JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 10 Number 2 | December 2025: 465-473
DOI: 10.15575/join.v10i2.1609

K-Means-Based Pseudo-Labeling Technique in Supervised Learning Models for Regional Classification Based on Types of Non-Communicable Diseases

Herison Surbakti[1], Tb Ai Munandar[2]
[1]Information Technology, Faculty of Science and Technology, Universitas Respati Yogyakarta, Indonesia
[2]Informatics, Faculty of Computer Science, Universitas Bhayangkara Jakarta Raya, Indonesia

Article Info

Article history:

Received April 17, 2025
Revised June 20, 2025
Accepted June 23, 2025
Published November 10, 2025

Keywords:

K-means
Non-communicable diseases
Pseudo-labeling
Regional classification
Semi-supervised learning

ABSTRACT

Non-Communicable Diseases (NCDs) pose a critical threat to global public health, with Indonesia experiencing significant challenges due to high mortality rates and uneven regional distribution. In Banten Province, limited access to labeled health data hampers effective, data-driven intervention strategies. This study proposes a semi-supervised learning approach to develop a regional classification model for NCDs. The methodology begins with K-Means clustering applied to data from 254 community health centers (Puskesmas) to generate pseudo-labels. Various cluster configurations (k=2 to 8) were evaluated, with the optimal result being two clusters based on a silhouette score of 0.735. These clusters were then used to create a semi-labeled dataset for supervised learning. Eight classification algorithms—CN2 Rule Inducer, k-Nearest Neighbor (kNN), Logistic Regression, Naïve Bayes, Neural Network, Random Forest, Support Vector Machine (SVM), and Decision Tree—were trained and compared. Among them, the Neural Network model achieved the highest performance, with an AUC of 0.999 and an MCC of 0.976, indicating excellent stability and predictive accuracy. The findings validate the effectiveness of semi-supervised learning for health classification tasks when labeled data is scarce. This approach can serve as a valuable decision-support tool for regional health planning and targeted interventions, enhancing the precision and efficiency of public health responses.

Corresponding Author:

Herison Surbakti,
Information Technology, Faculty of Science and Technology, Universitas Respati Yogyakarta
Jl. Laksda Adisucipto KM.6,3, Ambarukmo, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia
Email: herisonsurbakti@respati.ac.id

1. INTRODUCTION

Non-Communicable Diseases (NCDs) have emerged as a significant global health challenge, currently responsible for approximately 74% of all deaths worldwide [1], [2], a trend also observed in Indonesia [3], [4]. These diseases encompass a broad spectrum of chronic conditions, including cardiovascular diseases, diabetes, cancer, chronic respiratory illnesses, and other health disorders [5] - [7]. The rising prevalence of NCDs places considerable strain on healthcare systems and economic resources [8]. In Indonesia, Banten Province, as a developing region characterized by a large and heterogeneous population, is particularly affected by these challenges [9]. However, limited and

p-ISSN: 2528-1682
e-ISSN: 2527-9165

inadequately structured data concerning NCD distribution significantly hinder targeted health intervention strategies and effective public health responses.

A comprehensive understanding of disease distribution and associated regional risk factors is crucial in addressing the burden posed by NCDs [10]. Such insights can aid healthcare providers, researchers, and policymakers in more effectively allocating resources and implementing targeted preventive measures. Nevertheless, classifying regions based on NCD prevalence presents considerable complexity due to geographic variability in disease incidence, disparities in healthcare access, lifestyle variations across communities, and diverse environmental factors. Furthermore, the paucity of medical data remains a critical obstacle, especially in remote areas, exacerbated by infrastructural and technological limitations [11].

Advancements in technology have significantly enhanced opportunities for collecting, processing, and analyzing health data. Electronic health information systems and sophisticated medical devices now facilitate the acquisition of extensive and more accurate datasets [12] - [14]. Proper utilization of this data has the potential to yield valuable insights into regional NCD distribution patterns.

Numerous international studies have explored the distribution of NCDs. For instance, disparities in the prevalence of cardiovascular, digestive, and chronic respiratory diseases have been identified within the European Economic Area (EEA) [15]. In Bhutan, the persistence of NCD prevalence has been linked to factors such as inadequate dietary habits, alcohol consumption, hypertension, obesity, and diabetes, despite reductions in overall risk factors over the past decade [16] Similar disparities have been documented in South Africa [17], as well as in various other countries, including Burkina Faso, Bangladesh, Colombia, the Democratic Republic of Congo (DRC), Nigeria, Syria [18] Ethiopia [19], China [20] - [22], India [23], [24], and Japan [25].

To mitigate data scarcity issues, machine learning techniques have increasingly been employed for NCD classification. Several studies have integrated medical datasets with machine learning algorithms to predict conditions such as diabetes, cardiovascular diseases, chronic kidney disease, among others [26] - [35]. Research conducted by [27], [28], for example, demonstrated the effectiveness of machine learning models in accurately predicting diabetes and heart diseases. Additionally, machine learning has been successfully applied in detecting kidney diseases [29], intervening in acute pancreatitis [30], classifying medical conditions through imaging [31], and predicting diabetes risk [32] - [35].

