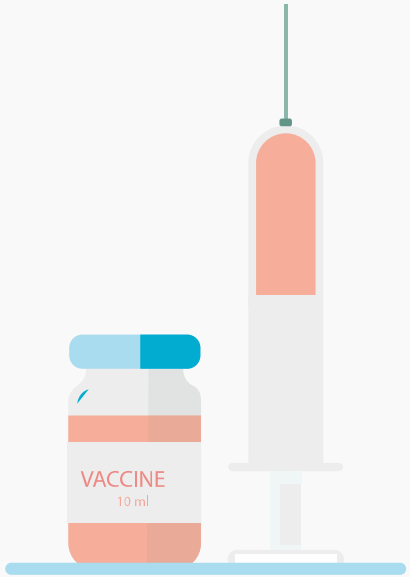
Prediction of COVID-19 diagnosis Based on symptoms

Data Mining Midterm Project

第四組

M104020042 陳亞琦
M104020052 張孫杰

目錄

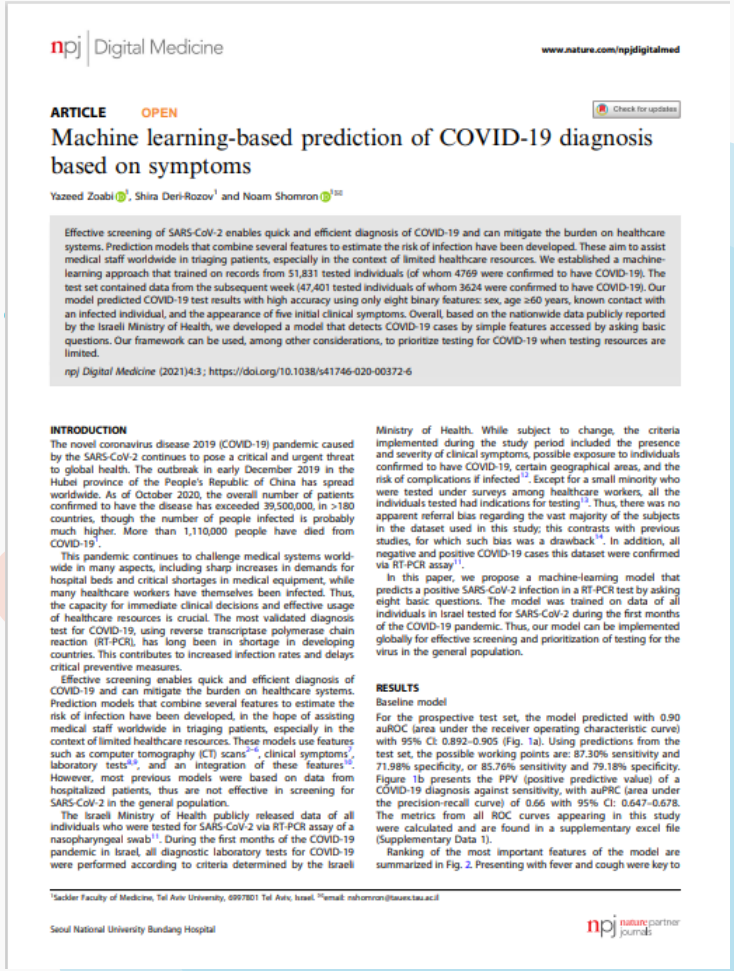


01

文獻探討

Literature review

Machine learning-based prediction of COVID-19 diagnosis based on symptoms



npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Machine learning-based prediction of COVID-19 diagnosis based on symptoms

Yazeed Zoabi[1], Shira Deri-Rozov[1] and Noam Shomron[1]✉

Effective screening of SARS-CoV-2 enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed. These aim to assist medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. We established a machine-learning approach that trained on records from 51,831 tested individuals (of whom 4769 were confirmed to have COVID-19). The test set contained data from the subsequent week (47,401 tested individuals of whom 3624 were confirmed to have COVID-19). Our model predicted COVID-19 test results with high accuracy using only eight binary features: sex, age ≥60 years, known contact with an infected individual, and the appearance of five initial clinical symptoms. Overall, based on the nationwide data publicly reported by the Israeli Ministry of Health, we developed a model that detects COVID-19 cases by simple features accessed by asking basic questions. Our framework can be used, among other considerations, to prioritize testing for COVID-19 when testing resources are limited.

npj Digital Medicine (2021)4:3 ; https://doi.org/10.1038/s41746-020-00372-6

INTRODUCTION

The novel coronavirus disease 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to pose a critical and urgent threat to global health. The outbreak in early December 2019 in the Hubei province of the People's Republic of China has spread worldwide. As of October 2020, the overall number of patients confirmed to have the disease has exceeded 39,500,000, in >180 countries, though the number of people infected is probably much higher. More than 1,110,000 people have died from COVID-19[1].

This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment, while many healthcare workers have themselves been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. The most validated diagnosis test for COVID-19, using reverse transcriptase polymerase chain reaction (RT-PCR), has long been in shortage in developing countries. This contributes to increased infection rates and delays critical preventive measures.

Effective screening enables quick and efficient diagnosis of COVID-19 and can mitigate the burden on healthcare systems. Prediction models that combine several features to estimate the risk of infection have been developed, in the hope of assisting medical staff worldwide in triaging patients, especially in the context of limited healthcare resources. These models use features such as computer tomography (CT) scans[2–4], clinical symptoms[5], laboratory tests[6,7], and an integration of these features[8]. However, most previous models were based on data from hospitalized patients, thus are not effective in screening for SARS-CoV-2 in the general population.

The Israeli Ministry of Health publicly released data of all individuals who were tested for SARS-CoV-2 via RT-PCR assay of a nasopharyngeal swab[11]. During the first months of the COVID-19 pandemic in Israel, all diagnostic laboratory tests for COVID-19 were performed according to criteria determined by the Israeli Ministry of Health. While subject to change, the criteria implemented during the study period included the presence and severity of clinical symptoms, possible exposure to individuals confirmed to have COVID-19, certain geographical areas, and the risk of complications if infected[12]. Except for a small minority who were tested under surveys among healthcare workers, all the individuals tested had indications for testing[12]. Thus, there was no apparent referral bias regarding the vast majority of the subjects in the dataset used in this study; this contrasts with previous studies, for which such bias was a drawback[14]. In addition, all negative and positive COVID-19 cases this dataset were confirmed via RT-PCR assay[11].

