computation

MDPI

Article

Machine Learning for Regional Tuberculosis Analysis in Ukraine: Regression Modeling and Clustering

Olena Kyrylenko [1], Kseniia Bazilevych [1], Yurii Parfeniuk [2], Ievgen Meniailov [2], Tetyana Chumachenko [3] and Dmytro Chumachenko [1,4,5,*]

[1] Mathematical Modelling and Artificial Intelligence Department, National Aerospace University "Kharkiv Aviation Institute", 61070 Kharkiv, Ukraine; o.y.kyrylenko@student.khai.edu (O.K.); ksenia.bazilevich@gmail.com (K.B.)
[2] V. N. Karazin Kharkiv National University, 61022 Kharkiv, Ukraine; parfeniuk@karazin.ua (Y.P.); evgenii.menyailov@gmail.com (I.M.)
[3] Epidemiology Department, Kharkiv National Medical University, 61022 Kharkiv, Ukraine; tatalchum@gmail.com
[4] Center for Digital Public Health in Emergencies, University College London, London WC1E 6BT, UK
[5] Isaac Newton Institute for Mathematical Sciences, University of Cambridge, Cambridge CB2 1TN, UK
* Correspondence: dichumachenko@gmail.com

Abstract

Regional tuberculosis (TB) surveillance requires analytical tools that distinguish outcome definitions, preserve temporal ordering, and characterize geographic heterogeneity. This study developed an information system integrating regression modeling and regional clustering for Ukrainian surveillance indicators. The primary predictive outcome was the quarterly regional count of patients whose multidrug-resistant or rifampicin-resistant TB (MDR/RR-TB) diagnosis was laboratory confirmed for the first time during the reporting quarter. Predictive performance was evaluated using rolling-origin one-quarter-ahead temporal validation, with Random Forest, k-nearest-neighbor regression, Gradient Boosting, a voting ensemble, and simple operational benchmarks compared on the same temporally held-out observations. Several leading approaches showed similar predictive performance, and paired region-level tests did not establish statistically significant superiority of a single model after correction for multiple comparisons. K-means clustering was used as a descriptive representation of regional heterogeneity; the selected three-group representation was stable for the illustrative 2021 analysis, although internal validation criteria did not uniquely identify three clusters as optimal. The study provides a reproducible workflow for temporally validated comparison of recorded surveillance counts together with descriptive regional grouping rather than a new learning algorithm or a causal model of TB transmission.

Keywords: epidemic model; tuberculosis; MDR/RR-TB; machine learning; temporal validation; Random Forest; KNN; clustering; surveillance; Ukraine

Check for updates

Academic Editor: Pier Luigi Martelli

Received: 16 September 2026
Revised: 23 October 2026
Accepted: 27 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Tuberculosis (TB) remains a major challenge for public health surveillance and health service planning. An estimated 10.7 million people developed TB worldwide in 2024, and the disease caused more than 1.2 million deaths [1]. Effective responses require information on how the recorded burden varies across regions and over time, as well as national estimates. Such differences are consequential for planning diagnostic capacity, treatment provision, and follow-up services. In a nationwide analysis of pulmonary TB notifications across

2851 Chinese counties, Hu et al. identified substantial geographic inequalities despite an overall decline in notifications [2]. Regional analysis is therefore necessary to determine whether national trends reflect the conditions experienced by individual administrative areas.

In Ukraine, this analytical need is intensified by the interaction between a historically high TB burden, the disruption associated with COVID-19, and Russia's full-scale invasion in February 2022. Using national surveillance reports for 2013–2023, Krokva et al. documented divergent regional patterns, including increased recorded incidence in central Ukraine and decreased rates in eastern Ukraine during 2022–2023 [3]. They linked the interpretation of these patterns to population displacement, disrupted surveillance, and difficulties obtaining data from occupied territories. Consequently, a decline in recorded cases in an affected region cannot, by itself, establish a reduction in disease burden. These conditions make Ukraine a challenging setting for computational analysis because changes in observations may reflect both epidemiological processes and shifts in the capacity to detect and report disease.

The interpretation of model outputs also depends on how the prediction target is defined. Incidence concerns new disease episodes during a specified interval, whereas prevalence concerns existing disease in a population. Notifications represent cases recorded by surveillance and are influenced by diagnostic access and reporting completeness [1,2]. Counts of newly confirmed multidrug-resistant or rifampicin-resistant TB additionally refer to a specific case category. For computational modeling, these outcomes differ in their units, denominators, and temporal meaning. A model trained on routinely reported indicators estimates patterns in those observations. Its predictions require additional evidence before they can be interpreted as estimates of the underlying, partly unobserved burden. This distinction is particularly relevant when comparing regional records across periods of disrupted health services [3].

Forecasting research offers several approaches to extracting predictive information from TB surveillance data. Autoregressive integrated moving average (ARIMA) models and their seasonal variants describe dependence on previous observations, while machine learning models can represent more flexible relationships between predictors and outcomes. In their review of TB incidence forecasting, Maipan-Uku and Cavus identified extensive use of conventional time series models and several neural network architectures, while highlighting limitations in predictive accuracy and geographic coverage [4]. The available literature consequently supports comparative evaluation in the intended-use setting. The suitability of a method depends on the temporal resolution, length, and completeness of the available records, as well as the outcome and prediction horizon.

Recent studies illustrate the range of modeling strategies and the specificity of their findings. Abade et al. compared statistical, machine learning, and deep learning methods for monthly TB and HIV coinfection series in Mato Grosso, Brazil, using the final 24 observations for validation. Bidirectional long short-term memory (LSTM) and combined convolutional and recurrent architectures performed best across the analyzed population strata [5]. Barman et al. subsequently developed an epidemic-guided deep learning framework that combines a networked transmission model with neural forecasting, evaluating it on regional TB data from Japan and mainland China and incorporating uncertainty quantification [6]. These studies demonstrate the potential of flexible and hybrid models, but their results do not establish which approach is appropriate for Ukraine's available regional indicators. Differences in surveillance coverage and data structure require an independent empirical assessment.

Model comparison must also distinguish agreement with historical observations from the ability to predict later observations. Hewamalage et al. showed how inappropriate data

partitioning, information leakage, unsuitable benchmarks, and the choice of error measures can produce misleading conclusions about forecasting performance [7]. A high coefficient of determination alone is therefore insufficient evidence of operational forecast accuracy. Absolute error measures describe the magnitude of discrepancies, while evaluation on observations unavailable during model development is necessary to assess temporal generalization. These considerations are especially relevant to short regional series spanning major disruptions, where a model may reproduce differences between regions without adequately anticipating subsequent changes. They motivate explicit attention to the scope of the evaluation when interpreting comparisons among regression models.

Regional grouping can complement predictive modeling by summarising similarities among administrative areas. Regression estimates the value of a specified indicator, whereas clustering organizes regions according to their observed profiles. Such groupings can help analysts examine variation obscured by a single national average; their interpretation depends on the selected indicators and the completeness of the records [2]. Integrating these functions into an information system also enables inspection of observed and predicted values, comparison of model errors, and exploration of regional patterns within the same analytical workflow. This practical emphasis is consistent with the broader perspective of Rodriguez et al., who situate data sources, modeling, and forecast use within an interconnected epidemic forecasting process [8].

The research problem addressed here is the integration and reproducible evaluation of complementary analytical functions for routinely reported TB statistics in Ukraine. Existing work demonstrates both regional heterogeneity and a broad range of forecasting methods, but differences in outcome definitions, surveillance coverage, and evaluation design prevent a universal ranking of models. The modeled outcome in the predictive analysis is therefore defined narrowly as a recorded quarterly MDR/RR-TB case count. It should not be interpreted as a direct measure of transmission, all incident disease, or individual risk.

Accordingly, this study evaluates an information system for regional TB surveillance using records from the Public Health Center of the Ministry of Health of Ukraine. The application-specific scientific contribution is a reproducible workflow that separates population-standardized TB rates from MDR/RR-TB case counts, evaluates conventional regressors under leakage-resistant rolling-origin temporal validation, compares them with simple operational benchmarks, and combines the resulting model assessment with descriptive regional clustering and visualization in one information-system workflow. The study does not propose a new learning algorithm; its added value is the consistent treatment of outcome definitions, temporal evaluation, benchmarking, and regional heterogeneity using the same Ukrainian surveillance records.

2. Current Research Analysis

Computational research on TB addresses distinct problems, including patient-level diagnosis, treatment outcome prediction, estimation of population burden, and forecasting of routinely reported indicators. These tasks differ in their observational units and evaluation requirements. For a regional surveillance information system, the most relevant evidence concerns the prediction of aggregate case counts or rates, the representation of geographic heterogeneity, and the reproducibility of the analytical workflow. The literature reviewed below is therefore assessed in relation to these functions, with particular attention to whether reported improvements reflect historical model fit, temporally held-out prediction, or descriptive spatial analysis.

2.1. Statistical and Hybrid Time Series Forecasting

Seasonal statistical models remain relevant comparators for TB forecasting because they explicitly represent temporal dependence and recurring variation in surveillance records. Kalizhanova et al. analyzed monthly TB data from Kazakhstan for 2014–2019 using seasonal autoregressive integrated moving average (SARIMA) and susceptible-infected-recovered (SIR) models [9]. They reported better predictive performance of SARIMA in both bacteriologically confirmed and unconfirmed categories of respiratory TB. However, the SARIMA forecasts concerned 2019, whereas the SIR evaluation covered 2018–2019. The reported difference should therefore be interpreted with attention to the unequal evaluation windows. Moreover, the performance of a basic SIR specification does not establish the limitations of mechanistic TB models in general, particularly models that explicitly represent latent infection and more detailed disease progression.

Hybrid approaches attempt to combine complementary representations of the same time series. Siamba et al. examined monthly notifications among children younger than 15 years in Homa Bay and Turkana Counties, Kenya, using records from 2012–2021 [10]. Their approach combined an ARIMA component with an artificial neural network fitted to its residuals and reported lower errors than the standalone seasonal ARIMA model. A rolling one-step procedure was used in selecting the ARIMA model. This study provides an example of residual modeling in a surveillance setting with constrained data availability. Nevertheless, its forecasts concern recorded childhood TB cases in two counties, and they cannot be transferred directly to all-age incidence, prevalence, or drug-resistant TB notifications in another country.

Signal decomposition provides another form of hybridization. Zhao et al. used monthly pulmonary TB notification rates in China from 2008–2018 to develop a model that combines empirical mode decomposition (EMD), autoregressive moving average (ARMA) models, and LSTM networks [11]. Models were fitted using 2008–2017 observations and evaluated over three-, six-, nine-, and twelve-month horizons in 2018. The EMD-ARMA-LSTM configuration produced lower reported errors than the comparison models. Its methodological relevance lies in assigning different forecasting components to different temporal structures. For operational use, however, decomposition must be performed using only information available at each forecast origin. Decomposing the complete series before evaluation can allow later observations to influence the representation of earlier data. Thus, a hybrid architecture requires evaluation of the complete preprocessing and prediction pipeline.

