**전자약 인수인계서**

**2024.07 ~ 08**

**서원형 실습생**

**1. 개요**

전자약은 AI 기술을 이용한 치매 진단 프로그램이다.

웹, 앱 서비스의 아이디어가 제시되었으며, 김민구 자문 위원님에게 논문을 받아 해당 논문은 구현은 완료한 상태이고, 현재 AI-HUB에 데이터를 신청하여 GPU 자원 할당 대기 중인 상태이다.

현재까지 진행되어온 과정은 다음과 같다.

|  |  |
| --- | --- |
| **날짜** | **내용** |
| 2024.07.11 | **최위원님 “전자약 개발 방안 탐구” 지시** |
| 2024.07.24 | **김민구 자문 위원님 지시 논문 구현 완료** |
| 2024.07.29 | **논문 內 CNN 추가 방안 강구** |
| 2024.08.08 | **최위원님 : 사용자 앱으로 사용자가 통화를 진행한 후, 자동으로 녹화를 진행해 일부 녹화 파일을 가지고 치매 진단을 제공 지시** |
| 2024.08.13 | **IRB 연구계획서 제출 및 심의** |
| 2024.08.19 | **IRB 연구계획서 제출 및 심의면제 확인 완료** |
| 2024.08.26 | **ai-hub 데이터 1차 신청 완료** |
| 2024.08.29 | **ai-hub 데이터 신청 반려 후 재신청 완료(연구계획서 불일치)** |

**2. 데이터 자료조사**

- 9가지의 정실질환 관련 데이터셋을 찾았고 활용 방안을 고안, 실현가능성 판단.

[개요]

전자약과 관련된 정부과제가 많이 생기고 있고, 인증된 데이터를 통해 비싼 양질의 서비스를 제공할 수 있을 것이나 구체화된 내역이 없어, 가장 중요한 데이터셋부터 조사하였음.

[조사 방법]

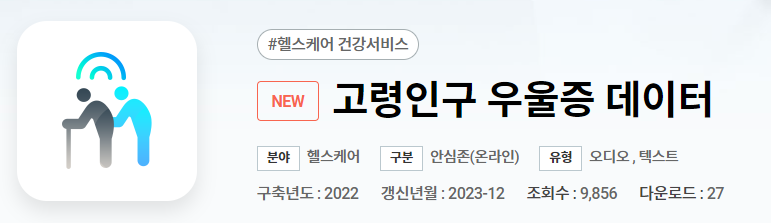
정실질환과 관련하여 우선 국내의 공공데이터 포털, AI-HUB를 조사하였고 총 5가지의 데이터를 가져올 수 있었음.



1. 노인 정신건강 영상 데이터(음성 데이터, 오각형 이미지, 얼굴표현 영상)



2 정신건강진단 및 예측을 위한 멀티모달 데이터(음성 데이터)



3 고령인구 우울증 데이터(음성 데이터, 텍스트)



4. 파킨슨병 및 관련 질환 진단 음성데이터(음성데이터)



5. 소아청소년 정실질환 진단 안저 데이터

추가적으로 Kaggle을 통해 총 4개의 데이터를 추가적으로 확인하였다.

1. 알츠하이머병을 감지하는 필기 데이터

2. 알츠하이머 MRI 전처리 데이터 세트

3. Synthetic Therapy Conversations

4.정신 건강 대화 데이터



위 9개 데이터 세트를 활용하는 방안은 위 표와 같다.

조금 더 응용된 활용방안으로는 모델을 2가지를 혼합하여 쓰면 난이도는 올라가지만 더 좋은 서비스로 발전 가능하다고 생각한다. (예시 : 노인의 정신 건강 영상 데이터를 활용하여 진단하고 LLM을 통하여 치료까지 연계해주는 서비스)

여러 요소를 고려해보았을 때, 2가지의 문제점이 크게 대두된다.

