify definitive outcomes. The method does not use precise variables, such as laboratory and imaging exams and it is not meant for diagnostics, COVIDuc must be used as a tool for surveillance and screening only. The employees from a tertiary hospital have used it as a surveillance tool since May 2020 voluntarily, with the occupational medicine department continually assessing the correspondence between the outcomes of COVIDuc and the employee's medical files. COVIDuc was developed for daily surveillance of HCWs, but it can be readily adapted for other environments or populational surveillance to assist public authorities in devising strategies to prevent the spread of the virus. It was first used in a study carried out by the Solidary Research Network to evaluate public policies during the COVID-19 pandemic. Furthermore, since the middle of August 2020, COVIDuc has been the screening tool to call outpatients to resume the medical activities on the same tertiary hospital after the peak of the pandemic period in the city of São Paulo.

After evaluation with laboratory data, we verified that COVIDuc correctly assigned higher risks to infected individuals. COVIDuc also tracked the proportion of HCWs with more severe symptoms and how fast the infection spread among HCWs daily. These results indicate that COVIDuc is calibrated to detect worse symptoms, not positive infections, as most SARS-CoV2 cases have lighter symptoms or no symptoms at all.

As future work, we will continue to support the monitoring platform, open collaborations to collect data from WHO GISRS and monitor other infections in the tropical and subtropical climates, such as Malaria, Zikavirus and possibly other neglected tropical diseases. Those future improvements will provide a complete dashboard for high-level decision support in public infection control.

ACKNOWLEDGMENTS

Zerbini Foundation and Foxconn Brazil supported this work as part of the research project “Machine Learning in Cardiovascular Medicine” (AIMED-CATI 030/2007 FOXCONN-001/2019). We also thank Dr. Diego Cardenas for his kind help in text revision.

ETHICAL APPROVAL

This study was approved by the IRB (CAAE:30404820.5.0000.0068).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

BATISTA, A. F. M. *et al*. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. **medRxiv**, preprint 2020.04.04.20052092, Jan. 2020. DOI 10.1101/2020.04.04.20052092

BENTLEY, J. L. Multidimensional Binary Search Trees Used for Associative Searching. **Commun ACM**, New York, v.18, n.9, p.509-517, 1975. DOI 10.1145/361002.361007

BREIMAN, L. *et al*. **Classification and Regression Trees**. Boca Raton: Chapman and Hall/CRC, 1984. 358p. DOI 10.1201/9781315139470

BREIMAN, L. Bagging predictors. **Mach Learn**, New York, v.24, n.2, p.123-140, 1996. DOI 10.1007/BF00058655

BREIMAN, L. Random forests. **Mach Learn**. New York, v.45, n.1, p.5-32, 2001. DOI 10.1023/A:1010933404324

CHOU, H. Y.; LIN, P. Y.; LIN, C. J. Dual coordinate-descent methods for linear one-class SVM and SVDD. In: SDM20, Newark, 2020. **Proc SIAM Int Conf Data Mining**. Newark: IEEE, 2020. p.181-189.

DOI 10.1137/1.9781611976236.21

DREW, D. A. *et al*. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. **Science**, Washington, v.368, n.6497, p.1362-1367, Jun. 2020. DOI 10.1126/science.abc0473

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Lect Notes Comput Sci**, Berlin, v.904, p.23-27, 1995. DOI 10.1007/3-540-59119-2_166

GEURTS, P.; ERNST, D.; WEHENKEL, L.. Extremely randomized trees**. Mach Learn**, New York, v.63, n.1, p.3-42, 2006. DOI 10.1007/s10994-006-6226-1

GRASSLY, N. *et al*. Comparison of molecular testing strategies for COVID-19 control: a mathematical modelling study. **Lancet Infect Dis**, London, v.20, n.12, p.1381-1389, Dec. 2020. DOI 10.1016/S1473-3099(20)30630-7

LIMA, D. M. *et al*. A COVID-19 surveillance platform to monitor risk of infection based on a machine learning model. **J Health Inform**, São Paulo, v.12, 2020. Apresentado no XVII Congresso Brasileiro de Informática em Saúde, 2020, Foz do Iguaçu, RS. Disponível em: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/839. Acesso em: 04 ago. 2021.

ICN - International Council of Nurses. High proportion of healthcare workers with COVID-19 in Italy is a stark warning to the world: protecting nurses and their colleagues must be the number one priority. **ICN**, Geneva, 2020. Disponível em: https://www.icn.ch/sites/default/files/inline-files/PR_09_COVID-19 - Italy.pdf. Acesso em: 27 jan. 2021.

