Forecasting Noncommunicable Diseases in Thailand: Evaluating the Predictive Power of Social Determinants of Health-Related Features

Peatiphat Bhoothookngoen* Nattapong Sanchan
School of Information Technology and Innovation, Bangkok University
9/1 Moo 5 Phaholyothin Road, Klong Nueng, Klong Luang, Pathumthani 12120
*Corresponding author Email: peatiphat.bhoo@bumail.net
(Received: May 3, 2024; Revised: March 27, 2024 ; Accepted: June 30, 2024)

ABSTRACT

In Thailand, non-communicable diseases (NCDs) presented a significant health and economic challenge. This study investigated machine learning (ML) for predicting NCDs prevalence using social determinants of health (SDHs). Two scenarios, baseline and inference (imputing missing values) were assessed. Monthly household expenditure and hospital counts appeared as pivotal features in the inference scenario. Model performance remained comparable between scenarios, with slight variations for specific NCDs. Random Forest (RF) showed slightly superior predictive power (RMSE: 1.53 -- 74.93, R Square: -0.11 -- 0.11), though interpretability remains a challenge. Addressing data limitations and enhancing interpretability are crucial for fully harnessing ML's potential in NCDs prediction and prevention. The study's findings underscore the importance of integrating ML and SDHs into public health policy to effectively combat NCDs in Thailand, potentially saving lives and fostering sustainable socio-economic development. The study revealed the intricate interplay between socio-economic factors and NCDs prevalence, emphasizing the need for targeted interventions. The exploration of machine learning algorithms in predicting NCDs prevalence provided valuable insights into model performance and highlighted the significance of features such as household expenditure and hospital counts. Moving forward, efforts to address data disparities and enhance model interpretability are essential to maximize the utility of predictive modeling in informing public health policies aimed at mitigating the impact of NCDs in Thailand.

Keyword: Noncommunicable diseases, predictive power, social determinants of health, machine learning, prevalence.

1. Introduction

In a report by the Ministry of Public Health of Thailand [MOPH], World Health Organization [WHO], United Nations Development Programme [UNDP], and United Nations Inter-Agency Task Force [UNIATF] (2021) [1], noncommunicable diseases (NCDs) exert a significant impact on Thailand, driven by factors such as cancers (CA),

cardiovascular diseases (CVD), diabetes (DM), and hypertension (HTN), collectively responsible for 74% of all deaths in the country, resulting in an annual toll of 400,000 lives. The economic consequences are substantial, with NCDs costing the Thai economy THB 1.6 trillion yearly, equivalent to 9.7% of its 2019 gross domestic product (GDP). This financial burden encloses THB 139 billion for NCDs treatment and THB 1.5 trillion in lost productive capacity due to absenteeism, "presenteeism," or early withdrawal from the labor force. NCDs not only adversely affect socioeconomic development but also challenge the long-term fiscal sustainability of the government and public services. Strategic investments in key clinical interventions for prevalent NCDs and policy measures targeting risk factors can both save lives and yield economic benefits. Allocating THB 211 billion to such interventions holds the potential to save 310,000 lives and generate THB 430 billion in economic benefits.

Social determinants of health (SDHs) played a pivotal role in both the prevalence and impact of NCDs, manifesting in diverse ways. NCDs exhibited a higher prevalence in low-income regions where poverty, recognized as a SDHs, significantly contributed to elevated incidence and mortality rates, as demonstrated by Rasesemola, Mmusi-Phetoe, and Havenga (2023) [2]. Their study also highlighted that the concentration of risk factors and poor health outcomes was notably more pronounced in economically disadvantaged communities compared to affluent ones. NCDs often coexisted with high levels of infectious diseases, particularly in impoverished communities. Interventions that focused solely on individual factors while neglecting the structural and commercial determinants of health exhibited limited effectiveness. Noteworthy, both individual and national poverty levels contributed to disparities in life expectancy, quality of life, and morbidity, as highlighted by Manderson and Jewett (2023) [3].

Enhancing the performance of predictive models can be achieved by incorporating SDHs features in certain vulnerable subgroups. However, the utility of incorporating Social Determinants of Outcomes (SDOs) features was contingent upon the specific characteristics of the cohort and the nature of the prediction task, as elucidated by Yang, Kwak, Pollard, Celi, & Ghassemi (2023) [4].