1. 데이터 활용 신청 및 선정의 복잡성

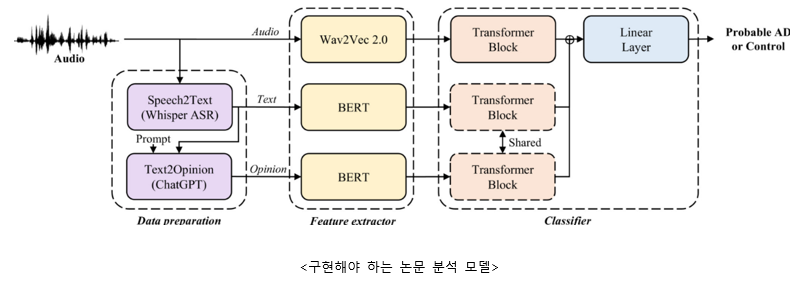
2. 모델 개발 환경의 부족

컴퓨팅 할 수 있는 GPU 자원의 부재이다. 텍스트 데이터는 가볍게 학습시킬 수 있지만, 이미지, 영상 데이터는 많은 용량으로 학습시간이 굉장히 오래걸린다. 이때, AI-HUB에서 컴퓨팅 자원을 30일간 무료로 대여해주지만, 신청일 기준 14일 이후에 이용가능하다는 점과 정확한 컴퓨팅 자원이 어느정도인지 공지되지 않았다. 이에 2가지 해결책이 있다. 첫번째로는 AWS 환경의 구축이다. 이때 SSD보다 빠른 속도를 지원하는 S3와 Sagemaker를 통한 빠른 모델 학습이 가능하다. 다만, 이때 AWS의 서버 비용이 일정하게 부과된다는 단점이 존재한다. 물론 Colab도 머신러닝에 이용이 가능하지만 Colab은 실현 가능성을 판단하기 위해 라이트한 모델을 사전학습 시키는 용도로 사용하고 실제로는 더 큰 GPU 리소스로 학습시켜야 한다 생각한다. 두번째로는 로컬 GPU 서버의 구축이다. 비용은 일시에 지불하지만 지속적으로 나가는 비용을 줄일 수 있다. 하지만, 최신 GPU가 계속해서 업데이트 되고 SSD 같은 추가 장비의 구축 필요성도 있기 때문에 첫번째 방법이 유용하다고 판단한다.

위와 같은 많은 어려움이 있음에도 불구하고 AI를 활용한 양질의 서비스를 개발한다는 것은 미래 먹거리를 빠르게 선점할 수 있다고 생각한다. 다만, 처음부터 올바른 계획을 세우고 차근 차근 협의된 프로세스대로 진행해야 할 것이다.