In this paper, we propose a machine-learning model that predicts a positive SARS-CoV-2 infection in a RT-PCR test by asking eight basic questions. The model was trained on data of all individuals in Israel tested for SARS-CoV-2 during the first months of the COVID-19 pandemic. Thus, our model can be implemented globally for effective screening and prioritization of testing for the virus in the general population.

RESULTS

Baseline model

For the prospective test set, the model predicted with 0.90 auROC (area under the receiver operating characteristic curve) with 95% CI: 0.892–0.905 (Fig. 1a). Using predictions from the test set, the possible working points are: 87.30% sensitivity and 71.98% specificity, or 85.76% sensitivity and 79.18% specificity. Figure 1b presents the PPV (positive predictive value) of a COVID-19 diagnosis against sensitivity, with auPRC (area under the precision-recall curve) of 0.66 with 95% CI: 0.647–0.678. The metrics from all ROC curves appearing in this study were calculated and are found in a supplementary excel file (Supplementary Data 1).

Ranking of the most important features of the model are summarized in Fig. 2. Presenting with fever and cough were key to

[1]Sackler Faculty of Medicine, Tel Aviv University, 6997801 Tel Aviv, Israel. ✉email: nshomron@tauex.tau.ac.il

npj nature partner journals

摘要

- 使用以色列衛生局的公開資料

  預測確診Covid-19的機率

- 在醫療資源有限時

  使用特徵篩選Covid-19患者

  可有效分配醫療資源

資料集

The following list describes each of the dataset's features used by the model:

A. Basic information:

1. Sex (male/female).
2. Age ≥60 years (true/false)

B. Symptoms:

3. Cough (true/false).
4. Fever (true/false).
5. Sore throat (true/false).
6. Shortness of breath (true/false).
7. Headache (true/false).

C. Other information:

8. Known contact with an individual confirmed to have COVID-19 (true/false).

資料來源

以色列衛生局公開資料

訓練集

本週51,831筆快篩資料
(4769筆確診)

測試集

下週47,401筆快篩資料
(3624筆確診)

Table 1. Characteristics of the dataset and the features used by the model in this study.

(#) Feature	Total $n = 99,232$		COVID-19 negative $n = 90,839$		COVID-19 positive $n = 8393$	
	n	%	*n*	%	*n*	%
(1) Sex						
Male	50,350	50.74	45,545	50.1	4805	57.2
Female	48,882	49.26	45,294	49.8	3588	42.7
(2) Age 60+						
True	15,279	15.4	13,619	14.9	1660	19.7
False	83,953	84.6	77,220	85	6733	80.2
(3) Cough						
True	14,768	14.88	10,715	11.8	4053	48.2
False	84,223	84.87	79,909	87.9	4314	51.4
(4) Fever						
True	8122	8.18	4387	4.83	3735	44.5
False	90,868	91.5	86,237	94.9	4631	55.1
(5) Sore throat						
True	1273	1.28	96	0.11	1177	14
False	95,062	95.8	88,059	96.9	7003	83.4
(6) Shortness of breath						
True	930	0.94	71	0.08	859	10.2
False	95,405	96.14	88,084	96.9	7321	87.2
(7) Headache						
True	1799	1.81	68	0.07	1731	20.6
False	94,536	95.27	88,087	96.9	6449	76.8
(8) Known contact with an individual confirmed to have COVID-19						
True	5507	5.55	1455	1.6	4052	48.2
False	93,725	94.45	89,384	98.4	4341	51.8

資料集

包含偏差資料

(5) Sore throat
(6) Shortness of breath
(7) Headache

不詳細的資料

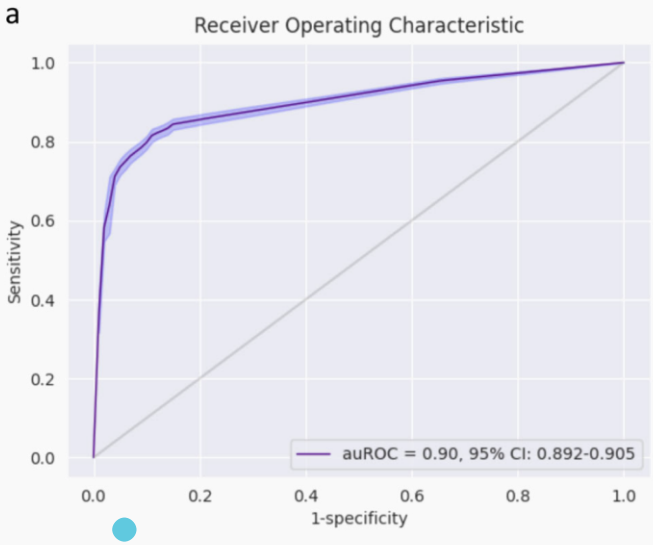
(8) Known contact with a person confirmed to have COVID-19