Effectively addressing the formidable challenge of NCDs in Thailand requires a comprehensive approach that recognizes the intricate interplay between SDHs and NCDs prevalence. The urgency of strategic interventions underscored by the profound impact of NCDs on both mortality rates and the economic landscape. Machine learning emerges as a transformative technology capable of predicting and understanding NCDs prevalence based on population-wide socio-economic factors, encompassed by SDHs. Integrating machine learning into policymaking provides invaluable insights into the determinants influencing NCDs, facilitating the formulation of targeted and effective public health policies. The adoption of

machine learning becomes a meaningful tool in the hands of policymakers, driving informed decision-making, enhancing preventive measures, and contributing to the well-being of the population. In the mission of a healthier future for Thailand, the synergy between machine learning and public health policy holds the promise of saving lives and fostering sustainable socio-economic development. This study aimed to test on the predictive power of machine learning algorithms in Thai population datasets utilizing the SDHs as the features.

2. Material and Method

Bhoothookngoen and Sanchan's (2023) [5] reviewed identifies existing gaps in NCDs prevalence prediction, specifically in the focus of prediction and the utilization of SDHs as features in models. Noteworthy, Hu, Qiu, Su, Shen, & Chen's (2020) [6] study was highlighted for its incorporation of some sets of features considered as SDHs for NCDs prediction not the prevalence.

2.1 Data Collection

In this experiment, the models from Hu, Qiu, Su, Shen, and Chen's (2020) [6] study were operationalised, substituted features with SDHs datasets in Thailand (population-based dataset) to predict NCDs prevalence in Thailand. The data were retrieved between September and December 2022 from internet access via various primary sources: the Open Data web portal by MOPH [7] for NCDs prevalence between 2023 and 2021, the Pollution Control Department [8] for fine particulate matter 2.5 (PM2.5), NSO's website [9] for all SDHs except for number of hospitals and MOPH's website [10] for number of hospitals.

The MOPH's Open Data web portal [7] emerged as a foundational repository, encapsulating deidentified datasets illuminating various health-related aspects of Thai population. This included information on health service access, healthcare providers, health status by some diseases (including NCDs), etiology of illnesses, Tuberculosis related activities, and diseases from occupational or environmental exposure. This repository served as the primary source of information for the number of NCDs patients for at least a decade preceding 2021.

In a report by MOPH et al., (2021) [1], prevalence rates of discrete NCDs, including CVD, CA (Cervical Cancer, Lung Cancer, and Breast Cancer), DM, and HTN were extracted from 2013 to 2021 because these four NCDs are the top of highest cost of illness (CA, DM, and CVD) and prevalence (HTN). Concurrently, hospital registries were cross-referenced with patient, describing correlations between NCDs incidence and healthcare facilities. The features as detailed in Table 1, painted a portrait of socio-economic and health-related dynamics across Thai provinces from 2013 to 2021. This encompassed key indicators including average household income, shedding light on economic variations. Environmental health angles were explored through the levels of Particulate Matter 2.5 (PM2.5), offering insights into air quality trends. Social behaviors were captured with data on smoking prevalence and the prevalence of alcohol consumption.

TABLE I

Social Determinants of Health Data (2013–2021)

Feature	SDHs Domain[λ]	Year
Average monthly household income by province.	Economic stability	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021.
Average Particulate Matter 2.5 (PM2.5) by province.	Neighborhoods and built environment	2013 – 2021.
The numbers of smokers or used to smoking by a whole country.	Social and community context, Neighborhoods and built environment	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021.
The number of alcoholic consumers by a whole country.	Social and community context, Neighborhoods and built environment	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021.
The number of educational institutions by Bangkok and upcountry group.	Education access and quality	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021
The educational year attained by a whole country	Education access and quality	2013 – 2021
Average monthly household expenditure by province.	Economic stability	2013 – 2021
Average annual household loan by province	Economic stability	2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021
Hospital counts by province by matching hospital code	Healthcare access and quality	2013 – 2021

[λ]Social Determinants of Health - Healthy People 2030 (2022).

Underline Imputation by mean for the missing years.

The educational landscape was subjected to analysis through the educational institution counts, the educational year attained, and regional disparities between Bangkok and upcountry locales. Economic dimensions were further illuminated through examining average household expenditures per month and annual loan disbursements. The healthcare infrastructure received with detailing the hospital counts in each province. These datasets provided an understanding of societal trends, serving as a valuable resource for policymakers and researchers aiming to navigate the regional development and well-being.

