Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 32−43
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7158

Integrating Support Vector Machines and Geospatial Analysis for Enhanced Tuberculosis Case Detection and Spatial Mapping

Miftahul Jannah*, Muhammad Jazman, M Afdal, Megawati Megawati

Faculty of Science and Technology, Information Systems, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1,*] 12150321893@students.uin-suska.ac.id, [2] jazman@uin-suska.ac.id, [3] m.afdal@uin-suska.ac.id, [4] megawati@uin-suska.ac.id
Corresponding Author Email: 12150321893@students.uin-suska.ac.id
Submitted: **26/03/2025**; Accepted: **31/05/2025**; Published**: 01/06/2025**

Abstract−Tuberculosis (TB) remains a significant global health problem, with Indonesia ranking third in the world in terms of TB burden. Riau Province recorded 13,007 notified TB cases in 2022 with a Case Notification Rate (CNR) of 138 per 100,000 population, still far from the national target. This study aims to develop a TB case classification system using Support Vector Machine (SVM) integrated with geospatial analysis to identify TB positive cases from screening data and visualize their spatial distribution in Riau Province. The research data was sourced from the Tuberculosis Information System (SITB) of the Riau Provincial Health Office for the period January-December 2024, covering 350 samples with demographic information, clinical symptoms, and patient risk factors. The research process includes data collection, preprocessing with Min-Max and Z-Score methods, feature extraction, modeling with SVM using various kernels (RBF, Linear, Polynomial, and Sigmoid), and geospatial visualization using Google Earth Engine (GEE). The results showed that the SVM model with Linear kernel achieved the highest accuracy of 80%, sensitivity of 100%, and specificity of 80% in detecting TB cases. Geospatial analysis successfully identified clusters of TB cases in several districts in Riau Province, with Pekanbaru City (112 cases) and Rokan Hulu (89 cases) as the main hotspots. The integration of machine learning and geospatial analysis proved effective in improving TB detection and providing a comprehensive understanding of disease spread patterns in Riau Province.

Keywords: Tuberculosis; Support Vector Machines; Geospatial; Google Earth Engine; Machine Learning

1. INTRODUCTION

Tuberculosis (TB) remains a significant global public health problem in the 21st century. Infection caused by the pathogen Mycobacterium tuberculosis continues to be a major cause of mortality globally, contributing to millions of deaths each year [1]. Despite declines in incidence and mortality rates in recent decades, TB remains a substantial public health burden, particularly in low- and middle-income countries. The World Health Organization (WHO) estimates that there were 10 million new cases and 1.5 million deaths from TB globally in 2020 [1]. The COVID-19 pandemic has exacerbated this situation through disruption of essential health services, including TB programs, resulting in decreased case detection, reduced accessibility to treatment, and hampering global efforts to achieve TB elimination targets. Most significantly, the WHO 2023 report indicated that for the first time in two decades, TB mortality increased by 14% during the COVID-19 pandemic, reflecting the severe impact of health service disruptions on disease control efforts [1].

Indonesia ranks third in terms of global TB burden after India and China, indicating the urgency for implementing comprehensive TB control programs at the national le vel. Data from the Ministry of Health of the Republic of Indonesia states that around 824,000 TB cases were notified in 2021, with a Case Notification Rate (CNR) reaching 301 per 100,000 population [2]. However, this notification figure remains below the estimated 969,000 cases, indicating a substantial gap between detected cases and actual estimates. This disparity is influenced by multiple factors such as limited accessibility to health facilities, low public health literacy regarding the urgency of TB examinations, and limitations in diagnostic methods. A recent investigation [3] revealed that financial and geographical barriers remain fundamental challenges in the accessibility of TB care in Indonesia, with 36% of TB patients experiencing catastrophic expenditures related to treatment.

In the specific context of Riau Province, with its accelerated economic growth, high population mobility, and environmental variability, TB problems persist with 13,007 notified cases and a CNR of 138 per 100,000 population in 2022 [4]. This figure shows a significant deviation from the national target of 245 per 100,000 population, highlighting the urgency of strengthening early detection strategies, increasing accessibility to TB services, and developing a comprehensive understanding of the determinants of TB spread in the region. Research [5] indicates that environmental factors such as residential density, inadequate ventilation, and indoor air pollution contribute significantly to TB transmission . Delays in TB diagnosis and initiation of therapy can result in fatal consequences, increasing the risk of complications, antimicrobial resistance, and communal transmission.

Early detection is an essential component of TB control strategies and a crucial step toward achieving the target of TB elimination. Conventional diagnostic methodologies, such as sputum microscopic examination, have limited sensitivity, especially in cases of negative Acid-Fast Bacteria (AFB) TB and pediatric TB, often resulting in false negatives. Studies [6] demonstrated that the sensitivity of sputum microscopic examination only reaches 60-70% in patients with pulmonary TB, with even lower sensitivity in cases of HIV co-infection. While mycobacterial culture is recognized as the gold standard in TB diagnosis, it requires several weeks to obtain results, thus causing delays in diagnosis and treatment initiation. Molecular diagnostic methods such as GeneXpert MTB/RIF have increased the sensitivity and speed of TB diagnosis, but the penetration of this technology remains limited in many regions of Indonesia, including Riau Province. These limitations in conventional diagnostic methodologies underscore the need for innovative approaches for more accurate, efficient, and spatially context-based TB detection to optimize public health interventions.

The evolution of machine learning technology provides new prospects for improving the accuracy and efficiency of TB detection. Several scientific investigations have demonstrated the potential of machine learning in TB diagnosis using diverse data, including radiological images, clinical data, and demographic information. Previous research on TB detection using machine learning approaches has shown promising results. Lopes and Valiati [7] implemented Support Vector Machine (SVM) with feature extraction from Convolutional Neural Networks (CNN) for TB detection from chest radiography images, achieving an accuracy of up to 87.4%. Wu et al. [8] proposed a segmentation methodology using Deep Convolutional Neural Network (DCNN) and SVM with accuracy of up to 91.2%. Vasquez-Morales et al. [9] demonstrated the effectiveness of SVM in predicting TB treatment outcomes with 89.7% accuracy using patient clinical data.