2.2. Machine Learning and Multivariate Prediction

The choice between conventional regressors and neural networks cannot be resolved by model complexity alone. In a benchmark of 45 tabular datasets, Grinsztajn et al. found that tree-based methods, including Random Forest and gradient boosting, remained strong competitors to the neural architectures evaluated, even before accounting for computational cost [12]. These findings support the inclusion of tree-based baselines when surveillance records are represented as structured tables. More recent work also shows that this ranking is not fixed. Hollmann et al. introduced TabPFN, a foundation model pretrained on synthetic tabular tasks, and reported improved classification and regression performance on small and medium-sized datasets [13]. Neither benchmark establishes performance on the Ukrainian TB time series. Together, they support empirical model selection rather than a general presumption that either trees or deep learning must be superior.

For regional surveillance, the representation supplied to a regressor is as important as its model family. Random Forest can capture interactions among included variables, but temporal dependence must be represented through suitable predictors. Similarly, KNN regression estimates outcomes from observations that are close under a specified

distance measure. Its behavior depends on feature scaling, the number of neighbors, and whether proximity in the selected feature space corresponds to comparable epidemiological conditions. Consequently, a lower ranking for a particular KNN configuration cannot, by itself, establish that the method is generally unsuitable for TB forecasting. Ensemble methods likewise require specification of their component learners and aggregation rule, and Random Forest is itself an ensemble, so the label alone does not identify a distinct algorithm.

TB-specific multivariate studies demonstrate how predictor selection interacts with temporal modeling. Yang et al. compared ARIMA, SARIMA, multivariate LSTM, and hybrid ARIMA-LSTM models for Liaoning Province, China, evaluating six-, twelve-, and twenty-four-month forecasting horizons [14]. They used LASSO and Random Forest for feature selection and applied SHapley Additive exPlanations (SHAP) to inspect model predictions. The preferred configuration varied by horizon: a multivariate LSTM performed well in shorter evaluations, while a hybrid model performed better in longer evaluations. This finding argues against reporting a single, context-independent ranking of forecasting methods. It also distinguishes the role of Random Forest as a feature selection component from its role as the final predictor. SHAP attributions describe the fitted model's dependence on its inputs, but they do not independently identify causal determinants of TB transmission.

Attention-based models provide a further point of comparison. Lv et al. analyzed 144 monthly observations from Changde City, China, with 2010–2019 used for training, 2020 for validation, and 2021 for testing [15]. Their self-attention model achieved a test-set RMSE approximately 7.41% lower than that of LSTM and 28.86% lower than that of ARIMA. The explicit chronological partition makes the setting of this comparison identifiable. Nevertheless, the final evaluation covers 12 observations from a single city and does not establish performance across regions or following other forms of surveillance disruption. The result, therefore, supports self-attention as a candidate forecasting approach, while leaving its relative value for shorter, differently aggregated Ukrainian records to be tested.

2.3. Spatial Modeling and Regional Stratification

Models that treat each geographic series independently may omit information shared between neighboring areas. Hao et al. developed a Spatial-CNN-Transformer-LSTM-Attention model that incorporated national coordinates, neighboring countries' TB incidence, and socioeconomic, demographic, and environmental variables [16]. Their framework produced multistep projections for 163 countries for 2024–2026. This approach extends forecasting beyond a single series by representing spatial information alongside temporal predictors. However, national projections and subnational notification forecasts address different scales. For prospective application, the availability of neighboring incidence values and other covariates must also be established at each prediction horizon. Incorporating a variable associated with TB does not make its future value observable, nor does predictive importance establish a causal mechanism.

Spatial regression addresses geographic variation in associations rather than only variation in future outcomes. Ma and Fan applied multiscale geographically weighted regression (MGWR) to TB indicators and potential explanatory variables in mainland China during 2014–2018 [17]. They reported an R^2 of 0.932 for MGWR, compared with 0.836 for geographically weighted regression and 0.466 for ordinary least squares. These are measures of model fit in the reported spatial analysis, not evidence that MGWR provides more accurate temporal forecasts. The study nevertheless demonstrates the value of allowing relationships to vary geographically and at different spatial scales. For Ukrainian regional analysis, this supports examining whether a pooled model conceals systematic differences between

administrative areas, without assuming that ecological associations describe individual level risk.

Grouping and visualization can make geographic heterogeneity easier to interpret. Lei et al. used a Geotree representation and multilevel modeling to examine TB incidence trajectories and socioeconomic development across 173 countries and territories [18]. Their framework organized geographic units by country type and development stage, making the grouping criteria explicit. Its relevance to an information system lies in its ability to connect a visual representation to a defined analytical structure. Such classifications differ from unsupervised partitions derived solely from observed TB indicators. A group defined by similar notification levels, for example, should not automatically be interpreted as sharing the same underlying socioeconomic conditions or disease transmission processes.

The spatial distribution of diagnostic access provides an additional layer of interpretation. Kunjok et al. studied 222 newly diagnosed, bacteriologically confirmed TB patients attending three facilities in Nairobi County, Kenya [19]. Their analysis combined spatial autocorrelation, local hotspot methods, and geographic assessments of access to diagnostic services. The reported concentration of diagnostic delays in particular areas illustrates why a geographic pattern in surveillance data may partly reflect the conditions under which cases are detected. Because the study recruited diagnosed patients at selected facilities, it does not directly measure the distribution of undiagnosed TB in the wider population. This distinction is important when using reported regional indicators to identify priorities for further investigation.

Unsupervised clustering also requires its own validation criteria. Schubert showed that the elbow heuristic for selecting the number of k-means clusters can produce misleading partitions, including apparent elbows in data without a meaningful cluster structure [20]. Alternative internal criteria, such as silhouette, Calinski–Harabasz, or Davies–Bouldin measures, address different aspects of compactness and separation and should be interpreted in light of the analytical purpose. When applied to regional TB records, the choice of scaling determines whether a partition primarily reflects differences in absolute burden or in temporal profiles. Furthermore, independently fitted annual partitions have arbitrary numeric labels, and changes in those labels cannot establish epidemiological transitions without matching the groups across years. A three group solution is therefore a descriptive representation whose stability and substantive meaning require examination.

2.4. Data Quality and Predictive Evaluation

The credibility of a model comparison depends on the full analytical workflow. The REFORMS recommendations developed by Kapoor et al. address study design, computational reproducibility, data quality, preprocessing, modeling, leakage, uncertainty, and generalizability [21]. For regional TB forecasting, their relevance is that transformations learned from data, including imputation, scaling, and feature selection, must be separated from the observations used for final evaluation. Reporting the algorithm name and aggregate errors is insufficient to reconstruct a result when the target, input features, hyperparameters, and evaluation observations are unspecified. These requirements apply equally to simple regression methods and more elaborate neural or hybrid systems.

Temporal and geographic generalization should also be distinguished. Predicting later observations for regions represented during training differs from predicting an entirely unobserved region. A random split of regional records can address neither objective adequately if closely related observations appear in both partitions. Chronological holdouts, as illustrated by Lv et al. [15], make the temporal boundary explicit, while evaluation across multiple forecast origins examines sensitivity to the selected period. Geographic transfer requires an additional design that includes relevant locations. For records spanning

substantial changes in population coverage or reporting practice, performance should also be examined by period; an overall error can conceal a model's failure during the interval most relevant to its intended use.

Performance measures answer different questions. MAE measures average absolute discrepancy in the target's units, whereas RMSE gives greater weight to larger errors. R^2 depends on the variability in the evaluation outcomes and can remain high when between-region differences are large. Accordingly, scores cannot be compared directly across case counts, population standardized rates, or different case definitions. Nor does a lower average error establish statistically reliable superiority when the evaluation set is small. The literature's use of different horizons, outcomes, and aggregation levels therefore precludes a defensible numerical ranking across these studies. A comparison within a single system is most informative when all methods are evaluated on the same observations and target, with uncertainty and region-specific errors reported where the data support them.

2.5. Implications for the Present Study

The reviewed evidence supports complementary analytical functions rather than a single universally preferred TB model. Time series and hybrid forecasting studies are retained as methodological context because they illustrate chronological validation, leakage control, forecast horizons, and benchmark selection; they are not presented as models implemented in this study. Cross-country findings are used qualitatively to motivate analytical choices, while interpretation of Ukrainian regional variation is anchored primarily in Ukrainian surveillance conditions.

Against this background, the present study compares established tabular-regression strategies and simple benchmarks for a clearly defined quarterly MDR/RR-TB surveillance count, while using K-means only as a descriptive representation of regional heterogeneity. This positioning avoids treating results from different countries, outcomes, units, or temporal resolutions as directly comparable and does not claim algorithmic novelty.

3. Background

TB surveillance in Ukraine during 2019–2023 reflects substantial regional heterogeneity alongside changes in the conditions under which cases were diagnosed and reported. The Public Health Center of the Ministry of Health of Ukraine provides national and regional statistics on notified diseases, active TB prevalence, and population subgroups [22]. These observations span the COVID-19 pandemic and the first two years following Russia's full-scale invasion in February 2022. Their interpretation, therefore, requires attention to diagnostic access and continuity of care, as well as to changes in the population covered by surveillance. Recent evidence on drug-resistant TB in conflict settings identifies displacement, disrupted diagnostic services, and interruptions to treatment delivery as important constraints on disease control [23].

The epidemiological indicators used in this context describe different quantities. Incident disease concerns new disease episodes arising over a specified period, whereas notifications record cases detected and reported to surveillance. The Ukrainian indicator conventionally described as TB incidence includes notified new and relapse cases and should not be equated with an estimate of all incident disease, including undiagnosed cases. In 2023, this notification rate was 48.4 per 100,000 population, compared with 45.1 in 2022, corresponding to a reported increase of 7.3% [22]. By contrast, active TB prevalence describes existing active disease within a defined population and reference period. It is distinct from both annual notifications and the prevalence of infection without active disease.

National averages conceal marked geographic differences. In 2023, the new and relapse TB rates reached 99.6 per 100,000 population in Odesa Oblast, 92.1 per 100,000 in Kirovohrad Oblast, and 86.1 per 100,000 in Dnipropetrovsk Oblast, compared with 15.2 per 100,000 in Kherson Oblast and 16.1 per 100,000 in Donetsk Oblast [22]. Recorded active TB prevalence declined nationally from 69.4 per 100,000 in 2019 to 40.3 in 2023. Nevertheless, the 2023 prevalence remained substantially higher in Odesa, Kirovohrad, and Dnipropetrovsk oblasts, at 112.2, 79.3, and 66.3 per 100,000, respectively [22]. The decline in recorded prevalence and the increase in annual notifications are not inherently contradictory, because these indicators measure different aspects of disease occurrence and duration. The prevalence of all forms of TB in 2019–2023 is presented in Figure 1.