The model framework, inspired by Hu et al. (2020) [6] study, embraced a diverse set of models—Linear Regression (LR) and Support Vector Regression (SVR), Random Forest (RF), Gradient Boosted Decision Trees (GBDT), and XGBoost. Additionally, embracing ensemble learning with the introduction of a Stacking Model (utilising Random Forest). To ensure a thorough evaluation, a 5-fold cross-validation approach was implemented, iteratively involved partitioning datasets into training and testing sets to enhance a robust assessment of model generalization. The Sequential Floating Forward Selection (SFFS), facilitated by the 'mlxtend' library, was employed during each fold of the cross-validation. This dynamic feature selection process, aimed to identify an optimal subset of features for model training.

Addressing missing values in model training, a mean-based imputation strategy using scikit-learn's 'SimpleImputer' was employed. This approach ensures a robust and consistent treatment of missing data, specifically tailored for the inference scenario. An alternative scenario was contemplated by creating datasets wherein instances with missing

values were omitted, thereby presenting an alternative scenario for further examination.

2.2 Model Implementation and Evaluation Framework

The model framework, inspired by Hu et al. (2020) [6] study, embraced a diverse set of models—Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosted Decision Trees (GBDT), and XGBoost. Additionally, ensemble learning was introduced with a Stacking Model (utilizing Random Forest). Figure 1 illustrated the complete workflow of the model implementation and evaluation framework.

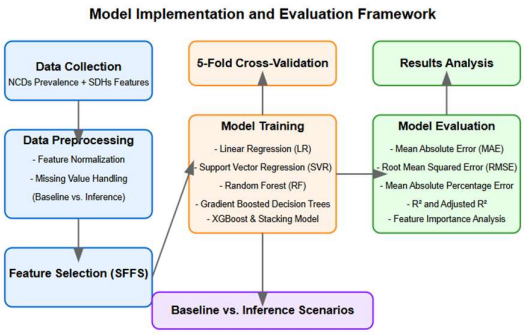


Figure 1 Workflow and Model Framework

2.2.1 Feature Relationship with NCDs

The SDHs features in our model had significant relationships with NCDs prevalence as followed.

Economic Stability Features (household income, expenditure, loans): Lower economic status often correlated with higher NCDs risk due to limited access to quality healthcare, healthier food options, and preventive services. Financial stress could also lead to unhealthy coping behaviors.

Environmental Features (PM2.5): Air pollution had been directly linked to increased risk of respiratory diseases, cardiovascular problems, and certain cancers.

Social Behavior Features (smoking, alcohol consumption): These were well-established risk factors for multiple NCDs including cancers, cardiovascular diseases, and liver disease.

Education Features (educational institutions, years attained): Higher education levels typically correlated with better health literacy, preventive health behaviors, and lower NCDs risk.

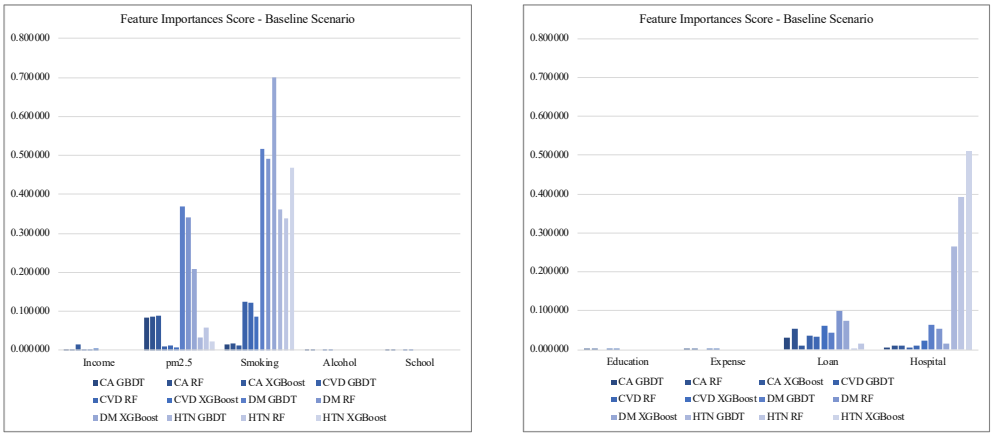
Healthcare Access Features (hospital counts): The availability of healthcare facilities directly impacted disease management, early detection, and treatment outcomes for NCDs.

