Epidemiology & Health Data Insights
(eISSN 3080-8111)

Methodological Paper

Methodological Note on Predicting One-Year Mortality for Chronic Diseases Using Administrative Data

Iliyar Arupzhanov[1], Aidar Alimbayev[2], Temirlan Seyil[1], Temirgali Aimyshev[1], Tilektes Maulenkul[1], Ainash Oshibayeva[3], Abduzhappar Gaipov[1]

[1]*School of Medicine, Nazarbayev University, Astana, Kazakhstan*
[2]*Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE*
[3]*Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan*

Received: Aug 27, 2025
Accepted: Sep 22, 2025

Corresponding author's email:
abduzhappar.gaipov@nu.edu.kz

Abstract:
Chronic diseases remain a leading cause of global mortality, underscoring the need for developing reliable models that predict mortality prediction to guide individualized treatments and optimize resource allocation. This methodological note presents a reproducible framework for predicting one-year mortality in chronic disease patients using large-scale administrative healthcare data. The approach employs retrospective cohort design, year-specific subcohorts, and a stratified 5-fold cross-validation using a broad range of machine learning models. Performance is assessed with multiple metrics, including AUC, sensitivity, specificity, and balanced accuracy, to account for class imbalance. Model interpretability is enhanced through SHapley Additive exPlanations (SHAP), enabling identification of key mortality predictors and their directional impact. The proposed framework is general and can be applied to different chronic diseases. It has already been successfully demonstrated in nationwide cohorts of patients with diabetes mellitus and chronic viral hepatitis in Kazakhstan, achieving AUC values of 0.74–0.80, comparable to international benchmarks despite relying on administrative data alone. The method is scalable and adaptable, allowing integration of laboratory and clinical data with feature selection to address high-dimensionality challenges. Its generalizability and clinical relevance, however, should be validated in practice using enriched datasets across additional chronic diseases and diverse populations.

Keywords: One-Year Mortality Prediction; Chronic Diseases; Machine Learning; SHAP Analysis

https://doi.org/10.63946/ehdi/17159

Introduction

Chronic diseases pose a significant challenge to global public health due to their prolonged duration, slow progression, and high mortality rates. Despite ongoing improvements in medical care and public awareness, mortality rates for such diseases remain high. According to the World Health Organization (WHO), chronic diseases were responsible for 43 million deaths in 2021, which is approximately 75% of global mortality, excluding those attributable to the COVID-19 pandemic [1].

Given the prolonged nature and relatively predictable progression of chronic diseases, one-year mortality prediction is especially valuable. Longer-term predictions, such as predicting 3-, 5-, 8-year mortality, tend to have lower accuracy and sensitivity due to increased uncertainty from disease trajectory change and external factors. In contrast, one-year mortality is an optimal choice, which provides predictive reliability and clinical utility. This shorter timeframe enables health professionals to develop individualized treatment strategies, make effective resource allocation and implement preventive measures in time. In this context, computational methods, especially machine learning (ML) models, have proven to be highly effective in accurately predicting mortality outcomes [2, 3].

The following methodological approach provides a clear and reproducible framework for predicting one-year mortality among patients with chronic diseases using large-scale administrative data. While the proposed framework was successfully implemented and validated on diabetes mellitus and chronic viral hepatitis patients using administrative data alone [4, 5]; the approach itself is not restricted to these conditions and can be generalized to other chronic diseases. Furthermore, it can be integrated with clinical notes, laboratory and imaging data where available. When enriched administrative data is available, the framework should be modified by introducing feature selection to address high-dimensionality challenges. This flexibility ensures that the approach can adapt to different data environments while maintaining methodological consistency. Nonetheless, its broader generalizability and clinical relevance should be validated in practice using enriched datasets across additional chronic diseases and diverse populations.