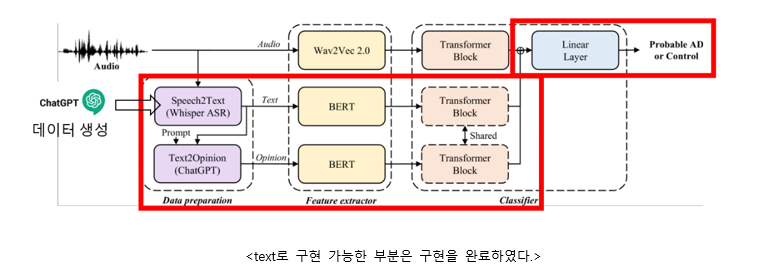
**3. 논문 구현**

과제로 주어진 논문의 LLM 모델(3종)을 구현함

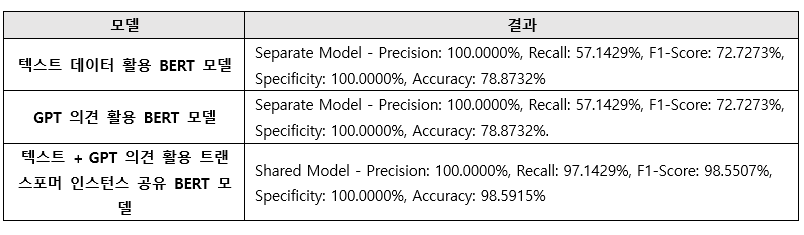


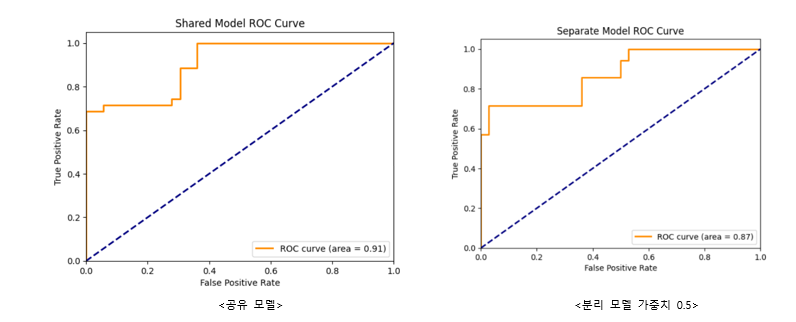
논문에 작성된 모델을 구현하기 위해서 데이터 선정, 개발환경 구축, 데이터 전처리, 모델 학습, 모델 평가, 모델 최적화 순서로 계획을 잡고 LLM 모델 구현을 실행하였다.

논문에서 사용된 데이터는 4년전 특정 대회에서 쓰여진 데이터로 쉽게 구할 수 없으며, 치매 포럼의 회원만이 다운로드가 가능했다. 따라서, Audio를 생성형 모델에게 치매환자처럼 만들어 달라고 할 수 없었으므로 가능한 text를 사용하여 아래 그림의 모델을 구현하였다.



개발환경은 google colab을 사용하였고, 유료 GPU를 사용할 수 없었기 때문에, 세션이 비정상적으로 종료되어 학습이 정상적으로 불가능한 경우도 있었지만, 모델의 성능보다는 모델 구현에 중심을 맞춰 파라미터를 조절하여(에폭시를 3으로 두는 등) 모델을 생성하였다.