To ensure a thorough evaluation, a 5-fold cross-validation approach was implemented, which iteratively involved partitioning datasets into training and testing sets to enhance a robust assessment of model generalization. The Sequential Floating Forward Selection (SFFS), facilitated by the 'mlxtend' library, was employed during each fold of the cross-validation. This dynamic feature selection process aimed to identify an optimal subset of features for model training.

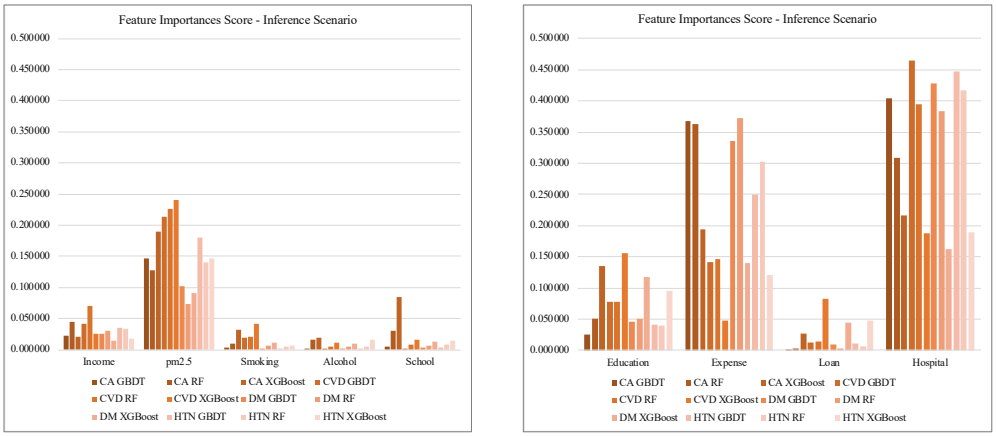
Addressing missing values in model training, a mean-based imputation strategy using scikit-learn's 'SimpleImputer' was employed. This approach ensured a robust and consistent treatment of missing data, specifically tailored for the inference scenario. An alternative scenario was contemplated by creating datasets wherein instances with missing values were omitted, thereby presenting a baseline scenario for further examination.

3. Result and Discussion

In this experiment, various machine learning models were utilized in two distinct scenarios. The baseline scenario, characterized by comparatively lower prognostications across NCDs, suggested a potential exclusion of instances with missing values. In contrast, the Inference scenario involves imputation of absent data points by their mean, aiming to elaborate the elevation of model predictions. Analysis of individual features using feature importance scores in GBDT, RF, and XGBoost revealed fluctuations in significance between scenarios, with monthly household expenditure and hospital counts emerging as crucial factors in the Inference scenario overall (Figure 2), followed by PM2.5, educational years attained, and others, respectively. Models like SVR, LR, and Stacking were excluded due to limitations in feature importance calculation.



(a) Feature Importances Score – Baseline Scenario



(b) Feature Importances Score – Inference Scenario

Figure 2 Feature Importance Score

Comparing two prediction scenarios based on model evaluation metrics (MAE, RMSE, MAPE) revealed that the Baseline scenario, where missing values were dropped, consistently outperformed the Inference scenario, where missing values were imputed. This suggested stronger predictive performance for Baseline, particularly for CA and HTN models. Conversely, in the cases of CVD and DM, improvements in metrics after imputation were somewhat higher. Across all NCDs datasets and scenarios as depicted in Figure 3, the performance of GBDT, LR, RF, Stacking, and XGBoost models remain consistently comparable as measured by MAE, RMSE, MAPE, R-squared, and adjusted R-squared between both scenarios. Some detectable deviations, specifically, compared to the baseline, MAPE for CA and HTN increased by approximately 20 points. Conversely, for DM, MAPE decreased by around 20 points in the Inference scenario, while CVD exhibits minimal change relative to other datasets. These findings warranted further investigation into the potential factors influencing model performance discrepancies across NCDs and scenarios.