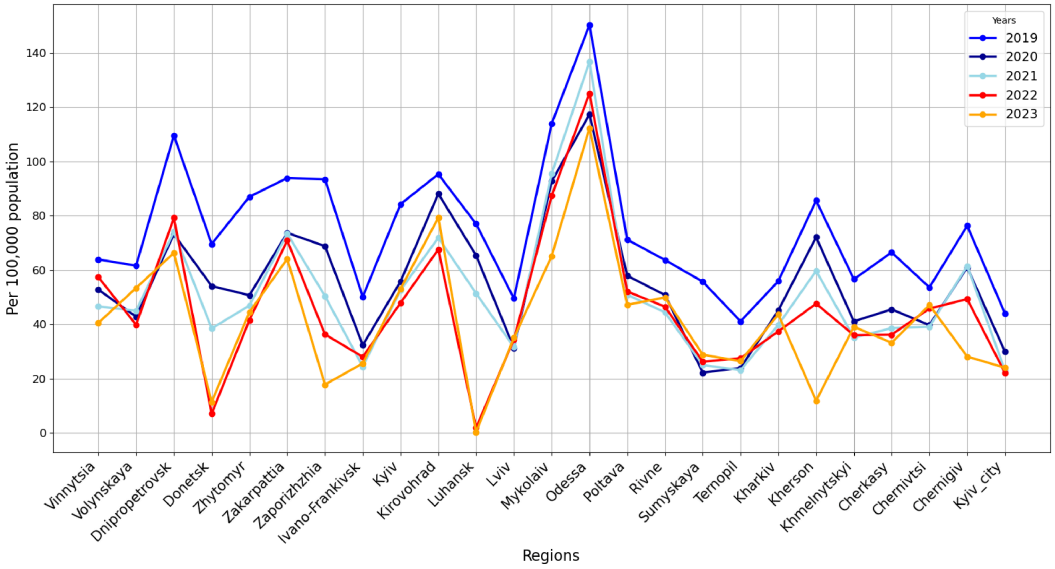

Figure 1. Prevalence of all forms of TB in 2019–2023.

Changes in recorded indicators cannot independently establish changes in transmission or the effectiveness of control measures. Lower notification or prevalence values in territories affected by occupation and hostilities may partly reflect restricted access to testing, population departure, or incomplete reporting. Displacement can also change the relationship between the population represented by recorded cases and the denominator used to calculate a regional rate. The importance of this distinction is illustrated by European surveillance. Stoycheva et al. reported an almost fourfold increase in TB notifications among people of Ukrainian origin in the European Union and European Economic Area in 2022 relative to the 2019–2021 annual average, without a corresponding increase in the calculated population-specific notification rate [24]. Accordingly, regional comparisons within Ukraine require explicit geographic coverage and consistent definitions of the numerator and denominator.

Pediatric TB warrants separate attention because its detection depends strongly on age-appropriate diagnostic pathways. National surveillance recorded 639 new and relapse cases among children aged 0–14 years in 2023, compared with 450 in 2022 [22]. This increase should be considered alongside changes in ascertainment. Diuzheva et al. described the introduction of stool-based molecular testing at 12 regional TB facilities in Ukraine from November 2021 to September 2022 [25]. Stool was the only specimen providing bacteriological confirmation of Mycobacterium tuberculosis in seven participants. The descriptive pediatric series shown in Figure 2 concerns the five regions with the highest recorded pediatric rates over 2019–2023 (Kirovohrad, Zaporizhzhia, Cherkasy, Dnipropetrovsk, and

Odesa). These pediatric data are contextual only and are not predictors or outcomes in the MDR/RR-TB regression analysis.

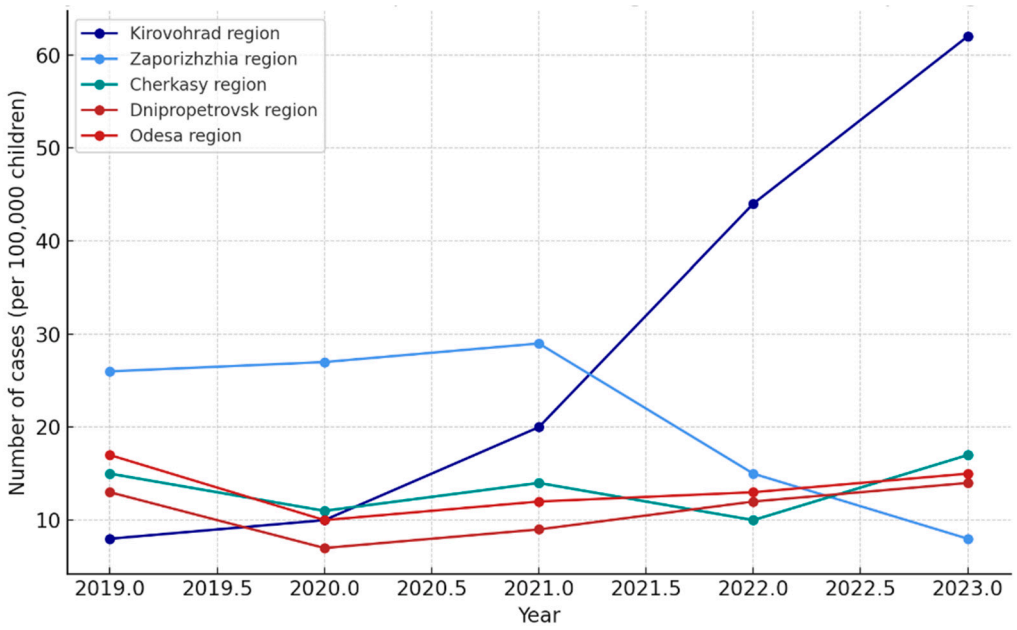

Figure 2. Recorded pediatric TB rates per 100,000 children in the five regions with the highest values over 2019–2023. These data are presented as descriptive context and were not used as model inputs.

The social distribution of notified cases provides descriptive context for service planning. Figure 3 shows the composition of newly diagnosed cases by recorded social group in 2023. These proportions are not population-specific risks and were not used as predictors in the regression or clustering analyses. Similarly, HIV coinfection and psychosocial factors discussed in the cited Ukrainian evidence are relevant to public health interpretation but are not operationalized in the supplied modeling dataset [26].

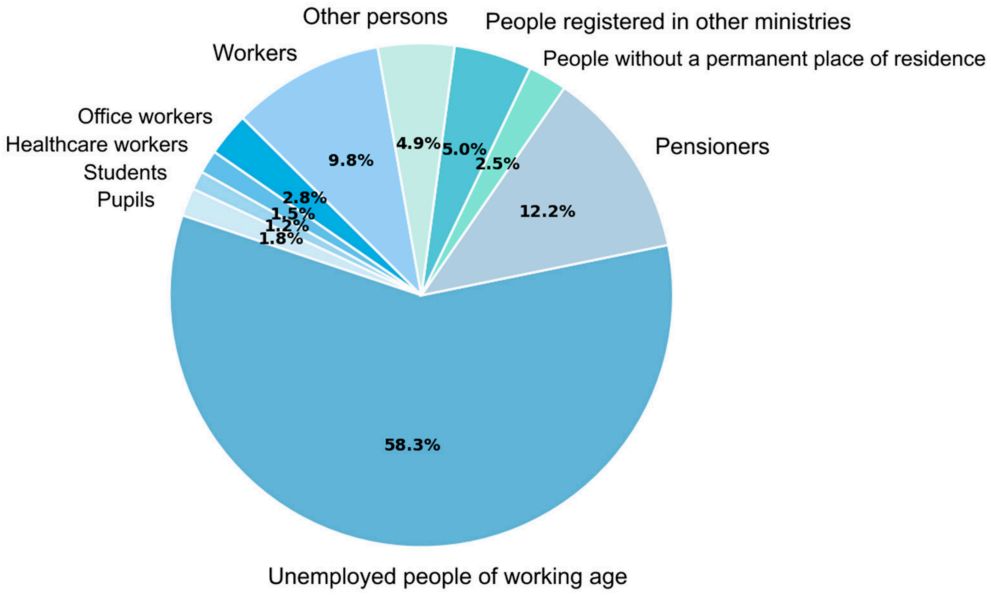

Figure 3. Distribution of newly diagnosed TB cases by recorded social group in 2023. Social-group variables were not included in the predictive or clustering models.

Drug resistance adds a clinically important dimension that aggregate all-form TB indicators do not capture. Multidrug-resistant TB denotes resistance to at least isoniazid and rifampicin, whereas rifampicin-resistant TB is defined by resistance to rifampicin, with or without resistance to other agents [23]. In Kharkiv, Konstantynovska et al. reported additional levofloxacin resistance in 23.1–31.2% of patients with multidrug-resistant or rifampicin-resistant TB during 2019–2023 [27]. These findings concern a defined drug-resistant subgroup in one region and should not be extrapolated to all notified TB cases nationally. They demonstrate why the number of recorded cases alone is insufficient to characterize treatment needs and why drug susceptibility categories must remain distinguishable in surveillance analyses.

Despite these pressures, Ukrainian studies also document the feasibility of maintaining effective treatment services with appropriate support. A prospective implementation cohort in Zhytomyr enrolled 300 patients with rifampicin-resistant or multidrug-resistant TB during 2019–2021 and evaluated all-oral regimens alongside multidisciplinary care [28]. Treatment success was achieved in 234 patients (78%), although post-treatment deaths and losses to follow-up remained important. As a single-arm study, it does not isolate the contribution of individual service components. It nevertheless shows that population level trends should be interpreted alongside evidence on diagnosis, treatment, and retention in care, rather than attributed solely to changes in transmission.

This epidemiological context motivates a regional analysis that preserves the meaning and limitations of the underlying observations. Annual notification rates, recorded active TB prevalence, pediatric indicators, social-group descriptions, and drug-resistant case counts serve different analytical purposes and are not pooled into one predictive target. The predictive analysis concerns quarterly MDR/RR-TB surveillance counts only; prevalence and other indicators are retained as descriptive epidemiological context.

4. Materials and Methods

A retrospective study was conducted to develop and evaluate an information system for regional TB surveillance in Ukraine. The analytical workflow combines descriptive examination of surveillance statistics, one-quarter-ahead regression of quarterly MDR/RR-TB counts, and unsupervised regional grouping. The historical MDR/RR-TB dataset covers 2018–2023. Predictive performance is evaluated using temporally held-out rolling-origin test quarters from Q1 2020 through Q4 2023, while the clustering analysis uses annual regional summaries derived from 2018–2023 quarterly counts.

The information system architecture is presented in Figure 4. Figure 4 summarizes the system logic from data preparation to analytical output. After harmonization, explicit outcome definition, and period selection, the regression branch constructs lag features and evaluates predictive models together with operational benchmarks using rolling-origin temporal validation. In parallel, the clustering branch aggregates annual regional values, standardizes the input, and evaluates K-means solutions using internal validation indices and stability analysis. The two branches remain analytically distinct and converge only at the visualization and reporting layer, where model performance summaries and regional groupings are presented within the same information-system workflow.