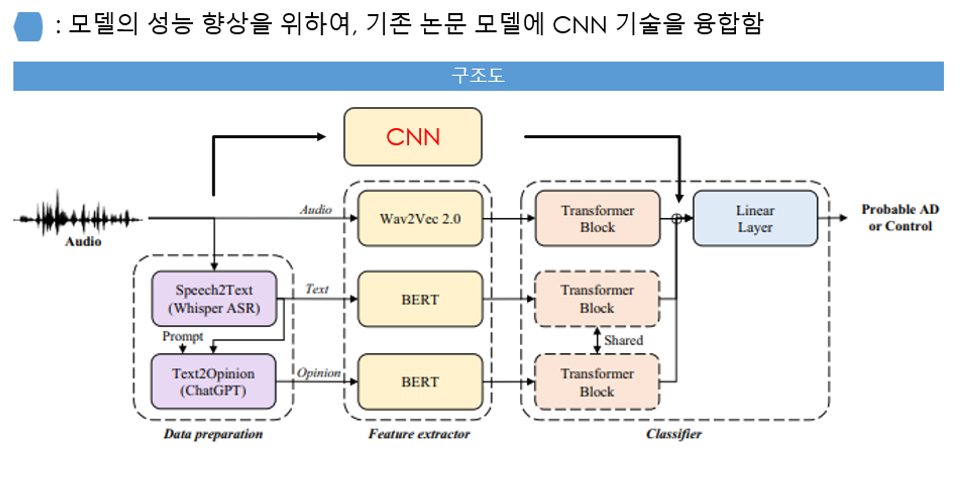




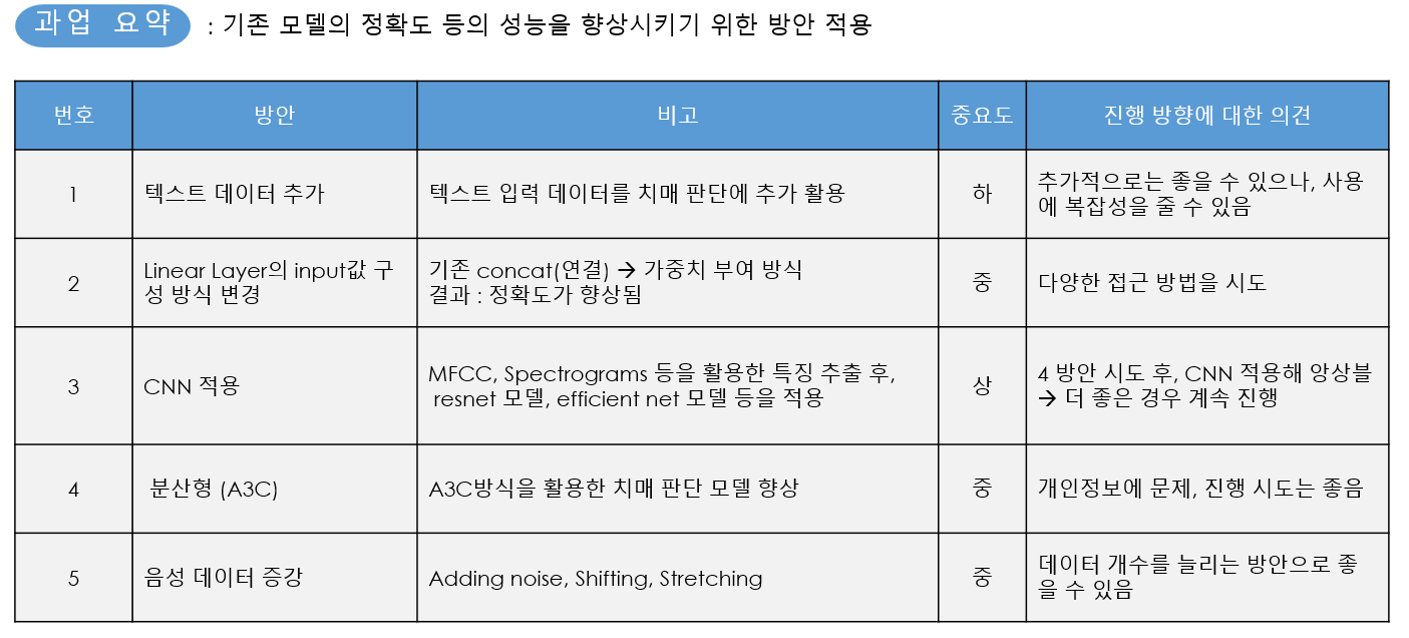
기존 논문의 정확도와 혼동행렬 지표를 뛰어넘은 것도 존재할만큼 유의미한 모델이 추출되었다. 또한, 데이터를 바꾸면 새로운 모델이 나올 수 있게 템플릿의 형태로 활용이 가능하다. 구글 드라이브에서 확인할 수 있으며, 직접 CRTL+SHIFT로 하나하나 셀을 실행시켜보는 것이 이해가 제일 빠를 것이다.

데이터가 정상적으로 모두 제공되지 않은점, GPT가 생성한 데이터이기 때문에 실제 알츠하이머 환자의 데이터와는 괴리감이 있을 수 있다는 점을 고려하여, 데이터를 확보한 후에 모델의 고도화가 가능하다고 생각한다. 해당 논문 데이터는 지도교수의 소속, 신분등을 치매 포럼에 제출 후 이용목적을 검증받는다. 이후 회원으로 인정이 되면 이용이 가능하다.

**5. 추가 구상안 적용**

****

- CNN 모델을 기존 모델에서 추가하였고, 추후 데이터가 확보되면 CNN을 포함하여 정확도 검증이 필요한 상황



- AI 모델에서 가장 중요한 정확도를 높이는 방안으로 여러 가지 대안이 있고 이를 잘 조합해서 최선의 정확도를 도출할 수 있도록 해야한다.