However, previous studies on TB detection exhibit significant methodological limitations that create critical research gaps. Most existing approaches have focused either on machine learning classifications without spatial context or on geospatial analysis without individual-level precision. Rahman et al. [10] achieved 85-95% accuracy using machine learning algorithms for TB detection but lacked the spatial context necessary for targeted interventions, while Shaweno et al. [11] employed geospatial analysis for TB epidemiology without the precision of individual-level classification. Many studies primarily analyze radiological images rather than comprehensive patient data that includes demographic and clinical variables, limiting holistic understanding of TB determinants. Sathitratanacheewin et al. [12] demonstrated that deep learning models for TB detection often suffer from dataset distribution shift, limiting their generalizability across different populations and healthcare settings. Furthermore, most previous research has not adequately addressed computational efficiency alongside accuracy metrics, a crucial consideration for real-world implementation in healthcare systems with limited infrastructure. Hansun et al. [13] highlighted the absence of comprehensive approaches that integrate clinical, demographic, and spatial data in a unified framework—a key limitation preventing more effective TB surveillance and control strategies. The limited integration between machine learning classification and geospatial analysis represents a critical gap in TB surveillance methodology, particularly in high-burden regions where both individual case detection and understanding of transmission dynamics are equally important.

This study addresses these limitations by developing a tuberculosis detection model using the Support Vector Machine (SVM) algorithm integrated with geospatial analysis. SVM was chosen because of its capability in handling high-dimensional and non-linear data, as well as its proven performance in classification applications. The research utilizes patient data from the Riau Provincial Health Office's Tuberculosis Information System (SITB), which contains comprehensive information related to demographics, clinical manifestations, therapeutic history, and patient risk factors[14]. A distinctive feature of this study is the integration of patient clinical data with spatial information processed using Google Earth Engine (GEE), a cloud computing platform for geospatial analysis[15]. This allows dynamic visualization of TB case distribution patterns across the 12 districts/cities in Riau Province[16]. The interactive nature of the developed GEE application enables health authorities to identify areas with high TB transmission risk and analyze factors contributing to case clusters in specific locations. By combining machine learning classification with geospatial analysis, this research aims to provide both accurate individual-level TB risk assessment and comprehensive population-level understanding of TB transmission dynamics[17]. This integrated approach offers significant advantages for TB surveillance and control by revealing geographic patterns of disease transmission, identifying environmental risk factors, and highlighting areas with limited healthcare access[18], all while maintaining the precision of individual case classification.

The purpose of this study is to develop a tuberculosis detection model using the SVM algorithm, complemented by geospatial visualization, by utilizing clinical patient data from the Riau Provincial Health Office alongside spatial mapping. The SVM model is expected to identify complex patterns from multiple clinical and demographic variables to produce more accurate and efficient classification of TB cases compared to conventional methods. The geospatial visualization component aims to map the distribution of detected TB cases across different districts, providing valuable context for decision-makers. Ultimately, this research aims to contribute to improving early detection of TB through the SVM algorithm while using spatial mapping to visualize TB risk areas, thereby optimizing resource allocation for TB control efforts in Riau Province. This dual approach addresses a critical need in public health surveillance and intervention planning by combining individual-level case detection with population-level spatial visualization.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The methodology of this research starts from the planning stage to the report preparation stage. The steps taken for this research can be seen in the image below:

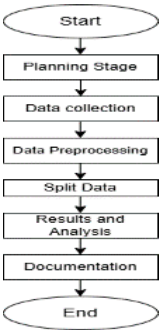


Figure 1. Research Flow Diagram

The systematic approach of this research is illustrated in Figure 1, which outlines the five main stages from planning to evaluation. The following is the research methodology:

a. Planning Stage: This initial phase involved defining research objectives, determining scope, and designing the study methodology. A comprehensive literature review was conducted to identify appropriate machine learning approaches for TB detection and geospatial analysis techniques.

b. Data Collection: The study utilized 350 TB screening records from the Riau Provincial Health Office's Tuberculosis Information System (SITB) for January-December 2024. The dataset included 14 predictor variables (demographic information, clinical symptoms, and risk factors) and one target variable indicating TB status. Spatial data including administrative boundaries of Riau Province was obtained from the FAO GAUL 2015 dataset.

c. Data Preprocessing: This critical phase involved several key steps:
   1. Data cleaning to handle missing values, particularly in the 'Duration of Cough' variable using median imputation
   2. Removal of duplicate records to ensure data integrity
   3. Data transformation using label encoding for categorical variables
   4. Data normalization using both Min-Max and Z-Score methods to scale values appropriately for the SVM algorithm

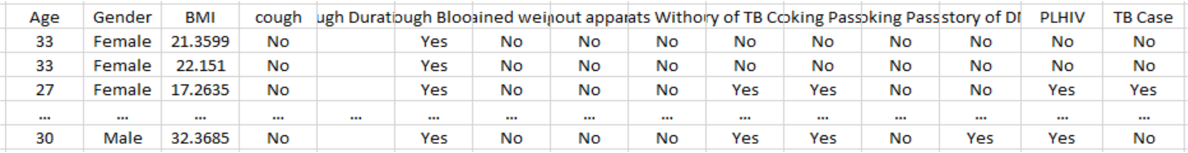
d. Split Data: The dataset was divided using both train-test split (with 70:30, 80:20, and 90:10 ratios) and K-fold cross-validation (with k=3 and k=5) to evaluate model performance and generalizability.

e. Results and Analysis: This phase encompassed:
   1. SVM model development with different kernel functions (RBF, Linear, Polynomial, and Sigmoid)
   2. Performance evaluation using various metrics (accuracy, precision, recall, F1-score)
   3. Geospatial analysis using Google Earth Engine to visualize TB case distribution
   4. Integration of classification results with spatial visualization to identify hotspots

f. Documentation: The final stage involved comprehensive documentation of the research findings, including model performance comparison, spatial distribution patterns of TB cases, identification of TB hotspots, and recommendations for TB control programs.