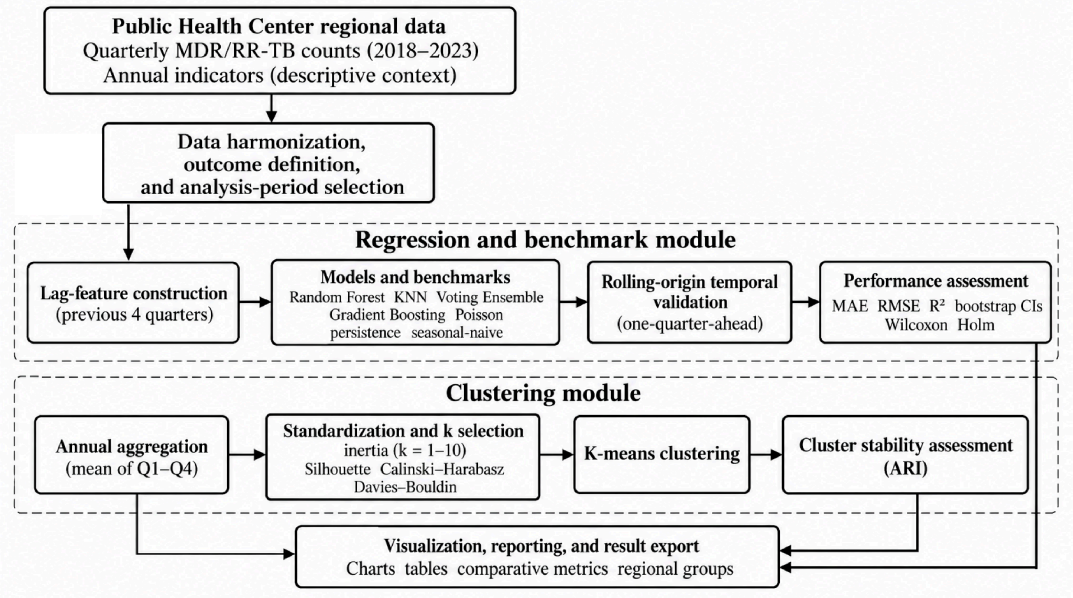

Figure 4. Information system architecture and analytical workflow. Regional surveillance data are harmonized before being routed to separate regression/benchmark and clustering modules; the two branches converge in a shared visualization, reporting, and result-export layer.

4.1. Data Sources and Epidemiological Variables

The study materials were publicly available statistical reports and regional tables published by the Public Health Center of the Ministry of Health of Ukraine [22,29]. The quarterly modeling file records patients whose MDR/RR-TB diagnosis was laboratory-confirmed for the first time in their lives during the reporting quarter, using information from the TB-04 laboratory register [29]. The modeling table contains 25 regional units (24 oblasts and Kyiv City); Crimea and Sevastopol are not represented. Records from Donetsk, Luhansk, Kherson, and Zaporizhzhia are retained when reported. Quantitative reporting-completeness indicators were not available in the supplied table.

The annual indicator conventionally termed TB incidence in national reports is a population-standardized notification rate of new and relapse cases per 100,000 population, whereas recorded active TB prevalence is also expressed per 100,000 population [22]. Neither is interchangeable with the quarterly absolute count of first laboratory-confirmed MDR/RR-TB cases. Throughout the manuscript, incidence rate is reserved for population-standardized annual rates, while MDR/RR-TB case count refers to the quarterly absolute surveillance outcome.

The primary predictive outcome was the quarterly regional MDR/RR-TB case count, and the unit of analysis was a region-quarter. For each target quarter, four numerical predictors were used: the same region's case counts in the four immediately preceding quarters (lags 1–4). Performance was evaluated using rolling-origin one-quarter-ahead validation from Q1 2020 through Q4 2023. The first development set contained 100 region-quarter observations (2019 Q1–Q4); the development set then increased by 25 observations at each forecast origin, reaching 475 observations for Q4 2023. Each test set contained 25 regions, yielding 400 out-of-sample predictions across 16 forecast origins. All preprocessing and fitting were restricted to observations available before the corresponding target quarter. A secondary 2023 summary comprised the final four test quarters ($n = 100$ predictions).

The available predictors do not directly encode passive versus active case finding, contact or risk-group screening, diagnostic delay, molecular testing coverage, treatment access, treatment success, loss to follow-up, death, or population displacement. The objective is therefore prediction and quality assessment of recorded surveillance observations rather

than causal inference about TB transmission or individual risk. A sensitivity analysis repeated the complete rolling-origin evaluation after excluding Donetsk, Luhansk, Kherson, and Zaporizhzhia from both development and test data.

4.2. Regression Models

The model set was selected to represent complementary tabular-regression strategies: a tree ensemble (Random Forest), a local distance-based method (KNN), a heterogeneous voting ensemble, and Gradient Boosting as an additional tree-based comparator. A graph convolutional network was not evaluated because the available dataset does not contain a validated graph of interregional mobility, referral, contact, or transmission links; constructing such a graph post hoc would add an unvalidated structural assumption.

4.2.1. Random Forest Regression

Random Forest combines regression trees constructed using resampled observations and randomized selection of candidate split variables [30]. Individual trees partition the predictor space into terminal regions and return a numerical estimate for each region. The forest prediction is the arithmetic mean of the tree predictions:

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x), \quad (1)$$

where B is the number of trees and $T_b(x)$ is the prediction from tree b for input vector x. Averaging allows the model to combine multiple partitions of the predictor space. The approach can represent nonlinear relationships and interactions without specifying a single global regression equation. Its empirical performance nevertheless depends on the available predictors, the tree configuration, and the evaluation design.

4.2.2. KNN Regression

KNN regression estimates an outcome using observations close to the query point in the predictor space [31]. The methodological formulation uses Euclidean distance to identify the k nearest observations. With uniform neighbor weights, the estimate is

$$\hat{f}_{KNN}(x) = \frac{1}{k}\sum_{j \in N_k(x)} y_j, \quad (2)$$

where $N_k(x)$ is the set of indices of the k nearest training observations, and y_j is the outcome of neighbor j. KNN used $k = 5$, uniform weights, and Euclidean distance. Because distance depends on input scale, StandardScaler was fitted separately within each development fold and then applied to that fold's test observations.

4.2.3. Ensemble Regression

Ensemble learning combines predictions from multiple fitted models, with its behavior depending on the component estimators and the dependence among their errors [32]. For an equally weighted regression ensemble, the aggregation rule is

$$\hat{f}_{ens}(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x), \quad (3)$$

where M denotes the number of component regressors and $f_m(x)$ their predictions. This averaging rule corresponds to the VotingRegressor formulation. The separate Ensemble model identifies an additional model configuration, rather than a methodological category that excludes Random Forest. Combining predictions does not, by itself, establish an improvement over the best component model.

Random Forest used 100 trees (random_state = 42). The VotingRegressor combined AdaBoost (100 estimators, learning_rate = 1.0, linear loss, random_state = 42), a DecisionTreeRegressor (squared-error criterion, random_state = 42), and an RBF-kernel SVR (C = 1.0, epsilon = 0.1, gamma = 'scale') with equal weights. Within the voting ensemble, only the SVR component was wrapped in a training-fold StandardScaler pipeline; the tree-based components were fitted on the original scale. GradientBoostingRegressor used 100 estimators, learning_rate = 0.1, max_depth = 3, and random_state = 42. PoissonRegressor was included as a count-oriented benchmark. Persistence (the previous-quarter value) and seasonal-naive prediction (the same quarter of the previous year) were included as simple operational benchmarks.

4.2.4. Evaluation of Regression Performance

Model-generated one-quarter-ahead estimates were compared with the corresponding held-out regional observations. Performance was summarized using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). For n observed-predicted pairs, with observed values y_i and predictions $\hat{y}_i$, MAE and RMSE are defined in Equations (4) and (5) [33]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{5}$$

MAE describes the average absolute discrepancy, whereas RMSE gives greater weight to larger errors. Both retain the units of the target; in the predictive analysis these units are MDR/RR-TB cases per quarter. R^2 describes agreement relative to the variability in the held-out observations and is not, by itself, evidence of temporal generalization.

The coefficient of determination was calculated relative to the mean of the observed values in each evaluation set [34]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{6}$$

where $\hat{y}_i$ denotes the model prediction and $\bar{y}$ denotes the mean observed value over the same evaluation observations. Uncertainty was summarized using a region-cluster bootstrap with 2000 resamples, preserving repeated quarterly observations from a sampled region together.

Paired model comparisons used two-sided Wilcoxon signed-rank tests on each region's mean absolute error across the 16 forecast origins, providing 25 paired regional observations rather than treating repeated quarters from the same region as independent. *p*-values were adjusted for multiple comparisons using the Holm method. These tests were interpreted as supportive rather than definitive because the number of geographic units is limited.

4.3. Regional Clustering

K-means clustering was used as a descriptive summary of similarity in recorded regional MDR/RR-TB counts. For the illustrative 2021 analysis, Q1–Q4 counts were averaged within each of the 25 regions to obtain one annual mean value per region. These 25 values were standardized using StandardScaler before clustering. K-means minimizes within-cluster squared Euclidean distances [35] as given in Equation (7); it does not impose geographic contiguity or identify transmission links.

$$J = \sum_{C=1}^{K}\sum_{r\in C_c}\|z_r - \mu_c\|^2, \tag{7}$$

where C_c is the set of regions assigned to cluster c, and μ_c is its centroid. The algorithm alternates between assigning each region to its nearest centroid and updating the centroid to the mean of the vectors assigned to it. The resulting partition describes similarity in the chosen numerical representation. It does not impose geographic contiguity or identify transmission links between regions.

Candidate solutions were examined for k = 1–10 using inertia. Internal validation for k = 2–10 additionally used the silhouette score, Calinski–Harabasz index, and Davies–Bouldin index. For 2021, stability of the k = 3 solution was assessed across 100 single-initialization random seeds against an n_init = 50 reference partition using the adjusted Rand index (ARI). The three-group solution is retained as an illustrative low/intermediate/high representation rather than a uniquely optimal epidemiological partition.

Cluster identifiers are arbitrary and have no intrinsic ordering. All assignments shown in Section 5.4 refer to 2021 only and should not be compared directly with 2023 notification or prevalence rankings. A low recorded value in a conflict-affected region, including Luhansk, is not interpreted as evidence of low underlying burden because surveillance access and reporting completeness are not quantified in the clustering inputs.

4.4. Information System

The information system integrates data loading, selection of an analysis period, regression and benchmark estimation, calculation of error summaries, K-means clustering, and visualization. The interface is restricted to years actually present in the MDR/RR-TB dataset (2018–2023). Users can inspect observed and one-quarter-ahead estimated values, model-comparison summaries, and regional groupings. The analytical workflow was executed in Python 3.13.5 using NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, SciPy 1.17.0, and Matplotlib 3.10.8.

The software workflow is portable to other regions, countries, and time intervals only after harmonizing outcome definitions, temporal aggregation, geographic coverage, and predictor availability. Scaling and model fitting must be repeated using the new development data, and predictive performance must be revalidated in the target setting. Geographic transfer to unseen regions would require a separate held-out geographic evaluation; the present temporal validation assesses later quarters for regions represented in development data.

5. Results

Rolling-origin validation generated 400 temporally held-out region-quarter predictions from Q1 2020 through Q4 2023. The leading point estimates were close across several models, and no single machine-learning configuration showed statistically established superiority after multiplicity correction. The clustering analysis remained a separate descriptive component and is reported for an illustrative 2021 partition.

5.1. Rolling-Origin Temporal Validation

Across all 16 one-quarter-ahead test origins, Gradient Boosting achieved MAE = 9.29, RMSE = 15.64, and R^2 = 0.741; Random Forest achieved 9.38, 15.56, and 0.744; KNN achieved 9.41, 15.22, and 0.755; persistence achieved 9.41, 15.28, and 0.753; and the Voting Ensemble achieved 9.74, 15.77, and 0.737. Seasonal-naive prediction was weaker, with MAE = 12.57, RMSE = 18.97, and R^2 = 0.619. Figure 5 shows observed versus temporally held-out Random Forest estimates across all 400 region-quarter observations.