Study Design and Cohort Definition

The proposed method adopts a retrospective cohort study design using administrative healthcare data [6]. While a complete, definitive list of all chronic diseases does not exist, the WHO identifies four key categories: 1) cardiovascular diseases (e.g., heart disease, stroke, and hypertension); 2) cancer; 3) diabetes; and 4) chronic respiratory diseases (e.g., chronic obstructive pulmonary disease, asthma) [1]. U.S. agencies such as National Institutes of Health (NIH) [7] and Department of Health and Human Services (HHS) [8] apply a broader definition that also includes conditions like HIV, chronic kidney disease, and chronic viral hepatitis.

Patients diagnosed with chronic disease can be identified using the International Classification of Diseases 10th Revision (ICD-10) codes relevant to that specific disease. The dataset must undergo preprocessing to merge different registries, eliminate duplicate records, and form a comprehensive patient cohort. Clinical and laboratory data can be incorporated where available, enhancing predictive accuracy.

To reflect the temporal aspect of chronic diseases, the dataset is divided into distinct year-specific sub-cohorts. For each observation year, patients who died before the start of that year or who were diagnosed within the year are excluded. This approach ensures that only patients with complete clinical data and confirmed survival status at the beginning of the observation period are included.

Figure 1 depicts six types of patient groups in a typical dataset. A triangle represents the date of disease diagnosis, while a circle represents the exit date, which indicates the patient's death. For illustration, we select the third year of observation (red). The subcohort for the third year consists of two patient groups:

- **Case group** – patients diagnosed before the beginning of the third year who died during that year.
- **Control group** – patients diagnosed before the beginning of the third year but who remained alive during that year (similar to cases 2 and 4, which are also highlighted by blue lines and markers).

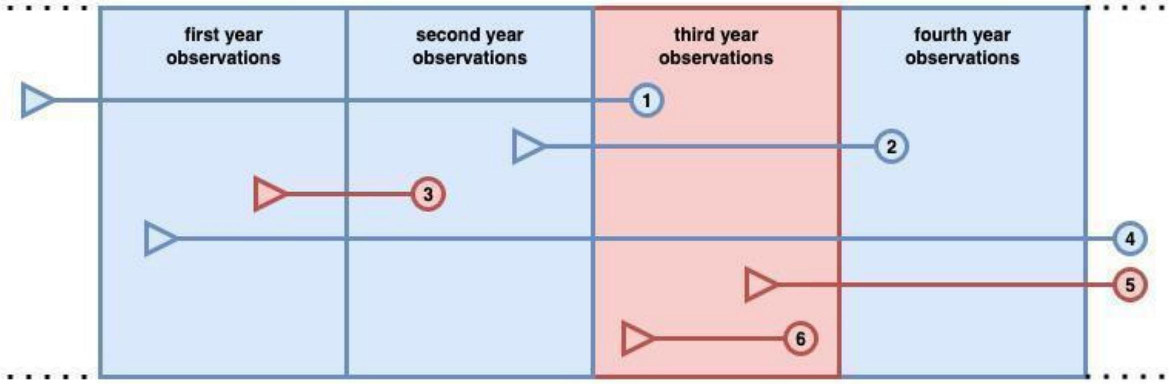

Figure 1. Description of subcohort selection

Patients who died before the start of the third year were excluded from the subcohort (e.g., case 3). Similarly, patients who were newly diagnosed during the third year were also excluded (e.g., cases 5 and 6). Only patients with available clinical information who were alive up to the end of the second year were included. Therefore, we selected subcohorts for the third year and predicted one-year mortality for all patients. It is also important to note that diseases may occur much earlier than the recorded diagnosis date.

Model Development and Validation

Selecting appropriate predictive models is critical for obtaining reliable mortality predictions. In our framework, we evaluate a diverse range of algorithms commonly used in health outcome prediction, including linear models, such as Logistic Regression (LR) [9] and Support Vector Machines with linear kernel (SVM) [10], Gaussian Naive Bayes (GNB) [11], K-Nearest Neighbors (KNN) [11], Discriminant Analysis (LDA, QDA) [12], Ensemble methods such as Random Forest (RF) [13], LightGBM (LGB) [14], XGBoost (XGB) [15], AdaBoost (ADB) [16], and Gradient Boosted Regression Trees (GBRT) [17].

