Development of AI in Tuberculosis Disease Diagnosis

By: Fenti Irnawati Supandi
For: A.I Competition DQLAB, April 2023

Rationale

The development of *Artificial Intelligence* (AI) is currently accelerating, with the emergence of various AI technologies that help solve various complex problems in various fields. One of the trending AI technologies is *GPT* (*Generative Pre-trained* Transformer) *chat* developed by *OpenAI*. This technology is capable of generating text similar to that produced by humans and is used in various applications, such as *chatbots* and text generation.

Seeing the very interesting potential of *GPT chat* technology, I became interested in further exploring the capabilities of this technology. One of the ideas that emerged was the utilization of *GPT chat* technology in the health sector, especially in the handling of tuberculosis (TB) in Indonesia. TB is an infectious disease that is still a public health problem in Indonesia, and the use of AI technology can help improve the accuracy of diagnosis and accelerate the treatment of TB patients.

Tuberculosis or TB remains a significant global health problem, including in Indonesia. Despite prevention and treatment efforts, the number of TB cases in Indonesia is still very high. According to data from the Indonesian Ministry of Health, in 2022 there were more than 700,000 TB cases in Indonesia, an increase of 61.98% compared to the previous year. Although there was a decrease in cases in 2020, the trend of TB cases in Indonesia still shows an increase in the last two years.

AI can be one of the solutions and innovations to help reduce the number of TB cases in Indonesia. By using AI technology, it is hoped that more effective and efficient TB prevention and treatment programs can be created. This will certainly help the government in achieving the 2030 TB elimination target.

As an individual who cares about public health issues, I believe that the utilization of AI in the health sector can have a major impact in addressing the TB problem in Indonesia. Therefore, cooperation from all parties is needed to optimally utilize AI technology to help achieve the 2030 TB elimination target.

AI Development Concept in TB Diagnosis

AI concept in TB with *supervised learning* method:

1. Introduction
    - Definition of TB
    - Definition of AI
    - Types of AI
    - *Supervised learning*
2. Data Collection
    - TB data source
    - *Data preprocessing*
    - Data labeling
3. Model Making
    - Feature selection
    - Algorithm selection
    - Model parameter initialization
4. Training Model
    - Train and test data split
    - Model training
    - Model validation
5. Model Evaluation
    - Model performance measurement
    - *Confusion matrix*
    - *Precision, recall, F1-score*
6. Model Usage
    - Model application on new data
    - TB patient monitoring
    - Providing proper care
7. Conclusion
    - Benefits of AI in TB management
    - Challenges and constraints in the use of AI in TB
    - Opportunities for AI development in other disease treatments

AI concept in TB with unsupervised learning or reinforcement learning method:

1. Introduction
    - Definition of TB
    - Definition of AI
    - Types of AI
    - *Unsupervised learning*; *Reinforcement learning*
2. Data Collection
    - TB data source
    - *Data preprocessing*
3. *Unsupervised Learning*
    - Clustering TB data
    - The k-means method
    - Cluster analysis
4. Reinforcement Learning
    - The concept of *reinforcement learning*
    - Modeling the TB environment
    - *Reward function* creation
5. Model Making
    - Feature selection
    - Algorithm selection
    - Model parameter initialization
6. Training Model
    - Model training
    - Model validation
7. Model Evaluation
    - Model performance measurement
    - *Confusion matrix*
    - *Precision, recall, F1-score*
8. Model Usage
    - Model application on new data
    - TB patient monitoring
    - Providing proper care
9. Conclusion
    - Benefits of AI in TB management
    - Challenges and constraints in the use of AI in TB
    - Opportunities for AI development in other disease treatments

Conceptual Discussion

Definition of TB

According to the Indonesian Ministry of Health, TB or *Tuberculosis* is an infectious disease caused by the bacteria *Mycobacterium tuberculosis* that can be transmitted through sputum droplets. Tuberculosis is not a hereditary disease or a curse and can be cured with regular treatment, supervised by Supervision of Taking Medicine (PMO). Most TB germs attack the lungs but can also affect other organs.

According to WebMD, *Tuberculosis* (TB) is a contagious infection that usually affects the lungs. However, it can also spread to other parts of the body, such as the brain and spine. A type of bacteria called *Mycobacterium tuberculosis* causes the disease.

