Abstract number: BA2025-1063

Using Big Data and Machine Learning to Develop Geospatial Distribution of High-Risk Type 2 Diabetes in Indonesia

Faris Rizky Andika[1], Josh Frederich[1], Aqsha Nur[2,3], Muhammad Rizal Khaefi[1], Esti Widiastuti Mangunadikusumo[4]

[1]GovtechHealth - Digital Transformation Office, Ministry of Health of the Republic of Indonesia, Jakarta, Indonesia
[2]Faculty of Public Health of the University of Indonesia, Jakarta, Indonesia
[3]London School of Hygiene and Tropical Medicine, London, the United Kingdom
[4]Directorate of Non-Communicable Diseases, Ministry of Health of the Republic of Indonesia, Jakarta, Indonesia

Background

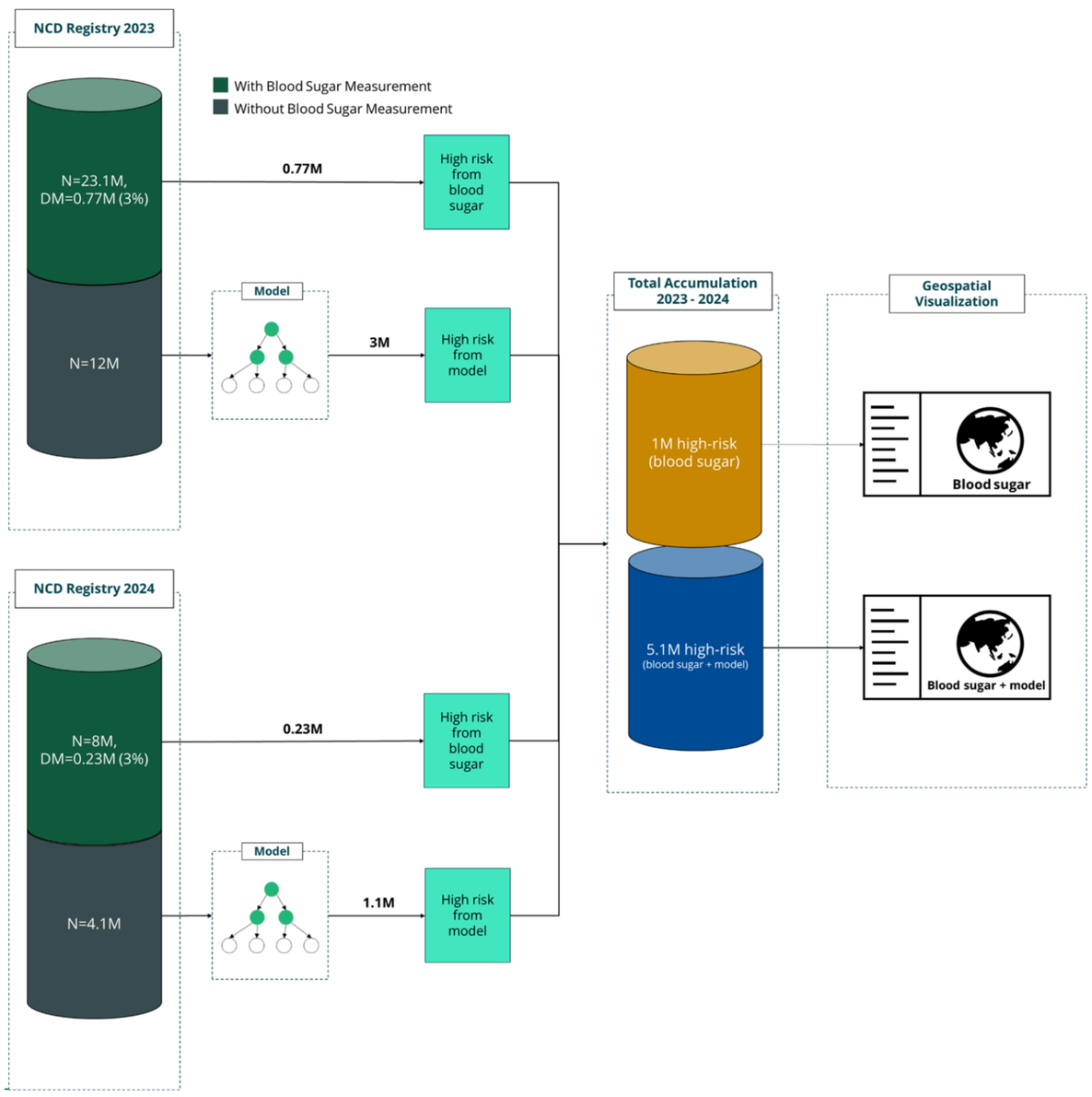
Indonesia, the fourth most populous country, ranks fifth in global diabetes cases. The 2023 Indonesian Health Survey reported a diabetes prevalence of 11.7% (95% CI: 11.1–12.4), with only about one-fifth of adults diagnosed. This reveals a major detection gap, especially in areas lacking targeted interventions. With many regions having limited resources to scale up diabetes screening, there is an urgent need to improve the visibility of high risk for type 2 diabetes mellitus (T2DM) for better resource allocation. Combining machine learning (ML) with screening data could help improve the visualization and support more focused, data-driven diabetes screening.