The choice of these models can be rationalized based on their proven efficacy in prior studies predicting chronic disease outcomes, mortality, or similar health-related events [2, 3, 18-20].

Each year-specific cohort is initially divided into a training set (80%) and a held-out test set (20%) using a stratified random split to preserve class distribution. Model selection and tuning are conducted via grid search with stratified 5-fold cross-validation:

1. The training set is split into five equally sized folds using a stratified random split, maintaining a class label ratio across each fold to mitigate imbalance.
2. In each iteration, one fold is set as the validation set, and the remaining four folds are combined as the training set.
3. Standardization is applied based on the training folds to ensure consistent scaling.
4. Models are trained on these folds using the predefined set of hyperparameters and evaluated on the validation fold using the Area Under the ROC Curve (AUC), a threshold-independent metric suited for imbalanced data.
5. This procedure repeats for every hyperparameter combination across all 5 folds, ensuring robust validation.
6. The optimal model is selected based on the highest average AUC score across the 5 folds
7. The final year-specific model is retrained on the entire training set and evaluated on the held-out test set.

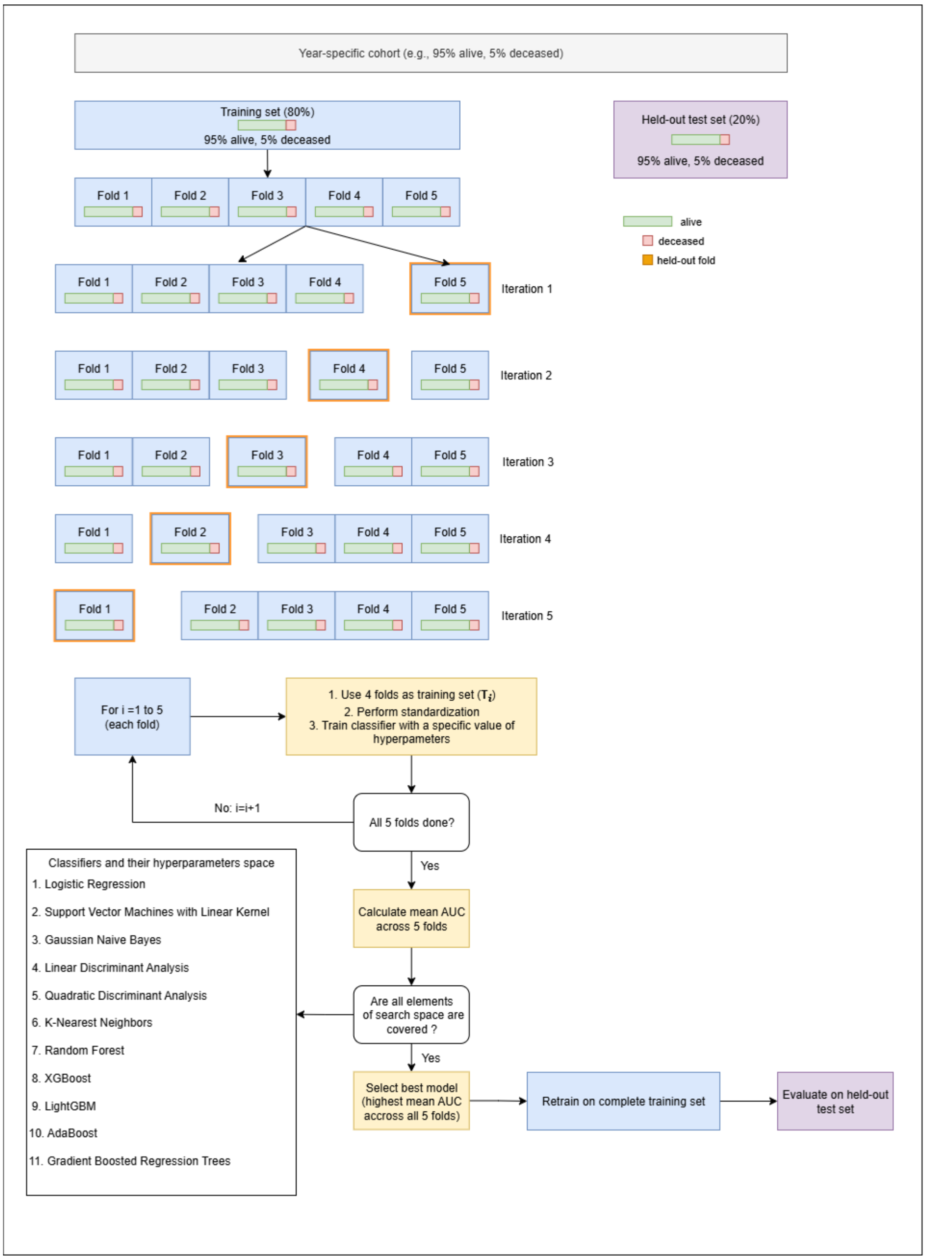
Year-specific cohort (e.g., 95% alive, 5% deceased)
Training set (80%)
95% alive, 5% deceased
Held-out test set (20%)
95% alive, 5% deceased
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
alive
deceased
held-out fold
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
For i =1 to 5
(each fold)
1. Use 4 folds as training set (T_i)
2. Perform standardization
3. Train classifier with a specific value of hyperparameters
No: i=i+1
All 5 folds done?
Yes
Calculate mean AUC across 5 folds
Are all elements of search space are covered ?
Yes
Select best model (highest mean AUC accross all 5 folds)
Retrain on complete training set
Evaluate on held-out test set
Classifiers and their hyperparameters space
1. Logistic Regression
2. Support Vector Machines with Linear Kernel
3. Gaussian Naive Bayes
4. Linear Discriminant Analysis
5. Quadratic Discriminant Analysis
6. K-Nearest Neighbors
7. Random Forest
8. XGBoost
9. LightGBM
10. AdaBoost
11. Gradient Boosted Regression Trees