Meanwhile, according to NCBI, *Tuberculosis* (TB) is a human disease caused by *Mycobacterium tuberculosis.* Typically, TB affects the lungs, making pulmonary disease the most common manifestation. Other commonly affected organ systems include the respiratory system, gastrointestinal (GI) system, *lymphoreticular system*, skin, central nervous system, musculoskeletal system, reproductive system, and liver.

From these three quotes, it can be concluded that TB or *Tuberculosis* is a contagious infectious disease caused by the bacteria *Mycobacterium tuberculosis*. The disease can affect various organs of the body, but the lungs are the most common manifestation. TB can be cured with regular medication and supervised by a supervisor (PMO).

AI in Healthcare

The concept of AI for healthcare encompasses the use of AI technology in applications and software to improve human health and well-being. There are many applications of AI in healthcare, including disease diagnosis, new drug development, drug inventory management, patient data management, public health management, and more.

AI applications in healthcare usually involve machine learning algorithms, which allow systems to learn and improve their performance over time. These algorithms can be used to analyze large health data, such as patient medical histories and medical records, as well as genetic, environmental, and behavioral data.

One of the major applications of AI in healthcare is disease diagnosis. Using AI technology, doctors and healthcare professionals can identify disease symptoms and diagnose medical conditions more accurately and efficiently. AI can also help doctors in planning the care and treatment of patients.

In addition, AI technology can also help in the development of new drugs. Machine learning algorithms can be used to analyze clinical and preclinical data for

identify more effective and safe drugs for the treatment of certain diseases.

However, the application of AI in healthcare also faces challenges, including concerns about data privacy, cybersecurity, and data misinterpretation. Therefore, the use of AI technology in healthcare must be done carefully and follow applicable regulations to ensure safety and optimal quality of healthcare services.

Types of AI

Supervised Learning:

Supervised learning is one of the most popular and widely used machine learning methods. In this method, AI models are trained using data that has been previously labeled or classified. These labels or classifications are used to help the AI model learn the patterns contained in the data so that it can make predictions or classifications on data that has not been labeled.

An example of the application of supervised learning in TB treatment is to use an AI model that is trained using TB patient data that has been labeled as a TB patient or not. This AI model can be used to assist doctors in making a diagnosis of TB in new patients based on symptoms experienced or laboratory test results.

Unsupervised Learning:

Unsupervised learning is one of the machine learning methods that does not require pre-labeled or classified data. In this method, AI models are trained using raw data without classification. The purpose of using this method is to identify patterns in the data without the help of labels or classification.

An example of the application of *unsupervised* learning in TB treatment is to use AI models to identify patterns in the laboratory test data of TB patients. By identifying these patterns, the AI model can assist doctors in making a diagnosis of TB in new patients based on the laboratory test results obtained.

Reinforcement Learning:

Reinforcement learning is a machine learning method that uses the concept of *reward* or *reward function* to help AI models learn patterns in data. In this method, the AI model is given a specific task or mission and will get a reward or penalty based on the performance shown in carrying out the task.

An example of the application of reinforcement learning in TB treatment is to use an AI model trained to identify TB patients who have a higher risk of TB drug side effects. This AI model will be *rewarded* when it successfully identifies patients with a higher risk of TB drug side effects, so that it can assist doctors in determining the right drug dose for these patients.

Data Collection

Data Collection: TB data collection is done from various sources, such as hospitals, clinics, health centers, or laboratories. Data collected can be in the form of medical data, such as medical records, laboratory test results, X-rays or CT scans, and other data related to TB. In addition, data can also be collected from other sources such as medical journals or scientific publications.

The importance of good data collection in TB management using AI is critical because the more and better quality data collected, the more accurate and *reliable* AI models can be generated. Therefore, it is important to ensure that the data collected is of good quality and appropriate to the needs.

Data Preprocessing: The collected data needs to be processed and prepared before it is used to train the AI model. This process is called *data preprocessing*. *Data preprocessing* includes several steps, such as *data cleaning*, *data transformation*, and *data reduction*.

Data cleaning is done to remove irrelevant data, incomplete data, or inaccurate data. *Data transformation is* done to convert raw data into a form that can be used by AI models, such as converting text data into vectors, or converting image data into a form that can be analyzed by AI models.