A comparative analysis between the baseline and Inference scenarios unveils refined shifts in model performance. In the Inference scenario for all NCDs except CVD, there was a minor signal of superior performance, indicated by lower MAE values. However, the superiority is premature to conclude. Contextualized exploration within specific scenarios reveals that RF, despite achieving the lowest MAE (0.58 points) for CVD in the Inference scenario, is accompanied by negative R Square (-0.07) and Adjusted R Square (-0.07), questioning its suitability. Conversely, the highest MAE (44.03 points) is observed in the Stacking model for CA in the baseline scenario with 0.04 R Square and 0.03 Adjusted R Square. For RMSE, the highest and lowest are found in the baseline scenario in SVR for CA (76.63 points) and GBDT for CVD (1.44 points), respectively. For MAPE, the highest and lowest are found in the Inference scenario, with 173.62 points in LR for CA and 22.70 points in RF for CVD, respectively. R-square and Adjusted R-square exhibit a similar pattern for the highest and lowest in XGBoost for CA in Inference (both 0.11) and in RF for CVD in the baseline scenario (-0.011 and -0.012, respectively).

Compared to other models, RF demonstrated slightly superior predictive power across diverse datasets referring to NCDs. With supportive results, on the CA dataset, RF exhibited a notably reduced MAPE of 166.88 when contrasted with alternative models during the Inference phase. This discernible trend persisted across additional datasets, such as CVD with a MAPE of 22.70, DM with a MAPE of 123.26, and HTN with a MAPE of 131.58.

The graphical representation in Figure 3 further accentuates the consistent outperformance of RF across a spectrum of scenarios. Another notably model, SVR consistently manifested accuracy in relation to alternative models, with MAPE exhibiting a pronounced variability across scenarios. This variability potentially signals challenges in the interpretability of SVR within specific contextual

settings. Additionally, inference scenario generally leads to better performance, except for CVD where model performance is relatively weak overall.

(a) Baseline Scenario (b) Inference Scenario

Figure 3 Comparative Analysis of Model Evaluation Results in Two Scenarios

This study drawn upon the algorithms and methodology form Hu et al. (2020) [6], advocating for the integration of SDHs into machine learning models for predicting NCDs prevalence through a decade-long dataset from the Ministry of Public Health's Open Data portal [7] and other Thai government bodies as outlined in Table 1. And also building on the findings of Bhoothookngoen and Sanchan [5], which highlighted the efficacy of various machine learning algorithms in NCDs prediction, including supervised techniques such as artificial neural networks (ANNs), SVM, decision trees, and RF, as well as unsupervised methods like K-means clustering and the Maximal Frequent Itemset Algorithm (MAFIA). Despite limited studies effectively incorporating SDH-related data as features, Hu et al. (2020) [6] offered a particularly applicable methodology, utilizing a wide range of data, including SDHs, as features.

This study delved deeper into the nuanced impact of data handling techniques on model performance. The features such as monthly household expenditure and environmental variables emerge as significant predictors when missing data is imputed in inference scenario as outlined in Figure 3. This study also compared the outcomes of a machine learning experiment with the original methodology proposed by Hu et al. (2020) [6]. Employing various machine learning models in two settings, the baseline and the inference scenario, revealed that the RF algorithm outperformed the stacking model as resulted in Hu et al. (2020)'s study [6]. This difference in performance is likely due to differences in the datasets used between the experiment and the study by Hu et al. (2020) [6]. Factors such as dataset size, quality, and feature composition strongly influence machine learning model performance.

While this experiment has yielded insights but not outperformed as compared to inspired algorithms and methodology from Hu et al. (2020) [6], its full potential remains hindered by several limitations that impeded the comprehensive utilization of the predictive power embedded within the explored features. Addressing these limitations is imperative to maximize the effectiveness of future research endeavors. One significant impediment is inconsistency, which posed a formidable obstacle to the result's integrity. The absence of guaranteed consistency across datasets spanning the entire feature space and anticipated time series introduced the risk of bias. This bias served as a distorting lens, potentially skewing the true relationships between features and predictive result.

Furthermore, granularity emerged as another critical limitation. The analysis of SDHs features, such as smoking prevalence, alcohol consumption, educational attainment, and the number of educational institutions, primarily operated at the national level, offering only a broad, population-wide perspective. This dearth of province-level data inhibited the capture of variations in these features across individual provinces which is on of SDHs category. Consequently, the generalizability of findings weakens, as conclusions derived from national data may not accurately reflect the unique demographic and risk profiles of each province,

potentially leading to incongruences with NCDs prediction.

The full potential of robust predictive modeling lied in overcoming current data limitations. A shift is therefore required, ensuring consistent time series for all features and establishing a reliable foundation for analysis. Incorporating province-level SDHs data is key, enabling a deeper understanding of geographic variations and enhancing the generalizability of findings. And the increasing the granularity of data for SDHs and other key features provided the necessary detail for nuanced analysis and facilitated deeper insights into predictive modeling efforts. This data metamorphosis holds the key for the future research endeuvor to reveal the full potential of SDHs as features for the prediction of NCDs prevalence in Thailand landscape.