2.2 Data Collection

Data were collected from two main sources. Primary data in the form of TB patient information was obtained from the Riau Provincial Health Office's Tuberculosis Information System (SITB) for the period January-December 2024. with a total of 350 samples consisting of 166 female patients (47.43%) and 184 male patients (52.57%). This dataset has a processed structure with 15 variables, consisting of 14 predictor variables and 1 target variable indicating the status of TB cases. Secondary data in the form of spatial data, such as administrative boundaries of districts/cities in Riau Province, were obtained from the FAO GAUL 2015 dataset. It's important to note that in this study, geospatial data was not directly integrated as a predictive feature in the SVM model. Rather, the SVM classification was performed based on patient clinical and demographic data, while geospatial analysis was conducted as a separate but complementary component. The results from the SVM classification (TB case detection) were then mapped geospatially to identify spatial patterns and clusters. This two-step approach allows for both individual case classification and population-level spatial analysis, providing a more comprehensive understanding of TB epidemiology in Riau Province.



Age	Gender	BMI	cough	ugh Durat	ough Bloo	ained wei	out appa	ats Witho	ry of TB C	oking Pass	oking Pass	story of DI	PLHIV	TB Case
33	Female	21.3599	No		Yes	No	No	No	No	No	No	No	No	No
33	Female	22.151	No		Yes	No	No	No	No	No	No	No	No	No
27	Female	17.2635	No		Yes	No	No	No	Yes	Yes	No	No	Yes	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30	Male	32.3685	No		Yes	No	No	No	Yes	Yes	No	Yes	Yes	No

Figure 2. Data Collection Results

2.2.1 Data Type and Variable Details

Table 1 shows the details of the data type of each attribute in the dataset. Label encoding will be used to convert categorical variables to numeric. As shown in Table 1, the dataset consists of 15 variables including demographic information, clinical symptoms, and risk factors:

Table 1. Data Type Details

Attribute	Number	Data Type	Information
Age	1	Numeric	Patient age in years.
Gender	2	Nominal	Patient gender (Male/Female). Will be encoded as 0/1.
IMT	3	Numeric	Patient Body Mass Index (Kg/ m^2)
Cough	4	Nominal	Cough present/absent. Will be encoded as 0/1.
Duration of Cough	5	Numeric	Duration of cough in weeks.
Coughing up blood	6	Nominal	Presence/absence of coughing up blood. Will be encoded as 0/1.
BB drops without any clear cause	7	Nominal	There is/is no weight loss. Will be encoded as 0/1.
Fever comes and goes for no apparent reason	8	Nominal	Fever present/absent. Will be encoded as 0/1.
Sweating at Night Without Activity	9	Nominal	Presence/Absence of night sweats. Will be encoded as 0/1.
TB Contact History	10	Nominal	Contact history present/absent. Will be encoded as 0/1.
Smoke	11	Nominal	Smoker/Non-Smoker. Will be encoded as 0/1.
Passive Smoker	12	Nominal	Presence/Not exposure to cigarette smoke. Will be encoded as 0/1.
DM History	13	Nominal	History of Diabetes Mellitus. Will be encoded as 0/1.
PLHIV	14	Nominal	HIV Presence/Absence. Encoded as 0/1.
TB Cases (Target Variable)	15	Nominal	Positive/Negative. Encoded as 0/1.

2.2.2 List of Regencies/Cities in Riau Province

The following table lists the districts/cities in Riau Province along with their geographic coordinates. This information complements the map and makes it easier to identify each region. Table 2 presents the comprehensive list of districts/cities in Riau Province along with their corresponding geographic coordinates, while Figure 3 provides a visual representation of the administrative boundaries.

Table 2. Regency/City of Riau Province

Regency	Latitude	Longitude
City of Pekanbaru	0.5071	101.4478
Dumai City	1.6677	101.4489
Regency Kampar	0.3089	101.2252
Regency Rokan Hulu	0.8527	100.6181
Regency Rokan Hilir	1.6866	100.9192
Regency Siak	0.9784	102.0305
Regency Indragiri Hulu	-0.4634	102,534
Regency Indragiri Hilir	-0.3273	103.0897
Regency Pelalawan	0.1028	102.1943
Regency Kuantan Singingi	-0.3244	101,537
Regency Bengkalis	1.4794	102.1347
Regency Meranti	0.9423	102.7231

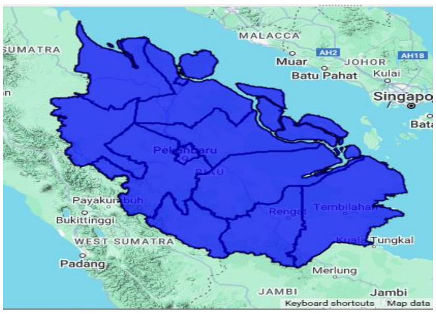


Figure 3. Administrative Map of Riau Province[19]

2.3 Data Preprocessing

The data preprocessing stage aims to prepare the data to be optimal for the SVM model training process. At this stage, data cleaning is carried out to remove noise to improve data quality [20].

