<div align="center">

|  |  |  |
| --- | --- | --- |
| **2019** | **Report – 2019/ 1월 3주** | **Seong-gu** |

</div>

**I. Plan for this week**

***Personal Research Subject***

* 논문 읽기
* NRF (미세먼지 과제) 결과보고서 및 성과보고서 작성

**II. Plan for next week**

* 미세먼지 예측 머신러닝 기반 모델링 최신 논문 계속 검토할 것.
* 바람의 영향에 의한 미세먼지 유입을 최적 고려한 모델링 찾아볼 것.

**III. Response on Prof. Shin's feedback last week**

* 현재는 시간에 따른 예측이지만, 누락된 공간적인 데이터도 잘 예측하는 방법 검토.
* Unsupervised learning이 될텐데 방법을 찾아보자.
* 랜드마크 등 지형, 지역적 특성 수기 추가
* Feature를 잘 선택 (데이터 프로세싱으로 정확도 개선)
* 지역은 한국 전범위 하지 말고 적당한 수준으로 결정함.
* 시간대 설정
* 효율성 및 정확도 면에서 트레이드 오프 분석 설계 방안을 제기
* 온도, 습도, 대기압, 바람세기, 풍향 데이터 외에 날씨(눈/비/안개 등)도 포함하면 좋을듯함

**IV. Milestone**

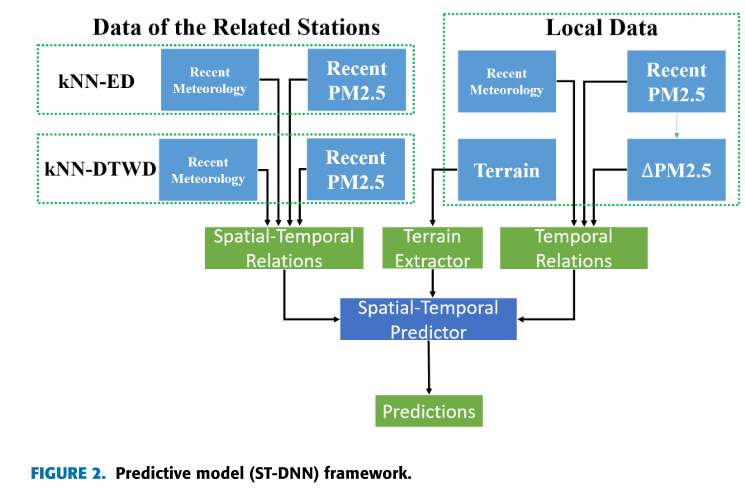
(모델링 설계 관련해 다시 계획을 짤 예정입니다.)

**V. Reports**

* 개인연구 진행사항
* **Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations 논문 리뷰**

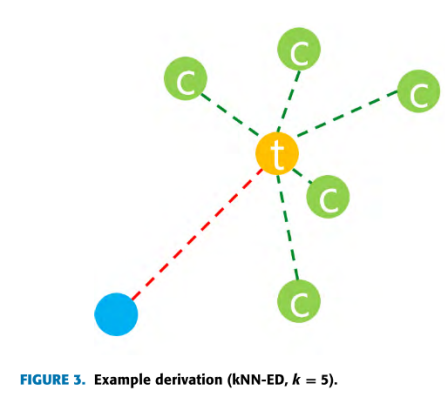
**목적 :** 이 논문은 시공간특성을 모두 고려하기 위해 CNN과 LSTM을 모두 활용하였으며, 특히 지역적 특성을 고려하기 위해 CNN을 사용한 논문입니다.