Data reduction is done to reduce the dimensions of very large data. One of the commonly used *data reduction* methods is *Principal Component Analysis* (PCA). This method reduces the dimension of data by reducing dimensions that have low variance so that only dimensions that have high variance are retained.

Good *data preprocessing* is essential to ensure the quality and accuracy of the resulting AI model. Therefore, before data is used to train AI models, make sure the data has gone through the correct *data preprocessing* process and according to the needs.

Supervised Learning

Data collection in the context of AI TB generally uses *supervised learning*, which is a machine learning method where algorithms learn patterns and rules from training data that has been labeled by humans.

In the context of TB, data collection is done by collecting data samples of TB patients who have been diagnosed through laboratory and radiology examinations. The data is then labeled in the form of TB status and the type of TB suffered by the patient, be it pulmonary TB or extra-pulmonary TB.

In addition, data can also be collected from the medical records and medical histories of patients who have been diagnosed with TB. This data includes information on symptoms, disease history, laboratory and radiology test results, as well as the type of medication and duration of treatment given to the patient.

Data collection is carried out continuously and periodically to update the data and improve the accuracy of the AI model in diagnosing and treating TB. The data is then processed using various techniques such as natural language processing, image analysis, and structured data processing to produce an AI model that can be used in diagnosing and treating TB.

Unsupervised Learning

Unsupervised Learning as explained earlier, *Unsupervised Learning* is a machine learning method that does not require pre-labeled or classified data. In this method, AI models are trained using raw data without classification. The purpose of using this method is to identify patterns in the data without the help of labels or classifications.

Clustering TB *Data* One of the *Unsupervised Learning* techniques commonly used in TB treatment is *clustering*. *Clustering* is a technique to group data into groups that have similarities based on certain characteristics. In TB treatment, data collected such as laboratory test results, patient symptoms, and treatment history data can be grouped into similar groups.

K-Means Method One of the *clustering* algorithms often used in TB treatment is K-Means. K-Means is a *clustering* algorithm that works by grouping data into k groups or *clusters*. This algorithm works by calculating the distance between each data and the centroid of each *cluster*. Data will be grouped into the *cluster* that has the closest *centroid*.

Cluster Analysis After the TB data has been successfully grouped using *clustering* techniques, the next step is to perform *cluster* analysis. *Cluster* analysis is performed to understand the characteristics or traits of each group or *cluster* formed. In TB treatment, cluster analysis can assist clinicians in identifying factors that contribute to the occurrence of TB and also help in developing more effective treatment strategies.

Reinforcement Learning

Reinforcement Learning is a machine learning method where the model learns from repeated interactions with the environment. In this method, the AI model is not given direct access to data or training examples. Instead, the AI model gains experience by performing actions in an environment and receiving feedback in the form of *rewards* or punishments.

There are several key concepts, namely *agent, environment, state, action, reward,* and *policy. An agent* is an entity that gains experience from interacting with the environment. *Environment* is the world in which the *agent* interacts. *State* is the condition of the environment at a certain time. Action is an action that can be performed by the *agent* in a *state. Reward* is the numerical value given to the *agent* after performing an action. *Policy* is the strategy used by the agent to choose the action to be performed in a *state*.

Modeling the TB Environment In the context of TB management, the environment can be viewed as a complex biological system consisting of human organisms and the TB-causing agent. In modeling the TB environment, clinical data, such as laboratory test results, treatment history, and geographic information, can be used to represent the state of the environment. Meanwhile, actions that can be taken by the *agent can* include examining patients, treating patients, or other actions deemed relevant to reduce the spread of TB.

Reward function is one of the important components in *Reinforcement* Learning. *Reward functions* are used to *reward* or punish the agent after performing an action in a *state*. In TB modeling, the *reward function* can be designed to give a high *reward* if the agent successfully prevents the spread of TB or successfully treats a patient. Conversely, the *reward function* can provide a high punishment if the *agent* fails to treat the patient or worsens the patient's condition.

Model Making

Model Building In this stage, a *machine learning* model is built to classify TB data. The model building process includes three important stages, namely feature selection, algorithm selection, and model parameter initialization.

Feature Selection Feature selection is the process of selecting the most relevant and significant features or variables in classifying data. In TB data, features that can be considered include clinical symptoms, laboratory test results, treatment history, and geographic information. Proper feature selection can help improve model performance and reduce *overfitting*.

