

This paper describes a study that builds a CNN model on nonimaging EHR data to predict incident nonmelanoma skin cancer (NMSC) using a large dataset consisting of Asian patients. The model achieves excellent discriminative performance with 5 fold cross validation. Predictive factors are identified with an ablation approach.

Here are some of my comments on this paper:

- As the authors have mentioned, the UV exposure had been reported as an important risk factor for NMSC by previous studies. However, the authors don't clarify the motivation for building the model without the information of UV exposure. Is such information harder to collect in specific subpopulations? Is it more noisy compared to other factors? Add an explanation on why the advantage of not using UV exposure information may make the paper more convincing.
- As the authors admitted, there is no external validation or even temporal validation conducted in this study and the dataset only consists of Asian people. The generalizability of the model would be a huge concern.
- This could be a common issue for most of the clinical/medical applications of machine learning models: the identified predictive factors are not causal. For example, one of the factors identified in this paper, metformin, could just be a proxy measure for aging rather than having causal relations with NMSC.