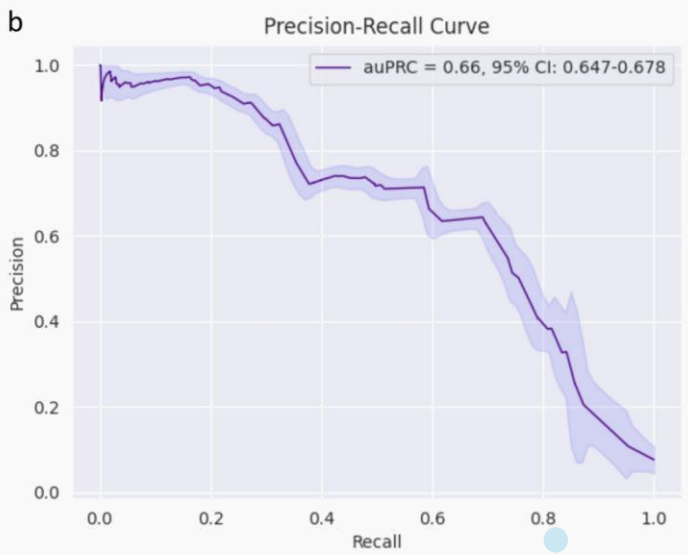
資料記載問題

- 患者自我回報的資料
- 沒有紀錄Covid-19的其他明顯症狀

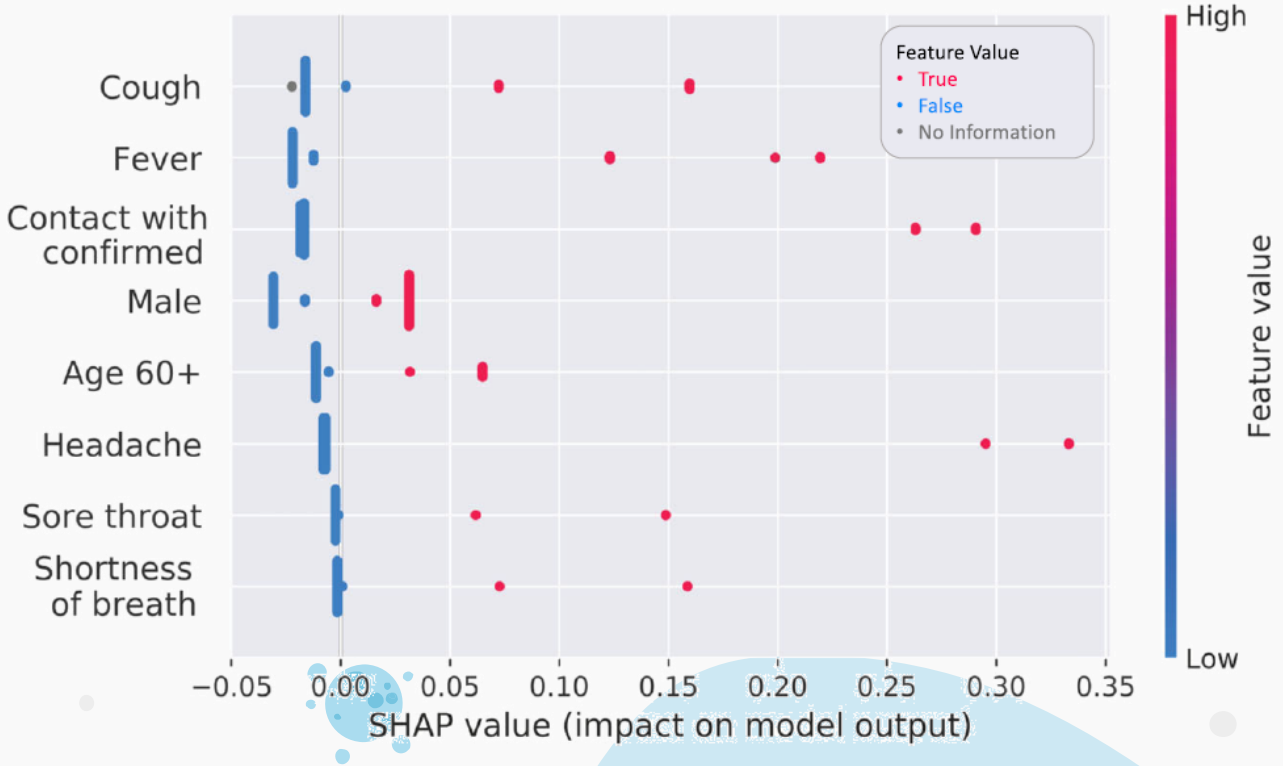
論文模型預測結果(bias)

Model	auROC	TPR	TNR
Gradient boosting(bias)	0.90	87.30%	71.98%



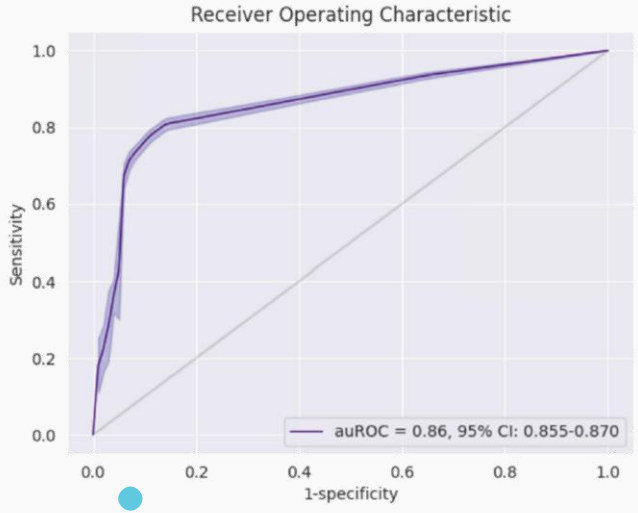


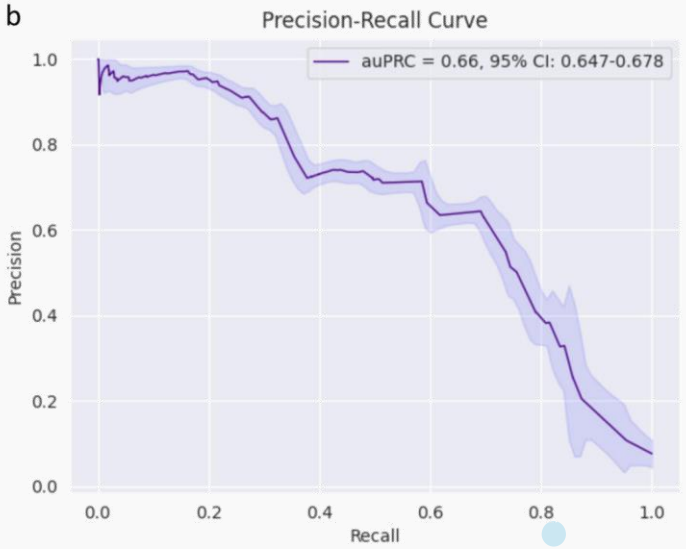
SHAP Value(bias)