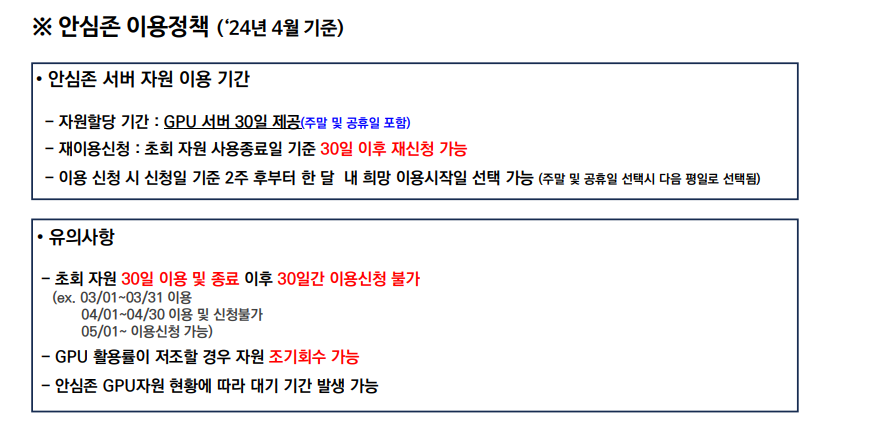
**6. 데이터 확보 이후 개발 방법**

GPU 자원을 AI-HUB것을 사용하기로 결정이 된다면, 대기자가 있으므로 GPU 자원 할당까지 대기가 필요하다. 단, GPU 서버를 무료로 높은 사양을 사용할 수 있다. GPU 할당을 받는 다면 CoLab 처럼 가상환경의 주피터 노트북에서 개발을 진행한다.

AI-HUB의 안심존 이용절차를 최종\_인수인계\_ 전자약.pptx에서 자세히 확인할 수 있고, 여기서는 절차만 다룬다.

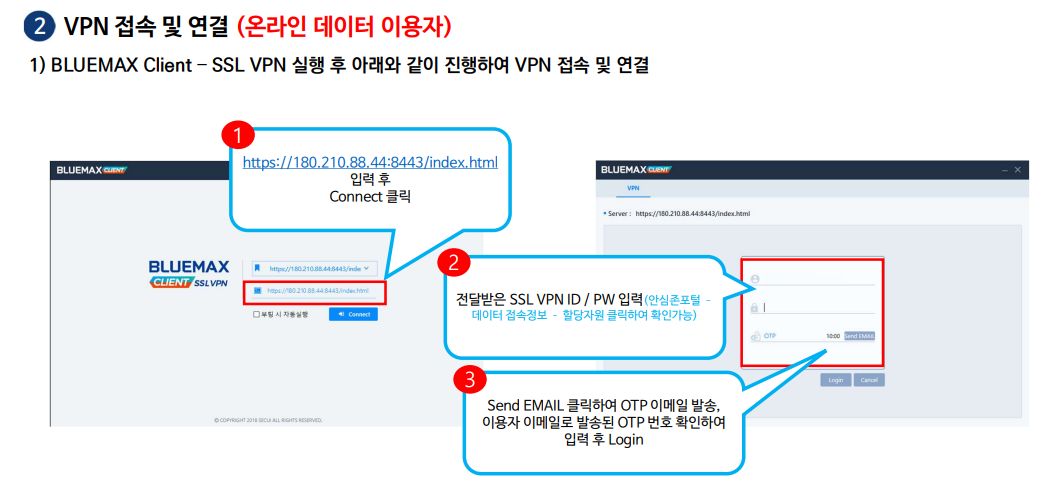
ID : [jk.choehan@gmail.com](mailto:jk.choehan@gmail.com)