The first step in Data Cleaning is to handle missing values and duplicate data. Missing values in the "Duration of Cough" variable are handled by imputation using the median value. Duplicate data, if found, will be removed to maintain data integrity. Next, data transformation is carried out, including coding categorical variables. This coding process uses Label Encoding, where each unique category in the variable is mapped to an integer. For example, the "Gender" variable will be changed to 0 for "Male" and 1 for "Female". This method was chosen because of its simplicity, but it should be noted that Label encoding can cause order interpretation problems in nominal variables. Finally, data normalization is performed using Min-Max Normalization to scale attribute values to the range 0 to 1. The formula for Min-Max Normalization is: x_norm = (x - min(x)) / (max(x) - min(x)), where x is the original value and x_norm is the normalized value [7]. This normalization process will produce new values as shown in the table below. After the data is cleaned and transformed, the data is ready to be used for model training. Tables 3-6 demonstrate the progressive transformation of the data through cleaning, transformation, and normalization stages:

Table 3. Empty Data

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
33	Woman	21,359	No		No	No	No	No	No	No	No	No	No	No
33	Woman	22.151	No		No	No	No	No	No	No	No	No	No	No
27	Women	17,263	No		No	No	No	No	No	No	No	No	No	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32	Man	19.140	No		No	No	No	No	No	Yes	No	No	No	No

Table 4. Data Cleaning

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
33	1	21,359	No		No	No	No	No	No	No	No	No	No	No
33	1	22.151	No		No	No	No	No	No	No	No	No	No	No
27	1	17,263	No		No	No	No	No	No	No	No	No	No	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32	0	19.140	No		No	No	No	No	No	Yes	No	No	No	No

Table 5. Data Transformation

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
33	1	21,359	0	0	0	0	0	0	0	0	0	0	0	0
33	1	22.151	0	0	0	0	0	0	0	0	0	0	0	0
27	1	17,263	0	0	0	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32	0	19.140	0	0	0	0	0	0	0	1	0	0	0	0

Table 6. Data Normalization Using *Min-Max*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.379747	1	0.016021	0	0	1	0	0	0	0	0	0	0	0	0
0.379747	1	0.017096	0	0	1	0	0	0	0	0	0	0	0	0
0.303797	1	0.010457	0	0	1	0	0	0	1	1	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.341772	0	0.030976	0	0	1	0	0	0	1	1	0	1	1	0

3. RESULTS AND DISCUSSION

3.1 Implementation and Testing of SVM Model

The SVM model is implemented using the scikit-learn library on the Google Colab platform. Testing is carried out with various scenarios to evaluate the model performance.

3.1.1 Testing the Effect of Using Preprocessing

In the first scenario, testing was conducted to determine the effect of using Min-Max and Z-Score preprocessing on all conditions. The test results can be seen in the table below. As illustrated in Figure 4 and detailed in Table 7, the comparison between Min-Max and Z-Score preprocessing methods reveals that Z-Score preprocessing slightly outperforms Min-Max with an average accuracy of 80.00% versus 79.52%.

Table 7. Results of the Preprocessing Effect Test

Preprocessing	Accuracy	Evaluation		
		Precision	Recall	F1-Score
Min-Max	77.14%	79.10%	96.36%	86.89%
	84.29%	84.29%	100.00%	91.47%
	77.14%	77.14%	100.00%	87.10%
Average	79.52%	80.18%	98.79%	88.48%
Z-Score	77.14%	79.10%	96.36%	86.89%
	85.71%	86.57%	98.31%	92.06%
	77.14%	77.14%	100.00%	87.10%
Average	80.00%	80.94%	98.22%	88.68%

Based on the test results, a comparison of performance between the Min-Max and Z-Score preprocessing methods in terms of accuracy, precision, recall, and F1-score can be seen. The Min-Max method produces an average accuracy of 79.52%, precision of 80.18%, recall of 98.79%, and F1-score of 88.48%. On the other hand, the Z-Score method produces an average accuracy of 80.00%, precision of 80.94%, recall of 98.22%, and F1-score of 88.68%. Although the difference is not too significant, the use of Z-Score preprocessing shows a slight increase in the average accuracy and F1-score, as well as a greater increase in the average recall compared to the Min-Max method. This indicates that the Z-Score method can identify positive cases of tuberculosis better.

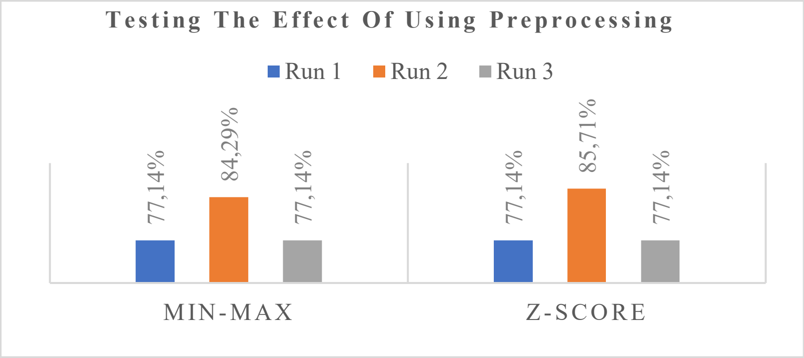


Figure 4. Testing the Effect of Using Preprocessing

3.1.2 Testing the Effect of Data Distribution Methods

In the second scenario, testing is conducted to determine the performance of the SVM model. Two commonly used data splitting methods, namely train-test split and K-Fold Cross-Validation, will be tested and compared. The goal is to find the method that provides the most accurate and robust model performance estimates, while minimizing bias and variance. The comparison of these two methods is important because a simple train-test split can produce performance estimates that depend on random data splits, while K-Fold Cross-Validation provides a more comprehensive evaluation by using all data for training and testing alternately.

a. Train Test Split

The performance of the SVM model was evaluated using the train-test split method with ratios of 70:30, 80:20, and 90:10. Each ratio was tested three times with different random conditions to ensure consistency and reduce bias. Model parameters (kernel, C, gamma) were optimized through Grid Search for each ratio and trial. Accuracy, precision, recall, and F1-score on the trial data were used for evaluation, which are recorded in Table 8, with the aim of finding the optimal ratio that produces the best performance and generalization, and avoids overfitting and underfitting. Table 8 summarizes the performance metrics across different train-test split ratios, with Figure 5 providing a visual comparison of these results.