Figure 2. Stratified 5-fold Cross-Validation

To mitigate overoptimism associated with the use of traditional accuracy as the primary performance metric in the presence of class imbalance, we report a range of performance metrics, including precision, specificity, sensitivity, balanced accuracy (BA), geometric mean of sensitivity and specificity (G-mean) and AUC. This comprehensive evaluation provides a comprehensive picture of the model performance, particularly when class imbalance is present.

picture of the model performance, particularly when class imbalance is present.

Explainability via SHAP Analysis

Understanding the underlying decision processes of ML models is critical in healthcare applications. To address this, we employ SHapley Additive exPlanations (SHAP) [21] to assess the contribution of each feature to model predictions. SHAP analysis achieves two primary objectives:

1. **Feature Importance:** assess the contribution of each predictor to the model's output.
2. **Directional Impact**: understand whether a feature increases or decreases mortality risk.

A SHAP value can be either positive or negative, indicating a corresponding positive or negative contribution to the model's prediction. In our case, a positive SHAP value for a predictor indicates that the predictor contributes to a higher probability of one-year mortality, while a negative value contributes toward survival. Given value can push the predicted outcome closer to 1 or 0, which indicates a higher or lower probability of one-year mortality, respectively. Not only does SHAP analysis help quantify the importance of features, but it also shows the direction of their influence on the outcome. This interpretability is crucial in supporting data-driven decision-making and improving communication between clinicians and patients.

Discussion

This is the first ML/AI based study in Kazakhstan as well as in Central Asia, which was conducted using big nationwide administrative healthcare data. The proposed framework was successfully implemented to predict one-year mortality in diabetes mellitus and chronic viral hepatitis patients using large-scale administrative data in Kazakhstan [4, 5]. These disease groups serve as demonstrations, but the framework is designed to be general and adaptable to additional chronic diseases.