**모델 프레임 워크 :**

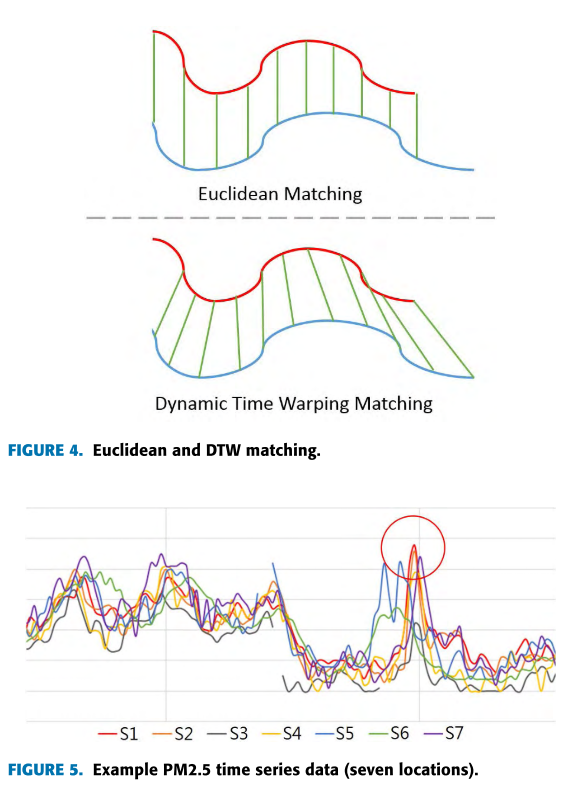


모델에는 예측을 원하는 지역에 대한 기상데이터 및 지형데이터 를 포함한 Local 데이터를 통해 시간특성과 지형특성을 뽑아내며, kNN-ED와 kNN-DTWD 방법을 이용해 주변의 데이터에서도 시공간 특성을 뽑아냅니다.

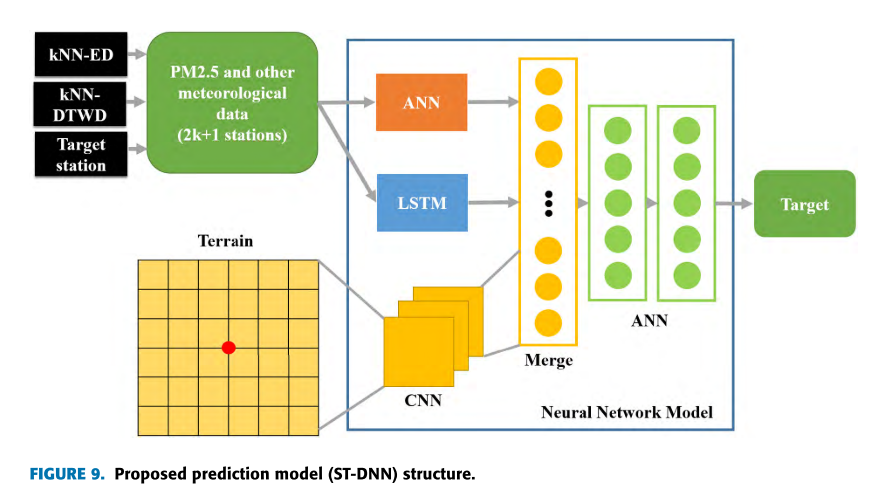
kNN-ED (k-Nearest neighbor by euclidean distance)는 유클리드 거리를 계산했을 때 k번째까지 가까운 주변 측정소 데이터를 가져오는 것입니다. 이 방법은 단순히 가까운 거리에 있는 측정소만을 활용하기 때문에 실제로 상관관계가 적은 측정소 데이터를 참조할 수 있습니다.



따라서, 추가적인 방법으로 kNN-DTWD (k-Nearest Neighbor by DTW Distance)를 사용하였습니다. 이 방법은 공간상의 거리가 아니라, 시간축에 따라 변화하는 미세먼지 농도 흐름의 유사성을 판별하여, 유사성이 높은 주변 측정소를 데이터로 활용하는 기법입니다.



**모델 구조:**



kNN-ED 및 kNN-DTWD를 통해 선정한 주변 측정소와 타겟 측정소 데이터를 ANN과 LSTM을 이용하여

시공간적 특성을 뽑아냅니다.