Table 8 Comparison of Model Performance with Various Train Test Split Methods

Train Test Split	**Accuracy**	**Evaluation**		
		Precision	**Recall**	**F1-Score**
10% - 90%	74.29%	75.76%	96.15%	84.75%
	82.86%	82.86%	100.00%	90.62%
	80.00%	80.00%	100.00%	88.89%
Average	**79.05%**	**79.54%**	**98.72%**	**88.09%**
20% - 80%	77.14%	79.10%	96.36%	86.89%
	84.29%	84.29%	100.00%	91.47%
	77.14%	77.14%	100.00%	87.10%

Average	**79.52%**	**80.18%**	**98.79%**	**88.48%**
30% - 70%	78.10%	81.44%	94.05%	87.29%
	81.90%	83.84%	96.51%	89.73%
	77.14%	77.14%	100.00%	87.10%
Average	**79.05%**	**80.81%**	**96.85%**	**88.04%**

Table 8 presents a comparison of the performance of the SVM models with various train-test split ratios. The 80:20 split ratio (20% test data and 80% training data) produced the best average performance with 79.52% accuracy, 80.18% precision, 98.79% recall, and 88.48% F1-score. The 70:30 and 90:10 split ratios produced comparable performance with an average accuracy of 79.05%. Although the 90:10 ratio (10% test data and 90% training data) had a slightly higher average recall, the 80:20 ratio showed a better balance between accuracy, precision, recall, and F1-score. These results indicate that the proportion of training and testing data affects the model performance, and in this case, the 80:20 ratio provided the most optimal results.

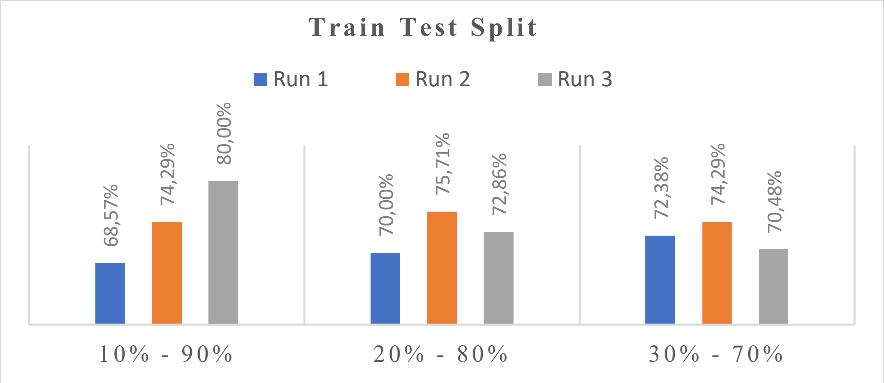


Figure 5. Comparison Train Test Split Method

b. K-Fold Cross-Validation
K-fold cross-validation is used for more robust model evaluation. The dataset is divided into k folds, the model is trained k times, each iteration uses k-1 folds for training and 1 fold for testing. This study uses k=3 and k=5 to compare the effect of the number of folds on performance (Table 9). As shown in Table 9 and visualized in Figure 6, k-fold cross-validation with k=3 produced slightly higher accuracy and F1-score compared to k=5.

Table 9. Test Results of the Effect of Using K-Fold

Train Test Split	Accuracy	Evaluation		
		Precision	Recall	F1-Score
K-Fold (k=3)	79.49%	79.49%	100.00%	88.57%
	81.03%	82.73%	96.81%	89.22%
	75.86%	75.86%	100.00%	86.27%
Average	**78.79%**	**79.36%**	**98.94%**	**88.02%**
K-Fold (k=5)	80.00%	80.00%	100.00%	88.89%
	78.57%	80.30%	96.36%	87.60%
	81.43%	83.82%	96.61%	89.76%
	70.00%	70.00%	100.00%	82.35%
	79.71%	81.82%	96.43%	88.52%
Average	**77.94%**	**79.19%**	**97.88%**	**87.43%**

Table 9 presents the results of k-fold cross-validation testing with k=3 and k=5. For k=3, the average accuracy, precision, recall, and F1-score are 78.79%, 79.36%, 98.94%, and 88.02%, respectively. While for k=5, the average accuracy, precision, recall, and F1-score are 77.94%, 79.19%, 97.88%, and 87.43%, respectively. The results show that k-fold with k=3 produces slightly higher accuracy and F1-score compared to k=5. This difference is relatively small, indicating that both k values provide fairly comparable performance. However, k=3 shows slightly higher recall values, indicating a slightly better ability to identify positive cases of tuberculosis. K=5, on the other hand, offers lower variance between folds, as seen from the more consistent evaluation metric values across iterations. This suggests that k=5 provides more stable and general performance estimates.