Algorithm selection Algorithm selection is the process of choosing the most suitable *machine learning algorithm* to solve the TB data classification problem. Some algorithms that can be considered in algorithm selection include *Decision Tree, Naive Bayes, Support Vector Machine (SVM),* and *Neural Network*. Proper algorithm selection can help improve model accuracy and efficiency.

Model Parameter Initialization Once an appropriate algorithm has been selected, the next step is model parameter initialization. Model parameters are values that are used in *machine learning* algorithms to optimize model performance. Some model parameters that need to be initialized include the number of *hidden layers* and *neurons* in *Neural Network*, kernel in SVM, and separation criteria in *Decision Tree. Proper* initialization of *model parameters can help improve model performance and prevent overfitting*.

In the entire modeling process, it is important to perform continuous model evaluation. This is done by validating the model on data that has never been seen before and optimizing the model parameters according to the validation results. In this way, the model can be continuously improved and optimized to achieve the best performance in TB data classification.

Training Model

In the *Training Model* stage, the *machine learning model* that has been created in the previous stage will be trained using *training data*. Model training is done by providing a number of training data into the model and optimizing the model parameters according to the model evaluation results.

Model Training consists of several stages such as:

- Model initialization: the model parameter values that have been selected and initialized in the Model Generation stage are used as the start for model training.
- *Forward propagation*: training data is fed into the model and output is generated from each *layer* until it reaches the *output layer*.
- *Loss function* calculation*: the* difference between actual *output* and predicted *output is* calculated and used as an evaluation of model performance at each *epoch* (training round).
- *Backward propagation: the* error value is generated from the *loss function* calculation and returned from the *output layer* to the *hidden layer* by updating the *weights* in each *layer.*
- Parameter optimization: after calculating the *loss function* and updating the weights at each *layer*, the model parameters are changed to improve the performance of the model. One method that is often used to optimize parameters is to use the *gradient descent* method.

Model Validation After model training, the next step is to validate the model on validation data. Validation data is data that is different from training data and is used to objectively measure model performance. Model validation is done by providing validation data into the model and calculating the accuracy, *precision, recall, and F1 score* values. Good model validation can provide an overview of how well the model can be used to classify TB data effectively and efficiently.

In the whole process of model training and validation, it is important to perform continuous model evaluation. This is done by optimizing the model parameters and testing the model performance on validation data that has not been seen before. In this way, the model can be continuously improved and optimized to achieve the best performance in TB data classification.

Model Evaluation

Model Evaluation After the model training and validation process, the next stage is Model Evaluation. At this stage, the model that has been trained and validated will be tested for performance on *test* data that has never been used in the training or validation process.

Model performance measurement can be done using various methods such as *accuracy*, *precision*, *recall, F1-score*, and *area under the curve (*AUC).

Confusion Matrix to measure model performance, one of the most common ways is to use *confusion matrix*. *Confusion matrix* is a table used to calculate the *true positive* (TP), *true negative* (TN), *false positive (*FP), and *false negative* (FN) values. From this *confusion matrix*, we can calculate various model evaluation matrices such as accuracy, precision, *recall,* and *F1-score*.

Precision, Recall, and F1-score Precision is the ratio of *true positives* to the total data predicted to be positive. *Recall* (sensitivity) is the ratio between *true positives* and total data that are actually positive. *F1-score* is the harmonic mean between precision and *recall*. These three metrics provide useful information about the model's performance in classifying TB data.

Once the model evaluation has been completed, the results can be used to determine whether the model can be used to classify TB data effectively and efficiently. If the performance of the model is still not good enough, then the model can be improved and optimized again by conducting more thorough training and validation stages of the model.

Model Usage

Explainable AI (XAI) is a method or technique in AI development that aims to help understand and explain the decisions made by AI models. By using XAI techniques, we can find out what factors influence the decisions made by AI models, so that it can help us understand the reasons behind the decisions made.

In TB treatment, XAI techniques can be used to provide a clear and transparent explanation of the rationale for the diagnosis or treatment recommendation given by the AI model. For example, we can use XAI techniques to explain what factors influence the AI model's decision to diagnose a patient with TB. In this way, we can ensure that the decision made by the AI model is based on relevant factors and can be justified.