PW: bluesol2024~

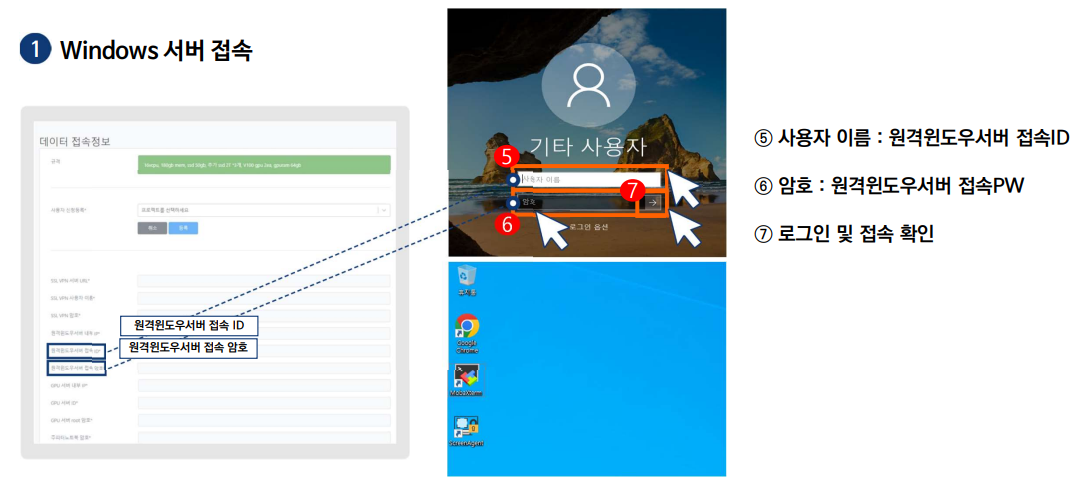


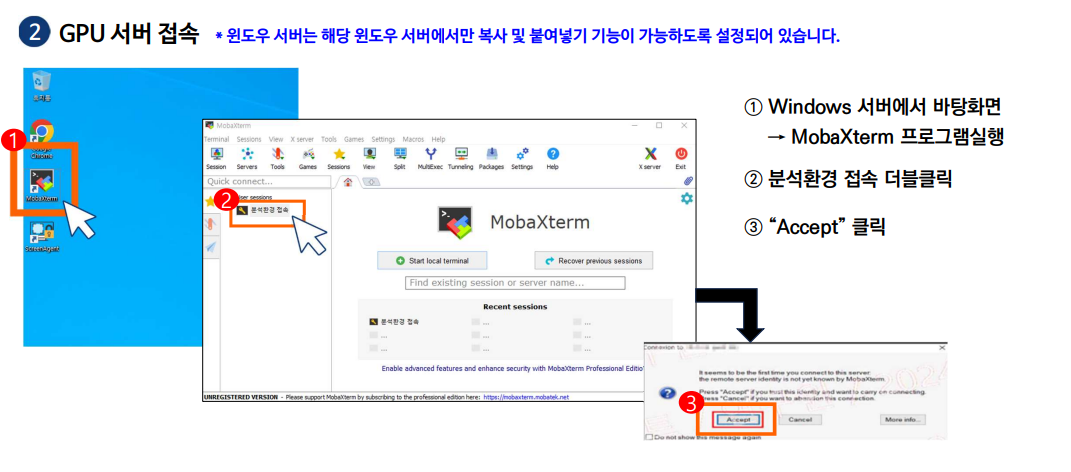


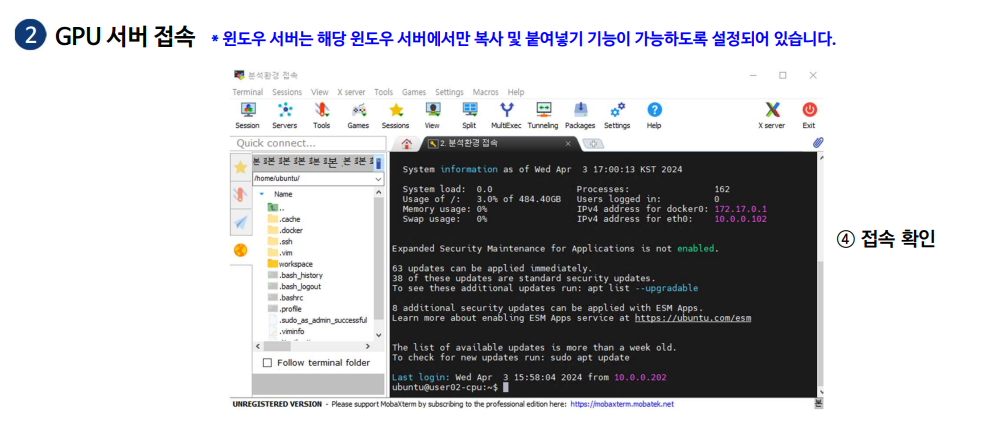


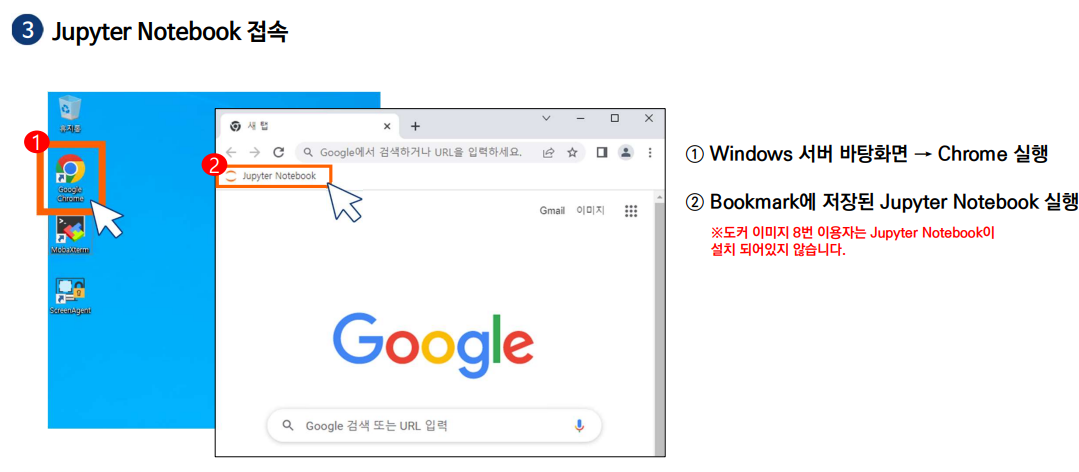