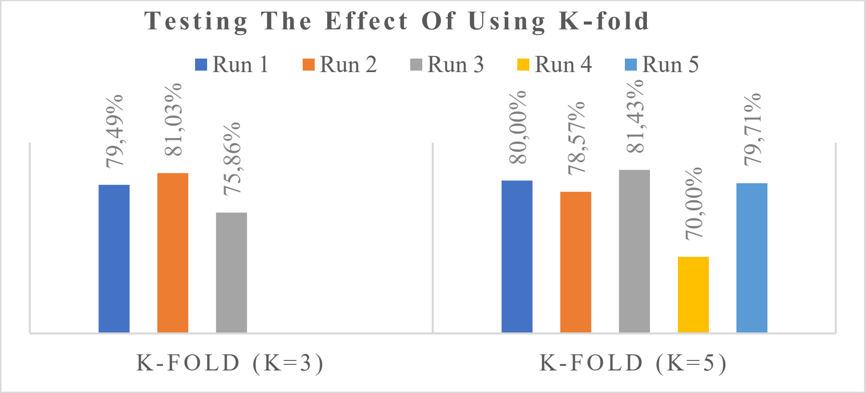


Figure 6. Testing the Effect of Using K-Fold

3.1.3 Testing the Effect of Kernel Usage

The third scenario aims to evaluate the effect of using different kernels on the performance of the SVM model in detecting tuberculosis. The four kernels tested were RBF, Linear, Polynomial, and Sigmoid. The dataset was divided using a train-test split, and Min-Max preprocessing was applied to equalize the testing conditions. Each kernel was tested three times to obtain more representative results. The test results can be seen in Table 10. Table 10 and Figure 7 present the comparative performance of different kernel functions, demonstrating that the Linear kernel achieved the highest average accuracy of 80.00%.

Table 10. Results of Testing the Effect of Kernel Usage

Train Test Split	Accuracy	Evaluation		
		Precision	Recall	F1-Score
RBF	77.14%	79.10%	96.36%	86.89%
	84.29%	84.29%	100.00%	91.47%
	77.14%	77.14%	100.00%	87.10%
Average	**79.52%**	**80.18%**	**98.79%**	**88.48%**
Linear	78.57%	78.57%	100.00%	88.00%
	84.29%	84.29%	100.00%	91.47%
	77.14%	77.14%	100.00%	87.10%
Average	**80.00%**	**80.00%**	**100.00%**	**88.86%**
Poly	77.14%	79.10%	96.36%	86.89%
	84.29%	84.29%	100.00%	91.47%
	77.14%	77.14%	100.00%	87.10%
Average	**79.52%**	**80.18%**	**98.79%**	**88.48%**
Sigmoid	71.43%	69.81%	71.43%	69.41%
	70.00%	76.60%	70.00%	72.55%
	70.00%	74.69%	70.00%	70.77%
Average	**70.48%**	**73.70%**	**70.48%**	**70.91%**

Table 10 shows the results of testing the effect of using different kernels on the SVM model. The Linear kernel showed the best average performance with an accuracy of 80.00%, a precision of 80.00%, a recall of 100.00%, and an F1-score of 88.86%. The RBF and Poly kernels showed similar performance with an average accuracy of 79.52%, a precision of 80.18%, a recall of 98.79%, and an F1-score of 88.48%. Although having slightly lower accuracy than the Linear kernel, both kernels showed high recall and F1-score values. The Sigmoid kernel, on the other hand, recorded the lowest performance among all tested kernels, with an average accuracy of 70.48%, a precision of 73.70%, a recall of 70.48%, and an F1-score of 70.91%. These results indicate that kernel selection has a significant effect on the performance of the SVM model. The performance differences between kernels are due to the characteristics of the data and how each kernel maps the data to a higher-dimensional space, which ultimately affects the model's ability to separate classes. The Linear kernel, in this case, seems to be the best fit for the data used. While the RBF and Poly kernels also produce good results, the Sigmoid kernel is less than optimal for this dataset.

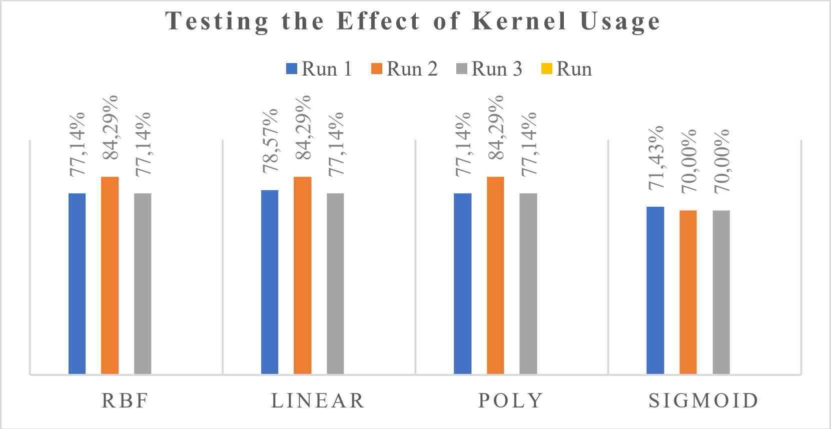


Figure 7. Testing the Effect of Kernel Usage

The Linear kernel demonstrated not only the highest accuracy but also the fastest training time, making it particularly suitable for deployment in resource-limited healthcare settings. The RBF and Polynomial kernels required approximately twice the computational resources while achieving slightly lower accuracy. The Sigmoid kernel showed both poorer performance and moderate computational requirements, making it the least favorable option for this application. For real-time TB screening applications, the Linear kernel's efficiency provides a significant advantage, particularly in regions with limited computational infrastructure.

3.1.4 Computational Performance Analysis

Beyond accuracy metrics, computational efficiency is an important consideration for implementing TB detection models in real-world healthcare settings. Table 11 presents the training time comparison for each kernel type. As detailed in Table 11, the Linear kernel not only demonstrated superior classification performance but also required significantly less computational resources compared to other kernels.

Table 11. Computational Performance Comparison of Different Kernels

Kernel Type	Average Training Time (seconds)	Relative Computational Load
Linear	0.15	Low
RBF	0.29	Medium
Polynomial	0.33	Medium
Sigmoid	0.27	Medium

The Linear kernel demonstrated not only the highest accuracy but also the fastest training time, making it particularly suitable for deployment in resource-limited healthcare settings. The RBF and Polynomial kernels required approximately twice the computational resources while achieving slightly lower accuracy. The Sigmoid kernel showed both poorer performance and moderate computational requirements, making it the least favorable option for this application. For real-time TB screening applications, the Linear kernel's efficiency provides a significant advantage, particularly in regions with limited computational infrastructure.