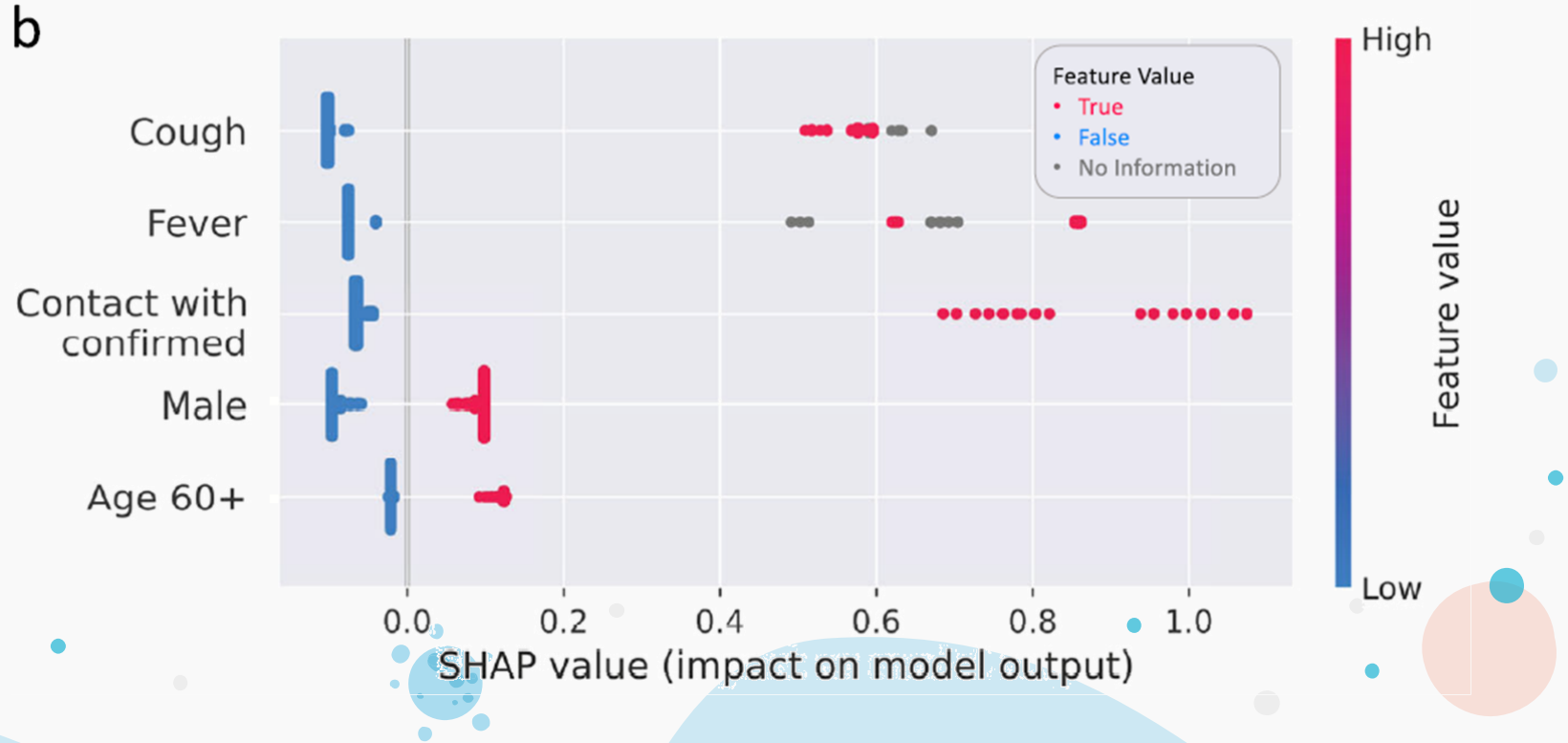
論文模型預測結果

Model	auROC	TPR	TNR
Gradient boosting	0.862	87.30%	71.98%





SHAP Value


b
Cough
Fever
Contact with confirmed
Male
Age 60+
Feature Value
True
False
No Information
High
Low
Feature value
0.0
0.2
0.4
0.6
0.8
1.0
SHAP value (impact on model output)

使用隨機測試集測試模型

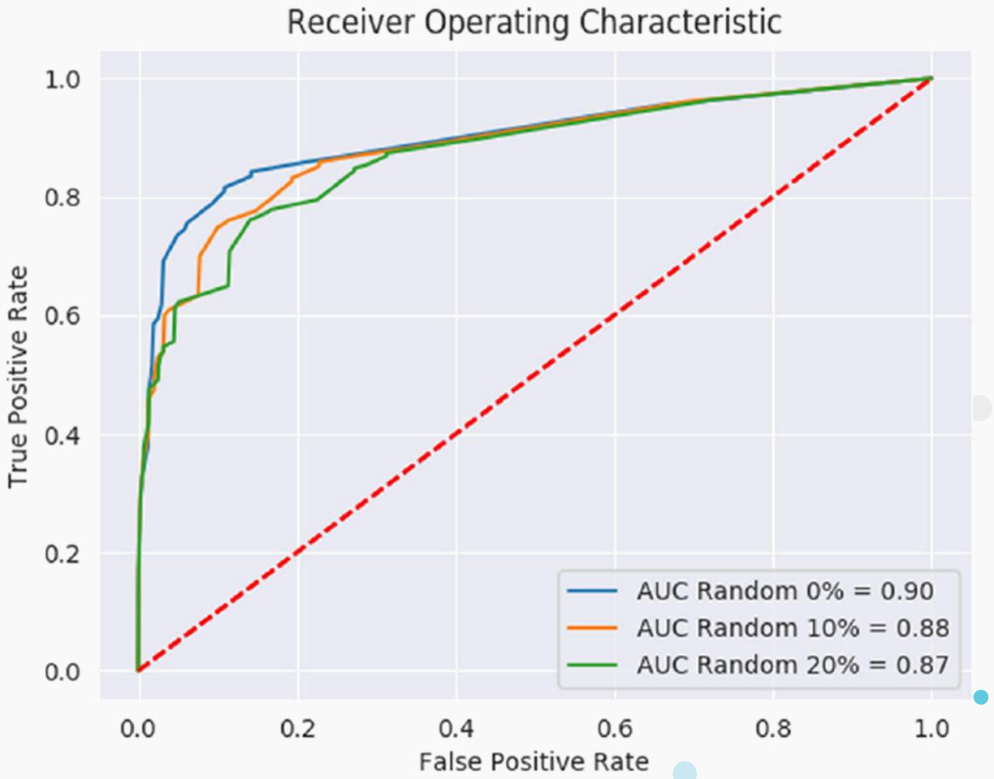
資料集	auROC
10% Random / 20% Random	0.88
移除結果值 (是否確診)的資料	0.87