The developed models using the proposed framework showed AUC values between 0.74 and 0.80, which are considered ‘fair’ and approaching ‘good’ as per the standard diagnostic metrics [22]. These studies indicate both feasibility and robustness of our approach. Importantly, the proposed framework was helpful in identifying key mortality predictors consistent with prior literature, such as age, sex, disease type and duration of disease, reinforcing the framework’s clinical relevance [23-26]. To illustrate this, Figure 3 shows SHAP analysis for the 2019-specific cohort of diabetes patients. The mean absolute SHAP values (Figure 3b) show that age, duration of diabetes and sex are the three top most important predictors (the longer bar shows a more important feature). In Figure 3a, reed dots with positive SHAP values indicate that higher feature values (e.g. older age or longer disease duration) increase the predicted risk of one-year mortality, whereas blue dots with positive SHAP values indicate an inverse relationship.

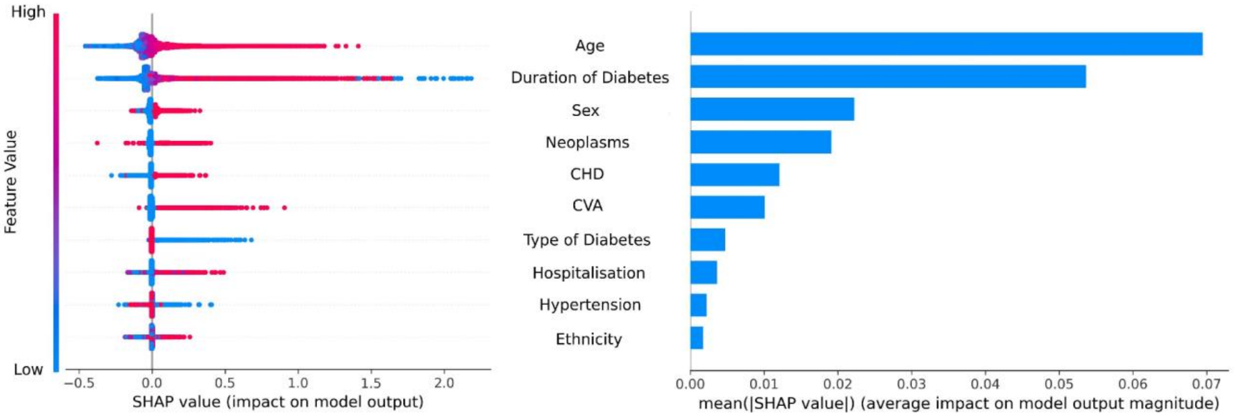

Figure 3. SHAP analysis of 2019-specific cohort of diabetes patients: (a) SHAP summary dot plot for the 2019-specific cohort. (b) The mean absolute SHAP value bar plots for the 2019-specific cohort. Reproduced from [4] under a Creative Commons Attribution 4.0 License.

Comparisons with international studies further highlight the strengths and limitations of our approach. A Chinese study predicting one-year mortality in older patients with coronary artery disease and impaired glucose tolerance or diabetes reported an AUC of 0.83 using GB [27]. In comparison, our study on predicting one-year mortality in patients with diabetes achieved slightly lower but comparable performance (AUC 0.78-0.80), despite being based solely on administrative data without laboratory data or medications use. However, the Chinese study was built on a relatively small set of 451 patients, whereas our model was trained on a larger nationwide cohort of 472,950 patients, offering generalizability.

Similar evidence exists for chronic viral hepatitis. In a study from Sultan Qaboos University Hospital (SQUH), LR models incorporating laboratory markers, genotype, and coinfections reached an AUC of 0.93 for one-year mortality prediction of patients with chronic viral hepatitis C [28]. Our Kazakhstan hepatitis models showed slightly lower results (AUC 0.74-0.80), which is expected given the absence of laboratory and treatment data. Similar to the Chinese study on diabetes, the study from SQUH was a single-center study with 702 patients, whereas our study on hepatitis was developed using large-scale administrative data, enhancing generalizability.