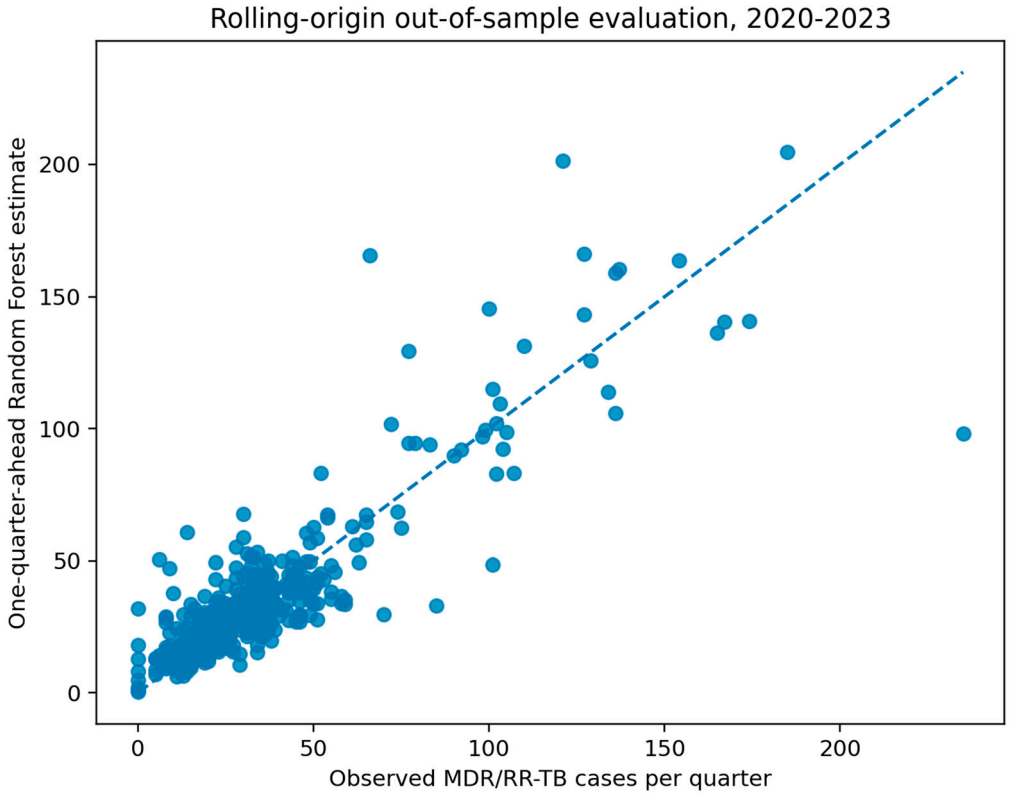

Figure 5. Observed MDR/RR-TB cases per quarter versus one-quarter-ahead Random Forest estimates under rolling-origin temporal validation, 2020–2023 (400 region-quarter test observations). The dashed diagonal indicates exact agreement.

Model ranking varied by calendar year. Persistence had the lowest MAE in 2020 (12.11), the Voting Ensemble in 2021 (7.77), Gradient Boosting in 2022 (9.00), and Gradient Boosting and Random Forest were nearly tied in 2023 (7.04 and 7.10, respectively). This variation argues against describing one algorithm as universally superior across periods.

5.2. Statistical Comparison and Sensitivity Analysis

Table 1 summarizes the primary rolling-origin results. Region-cluster bootstrap intervals were broad and overlapping. For Random Forest, the 95% bootstrap intervals were MAE 9.38 [7.23, 12.11], RMSE 15.56 [10.34, 20.89], and R^2 0.744 [0.428, 0.804].

Table 1. Rolling-origin one-quarter-ahead performance for quarterly regional MDR/RR-TB case counts, Q1 2020–Q4 2023 (n = 400 test observations).

Model	MAE	R^2	RMSE
Gradient Boosting	9.29	0.741	15.64
Random Forest	9.38	0.744	15.56
KNN (scaled)	9.41	0.755	15.22
Persistence	9.41	0.573	15.28
Voting Ensemble	9.74	0.737	15.77
Seasonal-naïve	12.57	0.619	18.97

Two-sided Wilcoxon tests on region-level mean absolute errors (25 paired regions), followed by Holm correction, did not establish a significant difference between Random Forest and KNN, the Voting Ensemble, Gradient Boosting, or persistence (all adjusted $p = 1.00$). Random Forest did outperform the seasonal-naive benchmark (adjusted $p < 0.001$). These

results support comparable performance among several leading configurations rather than a statistically established ranking.

A sensitivity analysis excluded Donetsk, Luhansk, Kherson, and Zaporizhzhia from both development and test data (21 regions; 336 test observations). Random Forest then achieved MAE/RMSE/R^2 = 9.16/15.28/0.773, KNN 9.17/15.40/0.770, Gradient Boosting 9.13/15.70/0.760, and persistence 8.95/14.84/0.786. The main conclusion therefore did not depend solely on the four conflict-affected regions, although this analysis does not establish equivalent reporting completeness across territories.

For the most recent year alone (Q1–Q4 2023; $n = 100$ held-out predictions), Gradient Boosting achieved MAE = 7.04, Random Forest 7.10, persistence 7.26, and KNN 7.63. These year-specific results are consistent with the broader conclusion that several leading methods perform similarly under temporally held-out evaluation.

5.3. Model Components and Scaling

The voting ensemble comprised AdaBoost, a decision tree, and SVR; Random Forest was not one of its components. Across the primary 2020–2023 test period, AdaBoost achieved MAE/RMSE/R^2 = 10.61/16.87/0.699, Decision Tree 11.60/19.77/0.586, SVR 12.34/22.41/0.468, and PoissonRegressor 11.60/17.78/0.665. Repeating the distance-sensitive methods with training-fold standardization produced only small changes: KNN changed from 9.48/14.99/0.762 unscaled to 9.41/15.22/0.755 scaled, and SVR from 12.35/22.40/0.469 to 12.34/22.41/0.468. Missing scaling therefore did not explain the relative ranking. The performance of voting-ensemble components is presented in Table 2.

Table 2. Performance of voting-ensemble components and the count-oriented Poisson benchmark under rolling-origin validation, Q1 2020–Q4 2023.

Model	MAE	R^2	RMSE
AdaBoost	10.61	0.699	16.87
Decision Tree	11.60	0.586	19.77
PoissonRegressor	11.60	0.655	17.78
SVR (scaled)	12.34	0.468	22.41

Equal-weight averaging of the three voting components did not improve on the strongest leading configurations. This result should be interpreted as a property of the fitted AdaBoost, decision-tree, and SVR predictions and their aggregation, not as dilution of a Random Forest component. The Poisson benchmark also did not outperform the leading machine learning models or persistence on this evaluation.

5.4. Regional Clustering

The illustrative 2021 analysis clustered the standardized annual mean of Q1–Q4 MDR/RR-TB counts for 25 regional units. The three-group representation separated Dnipropetrovsk and Odesa into the highest-valued group, with nine intermediate and fourteen lower-valued units (Figure 6; Table 3). Numeric cluster identifiers are arbitrary; the substantive labels refer only to relative recorded values in the 2021 partition.

The percentages in Table 3 describe the distribution of administrative units, not proportions of TB cases or population. The 2021 grouping must not be compared directly with 2023 notification or prevalence rankings. In particular, Kirovohrad's lower 2021 MDR/RR-TB annual mean is not inconsistent with its high 2023 population-standardized notification and prevalence rates, because the year and indicator differ.

Likewise, Dnipropetrovsk exceeds Odesa in the 2021 MDR/RR-TB clustering variable even though Odesa has the higher 2023 prevalence rate. Luhansk's low 2021 recorded value

is retained descriptively but is not interpreted as evidence of low underlying disease burden because reporting completeness and diagnostic access are not quantified.

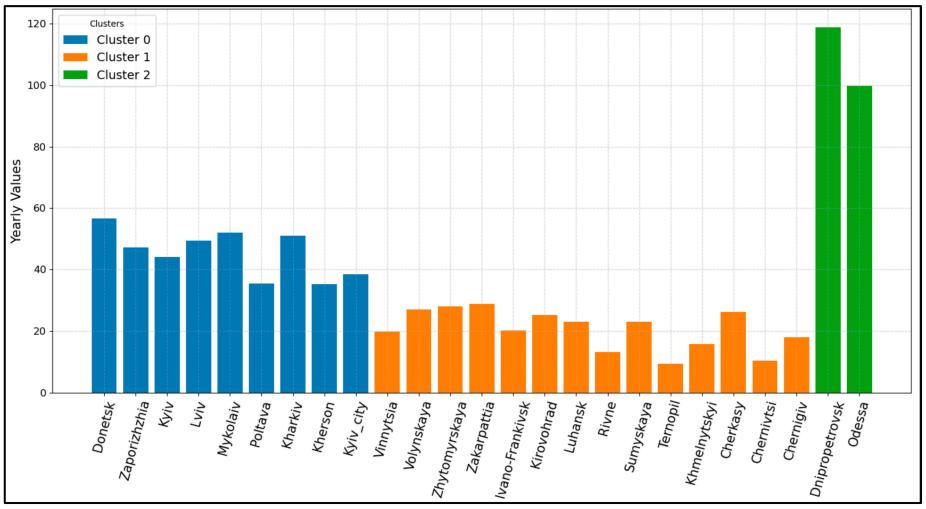

Figure 6. Illustrative regional K-means partition for 2021. The 25 displayed regional units are coloured by cluster identifier. In this plot, cluster 0 contains intermediate values, cluster 1 lower values, and cluster 2 the two highest regional values.

Table 3. Regional clustering for 2021.

ID	Relative Level	Units, *n* (%)	Regional Units
0	Intermediate	9 (36%)	Donetsk, Zaporizhzhia, Kyiv Oblast, Lviv, Mykolaiv, Poltava, Kharkiv, Kherson, Kyiv City
1	Lower	14 (56%)	Vinnytsia, Volyn, Zhytomyr, Zakarpattia, Ivano-Frankivsk, Kirovohrad, Luhansk, Rivne, Sumy, Ternopil, Khmelnytskyi, Cherkasy, Chernivtsi, Chernihiv
2	Higher	2 (8%)	Dnipropetrovsk, Odesa

5.5. Validation of the Three-Group Representation

The 2021 inertia curve decreased sharply through three clusters (Figure 7), but internal validation did not unanimously identify k = 3 as optimal. For k = 3, inertia was 1.857, silhouette = 0.648, Calinski–Harabasz = 137.09, and Davies–Bouldin = 0.396. For k = 2, silhouette was higher (0.776) and Davies–Bouldin lower (0.271), whereas the Calinski–Harabasz index favored larger k values as cluster number increased. The k = 3 partition was completely stable across 100 single-initialization random seeds relative to an n_init = 50 reference solution (mean and minimum ARI = 1.00), but stability was lower in several other years. Thus, three groups are retained as an interpretable descriptive representation rather than a uniquely optimal or persistent epidemiological structure. Internal validation of K-means solutions for the illustrative 2021 regional MDR/RR-TB annual-mean clustering are presented in Table 4.