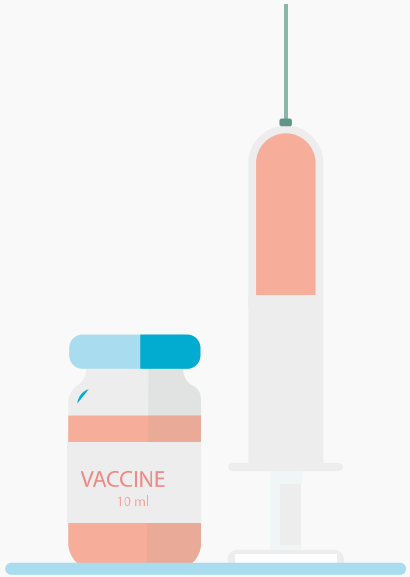
使用隨機測試集測試模型

Original Test set：藍色線

10% Random：橘色線

20% Random：綠色線





02

資料前處理

Literature review

Machine learning-based prediction of COVID-19 diagnosis based on symptoms

摘要

01

MERCURY

Mercury is the closest planet to the Sun

02

MARS

Despite being red, Mars is actually a cold place

03

JUPITER

It’s the biggest planet in the Solar System

04

VENUS

Venus has a beautiful name, but it’s terribly hot

05

SATURN

Saturn is the ringed one and a gas giant

06

NEPTUNE

Neptune is the farthest planet from the Sun

RESEARCH AND PUBLICATIONS

Venus has a beautiful name and is the second planet from the Sun

by VENUS

Mercury is the closest planet to the Sun and the smallest one in the Solar System

by MERCURY

Mars is full of iron oxide dust, which gives the planet its reddish cast

by MARS

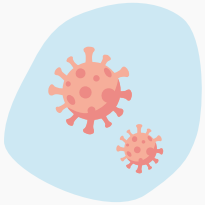
RESEARCH RESOURCES

- Here you can list your reference websites or publications.
- Here you can list your reference websites or publications.
- Here you can list your reference websites or publications.
- Here you can list your reference websites or publications.
- Here you can list your reference websites or publications.
- Here you can list your reference websites or publications.

FACTORS TO CONSIDER

VENUS

Venus has a beautiful name and is the second planet from the Sun



MERCURY

Mercury is the closest planet to the Sun and the smallest one

TRIAL TIMELINE

RESEARCH

Venus is the second planet from the Sun

PRECLINICAL

Jupiter is a gas giant and the biggest planet

EXPERIMENTATION

Despite being red, Mars is a cold place

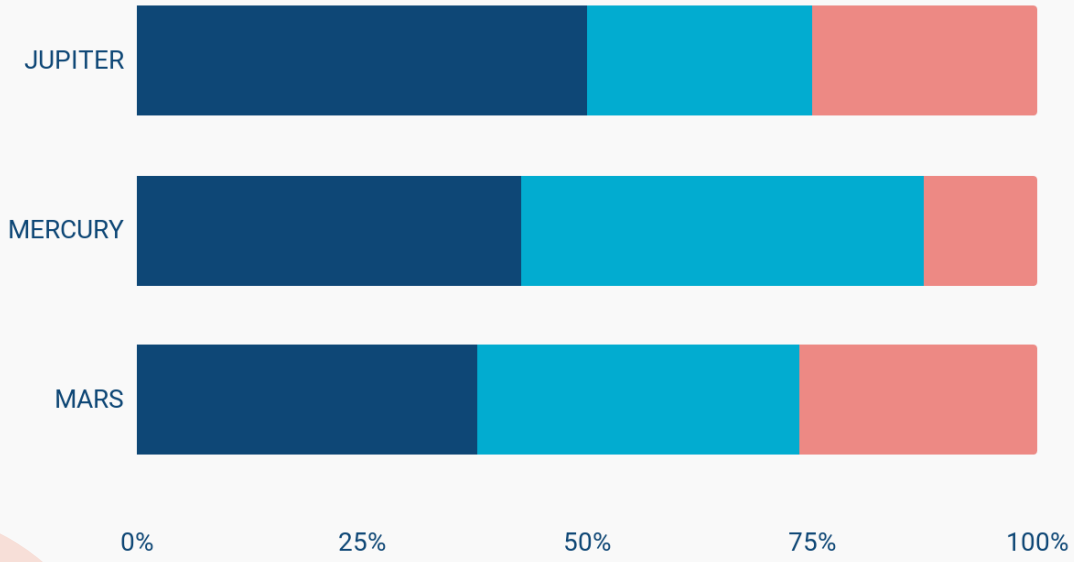
RESULTS

Saturn is a gas giant and has several rings

CONCLUSIONS

Mercury is the closest planet to the Sun

TENDENCY



MARS

Despite being red, Mars is a cold place

JUPITER

It’s the biggest planet in the Solar System

MERCURY

Mercury is the closest planet to the Sun

To modify this graph, click on it, follow the link, change the data and paste the resulting graph here

A PICTURE IS WORTH A THOUSAND WORDS

RESULTS

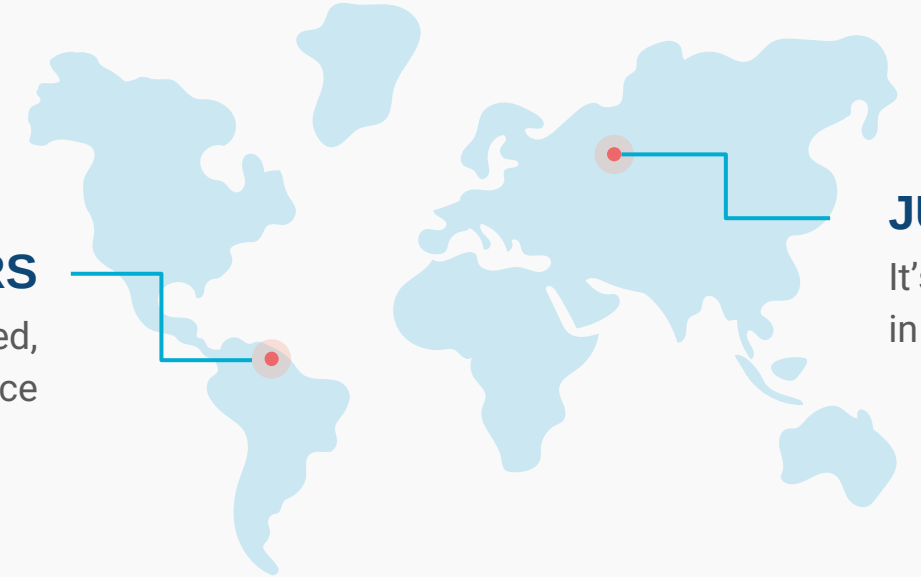
Experiment A

	OUTCOME		
TREATMENT	Test 1	Test 2	Test 3
Group 1	315	285	600
Group 2	210	390	600
Group 3	240	165	580