Limitations

This methodological note has several limitations. Results suggest that augmenting administrative data with laboratory measures and treatment information could potentially improve predictive performance. However, augmenting administrative data by laboratory data and clinical notes could introduce new statistical challenges due to high-dimensionality of data [11]. Therefore, the utility of appropriate feature selection techniques would seem crucial to mitigate those challenges. Another limitation is that suggested one year prediction model also may mediated and/or affected by unexpected life-threatening events like sudden death, CV-events, injuries and may differ than natural disease progression. Although, we considered the missing data and applied median imputation for numerical variables and mode imputations for categorical data, these simple approaches may not fully capture the underlying data patterns, and more advanced methods, such as multivariate imputation or k-nearest neighbors imputation, could provide greater robustness. Despite these limitations, the framework itself is designed to be scalable, reproducible, and adaptable, while its performance will need to be confirmed in enriched datasets across additional chronic diseases and diverse populations beyond Kazakhstan.

Conclusion

We propose a scalable and adaptable framework for predicting one-year mortality in chronic disease patients using large-scale administrative data. This approach introduces subcohort definition, model development, and comprehensive performance evaluation, ensuring both methodological robustness and clinical relevance. The inclusion of explainability via SHAP analysis helps healthcare professionals to understand not only which factors influence predictions but also the direction of their impact. Applied to real-world datasets in Kazakhstan, the framework achieved fair-to-good predictive performance, demonstrating its feasibility and reliability in practical settings.

While the framework performed well with administrative data alone, its design allows for integration with enriched datasets, such as laboratory results and clinical notes, which could potentially improve predictive accuracy. Future applications should incorporate feature selection techniques when handling high-dimensional data to maintain computational efficiency and generalizability. In addition, more advanced imputation techniques for handling missing data need to be considered in the future. Lastly, country-specific data quality highlights the need for external validation and application across diverse populations and disease groups. Overall, this methodology provides healthcare systems with a practical and interpretable tool for early mortality risk identification, supporting better-informed clinical decisions and targeted interventions.

Acknowledgements

Author Contributions: Conceptualization, I. A., A. A., and A. G.; methodology, I. A., A. A., T. S., T. A., and A. G.; validation, T. M. and A. O.; writing – original draft preparation, I. A. and A. A.; writing – review and editing, T. S., T. A, T. M., and A. O.; supervision, A. G. All authors have read and agreed to the published version of the manuscript.

Disclosures: There is no conflict of interest for all authors.

Funding: This study was conducted under the grant support from MSHE and under the consortium with Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan (Grant title: Development of innovative technologies and creation of modern infrastructure for the sustainable development of the South Kazakhstan region. Grant number: BR24992814).

References

1. World Health Organization. Noncommunicable diseases. 2024. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (accessed Aug 19, 2025).
2. Schwartz L, Anteby R, Klang E, Soffer S. Stroke mortality prediction using machine learning:

Systematic review. J Neurol Sci. 2023;444:120529. doi: 10.1016/j.jns.2022.120529

3. Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetol Metab Syndr. 2021;13(1):148. doi: 10.1186/s13098-021-00767-9
4. Alimbayev A, Zhakhina G, Gusmanov A, Sakko Y, Yerdessov S, Arupzhanov I, et al. Predicting 1-year mortality of patients with diabetes mellitus in Kazakhstan based on administrative health data using machine learning. Sci Rep. 2023;13(1):8427. doi: 10.1038/s41598-023-35551-4
5. Arupzhanov I, Syssoyev D, Alimbayev A, Zhakhina G, Sakko Y, Yerdessov S, et al. One-year mortality prediction of patients with hepatitis in Kazakhstan based on Administrative Health Data: A machine learning approach. Electron J Gen Med. 2024;21(6):em15747. doi: 10.29333/ejgm/15747
6. Gusmanov A, Zhakhina G, Yerdessov S, Sakko Y, Mussina K, Alimbayev A, et al. Review of the research databases on population-based registries of Unified Electronic Healthcare System of kazakhstan (UNEHS): Possibilities and limitations for epidemiological research and real-world evidence. Int J Med Inform. 2023;170:104950. doi: 10.1016/j.ijmedinf.2022.104950
7. National Institute of Mental Health. Understanding the Link Between Chronic Disease and Depression. https://www.nimh.nih.gov/health/publications/chronic-illness-mental-health (accessed Aug 19, 2025).
8. U.S. Department of Health & Human Services. Chronic conditions. 2019. https://www.hhs.gov/guidance/document/chronic-conditions (accessed Aug 19, 2025).
9. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. New York: Springer; 2009. doi: 10.1007/978-0-387-84858-7
10. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. doi: 10.1007/bf00994018
11. Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. Hoboken: John Wiley & Sons; 2001.
12. Anderson TW. Classification by multivariate analysis. Psychometrika. 1951;16(1):31–50. doi: 10.1007/bf02313425
13. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi: 10.1023/a:1010933404324
14. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. p. 3146-54.
15. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94. doi: 10.1145/2939672.2939785
16. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39. doi: 10.1006/jcss.1997.1504
17. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232. doi: 10.1214/aos/1013203451
18. Soladoye AA, Aderinto N, Popoola MR, Adeyanju IA, Osonuga A, Olawade DB. Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps. Int J Med Inform. 2025;203:106041. doi: 10.1016/j.ijmedinf.2025.106041

19. Tan KR, Seng JJ, Kwan YH, Chen YJ, Zainudin SB, Loh DH, et al. Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review. J Diabetes Sci Technol. 2023;17(2):474–89. doi: 10.1177/19322968211056917
20. Moulaei K, Sharifi H, Bahaadinbeigy K, Haghdoost AA, Nasiri N. Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis. Int J Med Inform. 2023;179:105243. doi: 10.1016/j.ijmedinf.2023.105243
21. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017.
22. Pines JM, Carpenter CR, Raja AS, Schuur JD. Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules. 2nd ed. Chichester: Wiley-Blackwell; 2013. doi: 10.1002/9781118482117
23. Tang O, Matsushita K, Coresh J, Sharrett AR, McEvoy JW, Windham BG, et al. Mortality implications of prediabetes and diabetes in older adults. Diabetes Care. 2020;43(2):382–8. doi: 10.2337/dc19-1221
24. Röckl S, Brinks R, Baumert J, Paprott R, Du Y, Heidemann C, et al. All-cause mortality in adults with and without type 2 diabetes: Findings from the National Health Monitoring in Germany. BMJ Open Diabetes Res Care. 2017;5(1):e000451. doi: 10.1136/bmjdrc-2017-000451
25. Bollerup S, Hallager S, Engsig F, Mocroft A, Krarup H, Madsen LG, et al. Mortality and cause of death in persons with chronic hepatitis B virus infection versus healthy persons from the general population in Denmark. J Viral Hepat. 2022;29(9):727–36. doi: 10.1111/jvh.13713
26. Montuclard C, Hamza S, Rollot F, Evrard P, Faivre J, Hillon P, et al. Causes of death in people with chronic HBV infection: A population-based Cohort Study. J Hepatol. 2015;62(6):1265–71. doi: 10.1016/j.jhep.2015.01.020
27. Li Y, Guan L, Ning C, Zhang P, Zhao Y, Liu Q, et al. Machine learning-based models to predict one-year mortality among Chinese older patients with coronary artery disease combined with impaired glucose tolerance or diabetes mellitus. Cardiovasc Diabetol. 2023;22(1):138. doi: 10.1186/s12933-023-01854-z
28. Al Alawi AM, Al Shuaili HH, Al-Naamani K, Al Naamani Z, Al-Busafi SA. A machine learning-based mortality prediction model for patients with chronic hepatitis C infection: An exploratory study. J Clin Med. 2024;13(10):2939. doi: 10.3390/jcm13102939