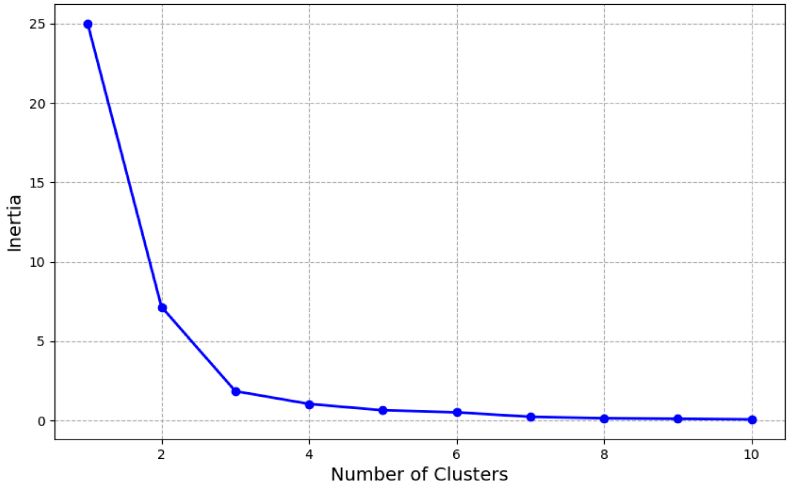

Figure 7. Within-cluster sum of squares (inertia) for candidate k-means solutions. The elbow motivated inspection of a three-group representation, which was subsequently evaluated with internal validation indices and initialization stability rather than treated as uniquely optimal.

Table 4. Internal validation of K-means solutions for the illustrative 2021 regional MDR/RR-TB annual-mean clustering.

K	Inertia	Silhouette	Calinski–Harabasz	Davies–Bouldin
2	7.162	0.776	57.28	0.271
3	1.857	0.648	137.09	0.396
4	1.024	0.563	163.86	0.435
5	0.647	0.573	188.12	0.405
6	0.366	0.519	255.62	0.318
7	0.234	0.519	317.94	0.330
8	0.142	0.489	425.17	0.393
9	0.102	0.506	488.48	0.320
10	0.065	0.481	636.13	0.339

6. Discussion

The analysis supports the feasibility of combining temporally validated regression, simple benchmarks, and exploratory regional grouping within an information system for TB surveillance in Ukraine. Across 400 held-out region-quarter observations, Gradient Boosting, Random Forest, KNN, persistence, and the Voting Ensemble produced relatively similar aggregate errors, and paired region-level tests did not establish the superiority of one leading machine-learning model after Holm correction. The 2021 clustering remained useful as a compact descriptive representation, but internal criteria did not uniquely support three clusters.

6.1. Interpretation of Model Performance

The three regression configurations provide a focused comparison of complementary approaches within the information system. Random Forest represents a tree-based ensemble capable of modeling nonlinear relationships and interactions in structured surveillance records [12,30]. KNN provides a contrasting, instance-based reference against which to examine whether similarity between observations in the supplied predictor space is sufficient to reproduce regional indicator values [31]. The VotingRegressor configuration combines AdaBoost, decision-tree, and support-vector regressors, enabling an assessment of whether averaging heterogeneous estimators improves agreement compared with the standalone

configurations [32]. These roles justify their inclusion as comparative baselines, although they do not establish that the selected models exhaust the approaches suitable for these data.

Gradient Boosting is a relevant alternative that could be evaluated using the same tabular representation [36]. Its exclusion reflects the restricted scope of the implemented comparison rather than an incompatibility with the surveillance data or evidence of inferior performance. The inclusion of AdaBoost in VotingRegressor does not replace a separate assessment of gradient boosted regression trees. Consequently, the present comparison cannot establish whether Random Forest would retain its advantage over a suitably configured Gradient Boosting model. Such an extension should compare models using identical evaluation observations, consistent availability of predictors, and comparable opportunities for hyperparameter selection.

GCNs introduce a distinct modeling requirement because they explicitly incorporate relationships among observations via a graph representation [37]. For regional TB analysis, this would require defining interregional connections, for example, through geographic adjacency, population mobility, or similarity in historical surveillance profiles. No such graph was specified in the present analytical workflow. Although an adjacency graph could be constructed, geographic proximity alone should not be assumed to represent the epidemiologically relevant relationships between regions. A graph-based extension would therefore require justification of the connectivity structure, examination of sensitivity to alternative graph definitions, and evaluation of whether relational information improves predictions beyond the corresponding tabular inputs.

Accordingly, the reported model ranking applies to the three evaluated configurations and does not establish superiority over Gradient Boosting or graph-based regression. Their omission limits the breadth of the comparison, while leaving the observed differences among the implemented configurations informative within the stated evaluation setting. A broader assessment should distinguish improvements attributable to the learning algorithm from those attributable to additional spatial information or a different input representation. For forecasting applications, all configurations would additionally require temporally ordered evaluation with preprocessing and model selection confined to the development observations.

Random Forest ranked first across the reported error measures in the 2023 comparison. However, RMSE and R^2 are mathematically related when calculated on the same observations and therefore do not provide independent evidence. The ranking concerns the specific configurations and observations evaluated, rather than the inherent superiority of any one algorithm family. A relevant counterpoint is provided by Tang et al., who modeled daily TB notifications in Changde, China: Random Forest achieved the best fit, whereas a back-propagation neural network performed best on the 2021 validation data [38]. More recently, Yang et al. reported that XGBoost outperformed ARIMA and Prophet on a chronologically later 2023–2024 test period for monthly pulmonary TB data in Fuzhou [39]. These differing findings favor a setting-specific evaluation. Their numerical errors cannot be directly compared with the present scores because the outcomes, units, temporal resolutions, predictors, and evaluation designs differ.

The weaker KNN result should therefore be interpreted as evidence about the tested implementation, not as proof that KNN regression is unsuitable for TB analysis. Its behavior depends on the neighborhood size, feature scaling, and the representation of regional and temporal information. Similarly, the additional Ensemble configuration did not improve on Random Forest, which is itself an ensemble of trees. Combining predictions can leave systematic errors unchanged or dilute a stronger component when the component predictions are poorly complementary. Neither explanation can be established here without the fitted

configurations and paired residuals. In contrast, Zhang et al. reported improved TB predictions using a deliberately constructed SARIMA-SVR-ELM hybrid with adjacent-region information [40].

The regional plots qualify the favorable aggregate scores. In the 2020 MDR/RR-TB case-count example, $R^2 = 0.97$ coexisted with a visible underestimation of the largest observed value in the Dnipropetrovsk region (Figure 4). A high R^2 can occur when between-region variability is large, even if some local discrepancies remain substantial. The 2023 prevalence plot likewise contained both overestimation and underestimation across regions (Figure 5). Such residual patterns matter operationally because underestimating a high recorded burden could misrepresent expected service demand. Conversely, agreement with the observed regional profile does not establish that undiagnosed disease has been captured. Regional residuals and their direction should consequently accompany aggregate error measures when the system is evaluated for planning purposes.

A further distinction is between reproducing observations within a selected year and predicting observations that were unavailable during model development. The horizontal axes in Figures 4 and 5 identify regions, not successive forecast origins. The plots, therefore, cannot independently demonstrate temporal generalization. Kapoor and Narayanan documented how information leakage can lead to overstated performance claims in scientific applications of machine learning [41]. Their findings justify scrutiny of the present data partitioning and preprocessing, but do not establish that leakage occurred in this study. Until the evaluated samples and their separation from model development are fully specified, the reported scores are most appropriately treated as descriptive evidence of agreement.

6.2. Temporal Variation and the Surveillance Observation Process

The annual prevalence results do not indicate uniformly stable performance. MAE increased from 1.50 in 2021 to 3.25 in 2023, and RMSE from 1.98 to 4.59, although both remained below their 2019 values. These changes could reflect altered outcome distributions, changing surveillance coverage, or differences in model evaluation. The scores do not identify their relative contributions. In particular, the favorable 2021 evaluation cannot be interpreted as evidence that TB transmission was better controlled that year. A related distinction appeared in the Yunnan RR-TB analysis by Yang et al.: the model reproduced historical seasonality but showed poorer agreement in 2023 as recorded incidence increased [42]. That study concerns a different outcome and setting, but reinforces the need to assess performance again when the observed epidemiological series changes.

In Ukraine, the relationship between recorded disease and underlying burden is especially important. Surveillance values depend on whether people can reach diagnostic services and whether detected cases enter the reporting system. Kannan et al. showed that open-source epidemic intelligence could supplement incomplete accessible reporting during the 2022 invasion, while also acknowledging ascertainment bias and limitations in the precision of extracted case counts [43]. For the present system, this implies that a low prediction error may indicate successful reproduction of an incomplete observation process. It does not resolve underdiagnosis or missing territorial coverage. Near-zero values in conflict-affected regions should therefore prompt examination of data completeness before they are interpreted epidemiologically or assigned to a lower-burden group.

Population displacement creates additional uncertainty in both regional case assignment and rate denominators. Nordstrand and Berild reported increasing TB notification rates among newly arriving Ukrainian refugees screened in Norway, with the increase concentrated among adult males [44]. Those observations describe a selected population of arrivals and cannot be extrapolated to all residents of Ukraine or all displaced people. Kashnitsky et al., in a qualitative study covering care access in six receiving countries,

identified treatment disruptions, language barriers, and administrative obstacles among people displaced from Ukraine [45]. Together, these studies identify relevant processes that a regional indicator alone cannot distinguish. Migration and disrupted access are plausible contextual influences, not effects estimated by the present regression or clustering analyses.

6.3. Meaning and Limitations of Regional Clustering

The 2021 partition offers a compact description of heterogeneity. The higher-valued group comprised Dnipropetrovsk and Odesa regions, while nine displayed units formed the intermediate group and fourteen formed the lower group (Figure 6; Table 3). Odesa region also had the highest recorded prevalence in the separate 2023 analysis. These observations justify closer examination of local surveillance and service requirements, but not an inference of stable cluster membership across years. The 8% of displayed units in the higher-valued group represents neither 8% of TB cases nor 8% of the population. Moreover, similarities in aggregate indicator values do not establish shared transmission pathways, equivalent diagnostic coverage, or comparable treatment needs.

The term cluster also requires methodological precision. K-means groups regions by similarity in the supplied numerical representation, but it does not require geographic adjacency or test whether neighboring areas have excess disease relative to their populations. Wang et al. used FleXScan with data from 103 counties in Hubei to identify statistically significant spatial concentrations of TB [46]. Their approach addresses a different question from the value-based partition examined here. The present groups should therefore be described as regional indicator groups rather than statistically confirmed transmission hotspots. If geographic concentration becomes an analytical objective, spatial dependence, population denominators, and uncertainty in regional risk estimates will need explicit treatment.

The reduction in inertia through three clusters supports a parsimonious display, but does not demonstrate that three is a uniquely appropriate partition. Schubert's methodological analysis explains why an apparent elbow is an unreliable stand-alone basis for choosing the number of K-means clusters [20]. For these data, sensitivity to scaling, initialization, and the two highest-valued observations warrants particular attention. Repeated fitting and resampling could establish whether the regional assignments are reproducible, while comparisons with alternative cluster counts could assess how much the interpretation depends on the selected representation. Longitudinal comparisons require consistent outcome definitions and explicit alignment of independently fitted cluster labels. The available 2021 display cannot support claims about year-to-year regional transitions.