Aim

This study aims to integrate ML predictions for high risk T2DM to create a geospatial picture of T2DM in Indonesia and highlight prevalence gaps to guide targeted interventions.

Methods

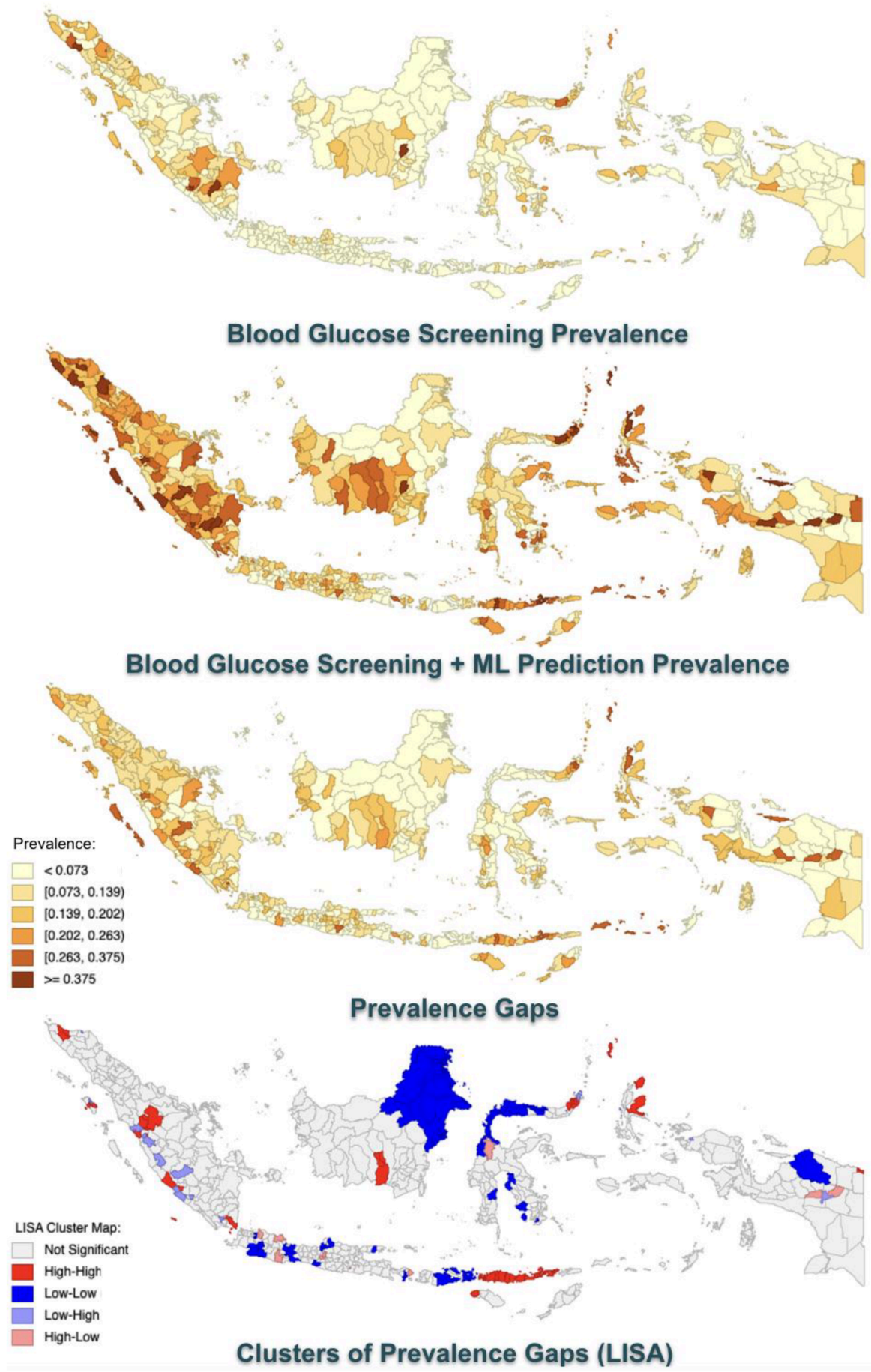
Trained ML model developed using Indonesia's National NCD Registry data from 2016–2022 (Aqsha et al., 2024)—with internal validation scores of 0.70 sensitivity, 0.88 specificity, 0.16 PPV, and 0.99 NPV—were applied to the National NCD Registries of 2023 (12 million records) and 2024 (4.1 million records) to predict additional high-risk diabetes cases. Prevalence estimates were calculated for (1) blood glucose screening and (2) screening plus ML predictions, and the prevalence gap (difference) was determined at the city/regency level. Geospatial risk mapping used Queen contiguity and K-nearest neighbours (k = 3), followed by spatial analysis (Univariate Moran's I and Local Indicators of Spatial Association, LISA) on the prevalence gap data. This analysis identified significant clusters and outliers in how much prevalence changed when ML outputs were added.



