Plan Overview

A Data Management Plan created using DMPTool

Title: COVID_19 Detection using machine learning

Creator: Yewoinhareg Girma

Affiliation: University of Gondar (uog.edu.et)

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

COVID-19 pandemic has rapidly affected our day to-day life disrupting the world trade and movements. The development of technology has a significant impact on every aspect of life, whether it is the medical industry or any other profession. By making decisions based on the analysis and processing of data, artificial intelligence has demonstrated promising outcomes in the field of health care. The most crucial action is an early detection of a life-threatening illness to stop its development and spread. Highly contagious COVID-19 is a disease that requires immediate attention as it has spread global. Due to its rapid speed of spreading comes the need for a system which can be used to detect the virus. With the increase in use of technology, lots of data about COVID-19 is readily available at our fingertips, which can be used to obtain important information about the virus. In this project, we compared the accuracies of different machine learning algorithms in predicting COVID-19 and used the most accurate one in the final model testing.

Start date: 06-20-2023

End date: 07-20-2023

Last modified: 06-17-2023

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

---

COVID_19 Detection using machine learning

Data Collection

What data will you collect or create?

for this project i will collect text based data that include symptom demographyc informaton and also test of result

How will the data be collected or created?

Clinical Data: This includes information such as patient demographics, symptoms, medical history, and laboratory test results. This data can be obtained from electronic medical records (EMRs) or other medical databases.

COVID-19 Test Results: Data from COVID-19 tests, such as RT-PCR tests, can be used to train the machine learning model to detect the virus.

Documentation and Metadata

What documentation and metadata will accompany the data?

To ensure the reproducibility and transparency of a project on COVID-19 detection using machine learning, it is important to include documentation and metadata with the data. Here are some examples of documentation and metadata that can accompany the data:

Data Description: A detailed description of the data, including the types of data collected or created, the source of the data, and any preprocessing or cleaning that was performed.
Data Format: Information on the format of the data, such as file type and structure, as well as any software or tools required to read or manipulate the data.
Data Quality: Information on the quality of the data, including any known limitations or issues with the data, and any quality control measures that were taken during data collection or creation.
Data Labeling: Information on how the data was labeled, including the criteria used for labeling and any inter-rater reliability measures that were taken.
Ethical Considerations: Documentation of any ethical considerations involved in collecting or creating the data, such as obtaining informed consent or ensuring patient privacy.
Code and Algorithms: Documentation of any code or algorithms used in data preprocessing, cleaning, or analysis, including any hyperparameters or tuning that was performed.
Citations: Proper citations for any sources of data, code, or algorithms used in the project

Ethics and Legal Compliance

How will you manage any ethical issues?

The development and implementation of a project on COVID-19 detection using machine learning must consider ethical issues to ensure that the project is conducted in a responsible and fair manner. Here are some approaches to managing ethical issues:

Informed Consent: If the project involves collecting data from human subjects, informed consent must be obtained from participants. This involves providing participants with information about the project, including its purpose, potential risks and benefits, and their rights as participants.
Privacy and Confidentiality: To protect patient privacy, sensitive data must be stored securely and anonymized where appropriate. Researchers must develop clear policies and procedures for managing and protecting data, including procedures for data sharing and storage.
Bias and Fairness: Machine learning models can be biased if they are trained on data that is not representative of the population. Researchers must be aware of the potential for bias and take steps to ensure that the data is diverse and representative.
Transparency and Interpretability: Machine learning models can be difficult to interpret, which can raise concerns about the fairness and accountability of the system. Researchers must develop methods for explaining how the model arrived at its decision and enable stakeholders to understand how the model works.
Ethical Review: The project should undergo ethical review to ensure that it meets ethical standards. This can involve obtaining approval from an institutional review board (IRB) or ethics committee.
Responsible Data Use: Researchers must ensure that the data is used responsibly and that the results of the project are communicated accurately and responsibly. This includes being mindful of the potential impact of the project on society and taking steps to mitigate any negative consequences.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

Legal Review: i should conduct a legal review to identify any potential copyright or IP/IPR issues related to the data, algorithms, or software used in the project. This can involve consulting with legal experts or conducting a thorough search of existing patents and publications.
Licensing: i must obtain the necessary licenses and permissions to use any copyrighted or patented materials involved in the project. This can involve negotiating with copyright holders or obtaining licenses through open-source or Creative Commons models.
Attribution: i must ensure that proper attribution is given to any copyrighted or patented materials used in the project. This can involve citing sources or providing acknowledgments in publications or presentations.
Ownership: i must be clear on who owns the intellectual property rights to the data, algorithms, or software used in the project. This can involve reviewing any agreements or contracts involved in the project and clarifying ownership and licensing arrangements.
Open Science: i can consider open science principles, such as open data and open-source software, to promote transparency and collaboration while also avoiding copyright and IP/IPR issues. This can involve making data and algorithms available through open-access repositories and using open-source software tools

Storage and Backup

How will the data be stored and backed up during the research?