Experiment B

	OUTCOME		
TREATMENT	Test 1	Test 2	Test 3
Group 1	189	285	474
Group 2	210	234	444
Group 3	367	123	396

RESULTS



MARS

Despite being red,
Mars is a cold place

JUPITER

It’s the biggest planet
in the Solar System

RESULTS ANALYSIS

MARS

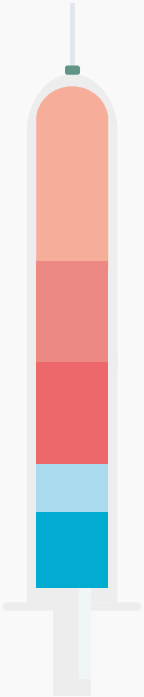
It’s actually a cold place

JUPITER

It’s the biggest planet

MERCURY

It’s a small planet



81 %

54 %

42 %

23 %

12 %

MERCURY

MARS

JUPITER

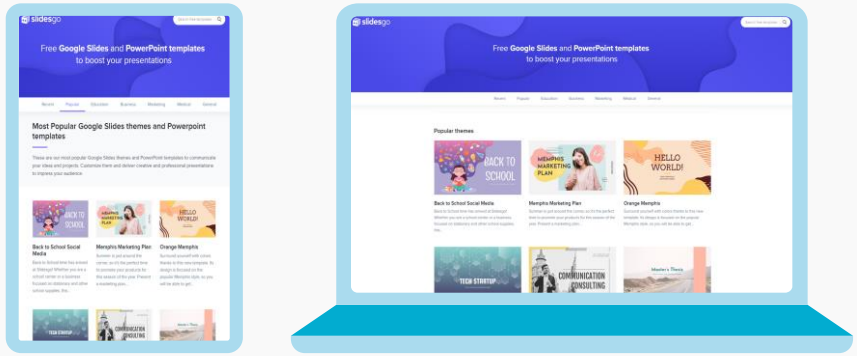
VENUS

NEPTUNE

10,000,000

Big numbers catch your audience’s attention

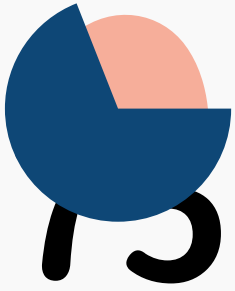
MULTIMEDIA



DESKTOP SOFTWARE

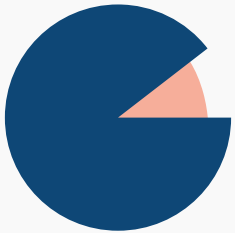
You can replace the image on the screen with your own work

SUCCESS RATE



75%

Patients cured



97%

Secondary effects

CONCLUSION

Mercury is the closest planet to the Sun and the smallest one in the Solar System—it’s only a bit larger than the Moon

THANKS!

Do you have any questions?
addyouremail@freepik.com
+91 620 421 838
yourcompany.com

facebook.com/Freepik

@Freepik_Vectors

company/freepik-company

Please keep this slide for attribution

ALTERNATIVE RESOURCES

RESOURCES

VECTOR

- Emergency ambulance van and person in hazmat suit
- Coronavirus vaccine development
- Coronavirus vaccine development concept
- Virus cure concept
- Mental health awareness and meditation concept
- Character wearing protection and holding a covid-19 test
- Science team trying to develop coronavirus cure
- Web responsive design

PHOTO

- Heart and medical dust mask copy space
- Front view of coronavirus concept with medical mask

Instructions for use

In order to use this template, you must credit **Slidesgo** by keeping the **Thanks** slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial projects.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute Slidesgo Content unless it has been expressly authorized by Slidesgo.
- Include Slidesgo Content in an online or offline database or file.
- Offer Slidesgo templates (or modified versions of Slidesgo templates) for download.
- Acquire the copyright of Slidesgo Content.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

Fonts & colors used

This presentation has been made using the following fonts:

Mukta
(https://fonts.google.com/specimen/Mukta)

Roboto
(https://fonts.google.com/specimen/Roboto)

#0e4776 #02acd0 #aadbee #f5ad9a #ed8984 #ec686a

Stories by Freepik

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out How it Works.