Methods in *Explainable AI* There are several methods that can be used in XAI techniques, such as *decision tree, feature importance, local interpretation,* and *global interpretation. Decision tree* can be used to show how AI models make decisions by dividing data into smaller groups. *Feature importance* can be used to show which factors most influence the decision made by the AI model. *Local interpretation can be* used to understand the decision made by the AI model on each input data individually, while *global interpretation* can be used to understand the decision made by the AI model as a whole on a problem.

In use, XAI methods should be selected according to the type of AI model being used and the type of problem at hand. In addition, the results of the XAI techniques should also be analyzed and evaluated to ensure that the explanations provided by the AI model can be understood and scientifically justified.

Conclusion

Ensemble learning is a technique in machine learning that utilizes multiple learning models to improve prediction accuracy. There are several methods in *ensemble learning*, including:

1. *Bagging*: The *bagging* technique *(bootstrap aggregating)* builds multiple independent models on a randomly drawn dataset with replacement. Each model is trained on a different subset of the data and then the prediction results from each model are combined into one final result. *Bagging* is suitable for large and complex datasets.

2. *Boosting*: The *boosting* technique builds the model incrementally by emphasizing the samples that were misclassified in the previous iteration. The next model is built to correct the errors of the previous model. *Boosting* is suitable for limited datasets.

3. *Stacking*: The *stacking* technique uses multiple learning models and a *meta-learner model* to combine the prediction results of the models. The *meta-learner* model is trained to combine the prediction results of the base models and produce the final result.

In TB treatment, *ensemble learning* can be used to improve the performance of diagnosis prediction or therapy effectiveness prediction in TB patients. In use, several models can be built using different techniques, then the prediction results from each model can be combined with techniques such as *voting* or *averaging* to produce a more accurate final result. This can assist healthcare workers in making decisions for the diagnosis and treatment of TB patients.

An additional note for the validation of the latent TB detection concept using AI is that further studies are needed to evaluate the accuracy and reliability of the algorithms used. Concept validation is an important step before implementing AI technology for latent TB detection on a wider scale.