Semi-supervised learning represents a particularly promising approach for contexts with limited labeled data. By combining labeled and unlabeled datasets, this method enhances classification performance [36], [37]. Applications in medical classification include the identification of Alzheimer's disease and mild cognitive impairment from unlabeled neuropsychological data [38], COVID-19 detection [39] - [41], diagnosing aortic stenosis using echocardiogram data [42], and medical image classification [43].

In response to these challenges, this study aims to develop a regional classification model for NCD types in Banten Province, Indonesia, employing a semi-supervised learning approach. By leveraging both labeled and unlabeled datasets, the proposed model intends to provide more accurate identification of NCD distribution patterns within the province. Enhanced regional mapping and classification are anticipated to facilitate the development of targeted, efficient preventive strategies and health interventions.

The novelty of this research lies in its innovative application of semi-supervised learning for regional NCD classification in Banten Province, representing an underexplored area within the Indonesian context. This study contributes methodologically to the fields of data science and public health, while also offering practical implications for regional health policy formulation. Ultimately, the improved understanding of NCD distribution patterns resulting from this study can support more precise decision-making, optimal resource allocation, and enhanced quality of life, thereby promoting sustainable health development in Banten Province.

2. METHOD

The methodology of this study begins with identifying critical issues related to regional classification based on the types of Non-Communicable Diseases (NCDs) in Banten Province. Subsequently, medical data is collected from 254 community health centers, which are distributed across eight administrative regions. Initially, the collected data undergoes a pre-processing phase aimed at ensuring data quality and suitability for subsequent analysis. This includes normalization of all numerical attributes using min-max scaling to ensure uniform feature ranges, which is a critical requirement for K-Means clustering due to its reliance on distance-based similarity measures. Following this preliminary processing, an unsupervised learning method utilizing the K-Means clustering algorithm is applied to categorize regions based on discernible data patterns. K-Means was selected due to its efficiency in clustering based on attribute similarity, ease of implementation [44], and proven effectiveness in health-related research [45], particularly in generating pseudo-labels from unlabelled datasets such as medical imagery [46] - [48]. Moreover, K-Means demonstrates strong computational performance and is well-suited to medium-sized, numerically scaled datasets such as those used in this study [49] The resulting clusters generated through this method serve as pseudo-labels or target classes for constructing the subsequent classification model.

Before proceeding to the supervised learning phase, an additional data pre-processing step is performed to align the dataset format with the newly assigned cluster labels. The classification model is then developed using a supervised learning approach, evaluating the performance of eight machine learning algorithms, specifically CN2 Rule Inducer, Random Forest, Neural Network, Naïve Bayes, k-Nearest Neighbor (kNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. Each algorithm's performance is rigorously assessed to identify the most effective model for accurately classifying regions according to NCD types.

The final stage involves deploying the best-performing classification model as a practical tool to facilitate enhanced health mapping and targeted intervention planning within Banten Province. All analytical processes in this research utilize Orange Data Mining software and the R programming language as the primary computational tools.

3. RESULT AND DISCUSSION

3.1. Result

This study utilizes case data on Non-Communicable Diseases (NCDs) collected from 254 community health centers (Puskesmas) across eight administrative regions in Banten Province, Indonesia, covering 147 subdistricts. The dataset includes seven attributes: two categorical (city and subdistrict) and five numerical attributes representing the number of cases per NCD category, namely EFC (Excessive Food Consumption-related diseases), ESC (Excessive Sugar Consumption-related diseases), ESUC (Excessive Salt Usage-related diseases), ETCS (Excessive Tobacco Consumption-related diseases), and SMOKE (number of smoking cases). Data were sourced from standardized health surveillance reports officially submitted by each Puskesmas to the provincial health authority during the latest complete reporting year. These reports follow uniform national guidelines issued by the Indonesian Ministry of Health, promoting consistency in data capture; however, minor reporting inconsistencies or underreporting may persist due to human and administrative variability. Nevertheless, the inclusion of all registered community health centers ensures that the dataset remains broadly representative of the province's public health conditions.

An initial pre-processing phase confirmed the dataset's completeness and consistency, with no missing values found and all numerical features formatted as int64, requiring no further conversion. Descriptive statistical analysis revealed notable regional disparities, as maximum values for efc, esc, and esuc exceeded 2,100 cases—substantially higher than the respective medians of 15–16 cases—indicating the presence of significant outliers. Figure 1 illustrates the distribution of NCD case categories across regions. It shows that EFC, ESC, and ETCS display a strong right-skewed distribution, with most subdistricts reporting low case counts but a few exhibiting extreme highs. SMOKE follows a similar trend with wider spread, while ESUC, although skewed, appears slightly less extreme. These disparities underscore the need for clustering and classification methods to uncover regional groupings with similar NCD profiles, thereby informing more targeted and equitable public health interventions.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