6.4. Implications for the Public Health Use

The practical contribution is the integration of model comparison, inspection of observed and estimated values, and regional grouping within a common interface. Related systems demonstrate that accessible interfaces can support substantially different analytical purposes. JSTMapp implements Bayesian joint spatiotemporal disease mapping and illustrates its use with bovine and human extrapulmonary TB data in Africa [47]. Tabby2 links a transmission-dynamic model to a web interface for exploring TB intervention scenarios, service use, and economic outcomes in the United States [48]. Compared with these applications, the present system is focused on examining Ukrainian regional surveillance indicators. Its contribution lies in this applied implementation and comparative analysis, rather than in a new regression algorithm or an established capacity to simulate intervention effects.

For surveillance analysts, the most defensible immediate use is to identify observations requiring review. A large residual may warrant investigation of an unusual regional change

or a reporting inconsistency, and a high recorded value may indicate substantial detected service demand. However, a region with low recorded values and restricted diagnostic access could also require additional attention. A prioritization rule based exclusively on model predictions or cluster membership could consequently favor areas with better ascertainment and overlook poorly observed populations. Operational displays should distinguish indicator values from indicators of reporting completeness and diagnostic activity. Any future allocation rule would need a specified decision objective and evaluation against actual service needs.

The availability of a digital system is not equivalent to evidence of benefit from its use. This distinction is illustrated by Jerene et al.'s four pragmatic cluster-randomized trials of TB digital adherence technologies, which found no evidence of reduced poor treatment outcomes in the participating countries, including Ukraine [49]. Those interventions differ from a surveillance analysis system, so their results do not estimate the effectiveness of the present application. They nevertheless show why technical functionality and health impact require separate assessment. The screenshots and performance summaries presented here do not measure analyst workload, decision timeliness, implementation costs, or improvements in diagnosis and treatment. Such benefits remain to be assessed in a subsequent implementation evaluation.

The system's analytical workflow could be adapted to other geographical contexts by replacing the Ukrainian regional records with surveillance data organized by the relevant administrative units and reporting periods. Portability would require harmonization of indicator definitions, population denominators, territorial coverage, and reporting frequency, followed by model retraining and validation in the target setting. Applications at finer spatial or temporal resolutions would increase data processing requirements and necessitate attention to sparse observations and spatial and temporal dependencies, whereas coarser aggregation could obscure local heterogeneity. Cluster assignments would also require re-estimation because they depend on the geographical units and indicator distributions analyzed. Computational scalability and predictive transferability were not benchmarked in this study and should therefore be assessed before deployment across larger datasets or different surveillance systems.

6.5. Study Limitations and Priorities for Further Evaluation

The simulation design limits inference to administrative-area observations. Regional associations cannot be translated into individual disease probabilities or attributed to particular social groups. The historical window also spans major disruptions, and observations from the same region and nearby periods are unlikely to provide fully independent information. The 25 units visible in the clustering example do not establish the sample size used to calculate the regression scores. Furthermore, the lack of a fully reconciled mapping between annual notification rates, recorded prevalence, and quarterly first-confirmed MDR/RR-TB counts limits comparisons between analytical outputs. Each target requires its own case definition, aggregation rule, units, and territorial coverage. These distinctions are essential for interpreting error magnitudes and determining the population to which an output applies.

The evaluation is also limited by the incomplete specification of the prediction inputs, preprocessing, and hyperparameters, as well as the separation of training, tuning, and assessment observations. No uncertainty intervals or repeated evaluation results accompany the model ranking. Consequently, the magnitude and reproducibility of the observed advantage cannot yet be established beyond the reported point estimates. The reporting principles of TRIPOD + AI and the quality and applicability considerations in PROBAST + AI provide useful guidance on documenting these aspects [50,51]. Both were developed for

clinical prediction research; their relevance here lies in transparent reporting and critical appraisal, rather than conferring validation on an aggregate surveillance model.

The next evaluation should match the intended use. Forecasting within known regions requires temporally ordered assessment at explicitly defined horizons, with all preprocessing and model selection confined to the development period. Transfer to other regions requires a separate geographically held-out assessment. A rolling-origin design would help determine whether the 2023 model ordering persists across forecast origins, rather than relying on a single comparison. Persistence and, where the temporal resolution permits, seasonal-naive forecasts would provide operationally meaningful benchmarks alongside statistical time series and machine learning alternatives. Error comparisons should use identical evaluation observations and account for dependence between repeated regional measurements.

Further development should prioritize data quality and uncertainty alongside model flexibility. Reporting completeness, diagnostic activity, and denominator provenance would help distinguish changes in detected burden from changes in observation. Prediction intervals should be evaluated for coverage on later data, and performance should be examined separately for high-burden and incompletely observed regions. Additional predictors, including lagged regional indicators and neighboring area information, should be included only when available at the intended forecast origin and demonstrate incremental value. Any model updating strategy should be evaluated on observations not used to choose that strategy. These extensions are proposed as the next steps, not analyses established by the current results.

Overall, the study provides an applied basis for comparing modeling choices and examining regional TB indicators in Ukraine. Random Forest is a reasonable candidate for further evaluation given its advantage in the reported 2023 comparison, while the clustering output provides a useful descriptive summary of regional variation. The strength of subsequent operational claims will depend on reproducible outcome definitions, independent evaluation, and explicit treatment of incomplete surveillance. With these requirements addressed, the system could support a more informative regional review; its contribution to better public health decisions must still be demonstrated.

7. Conclusions

The study developed an information system for analyzing regional TB surveillance indicators in Ukraine. Its application-specific scientific contribution is the integration of outcome-specific data handling, rolling-origin one-quarter-ahead model evaluation, simple operational benchmarks, uncertainty-aware comparison, and exploratory regional clustering within a common workflow.

Across Q1 2020–Q4 2023, Gradient Boosting, Random Forest, KNN, persistence, and the Voting Ensemble showed closely similar out-of-sample performance. Random Forest achieved MAE = 9.38, RMSE = 15.56, and $R^2 = 0.744$, while Gradient Boosting achieved 9.29, 15.64, and 0.741 and KNN 9.41, 15.22, and 0.755. Paired region-level tests did not establish statistically significant superiority of a single leading model after multiplicity correction. The evaluation therefore does not support a claim of universal superiority for Random Forest.

The 2021 K-means analysis provides an illustrative three-group representation of regional MDR/RR-TB annual mean counts. The partition was stable across random initializations, but internal validation indices did not uniquely favor $k = 3$; the groups should therefore be interpreted as descriptive regional indicator groups rather than persistent epidemiological categories or transmission hotspots.

https://doi.org/10.3390/computation14110269

The system can support structured review of recorded surveillance observations, but the models do not directly represent case-finding intensity, diagnostic access, treatment outcomes, displacement, or under-reporting. Application to new periods, regions, or countries requires harmonized data definitions, retraining, and independent validation. Future work should evaluate reporting completeness, prediction uncertainty, geographic transfer, and the effect of system use on actual public-health decisions.

Author Contributions: Conceptualization, K.B.; methodology, O.K., K.B. and D.C.; software, O.K., Y.P. and I.M.; validation, O.K., K.B., Y.P., I.M., T.C. and D.C.; formal analysis, K.B. and T.C.; investigation, O.K., K.B., Y.P., T.C. and D.C.; resources, D.C.; data curation, O.K. and K.B.; writing—original draft preparation, O.K., K.B. and D.C.; writing—review and editing, O.K., K.B., Y.P., I.M., T.C. and D.C.; visualization, O.K.; supervision, K.B.; project administration, K.B. and D.C.; funding acquisition, D.C. All authors have read and agreed to the published version of the manuscript.

Funding: The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2025.07/0270 on the topic "A Multidisciplinary Methodology for Modelling the Syndemic of War: Infectious Diseases, Information Manipulation, and Challenges of Incomplete Medical Data During Conflicts."

Data Availability Statement: The original contributions presented in this study are included in the article material. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5.6 Sol and Grammarly for the purposes of grammar check. The authors have reviewed and edited the output and take full responsibility for the content of this publication. DC wishes to thank the INI and LMS for the financial support and the Center for Digital Public Health in Emergencies at the University College London for their hospitality.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TB	Tuberculosis
COVID-19	Coronavirus disease 2019
ARIMA	Autoregressive integrated moving average
HIV	Human Immunodeficiency Virus
LSTM	Long short-term Memory
KNN	K-nearest-neighbor
SARIMA	Seasonal autoregressive integrated moving average
SIR	Susceptible-Infected-Recovered
ARMA	Autoregressive moving average
EMD	Empirical mode decomposition
TabPFN	Tabular Prior-data Fitted Network
LASSO	Least Absolute Shrinkage and Selection Operator
SHAP	SHapley Additive exPlanations
RMSE	Root Mean Squared Error
CNN	Convolutional Neural Network
MGWR	Multiscale geographically weighted regression
MAE	Mean Absolute Error
SVR	Support Vector Regression
ELM	Elaboration Likelihood Model
MDR-TB	Multidrug-resistant TB
RR-TB	Rifampicin-resistant TB
AI	Artificial Intelligence
PROBAST	Prediction model Risk Of Bias ASsessment Tool
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