****

****

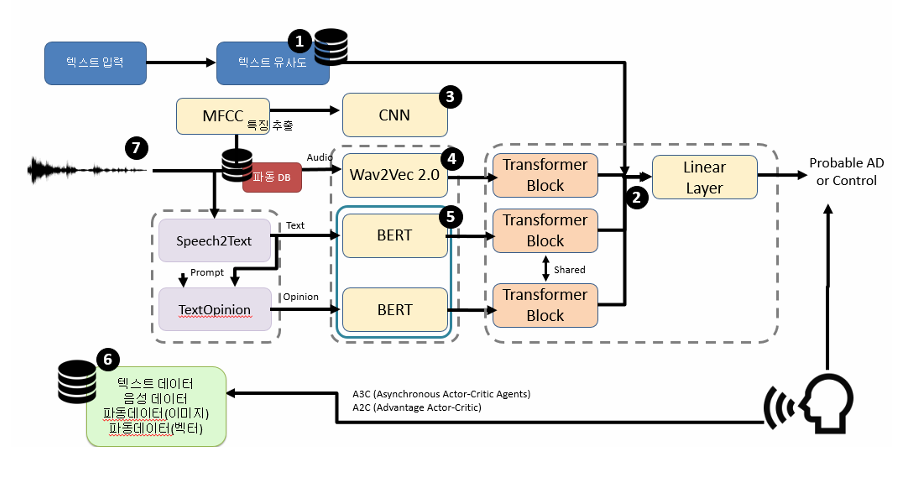
****

****

****

****

**<모델 반출입시 ai-hub에서 별도 신청이 필요합니다>**

<CNN 모델 추가한 전자약 구성도>

<기존 논문 전자약 구성도>