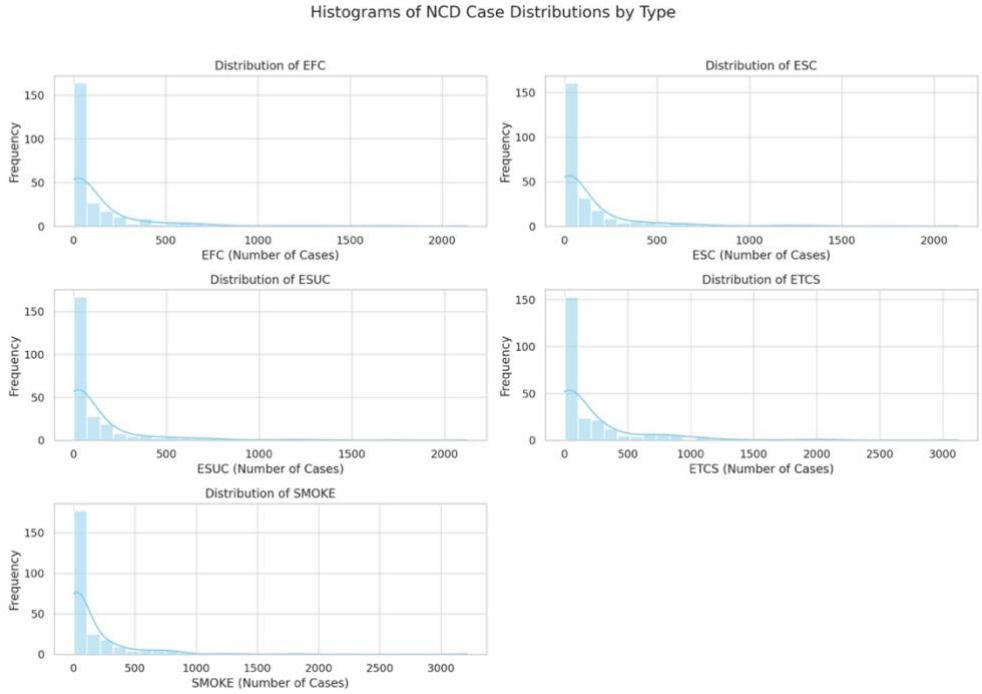

Figure 1. Distribution of Non-Communicable Disease Cases by Type

To explore these underlying patterns, an unsupervised learning technique using the K-Means algorithm was applied to the numerical attributes. The clustering process was conducted using Orange Data Mining software, testing cluster counts ranging from 2 to 8 with 10 repetitions and a maximum of 300 iterations. Random initialization was employed for initial centroid selection. Cluster quality was evaluated using silhouette scores, with the following results: 2 clusters (0.735), 3 clusters (0.608), 4 clusters (0.642), 5 clusters (0.618), 6 clusters (0.567), 7 clusters (0.556), and 8 clusters (0.539). The 2-cluster configuration was selected as optimal, achieving the highest silhouette score of 0.735, indicating well-separated and internally cohesive clusters. Figure 2 presents a visualization of the clustering output, depicting the resulting regional groupings.

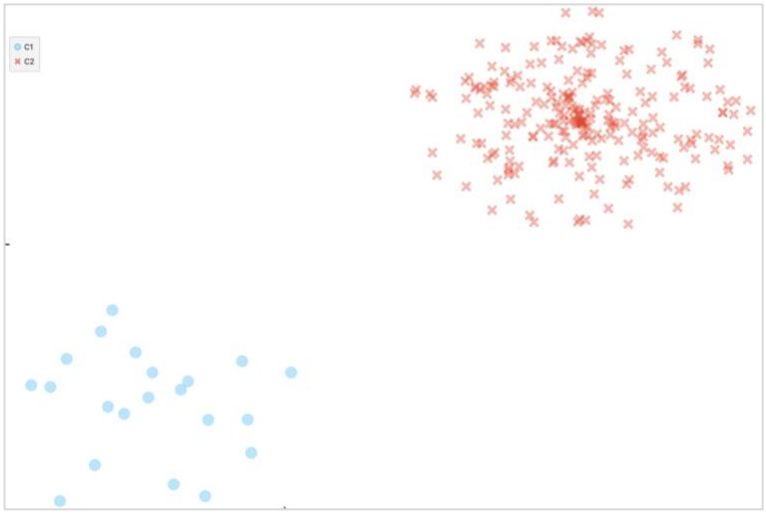

Figure 2. Visualization of K-Means Clustering Results

Analysis of the cluster characteristics showed that Cluster C1 is composed of subdistricts with significantly higher average values across all five indicators—efc (880.91), esc (852.41), esuc (817.05), etcs (1220.09), and smoke (377.05)—suggesting a high burden of NCDs likely associated with excessive consumption behavior and smoking prevalence. In contrast, Cluster C2 comprises subdistricts with much lower values, averaging in the 60s for efc, esc, and esuc, around 128 for etcs, and 122.61 for smoke. Based on these findings, Cluster C1 was interpreted as representing high-risk regions, while Cluster C2 represents relatively lower-risk regions. This clustering result was then used to enrich the dataset by adding a new column labeled "Cluster," effectively transforming it into a semi-labeled dataset. Each row in the dataset now includes a categorical label (C1 or C2), enabling the use of supervised machine learning algorithms for classification tasks. Table 1 presents selected entries from the updated dataset, showing the assigned cluster classes alongside their respective feature values.