Concept Overview

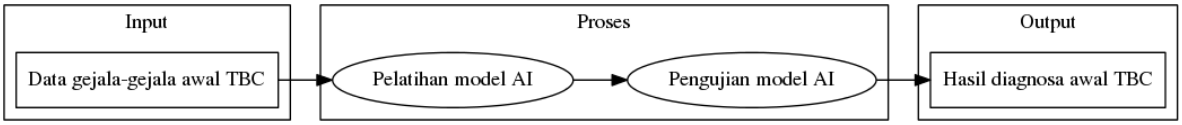

Figure 1: Flow Chart

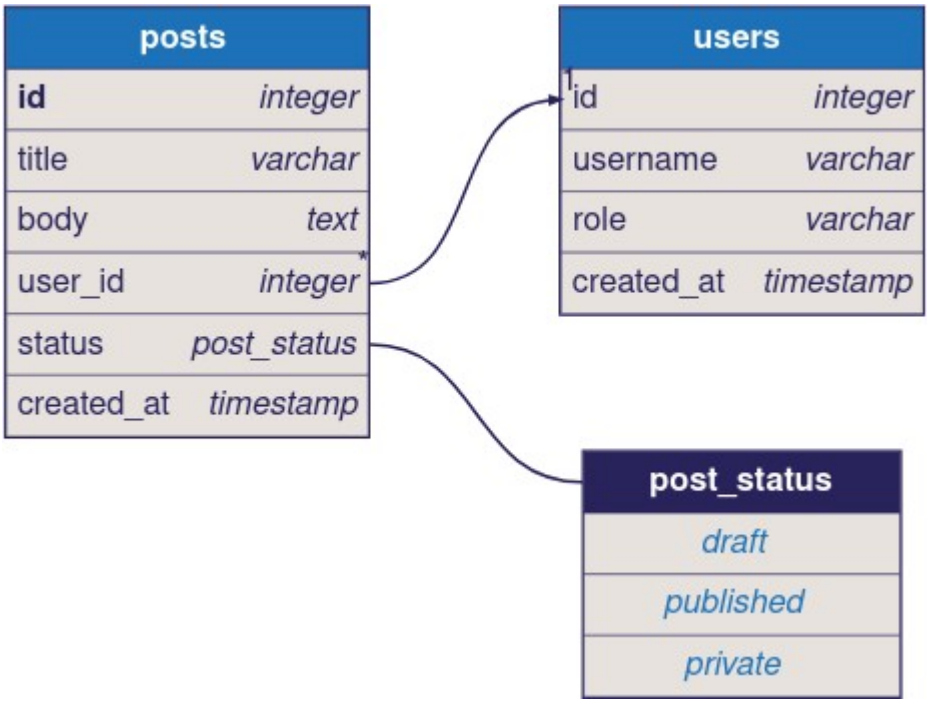

Figure 2: DBMS Flow

Stages of AI concept flow:

1. Symptom Input: Users can input the initial symptoms felt by the patient, such as persistent cough for more than two weeks, fever, loss of appetite, and weight loss of unknown cause, and other more detailed symptoms.

2. Output: The AI will generate information on the likelihood of the patient having TB based on the symptoms entered. It will also provide recommendations for further actions, such as sputum tests, chest X-rays, or further physical examinations.

3. Database: AI will utilize data on early TB symptoms and diagnosis results on patients previously diagnosed with TB to help improve diagnosis accuracy.

4. *Platform*: The AI can be accessed through *smartphone* apps*, websites*, or other online health *platforms,* allowing the public to easily conduct an initial TB diagnosis from anywhere.

5. Data Security: This AI will secure patient data and can only be accessed by authorized medical personnel.

With this AI concept, it is hoped that the community can easily diagnose the early symptoms of TB and take appropriate further action, so as to prevent the spread of TB and reduce the mortality rate from this disease in Indonesia.

Reading References

1. https://www.kemkes.go.id/article/view/23033100001/deteksi-tbc-capai-rekor-tertinggi-in-2022.html
2. https://www.kemkes.go.id/article/view/23033100001/deteksi-tbc-capai-rekor-tertinggi-in-2022.html
3. https://dataindonesia.id/ragam/detail/kasus-tbc-di-indonesia-melonjak-6198-pada-2022
4. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022
5. https://www.who.int/indonesia/news/campaign/tb-day-2022/fact-sheets
6. https://www.nhs.uk/conditions/tuberculosis-tb/
7. https://www.who.int/news-room/fact-sheets/detail/tuberculosis
8. https://www.webmd.com/lung/understanding-tuberculosis-basics
9. https://www.ncbi.nlm.nih.gov/books/NBK441916/
10. https://promkes.kemkes.go.id/?p=7439
11. https://www.cdc.gov/tb/topic/basics/default.htm
12. https://emedicine.medscape.com/article/230802-overview

Related Research References

1. Research by Aminah Ihsanawati, et al. (2019) which uses deep learning methods for TB detection from lung X-ray images.
2. Research by Ririn Andiyani, et al. (2019) which uses the decision tree method for TB detection from sputum examination results.
3. Research by Muhammad Iqbal, et al. (2020) which uses deep learning methods for TB detection from lung X-ray images.
4. Research by Ahmed Al-Jumaily (2018) from New Zealand who used the SVM method for TB detection from lung X-ray images.
5. Research by Jiang Xiau Zhang, et al. (2018) from China which uses deep learning methods for TB detection from lung X-ray images.
6. Research by Marta Mravec, et al. (2019) from Slovakia who used the decision tree method for predicting the effectiveness of therapy in TB patients.

Writing Application Reference

1. http://www.plantuml.com/plantuml/uml/RP1FQyCm3CNl-HGYznws7KC6-mDROR0oTYixg8rmr6cEiMns6FtkQwsIOiZ9VfBUyqyUs4KC8xp0NmCGqNyF7FktZ-wa_3GGR7N4G7saVKBlXBxE7Sqxi-SqxiqbQw9uKzm0WZkIlbtxa5FRbHAVxx2zT9un8JWweE3A3i1V5FXyV6fByd4X_RPEtyH9IIs-OYGH3cg408444u_evyStS-Ndv19uMBVph7VvmDMiLUhPPRwxckyzf-lwKn5px2ig2moCpa-wjBjzr-h-Yc_2Q8gmoAuvDly0

2. https://chat.openai.com/chat/2960c355-2ab4-44a1-bee8-ff59a6a0c72b

3. https://www.planttext.com/

4. https://kroki.io/