References

1. WHO. Global Tuberculosis Report 2025. Available online: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2025 (accessed on 15 July 2026).
2. Hu, M.; Feng, Y.; Li, T.; Zhao, Y.; Wang, J.; Xu, C.; Chen, W. Unbalanced risk of pulmonary Tuberculosis in China at the subnational scale: Spatiotemporal analysis. *JMIR Public Health Surveill.* **2022**, *8*, e36242. [CrossRef] [PubMed]
3. Krokva, D.; Mori, H.; Valenti, S.; Remez, D.; Hadano, Y.; Naito, T. Analysis of the impact of crises Tuberculosis incidence in Ukraine amid pandemics and war. *Sci. Rep.* **2025**, *15*, 17045. [CrossRef] [PubMed]
4. Maipan-Uku, J.Y.; Cavus, N. Forecasting Tuberculosis incidence: A review of time series and machine learning models for prediction and eradication strategies. *Int. J. Environ. Health Res.* **2024**, *35*, 645–660. [CrossRef] [PubMed]
5. Abade, A.; Porto, L.F.; Scholze, A.R.; Kuntath, D.; Barros, N.D.S.; Berra, T.Z.; Ramos, A.C.V.; Arcêncio, R.A.; Alves, J.D. A comparative analysis of classical and machine learning methods for forecasting TB/HIV co-infection. *Sci. Rep.* **2024**, *14*, 18991. [CrossRef] [PubMed]
6. Barman, M.; Panja, M.; Mishra, N.; Chakraborty, T. Epidemic-guided deep learning for spatiotemporal forecasting of Tuberculosis outbreak. *Mach. Learn.* **2025**, *114*, 213. [CrossRef]
7. Hewamalage, H.; Ackermann, K.; Bergmeir, C. Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Min. Knowl. Discov.* **2023**, *37*, 788–832. [CrossRef] [PubMed]
8. Rodríguez, A.; Kamarthi, H.; Agarwal, P.; Ho, J.; Patel, M.; Sapre, S.; Prakash, B.A. Machine learning for data-centric epidemic forecasting. *Nat. Mach. Intell.* **2024**, *6*, 1122–1131. [CrossRef]
9. Kalizhanova, A.; Yerdessov, S.; Sakko, Y.; Tursynbayeva, A.; Kadyrov, S.; Gaipov, A.; Kashkynbayev, A. Modeling Tuberculosis transmission dynamics in Kazakhstan using SARIMA and SIR models. *Sci. Rep.* **2024**, *14*, 24824. [CrossRef] [PubMed]
10. Siamba, S.; Otieno, A.; Koech, J. Application of ARIMA, and hybrid ARIMA models in predicting and forecasting Tuberculosis incidences among children in Homa Bay and Turkana counties, Kenya. *PLoS Digit. Health* **2023**, *2*, e0000084. [CrossRef] [PubMed]
11. Zhao, R.; Liu, J.; Zhao, Z.; Zhai, M.; Ren, H.; Wang, X.; Li, Y.; Cui, Y.; Qiao, Y.; Ren, J.; et al. A hybrid model for Tuberculosis forecasting based on empirical mode decomposition in China. *BMC Infect. Dis.* **2023**, *23*, 665. [CrossRef] [PubMed]
12. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520. [CrossRef]
13. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef] [PubMed]
14. Yang, E.; Zhang, H.; Guo, X.; Zang, Z.; Liu, Z.; Liu, Y. A multivariate multi-step LSTM forecasting model for Tuberculosis incidence with model explanation in Liaoning province, China. *BMC Infect. Dis.* **2022**, *22*, 490. [CrossRef] [PubMed]
15. Lv, Z.; Sun, R.; Liu, X.; Wang, S.; Guo, X.; Lv, Y.; Yao, M.; Zhou, J. Evaluating the effectiveness of self-attention mechanism in Tuberculosis time series forecasting. *BMC Infect. Dis.* **2024**, *24*, 1377. [CrossRef] [PubMed]
16. Hao, J.; Liang, Y.; Zhao, H.; Li, S.; Shen, J. Forecasting global Tuberculosis incidence rates using a multi-variate input, multi-region, multi-step output integrated model: SP-CTLA. *Trans. GIS* **2025**, *29*, e70060. [CrossRef]
17. Ma, Z.; Fan, H. Influential factors of Tuberculosis in mainland China based on MGWR model. *PLoS ONE* **2023**, *18*, e0290978. [CrossRef] [PubMed]
18. Lei, Y.; Wang, J.; Wang, Y.; Xu, C. Geographical evolutionary pathway of global Tuberculosis incidence trends. *BMC Public Health* **2023**, *23*, 755. [CrossRef] [PubMed]
19. Kunjok, D.M.; Mwangi, J.G.; Kairu-Wanyoike, S.; Kinyua, J.; Mambo, S. Spatial epidemiology of Tuberculosis diagnostic delays, healthcare access disparities, and socioeconomic inequities in Nairobi county, Kenya. *PLoS ONE* **2025**, *20*, e0329984. [CrossRef] [PubMed]
20. Schubert, E. Stop using the elbow criterion for K-means and how to choose the number of clusters instead. *ACM SIGKDD Explor. Newsl.* **2023**, *25*, 36–42. [CrossRef]
21. Kapoor, S.; Cantrell, E.M.; Peng, K.; Pham, T.H.; Bail, C.A.; Gundersen, O.E.; Hofman, J.M.; Hullman, J.; Lones, M.A.; Malik, M.M.; et al. REFORMS: Consensus-based recommendations for machine-learning-based science. *Sci. Adv.* **2024**, *10*, eadk3452. [CrossRef] [PubMed]
22. Public Health Centre of the Ministry of Health of Ukraine. *Tuberculosis in Ukraine: Analytical and Statistical Reference Report for 2023*; Public Health Centre of the Ministry of Health of Ukraine: Kyiv, Ukraine, 2024. (In Ukrainian)
23. Monedero-Recuero, I.; Wei, X.; Jouberton, F.; Duarte, R.; Rodríguez-Ortega, M.; Rosary, M.; Adekunle, A.O.; Daru, P.; Felker, I.; Conradie, F.; et al. Drug-resistant Tuberculosis in war and complex emergencies: Jeopardising progress towards TB elimination and antimicrobial resistance control—A scoping review and perspective. *BMJ Glob. Health* **2026**, *11*, e019011. [CrossRef] [PubMed]
24. Stoycheva, K.; Cristea, V.; Ködmön, C.; Rosales-Klintz, S.; Zenner, D.; Vasiliu, A.; van der Werf, M.; Lange, C. Tuberculosis in people of Ukrainian origin in the European Union and the European Economic Area, 2019 To 2022. *Eurosurveillance* **2024**, *29*, 2400094. [CrossRef] [PubMed]

25. Diuzheva, O.; Skoklyuk, L.; Zherebko, N.; Barbova, A.; Germanovych, M.; Klinkenberg, E.; Bogdanov, O.; Dravniece, G. Paving the way to innovative, child-friendly pediatric diagnostic methods for Tuberculosis: Introduction of stool-based testing in Ukraine. *Trop. Med. Infect. Dis.* **2024**, *9*, 209. [CrossRef] [PubMed]
26. Salnikova, A.; Makarenko, O.; Sereda, Y.; Kiriazova, T.; Lunze, K.; DeHovitz, J.; Ompad, D.C. Depression among people living with Tuberculosis and Tuberculosis/HIV coinfection in Ukraine: A cross-sectional study. *Glob. Health Action* **2025**, *18*, 2448894. [CrossRef] [PubMed]
27. Konstantynovska, O.; Synenko, T.; Honcharenko, A.; Volobuieva, O.; Liadova, T.; Reimann, M.; Lange, C.; Chesov, D. Fluoroquinolone resistance in drug-resistant Tuberculosis, Kharkiv, Ukraine, 2019–2023. *Emerg. Infect. Dis.* **2025**, *31*, 615–617. [CrossRef] [PubMed]
28. Fardhdiani, V.; Trush, O.; Lytvynenko, N.; Pylypchuk, S.; Terleeva, Y.; Malakyan, K.; Rupasinghe, P.; Wally, Y.; Duka, M.; Didyk, V.S.; et al. End of treatment and 12-month post-treatment outcomes in patients treated with all-oral regimens for Rifampicin-resistant Tuberculosis in Ukraine: A Prospective Cohort Study. *PLoS Glob. Public Health* **2025**, *5*, e0003983. [CrossRef] [PubMed]
29. Public Health Centre of the Ministry of Health of Ukraine. Analytical and Statistical Materials on Tuberculosis. Available online: https://phc.org.ua/kontrol-zakhvoryuvan/tuberkuloz/statistika-z-tb/analitichno-statistichni-materiali-z-tb (accessed on 15 July 2026). (In Ukrainian)
30. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
31. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185. [CrossRef]
32. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
33. Hodson, T.O. Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to use them or not. *Geosci. Model Dev.* **2022**, *15*, 5481–5487. [CrossRef]
34. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
35. Jain, A.K. Data clustering: 50 years beyond K-Means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
36. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
37. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907. [CrossRef]
38. Tang, N.; Yuan, M.; Chen, Z.; Ma, J.; Sun, R.; Yang, Y.; He, Q.; Guo, X.; Hu, S.; Zhou, J. Machine learning prediction model of Tuberculosis incidence based on meteorological factors and air pollutants. *Int. J. Environ. Res. Public Health* **2023**, *20*, 3910. [CrossRef] [PubMed]
39. Yang, J.; Pan, J.; Chen, J.; Xu, Y.; Zhang, X. Leveraging machine learning for accurate forecasting of pulmonary Tuberculosis epidemics in a Coastal City in China. *Trop. Med. Health* **2026**, *54*, 63. [CrossRef] [PubMed]
40. Zhang, Y.; Song, H.; Zhang, S.; Wang, X.; Tang, J. A hybrid machine learning model for pulmonary Tuberculosis forecasting of Chongqing with adjacent-region data. *PLoS ONE* **2025**, *20*, e0339453. [CrossRef] [PubMed]
41. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
42. Yang, Y.; Liu, L.; Chen, J.; Li, L.; Qiu, Y.; Wu, W.; Xu, L. Predicting the incidence of Rifampicin resistant Tuberculosis in Yunnan, China: A seasonal time series analysis based on routine surveillance data. *BMC Infect. Dis.* **2024**, *24*, 835. [CrossRef] [PubMed]
43. Kannan, A.; Chen, R.; Akhtar, Z.; Sutton, B.; Quigley, A.; Morris, M.J.; MacIntyre, C.R. Use of open-source epidemic intelligence for infectious disease outbreaks, Ukraine, 2022. *Emerg. Infect. Dis.* **2024**, *30*, 1865–1871. [CrossRef] [PubMed]
44. Nordstrand, K.; Berild, J.D. Increase in Tuberculosis notification rates among newly arriving male Ukrainian refugees to Norway, 2022 To 2024. *Eurosurveillance* **2025**, *30*, 2500633. [CrossRef] [PubMed]
45. Kashnitsky, D.; Vyatchina, M.; Ariabinska, A.; Basquin, P.; Kasianczuk, M.; Kvinikadze, G.; Landeau, E.; Rabinciuc, V.; Rivera, K.; Shishniashvili, I.; et al. From barriers to solutions: A qualitative study of access to HIV and TB care for forced migrants from Ukraine. *BMC Health Serv. Res.* **2026**, *26*, 253. [CrossRef] [PubMed]
46. Wang, L.; Li, Y.; Chen, J.; Zhang, D.; Zhang, Z.; Li, X. Spatiotemporal distribution and detection of spatial clusters of Tuberculosis in Hubei Province, China using FleXScan (2017–2023). *BMC Public Health* **2025**, *25*, 2758. [CrossRef] [PubMed]
47. Ngwira, A.; Manda, S.O.M.; Karimuribo, E.D.; Kimera, S.I. JSTMapp: A web-based joint spatiotemporal modelling and mapping application for epidemiologists. *PLoS ONE* **2025**, *20*, e0332607. [CrossRef] [PubMed]
48. Swartwood, N.A.; Testa, C.; Cohen, T.; Marks, S.M.; Hill, A.N.; Asay, G.B.; Cochran, J.; Cranston, K.; Randall, L.M.; Tibbs, A.; et al. Tabby2: A user-friendly web tool for forecasting state-level TB outcomes in the United States. *BMC Med.* **2023**, *21*, 331. [CrossRef] [PubMed]
49. Jerene, D.; van Kalmthout, K.; Levy, J.; Alacapa, J.; Deyanova, N.; Dube, T.; Mganga, A.; Tasca, B.; Bogdanov, A.; Efo, E.; et al. Effect of digital adherence technologies on treatment outcomes in people with drug-susceptible Tuberculosis: Four pragmatic, cluster-randomised trials. *Lancet* **2025**, *405*, 1155–1166. [CrossRef] [PubMed]

50. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, q902. [CrossRef] [PubMed]
51. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Andaur Navarro, C.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **2025**, *388*, e082505. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.