Table 1. Sample of the New Dataset with Target Class Labels from K-Means Clustering

no.	city	subdistrict	efc	esc	esuc	etcs	smoke	Class
1	Pasar Kemis	Kab. Tangerang	98	199	183	344	3212	C2
2	Pondok Aren	Kota Tangerang Selatan	2135	2129	2121	2002	9	C1
3	Rajeg	Kab. Tangerang	3	3	3	800	1806	C2
4	Curug	Kab. Tangerang	82	82	82	82	71	C2
5	Ciputat	Kota Tangerang Selatan	4	1	1	5	0	C2
6	Pamulang	Kota Tangerang Selatan	747	839	853	856	759	C1
7	Pondok Aren	Kota Tangerang Selatan	827	1272	745	3121	93	C1
8	Kelapa Dua	Kab. Tangerang	411	411	411	440	2	C2
9	Serpong Utara	Kota Tangerang Selatan	159	162	159	251	363	C2
10	Cikupa	Kab. Tangerang	9	7	0	338	0	C2
11	Teluknaga	Kab. Tangerang	1	0	0	0	893	C2
12	Tigaraksa	Kab. Tangerang	267	123	119	398	858	C2
13	Serpong Utara	Kota Tangerang Selatan	153	147	122	580	35	C2
14	Pamulang	Kota Tangerang Selatan	207	206	200	233	11	C2
15	Ciputat	Kota Tangerang Selatan	81	79	79	95	678	C2

With the newly labeled dataset, eight supervised learning algorithms were trained and evaluated using Orange Data Mining: CN2 Rule Inducer, k-Nearest Neighbors (kNN), Logistic Regression, Naïve Bayes, Neural Network, Random Forest, Support Vector Machine (SVM), and Decision Tree. Performance evaluation used key classification metrics: Area Under the Curve (AUC), Classification Accuracy (CA), F1-score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Figure 3 presents a comparative visualization of the models' performance. Most models achieved outstanding AUC scores above 0.98, indicating excellent discriminatory capacity. The Neural Network and kNN models yielded the highest overall performance, both scoring 0.996 across CA, F1, Precision, and Recall. Logistic Regression also performed strongly with an AUC of 1.000 and an MCC of 0.950, though it was slightly less robust in class balance than the top two.

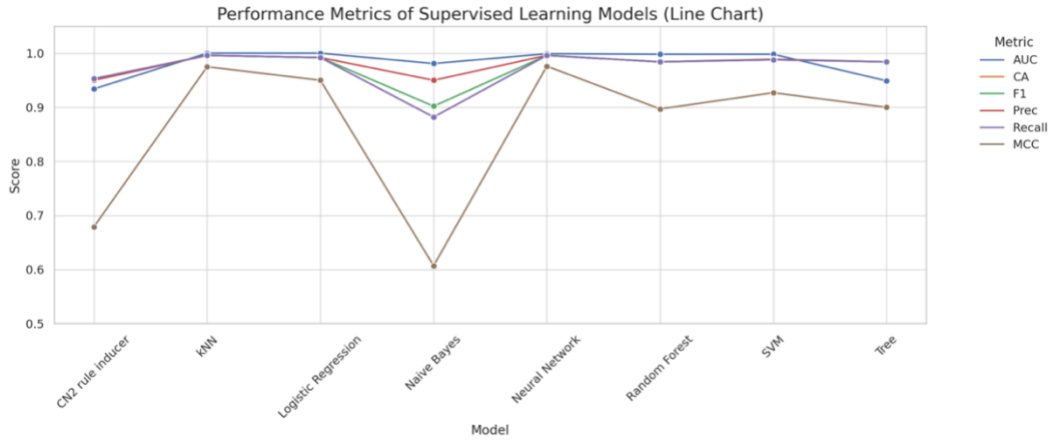

Figure 3. Performance Metrics of Supervised Learning Models

p-ISSN: 2528-1682
e-ISSN: 2527-9165

To address the potential risk of overfitting due to the relatively small sample size (254 observations), 10-fold cross-validation was implemented throughout the training and evaluation processes. This method ensured that each instance contributed to both training and validation phases, enhancing the robustness and generalizability of the models. Furthermore, for the Neural Network model, dropout regularization was applied during training, effectively reducing the risk of overfitting by preventing the model from relying excessively on specific features or pathways within the network.