ISCIII - Instituto de Salud Carlos III. Informe no. 22 - situación de COVID-19 en España a 13 de abril de 2020. **ISCIII**, Madrid, 2020. Disponível em : https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/Informe%20n%C2%BA%2022.%20Situaci%C3%B3n%20de%20COVID-19%20en%20Espa%C3%B1a%20a%2013%20de%20abril%20de%202020.pdf. Acesso em: 27 jan. 2021.

KLUYTMANS-VAN DEN BERGH, M. F. Q. *et al*. Prevalence and Clinical Presentation of Health Care Workers With Symptoms of Coronavirus Disease 2019 in 2 Dutch Hospitals During an Early Phase of the Pandemic. **JAMA Netw Open**, Chicago, v.3, n.5, e209673, 2020. DOI 10.1001/jamanetworkopen.2020.9673

MA, C. *et al*. Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis. **medRxiv**, preprint 2020.03.17.20037572, Jan. 2020. DOI 10.1101/2020.03.17.20037572

MCALONAN, G. M. *et al*. Immediate and Sustained Psychological Impact of an Emerging Infectious Disease Outbreak on Health Care Workers. **Can J Psychiatry**, Thousand Oaks, v.52, n.4, p.241–247, Apr. 2007. DOI 10.1177/070674370705200406

MCCALLUM, A.; NIGAM, K. A comparison of event models for Naive Bayes text classification. In: ICML/AAAI-98 Workshop on Learning for Text Categorization, Madison, 1998. **Proc Natl Conf Artif Intell**. Menlo Park: AAAI Press, 1998. Disponível em: http://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf. Acesso em: 27 jan. 2021.

MELLO, L. E. *et al*. Opening Brazilian COVID-19 patient data to support world research on pandemics. **Zenodo**. São Paulo, FAPESP, 2020. DOI 10.5281/zenodo.3966427. Base de dados.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **J R Stat Soc Ser A**, London, v.135, n.3, p.370-384, Jan. 1972. Disponível em: http://www.jstor.org/stable/2344614. Acesso em: 27 jan. 2021.

NOGUEIRA, J. M. R.; SILVA, L. O. P. Diagnóstico laboratorial da COVID-19 no Brasil. **Rev Bras Anal Clin**, Rio de Janeiro, v.52, n.2, p.117-121, 2020. DOI 10.21877/2448-3877.20200007

OMOHUNDRO, S. M. **Five balltree construction algorithms**. Berkeley: International Computer Science Institute, 1989. Disponível em: https://www.icsi.berkeley.edu/icsi/node/2291. Acesso em: 05 ago. 2021.

PAN, A. *et al*. Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. **JAMA**, Chicago, v.323, n.19, p.1915-1923, May 2020. DOI 10.1001/jama.2020.6130

QUINLAN, J. R. Induction of Decision Trees. **Mach Learn**, New York, v.1, n.1, p.81-106, 1986. DOI 10.1007/BF00116251

RIVETT, L. *et al*. Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. **eLife**, Cambridge, v.9, e58728, 2020. DOI 10.7554/eLife.58728

RKI - Das Robert Koch Institut. Coronavirus Disease 2019 (COVID-19) Daily Situation Report of the Robert Koch Institute. **RKI**, Berlin, 2020. Disponível em: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/2020-04-13-en.pdf. Acesso em: 27 jan. 2021.

ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. **Psychol Rev**, Washington, v.65, n.6, p.386-408, 1958. DOI 10.1037/h0042519

ROSSMAN, H. *et al*. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. **Nat Med**, New York, v.26, n.5, p.634–638, 2020. DOI 10.1038/s41591-020-0857-9

TANG, J. W.-T. *et al*. Differing Symptom Patterns in Early Pandemic vs Seasonal Influenza Infections. **Arch Intern Med**, Chicago, v.170, n.10, p.861-867, May 2010. DOI 10.1001/archinternmed.2010.108

WU, Z.; MCGOOGAN, J. M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. **JAMA**, Chicago, v.323, n.13, p.1239-1242, Apr. 2020. DOI 10.1001/jama.2020.2648

ÍNDICE REMISSIVO

COLEÇÃO

DESAFIOS DAS ENGENHARIAS:

ENGENHARIA DE COMPUTAÇÃO 3

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena
Editora
Ano 2021

COLEÇÃO

DESAFIOS DAS ENGENHARIAS:

ENGENHARIA DE COMPUTAÇÃO 3

www.atenaeditora.com.br

contato@atenaeditora.com.br

@atenaeditora

www.facebook.com/atenaeditora.com.br

Atena Editora

Ano 2021