References

1. Lusiana, V., & Alfana, M. a. F. (2025). Spatial autocorrelation analysis of non-communicable diseases: Unveiling hidden patterns and hotspots of hypertension in the Yogyakarta Special Region. E3S Web of Conferences, 605, 02003.
2. Wang, Z., Dong, W., & Yang, K. (2022). Spatiotemporal Analysis and Risk Assessment Model Research of Diabetes among People over 45 Years Old in China. International Journal of Environmental Research and Public Health, 19(16), 9861.

Results and Discussions



Mapping the prevalence gap—the difference between screening-only and ML-augmented estimates—highlights distinct zones of elevated diabetes risk. Notable gaps occur in East Nusa Tenggara (especially Flores), North Maluku, northern Sulawesi, parts of Papua, Central Kalimantan, and southwestern Sumatra, while Java and Bali show comparatively modest gaps. Factoring in ML predictions elevates nationwide prevalence calculated from National NCD Registries from **3% to 11%,** which is close to 2023 Indonesian Health Survey and suggest a more accurate reflection.

A Moran's I of 0.214 indicates that these city/regency-level gaps are moderately clustered across the country rather than randomly distributed. The LISA analysis ($p < 0.05$) clarifies these spatial patterns. High-High (HH) clusters—indicating statistically larger jumps in prevalence—form in East Nusa Tenggara, North Maluku, North Sulawesi, Central Kalimantan, and western Sumatra. High-Low (HL) clusters emerge in central Papua, West Java, eastern Bali, and Central Sulawesi, signaling unusually large prevalence gaps located among lower-gap neighbors. Low-High (LH) clusters lie near HH regions (e.g., western Sumatra, northern Sulawesi), reflecting comparatively low local gaps adjacent to higher-gap areas.

Conclusion

Integrating ML-based predictions into screening data not only elevates the estimated overall prevalence but also highlights spatial clustering where high risk cases may be overlooked. High-gap clusters represent potential hotspots for more focused interventions, whereas HL and LH outliers warrant further investigation to understand local anomalies in the neighborhood patterns.