Predictive Modeling of Non-Communicable Diseases Using Social Determinants of Health as Features: A Review of Existing Approaches

Peatiphat Bhoothookngoen Nattapong Sanchan
[1]School of Information Technology and Innovation, Bangkok University, Main Campus 9/1 Village No.5, Phaholyothin Road, Klong Nueng Sub-district, Klong Luang District, Pathumthani Province 12120, Thailand
*Corresponding author Email: peatiphat.bhoo@bumail.net

(Received: June 12, 2023; Revised: August 5, 2023 ; Accepted: December 18, 2023)

ABSTRACT

The paper presents a comprehensive review of the current state of predictive models for non-communicable diseases (NCDs) prevalence, specifically focusing on utilising social determinants of health (SDHs) as features for model training. The review's search strategy employed a thorough screening process to select sixteen studies for inclusion. These studies used supervised, unsupervised, and other algorithms to forecast NCDs' burden; the most frequently applied attributes were age, gender, Fasting Blood Sugar (FBS), physical inactivity, obesity, and smoking. The evaluation methods for the models included a range of metrics, such as Percent Accuracy, Receiver Operating Characteristic (ROC), and Hamming loss. The review concludes that predictive models have the potential to forecast NCD prevalence accurately and highlights the need for further research that focuses on incorporating SDH-related factors as features for model training.

Keyword: Non-communicable diseases, social determinants of health, prediction, machine learning.

1. Introduction

Non-communicable diseases (NCDs) are a significant public health issue that accounts for approximately 71% of global deaths, with low- and middle-income countries being disproportionately affected [1]. The burden of NCDs is increasing, primarily due to ageing populations, changes in lifestyle, and environmental factors such as air pollution and unhealthy diets. The prevention and management of NCDs require a comprehensive approach addressing the complex interplay of factors contributing to their development.

The social determinants of health (SDHs) are crucial in developing and managing NCDs. These determinants encompass a wide range of non-medical factors influencing health outcomes, including income, education, employment, food security, housing, and access to healthcare. Individuals with lower socioeconomic positions have a higher risk of developing NCDs and experiencing worse health outcomes. For example, low-income individuals may face barriers to accessing healthy foods and adequate medical care, increasing the risk of developing chronic conditions such as diabetes or heart disease [2]. Research suggests that addressing SDHs is crucial for reducing the burden of NCDs and improving health equity. SDHs can impact health outcomes more significantly than lifestyle choices or healthcare, accounting for up to 55% of health outcomes [1]. Thus, a comprehensive approach to managing NCDs must address the underlying social determinants of health to reduce health disparities.

Forecasting models can be valuable in predicting NCDs prevalence and mortality rates, as they can consider the complex interplay of various factors, including SDHs. Such models can help policymakers and healthcare providers allocate resources and develop targeted interventions to manage NCDs effectively. Accurate forecasting models incorporating SDHs can help identify populations at higher risk of developing NCDs, allowing for the implementation of timely interventions and preventive measures. Managing NCDs requires a comprehensive approach that considers the complex interplay of factors contributing to their development, including social determinants of health. Forecasting models can be useful in predicting NCDs prevalence and mortality rates, especially when incorporating SDHs as an essential feature. Addressing the social determinants of health and developing accurate predictive models should help reduce health disparities and effectively manage NCDs.

This review aims to investigate the current state of predictive models in NCDs prevalence, specifically focusing on using SDHs as features for training.

2. Related Works

There is considerable evidence linking SDHs and NCDs, as demonstrated by several studies. Braveman [3] found that individuals residing in neighbourhoods with poor socio-economic conditions had higher rates of NCDs than those in more affluent neighbourhoods. Similarly, Vallejo-Torres and Morris [4] used the corrected concentration index to measure inequality across time and areas of England and found that smoking and obesity contribute to income-related inequality in health. In another study, Stringhini [5] observed that low socioeconomic status was associated with greater mortality, with smoking having the highest population-attributable fraction, followed by physical inactivity and socioeconomic status. Moreover, Hosseinpoor [6] found that wealth and education were inversely associated with the prevalence of certain NCDs, such as angina, arthritis, asthma, depression, and comorbidity, with the strongest inequalities reported for angina, asthma, and comorbidity.