In contrast, the Naïve Bayes model showed comparatively weaker performance, achieving only 0.882 in accuracy and an MCC of 0.607, likely due to its assumption of feature independence, which may not hold for NCD data. The CN2 Rule Inducer, while achieving a fair accuracy of 0.953, scored a lower MCC of 0.679, suggesting a potential imbalance in prediction between classes. The remaining models—Random Forest, SVM, and Decision Tree—performed well with MCC values above 0.89. Overall, the results demonstrate that the dataset labeled using clustering is well-suited for supervised learning, particularly when using instance-based and deep learning methods. Among them, the Neural Network model emerged as the most effective classifier.

The Neural Network model was selected as the best-performing model based on its highest MCC value of 0.976 and consistently strong scores across all performance metrics. This model also demonstrated excellent generalization capabilities, as confirmed by the cross-validation outcomes, which reflected consistently high precision and recall values while maintaining a low rate of false positives and false negatives. Its strong performance can be attributed to its capacity to model complex, non-linear patterns in the data, rendering it particularly suitable for heterogeneous public health datasets. These findings affirm the efficacy of integrating clustering and supervised learning in a semi-supervised pipeline. The final model holds significant potential as a decision-support tool for identifying at-risk regions for NCDs in Banten Province and can inform data-driven health policy and planning.

3.2. Discussion

The findings of this study clearly illustrate that employing a semi-supervised learning methodology—initiating with K-Means clustering followed by dataset labeling—effectively established a robust foundation for developing a regional classification model based on Non-Communicable Disease (NCD) case data. Utilizing Orange Data Mining significantly streamlined analytical tasks, particularly in data exploration, model development, and performance evaluation phases. The initial clustering yielded two clusters with an optimal silhouette score of 0.735, denoting strong inter-cluster separation. These clusters, specifically Cluster C1 (regions with high disease prevalence) and Cluster C2 (regions with lower disease prevalence), subsequently served as pseudo-labels for training the supervised learning model. Although this pseudo-labeling approach offers a practical solution in the absence of ground-truth labels, it also introduces potential limitations, such as the risk of inaccurate grouping due to reliance on purely statistical similarity rather than domain-expert validation.

During the supervised learning stage, eight distinct machine learning algorithms were evaluated to determine the most effective classification model. The majority of tested models demonstrated excellent performance, as evidenced by Area Under the Curve (AUC) values exceeding 0.98, reflecting robust discriminative capabilities. Among these, the Neural Network and k-Nearest Neighbor (kNN) models stood out prominently, achieving nearly perfect scores in key evaluation metrics such as Classification Accuracy (CA), F1-score, Precision, and Recall. Both models also recorded exceptionally high Matthews Correlation Coefficient (MCC) scores, reinforcing their reliable classification performance, especially significant given potential data imbalances.

Nonetheless, it is important to acknowledge that high performance on a small dataset can be susceptible to overfitting. To mitigate this, 10-fold cross-validation was utilized to validate model generalizability. In addition, dropout regularization was employed in training the Neural Network model to prevent co-adaptation of neurons, thereby enhancing the model's capacity to generalize across varying data instances. These methodological safeguards were critical in ensuring that the models' performance metrics were not merely artifacts of memorization or spurious patterns in the training data.

Ultimately, the Neural Network model emerged as the best-performing classifier, registering the highest MCC value of 0.976 and consistently strong outcomes across all metrics. Although the kNN model closely matched these results (with a marginal MCC difference of 0.001), the Neural Network was preferred due to its superior ability to capture complex, non-linear data relationships typically found in public health datasets. Furthermore, the Neural Network displayed notable resilience against overfitting and consistently maintained high accuracy despite significant variability across regional data. Logistic Regression and Support Vector Machine (SVM) models also showed commendable performances but slightly underperformed concerning MCC consistency and overall classification precision.

These conclusions are further substantiated by visual analyses, including model performance graphs and average metric evaluations (Figure 3). Comparative visual assessments highlighted AUC and F1-score as having notably high average scores across all models, indicating robust sensitivity and precision. Nonetheless, MCC proved pivotal in model selection, as it provided a balanced assessment of predictive performance across positive and negative classifications, particularly relevant in potentially imbalanced data scenarios. Thus, the visual analysis underscores that the Neural Network model excels not only in classification accuracy but also in stability and predictive reliability.

In summary, the outcomes of this study underscore the efficacy of a semi-supervised learning approach in overcoming challenges associated with limited labeled data in public health contexts. The use of clustering-generated pseudo-labels significantly facilitated the development of an accurate and practical classification model. The Neural Network model, identified as the optimal classifier, holds considerable promise as a decision-support tool for local governments aiming to map regions at high risk for NCDs. Its integration could inform data-driven public health policies, guide resource prioritization, and enhance the precision of intervention strategies throughout Banten Province.

4. CONCLUSION

Based on the research findings, it can be concluded that the semi-supervised learning approach, combining K-Means clustering for pseudo-label generation and supervised learning for classification, effectively distinguishes regions in Banten Province based on Non-Communicable Disease (NCD) prevalence. The utilization of clustering-derived labels addresses limitations in labeled data availability, enabling the development of a high-performing classification model. Among the evaluated algorithms, the Neural Network model demonstrated the most consistent and reliable performance, evidenced by a Matthews Correlation Coefficient (MCC) of 0.976, affirming its predictive strength and potential as a decision-support tool in formulating data-driven public health interventions.