3.2 Geospatial Analysis Results

Geospatial analysis was conducted using Google Earth Engine (GEE) to visualize and analyze the spatial distribution of TB cases in Riau Province. This approach allows for hotspot identification, correlation with environmental and demographic factors, and a more comprehensive understanding of TB transmission dynamics. TB case data in 2024 from the Riau Provincial Health Office's Tuberculosis Information System (SITB) is integrated with district/city administrative boundaries on the GEE platform. The following table shows TB case data in Riau Province for the period January - December 2024. Figure 8 visualizes the distribution of TB cases across Riau Province, with Table 12 providing the corresponding numerical data for each district/city.

Table 12. TB Case Data in Riau Province 2024

Regency	**TB cases**
City of Pekanbaru	112
Dumai City	0
Regency Kampar	1
Regency Rokan Hulu	89
Regency Rokan Hilir	0
Regency Siak	1
Regency Indragiri Hulu	1

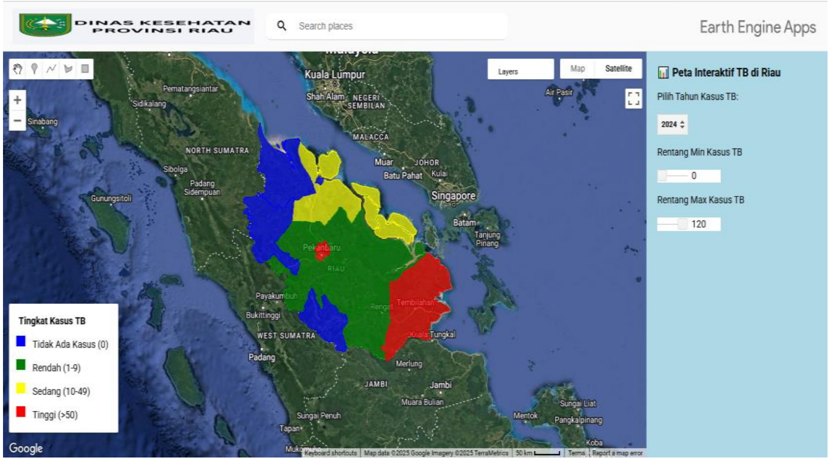
Regency Indragiri Hilir	50
Regency Pelalawan	3
Regency Kuantan Singingi	0
Regency Bengkalis	13
Regency Meranti	0

3.2.1 Interactive Map of TB Cases in Riau Province

An interactive map developed in GEE (see Figure 8) allows for dynamic exploration of TB case data. The map uses a tiered color scheme to represent the level of TB cases in each district/city:

a. Blue: No Cases (0)
b. Green: Low (1-9 cases)
c. Yellow: Moderate (10-49 cases)
d. Red: High (>50 cases)

Users can filter data by year (currently only 2024) and case range (5-120) for more focused analysis. This interactive feature facilitates identification of TB hotspots, monitoring of case trends, and evaluation of the effectiveness of intervention programs.



Peta Interaktif TB di Riau

Pilih Tahun Kasus TB: 2024

Rentang Min Kasus TB 0

Rentang Max Kasus TB 120

Figure 8. WebApss Interactive Map of TB Cases in Riau Province

Based on Figure 8, the distribution of TB cases in Riau Province in 2024 shows significant variation between districts/cities. Pekanbaru City and Rokan Hulu stand out as TB hotspots, marked in red indicating the number of cases above 50. Pekanbaru City, as an urban center, recorded the highest number of cases (112 cases), indicating a potential correlation between population density and TB transmission. The high number of cases in Rokan Hulu (89 cases), despite its rural characteristics, indicates the need for further investigation into contributing local factors, such as limited access to health facilities, community awareness, and socio-economic conditions. Indragiri Hilir also showed a relatively high number of cases (50 cases), while Bengkalis was in the moderate category (13 cases). Other districts showed a low number of cases (1-9 cases), and several districts such as Dumai City, Rokan Hilir, Kuantan Singingi, and Kepulauan Meranti did not report any TB cases in 2024.

These spatial patterns provide valuable information for the Riau Provincial Health Office in planning and evaluating TB control programs. The interactive map facilitates the identification of priority areas for intervention, such as Pekanbaru and Rokan Hulu. Further analysis with GEE can integrate demographic, environmental (e.g., air pollution, sanitation), and health service accessibility data to understand risk factors that influence the spatial distribution of TB and develop more effective and targeted management strategies. Monitoring TB case trends through this platform also allows for ongoing evaluation of the impact of implemented TB control programs.

3.3 Integration of SVM Classification and Geospatial Analysis

While the SVM model provides individual-level TB classification based on patient characteristics, the geospatial analysis offers population-level insights into TB distribution patterns. This dual approach creates synergistic value for TB control programs.

The SVM model with Linear kernel achieved 80% accuracy and 100% recall in detecting TB cases, making it effective for individual patient screening. When these classified cases were mapped geospatially, clear hotspots emerged in Pekanbaru City (112 cases) and Rokan Hulu (89 cases), enabling targeted interventions. The integration of these approaches revealed important patterns: while the SVM model identified clinical risk factors at the individual level (with cough duration, BMI, and TB contact history being the most significant predictors), the geospatial analysis revealed that population density and urbanization correlate strongly with TB clusters. Notably, the SVM model showed similar performance patterns across different regions, but case concentration varied significantly, highlighting the importance of combining both analytical methods. This integrated approach allows health authorities to both screen individual patients effectively and allocate resources strategically to high-burden areas. For example, the high accuracy of the SVM model suggests it could be deployed for screening in healthcare facilities across Riau Province, while the geospatial analysis indicates that intensified case finding efforts should be prioritized in Pekanbaru City and Rokan Hulu.