Despite the evidence linking SDHs and NCDs, it remains to be seen whether predictive models have incorporated SDHs as features in their training. This is a gap in the literature, as SDHs could be essential in understanding and predicting the risk of NCDs. Furthermore, by ignoring SDHs, predictive models may only notice critical features that could significantly impact the development of NCDs, resulting in less effective interventions and policies. Therefore, it is essential to investigate and incorporate SDHs in predictive models to enhance our understanding of NCDs.

3. Methodology

A systematic search strategy identified relevant literature from several databases, including PubMed, Scopus, Institute of Electrical and Electronics Engineers (IEEE), and Random Search. Appropriate keywords and Medical Subject Headings (MeSH) terms related to NCDs, SDHs, predictive modelling, and review articles will be used. The search strategy will be refined using Boolean operators (AND, OR, NOT) to filter the results. Additionally, the reference lists of relevant articles will be examined to identify additional sources.

A screening process was employed to select articles for inclusion in this review. The first stage will involve screening the titles and abstracts of the articles, followed by a full-text review of the selected articles in the second stage. This review's inclusion criteria include peer-reviewed articles written in English and addressed the research question. The exclusion criteria include articles irrelevant to the topic, not primary research articles or limited scope.

A structured data extraction form was developed to extract relevant data from the selected articles. The extracted data included the author, year, study objectives, SDHs used, modelling approach, model evaluation, and outcome measures. The extracted data were synthesised according to themes such as the features used, the modelling techniques employed and the limitations and challenges of the models. The review's findings were summarised, and conclusions were drawn based on the evidence.

4. Results

The escalating incidence of NCDs has led to an increased focus on developing predictive models that accurately forecast their occurrence. This review presents a literature review to consolidate criteria-met predictive model studies for forecasting NCDs prevalence. The systematic search yielded thirty-two studies, of which fourteen were excluded for not meeting the model criteria or not being applied in NCDs studies. Two were excluded for not being machine learning modelling studies.

The remaining sixteen literature sources encompass various algorithms, encompassing supervised, unsupervised, and other methodologies (Table I). Each of these sources specifically examines the prediction of NCDs, either from a diagnostic or an epidemiologic standpoint. In these studies, the algorithms and attributes employed for developing NCDs prediction were thoroughly reviewed and extracted.

TABLE I

Algorithms Used by Category

Category	Algorithms
Supervised: Labeled training data to make predictions	ANN, SVM, LSTM, Logistic Regression, Decision Tree, Naive Bayes, Random Forest, KNN, AdaBoost
Unsupervised: Exploring unlabeled data to uncover patterns	K-means clustering, MAFIA Binary Relevance (BR), Classifier Chains (CC), RAkEL, ML-KNN
Other: Algorithms that might not fall into supervised or unsupervised	DeepSHAP, GBDT, GAMM, Fuzzy Logic IF-THEN rules, Dynamic population model, Combination of evolution tree model and Multilevel Modelling
Evaluation Method	% Accuracy, Algorithms comparison, 95% CI, Kappa statistics, RMSE, Precision, Recall, F-measure, ROC, and Hamming loss

These sixteen studies highlight the effectiveness of supervised learning techniques in predicting NCDs, emphasising the importance of algorithm selection and the potential for tailored interventions based on these predictions. Several studies have utilised supervised learning techniques to predict and model NCDs. For example, Ngom [7] employed artificial neural networks (ANN), support vector machines (SVM), decision trees, naive Bayes, logistic regression, and random forest. Equivalently, Keerthi Samhitha [8] and Mohan [9] used decision trees, K-nearest neighbour (KNN), K-means clustering, AdaBoost, and logistic regression in their supervised learning models. Hu [10] utilised gradient-boosting decision trees to predict non-communicable diseases and improve intervention programs in Bangladesh. Hu [11] employed a stacking ensemble model that combined linear regression, support vector regression, extreme gradient boosting, random forest, and gradient-boosting decision trees to predict daily hospital admissions for cardiovascular diseases. These studies demonstrate the effectiveness of supervised learning techniques in predicting and modelling NCDs.