For future research, several specific directions are proposed to enhance both the robustness and practical utility of the model. First, incorporating additional contextual variables—such as age distribution, income levels, healthcare access, and environmental exposures—could improve the model's sensitivity to factors influencing NCD risk. Second, longitudinal data integration is suggested to capture temporal patterns and predict future outbreaks, enabling proactive health management. Third, external validation using datasets from other provinces or national-level health records is necessary to assess the generalizability and scalability of the proposed framework. Fourth, hybrid modeling strategies such as ensemble deep learning or semi-supervised GANs could be explored to further optimize classification accuracy under label-sparse conditions. Finally, embedding the model within a GIS-based public health dashboard would enhance spatial decision-making, supporting geographically targeted interventions and more efficient resource allocation.

REFERENCES

[1] Kesehatan Kementerian, "Laporan Kinerja Instansi Pemerintah Kementerian Kesehatan RI untuk Tahun Anggaran 2021," Jakarta, Feb. 2022. Accessed: Oct. 18, 2024. [Online]. Available: https://ppid.kemkes.go.id/wp-content/uploads/2022/06/lakip_2022.pdf

[2] H. Arifin *et al.*, "Analysis of Modifiable, Non-Modifiable, and Physiological Risk Factors of Non-Communicable Diseases in Indonesia: Evidence from the 2018 Indonesian Basic Health Research," *J Multidiscip Healthc*, vol. Volume 15, pp. 2203–2221, Sep. 2022, doi: 10.2147/JMDH.S382191.

[3] Indriani and V. Fatmawati, "The Identification of Non-Communicable Diseases (NCDS) Risk Factors in Yogyakarta, Indonesia," 2023, pp. 165–174. doi: 10.2991/978-94-6463-190-6_21.

[4] World Health Organization, "NONCOMMUNICABLE DISEASES COUNTRY PROFILES 2018." Accessed: Oct. 18, 2023. [Online]. Available: https://www.who.int/docs/default-source/ncds/9789241514620-eng.pdf?sfvrsn=48f7a45c_2

[5] A. Afif, "Analisis Cluster Ward Pada Pengelompokan Wilayah Puskesmas Di Kota Kediri Berdasarkan Penyakit Tidak Menular," *Unisda Journal of Mathematics and Computer Science (UJMC)*, vol. 8, no. 2, pp. 39–44, Dec. 2022, doi: 10.52166/ujmc.v8i2.3567.

[6] R. Ferdousi, M. A. Hossain, and A. El Saddik, "Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS," *IEEE Access*, vol. 9, pp. 96823–96837, 2021, doi: 10.1109/ACCESS.2021.3094063.

[7] C. Wu, T. Zhou, Y. Tian, J. Wu, J. Li, and Z. Liu, "A method for the early prediction of chronic diseases based on short sequential medical data," *Artif Intell Med*, vol. 127, p. 102262, May 2022, doi: 10.1016/j.artmed.2022.102262.

[8] B. Legetic, A. Medici, M. Hernández-Avila, G. Alleyne, and A. Hennis, *DISEASE CONTROL PRIORITIES • THIRD EDITION Economic Dimensions of Noncommunicable Diseases in Latin America and the Caribbean*. 2016. Accessed: Nov. 06, 2025. [Online]. Available: www.paho.org/permissions

[9] A. Taher *et al.*, "Comprehensive Efforts to Accelerate Non-Communicable Disease Services in the Era of COVID-19 in Indonesia's Suburban Area," *ASEAN Journal of Community Engagement*, vol. 6, no. 1, pp. 152–68, Jul. 2022, doi: 10.7454/ajce.v6i1.1167.

[10] A. Budreviciute *et al.*, "Management and Prevention Strategies for Non-communicable Diseases (NCDs) and Their Risk Factors," *Front Public Health*, vol. 8, Nov. 2020, doi: 10.3389/fpubh.2020.574111.

[11] L. Handayani and L. Kristiana, "Faktor-Faktor Yang Memengaruhi Keterjangkauan Pelayanan Kesehatan Di Puskesmas Daerah Terpencil Perbatasan Di Kabupaten Sambas (Studi Kasus di Puskesmas Sajingan Besar)", Accessed: Nov. 06, 2025. [Online]. Available: https://media.neliti.com/media/publications-test/21346-faktor-faktor-yang-memengaruhi-keterjang-cdf92541.pdf

[12] L. C. S. Edmund, C. K. Ramaiah, and S. P. Gulla, "Electronic Medical Records Management Systems: An Overview," *DESIDOC Journal of Library & Information Technology*, vol. 29, no. 6, pp. 3–12, Nov. 2009, doi: 10.14429/djlit.29.273.