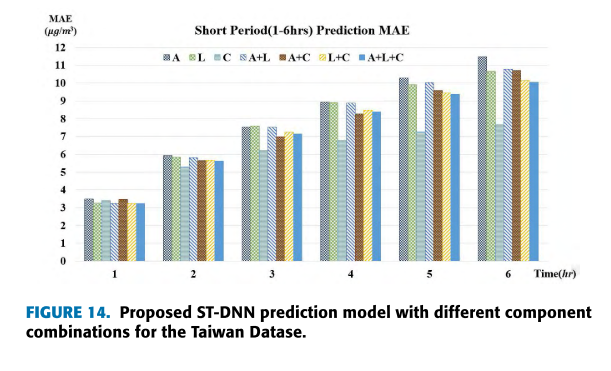
특별히 LSTM은 과거 데이터를 포함한 시퀀스 데이터를 활용하여 시간적 특성을 뽑아내기 때문에 현재 trand에 따른 예측이 가능하고, ANN은 현재 데이터만 활용하여 시간적 특성을 뽑아내기 때문에, 갑자기 변화하는 특성을 잡아낼 수 있도록 설계했다고 합니다.

타켓 측정소를 중심으로 11x11의 격자를 형성하고 각 격자는 IDW(Inverted distance weight)를 이용해 지형에 의한 주변 미세먼지 데이터 값을 형성하며, 이를 CNN을 통해 지형적 특성을 뽑아 냅니다.

ANN, LSTM, CNN에서 뽑아낸 특성은 마지막 ANN 레이어를 통해 결합됩니다.

**결과:**

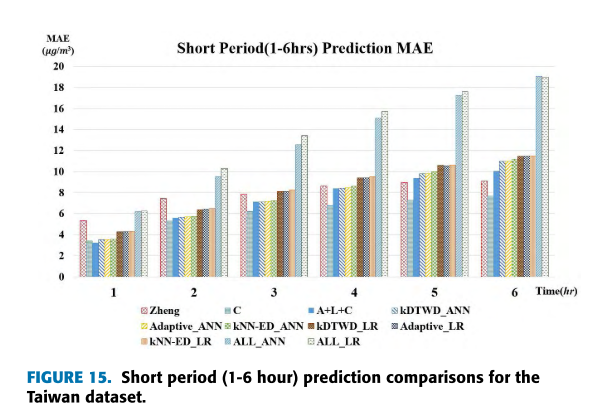
**만든 모델간 비교**

****

모델에서 A:ANN, L:LSTM, C:CNN을 나타내며, 다음 1시간 예측에서는 모든 레이어(A+L+C)를 쓴 모델이 가장 좋은 결과를 나타냈으나, 2시간 이후는 C만 쓴 모델이 압도적으로 좋은 결과를 보였습니다.

장기 예측 시에는, CNN이 주변 지역에 의한 미세먼지 분포의 공간 특성을 분석하므로 합리적인 예측 결과를 보인다고 결론 지었습니다.

**다른 모델간 비교:**

****

다른 모델과 비교하더라도, CNN만을 사용한 모델이 훨씬 나은 결과를 보여주었습니다. 추가적인 정보 (ANN+LSTM)가 없이 오히려 CNN만 사용했을 때 성능이 훨씬 낫게 나왔다는 점은 이례적이라고 볼 수 있습니다.

**결론**

본 논문에서는 KNN-ED, KNN-DTWD 와 같은 기법을 통해 주변 측정소 데이터를 선정하고, LSTM 및 ANN을 통해 미세먼지 농도의 트렌드 및 갑작스러운 변화를 예측하도록 설계하였습니다. 그러나 이런 설계가 결과적으로 2시간 이후 예측의 성능개선에는 오히려 악영향을 미쳤으며, CNN을 이용해 주변 지역의 미세먼지 분포 공간 특성을 분석하는 것이 장기 예측에는 좋은 결과를 보였다는 점은 참고할 만합니다.