Banu and Gomathy [12] utilised various unsupervised learning techniques, including K-means clustering, Maximal Frequent Itemset Algorithm (MAFIA), and C4.5 algorithm (supervised), to develop a disease forecasting system. Equivalently, Sangkatip and Phuboon-ob [13] employed multiple techniques, such as binary relevance (BR), classifier chains (CC), random k-labelsets (RAkEL), and multi-label k-nearest neighbour (ML-KNN), to classify non-communicable diseases. In contrast, Davagdorj [14] utilised a combination of both supervised and unsupervised learning techniques, including hybrid feature selection, eXtreme Gradient Boosting (XGBoost), logistic regression (supervised), random forest (supervised), KNN (supervised), Support Vector Machine - Recursive Feature Elimination (SVM-RFE) (supervised), Multi-

Layer Perceptron (MLP) (supervised), Neural Network (NN) (supervised), and random forest-based feature selection, in their learning models to predict smoking-induced NCDs.

And other various machine-learning techniques have been used to investigate NCDs. For example, George and Thomas [15] developed fuzzy logic-based IF-THEN rules to forecast peak demand days of chronic respiratory diseases. Hu [10] and Hu [11] employed machine learning techniques to examine NCDs. Hastings [16] utilised a dynamic population model with regression to project new-onset cardiovascular disease by socioeconomic group in Australia. Davagdorj [14] used an Explainable Artificial Intelligence Based Framework for NCDs Prediction, incorporating Deep Shapley Additive Explanations (DeepSHAP) to enhance interpretability. Wang and Wang [17] combined the evolution tree model and Multilevel Modelling (MLM) to model and predict global NCDs. Lastly, Stringhini [5] applied a generalised additive mixed model (GMM) to study NCDs risk factors in older adults. These studies collectively demonstrate the diverse machine-learning techniques employed in investigating NCDs.

Based on the reviewed literatures, the evaluation methods for the models included Percent Accuracy, Algorithms comparison, 95% Confident Interval (CI), Kappa statistics, Root Mean Square Error (RMSE), Precision, Recall, F-measure, Receiver Operating Characteristic (ROC), and Hamming loss. Five studies (table II) used non-individual factors as attributes, eleven used individual factors (Non-SDHs) [18][19][20][21], and the rest were excluded due to unidentified attributes [22]. The studies that used individual factors proposed the outcome as an individual diagnostic result based on individual input factors rather than SDHs-related factors. Examples of studies that used predictive models for forecasting NCDs include those by Wang and Wang [20], Stringhini [18], George and Thomas [4], Hastings [9], and Hu [12]. These studies employed different algorithms, attributes, and evaluation methods to obtain the desired outcomes.

Numerous recent studies have been conducted to develop predictive models for NCDs using a variety of attributes. Eleven of the nineteen studies reviewed selected individual factors as the primary training attribute, while five opted for non-individual factors (SDHs). The attributes used in the remaining two studies were not explicitly specified. Among the five studies that explored non-individual factors, Wang and Wang [20] investigated a predictive model for global NCDs deaths, incorporating socioeconomic factors, country development level, income at the country level, and the number of NCDs deaths. Their study proposed a novel algorithm that combined the evolution tree model and the Multilevel model (MLM) and evaluated its accuracy in fitting the data. Comparing it to linear regression (LR), the proposed algorithm achieved an R-squared value of 0.7932, while LR yielded 0.7005. These findings underscore the notable association between socioeconomic factors and NCD-related mortality.

In another study, Stringhini [18] examined the relationship between low socioeconomic status and NCDs risk factors, such as diabetes, high alcohol intake, high blood pressure, obesity, physical inactivity, and smoking among older individuals in multi-cohort populations from 24 countries. The study used generalised additive mixed models (GAMM) for analysis and found an association between socioeconomic status and physical functioning. George and Thomas [4] developed a model for forecasting the peak demand days of chronic respiratory diseases using Fuzzy logic and applied environmental factors to predict the peak demand day. Hastings [9] deployed a dynamic population model to determine the new-onset cardiovascular disease (CVD) by socioeconomic group in Australia, which included parameters such as population, risk of new-onset CVD by socioeconomic quintile, and utility and found that 8.4% of people in the most disadvantaged quintile were at high risk of CVD. Finally, Hu [12] developed a predictive model for the number of CVD admissions using air quality, hospital admission, and meteorological data. The authors used the stacking model and Sequential Forward Floating Selection (SFFS) for feature selection in training the model. Their study found that the stacking model outperformed RF regarding MAE, RMSE, MAPE, and R square.