[13] N. N. Basil, S. Ambe, C. Ekhator, and E. Fonkem, "Health Records Database and Inherent Security Concerns: A Review of the Literature," *Cureus*, Oct. 2022, doi: 10.7759/cureus.30168.

[14] I. Silva, D. Ferreira, H. Peixoto, and J. Machado, "A Data Acquisition and Consolidation System based on openEHR applied to Physical Medicine and Rehabilitation," *Procedia Comput Sci*, vol. 220, pp. 844–849, 2023, doi: 10.1016/j.procs.2023.03.113.

[15] C. A. S. Andrade *et al.*, "Inequalities in the burden of non-communicable diseases across European countries: a systematic analysis of the Global Burden of Disease 2019 study," *Int J Equity Health*, vol. 22, no. 1, p. 140, Jul. 2023, doi: 10.1186/s12939-023-01958-8.

[16] S. Pengpid and K. Peltzer, "Trends in behavioral and biological risk factors for non-communicable diseases among adults in Bhutan: results from cross-sectional surveys in 2007, 2014, and 2019," *Front Public Health*, vol. 11, Aug. 2023, doi: 10.3389/fpubh.2023.1192183.

[17] R. A. Roomaney, B. van Wyk, A. Cois, and V. Pillay-van Wyk, "Inequity in the Distribution of Non-Communicable Disease Multimorbidity in Adults in South Africa: An Analysis of Prevalence and Patterns," *Int J Public Health*, vol. 67, Aug. 2022, doi: 10.3389/ijph.2022.1605072.

[18] J. Shu and W. Jin, "Prioritizing non-communicable diseases in the post-pandemic era based on a comprehensive analysis of the GBD 2019 from 1990 to 2019," *Sci Rep*, vol. 13, no. 1, p. 13325, Aug. 2023, doi: 10.1038/s41598-023-40595-7.

[19] A. Mohammed, "The effects of COVID-19 on Non-Communicable Disease : A Case Study of Six Countries (COVID-19 Situational Analysis Project)".

[20] T. T. Alamnia, G. M. Sargent, and M. Kelly, "Patterns of Non-Communicable Disease, Multimorbidity, and Population Awareness in Bahir Dar, Northwest Ethiopia: A Cross-Sectional Study," *Int J Gen Med*, vol. Volume 16, pp. 3013–3031, Jul. 2023, doi: 10.2147/IJGM.S421749.

[21] X.-F. Pan, J. Yang, Y. Wen, N. Li, S. Chen, and A. Pan, "Non-Communicable Diseases During the COVID-19 Pandemic and Beyond," *Engineering*, vol. 7, no. 7, pp. 899–902, Jul. 2021, doi: 10.1016/j.eng.2021.02.013.

[22] Q. Zeng *et al.*, "The Epidemiological Characteristics of Noncommunicable Diseases and Malignant Tumors in Guiyang, China: Cross-sectional Study," *JMIR Public Health Surveill*, vol. 8, no. 10, p. e36523, Oct. 2022, doi: 10.2196/36523.

[23] W. Peng *et al.*, "Trends in major non-communicable diseases and related risk factors in China 2002–2019: an analysis of nationally representative survey data," *Lancet Reg Health West Pac*, p. 100809, Jun. 2023, doi: 10.1016/j.lanwpc.2023.100809.

[24] G. R. Menon, J. Yadav, and D. John, "Burden of non-communicable diseases and its associated economic costs in India," *Social Sciences & Humanities Open*, vol. 5, no. 1, p. 100256, 2022, doi: 10.1016/j.ssaho.2022.100256.

[25] A. K. Yadav, K. R. Paltasingh, and P. K. Jena, "Incidence of Communicable and Non-communicable Diseases in India: Trends, Distributional Pattern and Determinants," *The Indian Economic Journal*, vol. 68, no. 4, pp. 593–609, Dec. 2020, doi: 10.1177/0019466221998841.

[26] S. Nomura, H. Sakamoto, C. Ghaznavi, and M. Inoue, "Toward a third term of Health Japan 21 – implications from the rise in non-communicable disease burden and highly preventable risk factors," *Lancet Reg Health West Pac*, vol. 21, p. 100377, Apr. 2022, doi: 10.1016/j.lanwpc.2021.100377.

[27] F. Mbonyinshuti, J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare, "Application of random forest model to predict the demand of essential medicines for noncommunicable diseases management in public health facilities," *Pan African Medical Journal*, vol. 42, 2022, doi: 10.11604/pamj.2022.42.89.33833.

[28] A. S. Abdalrada, J. Abawajy, T. Al-Quraishi, and S. M. S. Islam, "Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study," *J Diabetes Metab Disord*, vol. 21, no. 1, pp. 251–261, Jan. 2022, doi: 10.1007/s40200-021-00968-z.

[29] Q. Liu *et al.*, "Predicting the Risk of Incident Type 2 Diabetes Mellitus in Chinese Elderly Using Machine Learning Techniques," *J Pers Med*, vol. 12, no. 6, p. 905, May 2022, doi: 10.3390/jpm12060905.