Pana

Amico

Bro

Rafiki

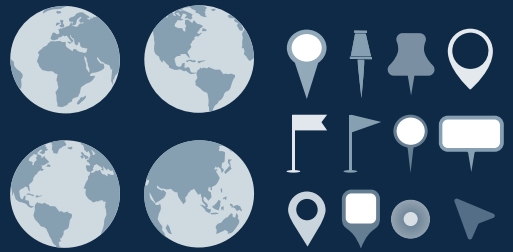
Cuate

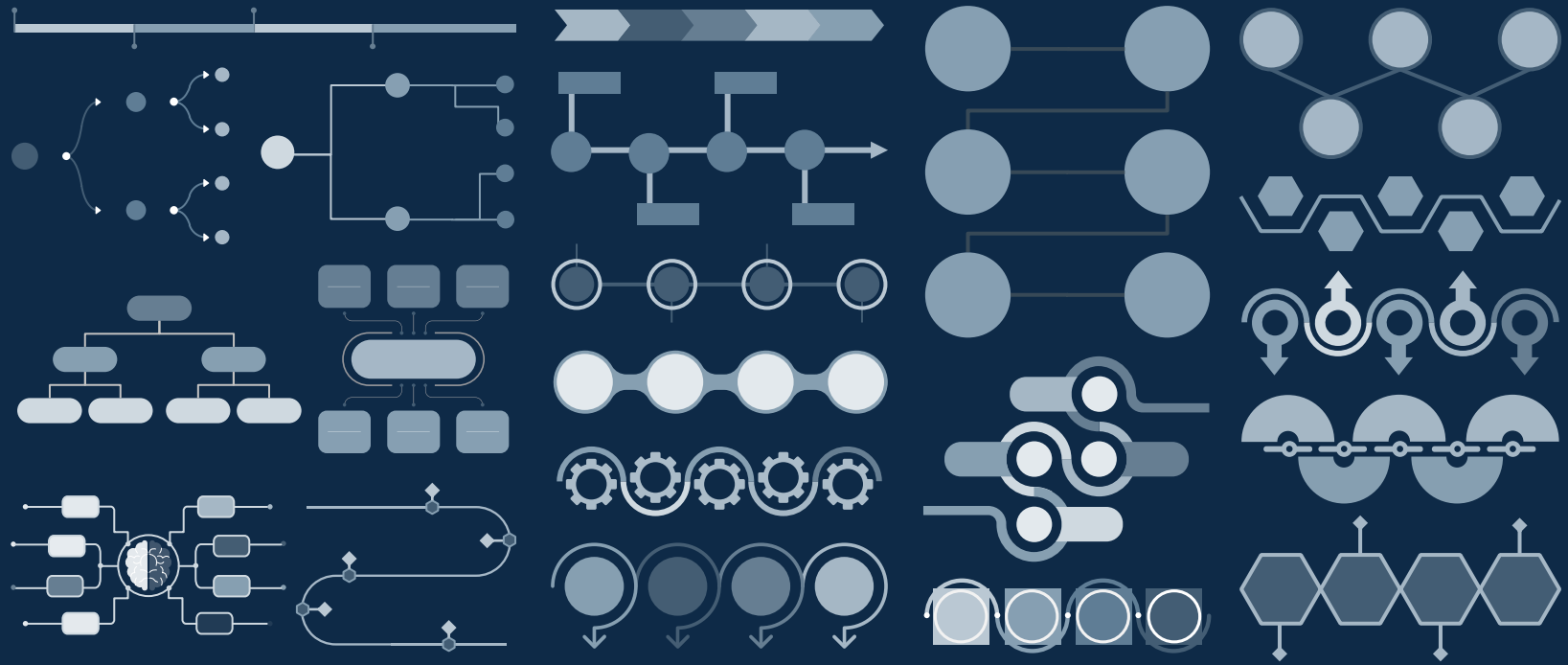
Use our editable graphic resources...

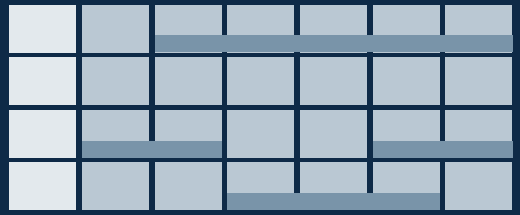
You can easily resize these resources without losing quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done. You can also look for more infographics on Slidesgo.