Among the various attributes, certain factors were frequently considered in the analysis. Age emerged as the most applied attribute, appearing seven times. Fasting Blood Sugar (FBS) and gender followed closely behind, each being utilised six times. Other attributes that garnered significant attention included Alcoholic habits, Blood pressure, Exercise-induced angina, Family history, Maximum heart rate achieved, Obesity, Physical inactivity, the slope of peak exercise ST segment, Smoking, and Thalassemia, all being taken into account four times. Several attributes were also considered moderate, appearing three times in the analysis. These included Serum Cholesterol, resting blood pressure, psychological stress, and Diabetes Mellitus (DM). Additionally, a few attributes were considered two times. All of these attributes were considered individual factors, risk factors for NCDs but not SDHs. While individual factors were the most commonly used attributes in the reviewed studies, insights from the five studies (Table II) using non-individual factors highlight the association between socioeconomic and environmental factors and NCDs. Therefore, crucial to consider various attributes when developing NCD predictive models. Hu's [12] study is particularly relevant to SDHs, and its model development method will be modified to be used in this experiment due to the similarity in study design and selected attributes.

5. Conclusion and Discussion

The objective of this review is to synthesise studies focusing on predictive models that forecast the prevalence of NCDs by utilising SDHs (non-individual factors) as features. A systematic search strategy was employed to identify relevant studies, resulting in a total of thirty-two studies. After applying inclusion and exclusion criteria, sixteen studies were considered eligible for inclusion in this review. Out of these, only five studies incorporated relevant SDHs as features in their models.

TABLE II
Summary of Forecast Model Literature Review for Studies Deploying Non-Individual Factors as Features

Authors	Features
Wang, Y., & Wang, J. (2020) [20]	Country type [income] and Country development stage, NCD death, Socio-economic status
Stringhini, S., et al. (2018) [18]	High alcohol intake, Low socioeconomic status, Current smoking, Hypertension, Diabetes Mellitus, Obesity, Physical inactivity
George, N., & Thomas, J. (2018) [4]	Nitrogen Dioxide, Outdoor temp, Particle matter, Relative humidity, Sulphur Dioxide, Wind speed
Hu, Z., et al. (2020) [12]	Air quality, Hospital admission, Meteorological
Hastings, K., et al. (2022) [9]	Population, Risk of new-onset CVS by socioeconomic quintile, Utility

The identified studies employed a variety of supervised, unsupervised, and other algorithms, with individual factors (Non-SDHs) being the most frequently used attributes. The studies also utilised various evaluation methods, including Percent Accuracy, Algorithms comparison, 95% CI, Kappa stat, RMSE, Precision, Recall, F measure, ROC, and Hamming loss. The review demonstrates the growing interest in utilising machine learning algorithms for predicting and diagnosing NCDs. Examples of studies that employed predictive models for forecasting NCDs include those by Wang and Wang [20], Stringhini [18], George and Thomas [4], Hastings [9], and Hu [12]. These studies employed various algorithms, attributes, and evaluation methods to obtain the desired outcomes. The review also reveals the effectiveness of supervised learning techniques, such as ANN, SVM, decision trees, naive Bayes, logistic regression, and random forest, in predicting and modelling NCDs. Moreover, unsupervised learning techniques, including K-means clustering, MAFIA, and the C4.5 algorithm, were employed in some studies to develop disease forecasting systems. From the results of this review, it is evident that the features considered SDHs have rarely been used. A few studies integrated SDHs (Table II) into their models in conjunction with non-SDH factors. However, no study exclusively employed SDHs to predict NCDs' prevalence directly. This perspective could be attributed to the prevailing focus of model developers on training models to predict NCDs diagnoses primarily based on individual factors, such as laboratory testing or other disease risk factors, rather than considering the broader epidemiological context.

The results of the study showed that the SDHs categorisation identifies some categories deployed as features in NCDs prediction (see Table II). However, several categories [21] still have not yet been deployed, especially for NCDs prevalence prediction. These categories include education, social protection, unemployment and job insecurity, working life conditions, food insecurity, social inclusion, and non-discrimination.

Determining the accuracy and effectiveness of either comparing SDHs as features amongst themselves or pitting SDHs as features against others is a premature endeavour, supported by the following reasons. Firstly, several categories of SDHs have yet to be utilised as features. Secondly, the variations among studies encompass diverse deployed algorithms and selected features. Thirdly, the outcomes of this review indicate a scarcity of studies that have tested the approach of employing SDHs as features for predicting NCDs prevalence.

This study provided insights into the methods employed and outcomes achieved in recent studies focused on developing predictive models for NCDs. The findings of this review can inform future studies aimed at developing more accurate and effective predictive models for NCDs prevalence.