4. Conclusion

NCDs pose a global threat, demanding innovative solutions through data-driven approaches. This study investigated machine learning models for predicting NCDs occurrence, focusing on algorithmic diversity and data dynamics, comprising cancer, cardiovascular diseases, diabetes, and hypertension, represent a significant menace to Thailand's public health and economic stability. As highlighted in a collaborative report from the Ministry of Public Health, the World Health Organization, and other UN agencies (MOPH et al., 2021), these conditions claim numerous lives and impose substantial economic burdens on the nation.

This study emphasized the pivotal role of SDHs in shaping the prevalence and impact of NCDs. Factors such as poverty, education, environmental quality, and healthcare access exert significant influence over the distribution and severity of NCDs within populations. Addressing NCDs effectively necessitates a profound understanding of these social determinants and their intricate connections with health outcomes.

Machine learning emerged as a potent tool in the battle against NCDs. Through the analysis of vast socio-economic datasets, including SDHs, machine learning algorithms can predict and elucidate patterns of NCDs prevalence, providing policymakers with invaluable insights. This integration of machine learning and public health policy holds promise for saving lives and fostering sustainable development in Thailand. The results of the study shed light on the performance of various machine learning algorithms in predicting NCDs prevalence within the Thai population. Through evaluation and comparison, certain algorithms, such as RF, exhibit promising predictive power across diverse datasets. Particularly, RF demonstrated the reduced MAPE values, suggesting its efficacy in NCDs prediction compared to alternative models with underperformed as compared to the inspired algorithms and methodology from Hu et al. (2020).

Despite promising results from certain machine learning models, limitations such as data disparities and the absence of provincial-level details hinder a comprehensive analysis. Moving

forward, it is imperative to address these limitations by establishing consistent time series data, integrating province-level SDHs, and expanding data collection efforts on environmental factors like air pollution. By doing so, Thailand can develop a more nuanced understanding of NCDs prevalence and its determinants, paving the way for targeted interventions that enhance public health outcomes and promote the well-being of its citizens.

5. Acknowledgement

This work was completed independently and does not necessitate specific acknowledgments.

6. References

[1] Ministry of Public Health of Thailand, World Health Organization, United Nations Development Programme, and United Nations Inter-Agency Task Force (2023). Prevention and Control of Noncommunicable Diseases in Thailand – The Case for Investment, 2021 [Online]. Available: https://www.who.int/thailand/activities/NCDs_Investment_Case_Report

[2] R. M. Rasesemola, R. Mmusi-Phetoe and Y. Havenga, "Social determinants of health in non-communicable diseases prevention policies in South Africa," *Curationis*, vol. 46, no. 1, a2387, 2023.

[3] L. Manderson and S. Jewett, "Risk, lifestyle and non-communicable diseases of poverty," *Globalization and Health*, vol. 19, no. 1, 2023. [Online]. Available: https://doi.org/10.1186/s12992-023-00914-z

[4] M. Y. Yang, G. H. Kwak, T. Pollard, L. A. Celi and M. Ghassemi, "Evaluating the impact of social determinants on health prediction in the Intensive Care Unit," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES '23)*, Montréal, QC, Canada, 2023, pp. 333–350.

[5] P. Bhoothookngoen and N. Sanchan, "Predictive modeling of non-communicable diseases using social determinants of health as features: A review of existing approaches," *Srinakharinwirot Univ. Eng. J.*, vol. 19, no. 1, pp. 79–88, 2023.

[6] Z. Hu, H. Qiu, Z. Su, M. Shen and Z. Chen, "A stacking ensemble model to predict daily number of hospital admissions for cardiovascular diseases," *IEEE Access*, vol. 8, pp. 138719–138729, 2020.

[7] Open Government Data of Thailand [Online]. Available: https://opendata.data.go.th/en/group/public-health

[8] Pollution Control Department (2022, Oct. 22). Air4Thai [Online]. Available: http://air4thai.pcd.go.th/webV2/region.php?region=0.

[9] National Statistical Office Thailand [Online]. Available: http://www.nso.go.th/

[10] Ministry of Public Health of Thailand, *Hospitals Code*, [Online]. Available: https://hcode.moph.go.th/code/