To ensure the security and accessibility of the data during the research, it is important to have a proper data storage and backup strategy. Here are some approaches to storing and backing up data during the research:

Secure Storage: Data should be stored securely to prevent unauthorized access or loss. This can involve using secure cloud services or local servers with access controls, firewalls, and encryption.
Data Backup: Regular backups of the data should be created to prevent loss due to hardware failure, natural disasters, or cyber attacks. This can involve creating backups on external hard drives, cloud storage, or other secure off-site locations.

How will you manage access and security?

Managing access and security is an important consideration for any project, including COVID-19 detection using machine learning. Here are some approaches to managing access and security:

Access Controls: Researchers must implement access controls to ensure that only authorized personnel can access the data and systems involved in the project. This can involve using passwords, multi-factor authentication, or role-based access controls.
Data Encryption: Data should be encrypted to prevent unauthorized access during storage, transfer, and processing. This can involve using industry-standard encryption algorithms such as AES or RSA.

Data Anonymization: Sensitive data should be anonymized or de-identified to prevent unauthorized access or disclosure. This can involve removing personally identifiable information or using techniques such as differential privacy.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

For a project on COVID-19 detection using machine learning, the following types of data are of long-term value and should be retained, shared, and/or preserved:

Raw Data: The raw data used to train the machine learning model, such as medical images, clinical data, and COVID-19 test results, should be retained and preserved. This can enable future researchers to re-analyze the data using newer algorithms or techniques.
Preprocessed Data: The preprocessed data used to create the machine learning model, such as augmented data or feature engineering, should also be retained and preserved. This can enable future researchers to understand the preprocessing techniques used and replicate the results.
Trained Model: The trained machine learning model should be retained and preserved, along with any relevant documentation or code. This can enable future researchers to use the model for similar applications or improve upon the model using newer data or techniques.
Evaluation Metrics: The evaluation metrics used to assess the performance of the machine learning model, such as accuracy, sensitivity, and specificity, should be retained and preserved. This can enable future

researchers to compare the results with newer models or techniques.
Metadata: Metadata, such as data descriptions, data formats, and data quality assessments, should be retained and preserved. This can enable future researchers to understand the context and quality of the data used in the project

What is the long-term preservation plan for the dataset?

A long-term preservation plan for the dataset in a project on COVID-19 detection using machine learning should ensure that the data is available and accessible for future use. Here are some key steps to include in a preservation plan:

Selection of Preservation Format:

Data Documentation

Data Storage

Data Backups

Data Access

Data Retention

Data Licensing

Data Sharing

How will you share the data?

Sharing the data in a project on COVID-19 detection using machine learning can promote transparency, collaboration, and accelerate research. Here are some approaches to sharing the data:

Open-Access Repositories: The data can be shared through open-access repositories, such as Zenodo, Dryad, or Figshare. These repositories provide a platform for researchers to share their data openly and provide a persistent identifier to ensure that the data is easily discoverable and citable

Are any restrictions on data sharing required?

Yes, there may be restrictions on data sharing required for a project on COVID-19 detection using machine learning. These restrictions may be necessary to protect the privacy and confidentiality of individuals, comply with legal and ethical requirements, or protect the intellectual property rights of the data creators

Responsibilities and Resources

Who will be responsible for data management?

In a project on COVID-19 detection using machine learning, data management is a critical task that requires a designated individual or team responsible for overseeing the data throughout the project's lifecycle. Here are some potential roles and responsibilities for data management:

Data Manager: A data manager can be assigned to oversee all aspects of data management, including data collection, preprocessing, storage, backup, sharing, and preservation. The data manager can be responsible for ensuring that the data is

What resources will you require to deliver your plan?

To deliver a plan for COVID-19 detection using machine learning, the following resources may be required

Hardware,Software,Data Storage,Data Management Software,Data Management Software,Funding,and Institutional Support

Planned Research Outputs

Model representation - "covid _19 detection using machine learning"

The research output for a project on COVID-19 detection using machine learning can take various forms, depending on the research question, methodology, and data used. Here are some potential research outputs:

Machine Learning Models: The primary research output may be machine learning models that can accurately detect COVID-19 in medical images or clinical data. These models can be used to guide diagnosis and treatment decisions and improve patient outcomes.
Performance Metrics: The project may produce performance metrics, such as accuracy, sensitivity, specificity, and area under the curve (AUC), that demonstrate the effectiveness of the machine learning models. These metrics can be used to compare different models or techniques and guide future research.
Data Analysis: The project may produce data analysis that provides insights into the characteristics of COVID-19 and its impact on patients. This can include statistical analyses, data visualizations, and other exploratory data analysis techniques.
Data Resources: The project may produce data resources, such as datasets, annotations, or preprocessing techniques, that can be used by other researchers to develop or improve machine learning models for COVID-19 detection.
Research Papers: The project may produce research papers that describe the methodology, results, and implications of the project. These papers can be published in peer-reviewed journals and presented at conferences to disseminate the research findings to the wider scientific community.
Software: The project may produce software tools, such as open-source libraries or applications, that enable researchers to develop or apply machine learning models for COVID-19 detection

---

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
covid _19 detection using machine learning	Model representation	2023-07-28	Open	Academic Torrents		None specified	CERIF (Common European Research Information Format)	No	No