6. References

[1] World Health Organization (2022). Non communicable diseases, World Health Organization [online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[2] World Health Organization (2019). Social Determinants of Health, World Health Organization. Available: https://www.who.int/health-topics/social-determinants-of-health

[3] P. Braveman, S. Egerter and D.R. Williams, "The Social Determinants of Health: Coming of age", *Annual Review of Public Health*, vol. 32, no. 1, pp. 381–398, 2011.

[4] L. Vallejo-Torres and S. Morris, "The contribution of smoking and obesity to income-related inequalities in health in England", Social Science & Medicine, vol. 71, no. 6, pp. 1189–1198, 2010.

[5] S. Stringhini, C. Carmeli, M. Jokela, M, Avendaño, P. Muennig, F. Guida et al., "Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: A multicohort study and meta-analysis of 1·7 million men and women", *The Lancet*, vol. 389, no. 10075, pp. 1229–1237, 2017.

[6] A.R. Hosseinpoor, N. Bergen, S. Mendis, S. Harper, E. Verdes, A. Kunst and S. Chatterji, "Socioeconomic inequality in the prevalence of noncommunicable diseases in low- and middle-income countries: Results from the World Health Survey", *BMC Public Health*, vol.12, article number 474, 2012.

[7] F. Ngom, I. Fall, M. S. Camara and A. Bah, "A study on predicting and diagnosing non-communicable diseases: case of cardiovascular diseases," in *International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco, 2020, pp. 1-8.

[8] B. Keerthi Samhitha, M.R. Sarika Priya, C. Sanjana, S.C. Mana and J. Jose, "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms," in *International Conference on*

Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1326-1330.

[9] N. Mohan, V. Jain and G. Agrawal, "Heart Disease Prediction Using Supervised Machine Learning Algorithms," in *5th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2021, pp. 1-3.

[10] M. Hu, Y. Nohara, Y. Wakata, A. Ahmed, N. Nakashima and M. Nakamura, “Machine learning based prediction of non-communicable diseases to improving intervention program in Bangladesh”, *European Journal for Biomedical Informatics*, vol.14, no.4, 2018.

[11] 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.

[12] M. A. N. Banu and B. Gomathy, "Disease Forecasting System Using Data Mining Methods," in *International Conference on Intelligent Computing Applications*, Coimbatore, India, 2014, pp. 130-133.

[13] W. Sangkatip and J. Phuboon-Ob, "Non-Communicable Diseases Classification using Multi-Label Learning Techniques," in *5th International Conference on Information Technology (InCIT)*, Chonburi, Thailand, 2020, pp. 17-21.

[14] K. Davagdorj, J. -W. Bae, V. -H. Pham, N. Theera-Umpon and K. H. Ryu, "Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction," *IEEE Access*, vol. 9, pp. 123672-123688, 2021.

[15] N. George and J. Thomas, "Forecasting the Peak Demand Days of Chronic Respiratory Diseases with Fuzzy Logic," in *International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, Kottayam, India, 2018, pp. 1-5.

[16] K. Hastings, C. Marquina, J. Morton, D. Abushanab, D. Berkovic, S. Talic, E. Zomer , D. Liew, Z. Ademi, “Projected new-onset cardiovascular disease by Socioeconomic Group in Australia”, *Pharmacoeconomics*, vol.40, no.4, pp. 449–460, 2022.

[17] Y. Wang and J. Wang, “Modelling and prediction of Global Non-communicable diseases”, *BMC Public Health*, vol.20, no.1, 2020.

[18] M. A. Alim, S. Habib, Y. Farooq and A. Rafay, "Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model," in *3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, 2020, pp. 1-5.

[19] K. Davagdorj, J. -W. Bae, V. -H. Pham, N. Theera-Umpon and K. H. Ryu, "Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction," *IEEE Access*, vol. 9, pp. 123672-123688, 2021.

[20] R. Ferdousi, M. A. Hossain and A. E. Saddik, "Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS," *IEEE Access*, vol. 9, pp. 96823-96837, 2021.

[21] S. Islam, N. Jahan and M. E. Khatun, "Cardiovascular Disease Forecast using Machine Learning Paradigms," in *Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020, pp. 487-490.

[22] M. Marmot and R. Bell, “Social Determinants and non-communicable diseases: Time for Integrated Action”, *thebmj*, BMJ 2019; 364:l251.