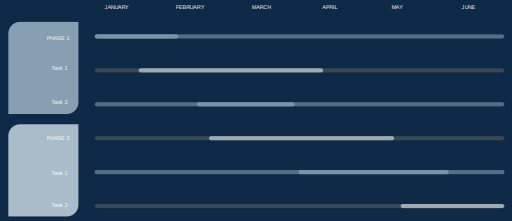


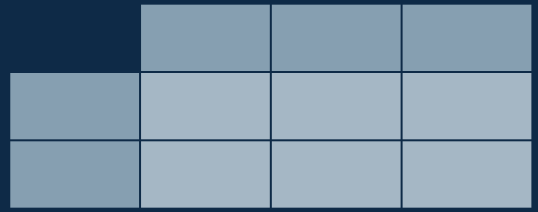


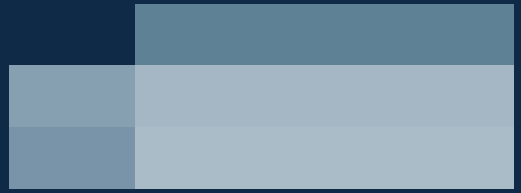



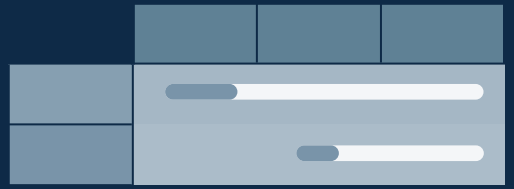


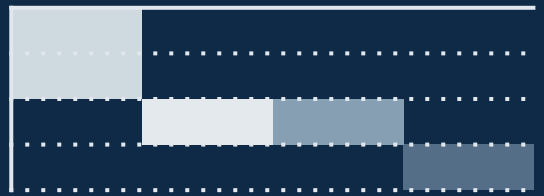


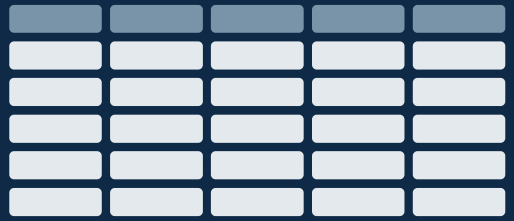

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

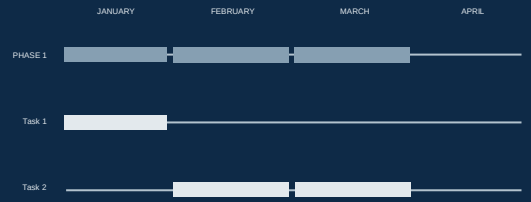


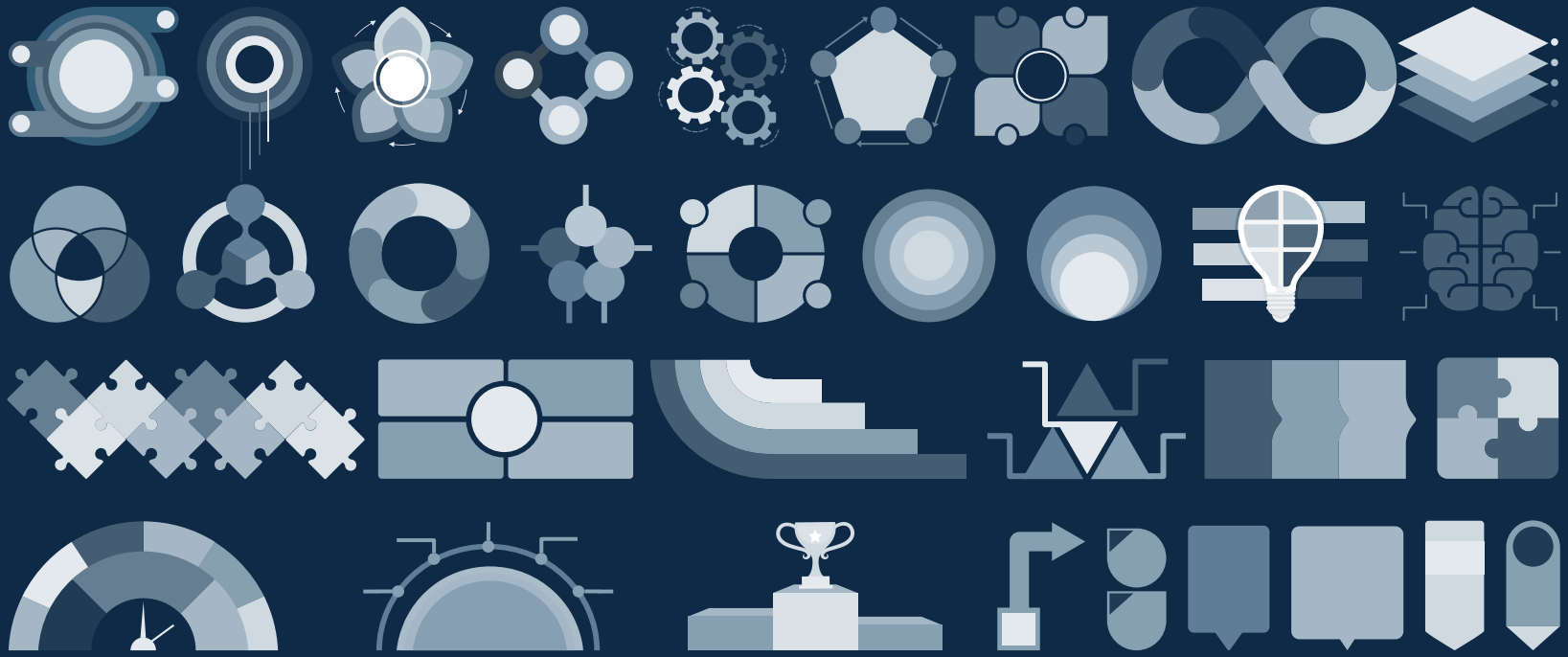









JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2



...and our sets of editable icons

You can resize these icons without losing quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

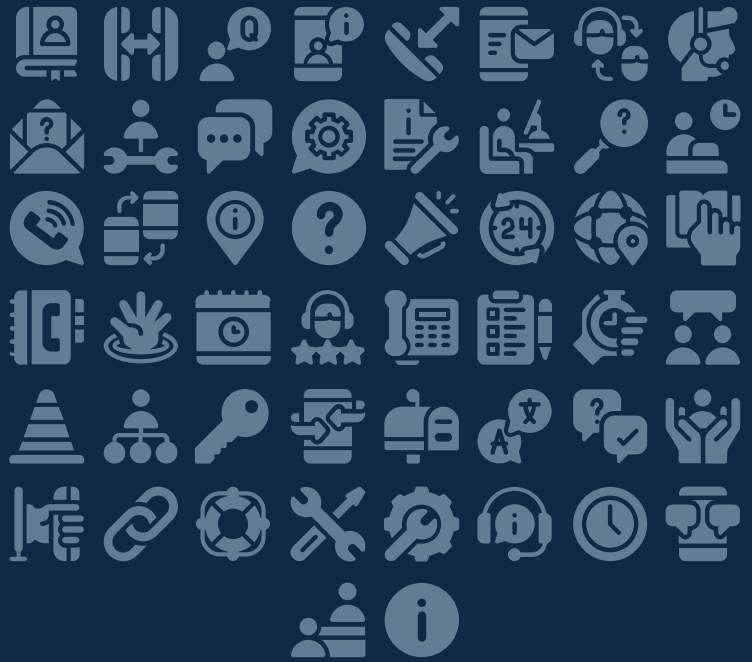
Educational Icons

Medical Icons

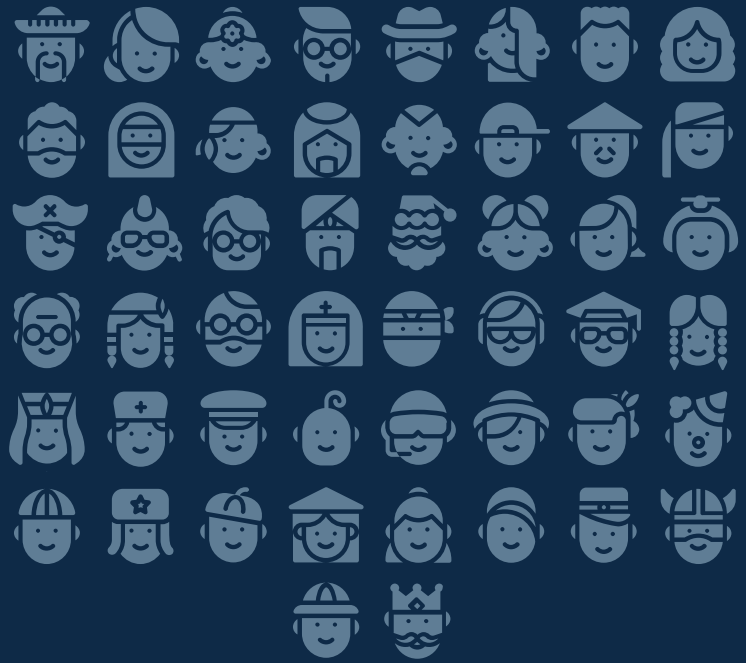
Business Icons

Teamwork Icons

Help & Support Icons



Avatar Icons



Creative Process Icons

Performing Arts Icons

Nature Icons

SEO & Marketing Icons


slidesgo