[30] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *J Big Data*, vol. 9, no. 1, p. 109, Nov. 2022, doi: 10.1186/s40537-022-00657-5.
[31] N. Shi *et al.*, "Predicting the Need for Therapeutic Intervention and Mortality in Acute Pancreatitis: A Two-Center International Study Using Machine Learning," *J Pers Med*, vol. 12, no. 4, p. 616, Apr. 2022, doi: 10.3390/jpm12040616.
[32] J. Zhang, R. Han, G. Shao, B. Lv, and K. Sun, "Artificial Intelligence in Cardiovascular Atherosclerosis Imaging," *J Pers Med*, vol. 12, no. 3, p. 420, Mar. 2022, doi: 10.3390/jpm12030420.
[33] K. Al Sadi and W. Balachandran, "Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers," *Applied Sciences*, vol. 13, no. 4, p. 2344, Feb. 2023, doi: 10.3390/app13042344.
[34] G. Özsezer and G. Mermer, "Diabetes Risk Prediction with Machine Learning Models," *Artificial Intelligence Theory and Applications*, vol. 2, no. 2, pp. 1–9, 2022.
[35] O. A. Ebrahim and G. Derbew, "Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia 2021," *Sci Rep*, vol. 13, no. 1, p. 7779, May 2023, doi: 10.1038/s41598-023-34906-1.
[36] J. J. Boutilier, T. C. Y. Chan, M. Ranjan, and S. Deo, "Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis," *J Med Internet Res*, vol. 23, no. 1, p. e20123, Jan. 2021, doi: 10.2196/20123.
[37] Y. C A Padmanabha Reddy, P. Viswanath, and B. Eswara Reddy, "Semi-supervised learning: a brief review," *International Journal of Engineering & Technology*, vol. 7, no. 1.8, p. 81, Feb. 2018, doi: 10.14419/ijet.v7i1.8.9977.
[38] M. F. A. Hady and F. Schwenker, "Semi-supervised Learning," 2013, pp. 215–239. doi: 10.1007/978-3-642-36657-4_7.
[39] Y. Wang, X. Gu, W. Hou, M. Zhao, L. Sun, and C. Guo, "Dual Semi-Supervised Learning for Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Neuropsychological Data," *Brain Sci*, vol. 13, no. 2, Feb. 2023, doi: 10.3390/brainsci13020306.
[40] M. U. Alam and R. Rahmani, "Federated Semi-Supervised Multi-Task Learning to Detect COVID-19 and Lungs Segmentation Marking Using Chest Radiography Images and Raspberry Pi Devices: An Internet of Medical Things Application," *Sensors*, vol. 21, no. 15, p. 5025, Jul. 2021, doi: 10.3390/s21155025.
[41] Y. Zhang, L. Su, Z. Liu, W. Tan, Y. Jiang, and C. Cheng, "A semi-supervised learning approach for COVID-19 detection from chest CT scans," *Neurocomputing*, vol. 503, pp. 314–324, Sep. 2022, doi: 10.1016/j.neucom.2022.06.076.
[42] C. H. Han, M. Kim, and J. T. Kwak, "Semi-supervised learning for an improved diagnosis of COVID-19 in CT images," *PLoS One*, vol. 16, no. 4, p. e0249450, Apr. 2021, doi: 10.1371/journal.pone.0249450.
[43] Z. Huang, G. Long, B. Wessler, and M. C. Hughes, "A New Semi-supervised Learning Benchmark for Classifying View and Diagnosing Aortic Stenosis from Echocardiograms," 2021. [Online]. Available: https://github.com/tufts-ml/ssl-for-echocardiograms
[44] H. Wu, J. Sun, and Q. You, "Semi-Supervised Learning for Medical Image Classification Based on Anti-Curriculum Learning," *Mathematics*, vol. 11, no. 6, p. 1306, Mar. 2023, doi: 10.3390/math11061306.
[45] S. Lim, J. Park, M. Lee, and H. Lee, "Unsupervised object discovery with pseudo label generated using K-means and self-supervised transformer," *Neurocomputing*, vol. 545, p. 126326, Aug. 2023, doi: 10.1016/j.neucom.2023.126326.
[46] L. Chen *et al.*, "Making Your First Choice: To Address Cold Start Problem in Medical Active Learning," 2023. [Online]. Available: https://github.com/cliangyu/CSVAL.
[47] F. H. Awad, M. M. Hamad, and L. Alzubaidi, "Robust Classification and Detection of Big Medical Data Using Advanced Parallel K-Means Clustering, YOLOv4, and Logistic Regression," *Life*, vol. 13, no. 3, p. 691, Mar. 2023, doi: 10.3390/life13030691.
[48] K. Liu, X. Ning, and S. Liu, "Medical Image Classification Based on Semi-Supervised Generative Adversarial Network and Pseudo-Labelling," *Sensors*, vol. 22, no. 24, p. 9967, Dec. 2022, doi: 10.3390/s22249967.
[49] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, vol. 11, pp. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.