4. CONCLUSION

This study successfully developed an integrated approach for tuberculosis detection in Riau Province using Support Vector Machine (SVM) classification and geospatial analysis. The SVM model serves as a patient-level classification tool, while the geospatial component provides population-level surveillance and visualization. From the implementation and testing results, we can draw several key conclusions: First, the SVM model with Linear kernel demonstrated optimal performance with 80% accuracy, 80% precision, 100% recall, and 88.86% F1-score. The perfect recall score indicates that the model successfully identified all actual TB cases without any false negatives, which is crucial in a clinical context where missing positive cases could have serious health consequences. The Linear kernel's superior performance suggests that the TB classification problem exhibits linear separability in our feature space. Additionally, the Linear kernel demonstrated the shortest training time (0.15 seconds on average), making it well-suited for deployment in resource-constrained healthcare settings. Z-Score preprocessing slightly outperformed Min-Max normalization (80% vs. 79.52% average accuracy), indicating that standardization to account for feature variability is beneficial for this dataset. For data partitioning methods, the 80:20 train-test split ratio yielded optimal results, though the differences compared to K-Fold Cross-Validation were not substantial. The geospatial analysis using Google Earth Engine (GEE) effectively visualized TB case distribution throughout Riau Province, identifying significant hotspots in Pekanbaru City (112 cases) and Rokan Hulu (89 cases). This spatial pattern revealed considerable variation between districts/cities, suggesting the influence of local factors such as population density, healthcare accessibility, and socioeconomic conditions on TB prevalence. The integration of SVM classification with geospatial analysis provided complementary insights: while the SVM model identified individual-level risk factors, the spatial analysis revealed population-level patterns and environmental determinants. This combined approach offers significant advantages for public health decision-making, enabling both effective individual screening and strategic resource allocation to high-burden areas. The limitations of this study include the relatively small sample size (350 samples), the limited data collection period (January-December 2024), and the separate rather than integrated implementation of machine learning and geospatial components.

REFERENCES

[1] "WHO,2021." Accessed: Jan. 31, 2025. [Online]. Available: https://www.who.int/indonesia/news/campaign/tb-day-2022/fact-sheets

[2] "Repository - Aplikasi Repository Kementrian Kesehatan Republik Indonesia." Accessed: Apr. 12, 2025. [Online]. Available: https://repository.kemkes.go.id/book/1288

[3] A. Fuady, T. A. J. Houweling, M. Mansyur, E. Burhan, and J. H. Richardus, "Cost of seeking care for tuberculosis since the implementation of universal health coverage in Indonesia," *BMC Health Serv Res*, vol. 20, no. 1, p. 502, Dec. 2020, doi: 10.1186/s12913-020-05350-y.

[4] Dinas Kesehatan Provinsi Riau, " Dinas Kesehatan Provinsi Riau (2023), Profil Kesehatan Provinsi Riau Tahun 2022. Pekanbaru: Dinkes Provinsi Riau. - Penelusuran Google." Accessed: Mar. 20, 2025. [Online]. Available: https://www.google.com/search?q=Dinas+Kesehatan+Provinsi+Riau.+(2023).+Profil+Kesehatan+Provinsi+Riau+Tahun+2022.+Pekanbaru%3A+Dinkes+Provinsi+Riau.&oq=Dinas+Kesehatan+Provinsi+Riau.+(2023).+Profil+Kesehatan+Provinsi+Riau+Tahun+2022.+Pekanbaru%3A+Dinkes+Provinsi+Riau.&gs_lcrp=EgZjaHJvbWUqBggAEEUYOzIGCAAQRRg70gEHNzQ4ajBqNKgCALACAQ&sourceid=chrome&ie=UTF-8

[5] E. A. Wikurendra, G. Nurika, Y. G. Tarigan, and A. A. Kurnianto, "Risk Factors of Pulmonary Tuberculosis and Countermeasures: A Literature Review," *Open Access Maced J Med Sci*, vol. 9, no. F, pp. 549–555, Nov. 2021, doi: 10.3889/oamjms.2021.7287.

[6] S. E. Dorman *et al.*, "Xpert MTB/RIF Ultra for detection of Mycobacterium tuberculosis and rifampicin resistance: a prospective multicentre diagnostic accuracy study," *The Lancet Infectious Diseases*, vol. 18, no. 1, pp. 76–84, Jan. 2018, doi: 10.1016/S1473-3099(17)30691-6.

[7] U. K. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in Biology and Medicine*, vol. 89, pp. 135–143, Oct. 2017, doi: 10.1016/j.compbiomed.2017.08.001.

[8] W. Wu *et al.*, "An Intelligent Diagnosis Method of Brain MRI Tumor Segmentation Using Deep Convolutional Neural Network and SVM Algorithm," *Computational and Mathematical Methods in Medicine*, vol. 2020, pp. 1–10, Jul. 2020, doi: 10.1155/2020/6789306.

[9] G. R. Vasquez-Morales, S. M. Martinez-Monterrubio, P. Moreno-Ger, and J. A. Recio-Garcia, "Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning," *IEEE Access*, vol. 7, pp. 152900–152910, 2019, doi: 10.1109/ACCESS.2019.2948430.

[10] T. Rahman *et al.*, "Reliable Tuberculosis Detection Using Chest X-Ray With Deep Learning, Segmentation and Visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020, doi: 10.1109/ACCESS.2020.3031384.

[11] D. Shaweno *et al.*, "Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review," *BMC Med*, vol. 16, no. 1, p. 193, Dec. 2018, doi: 10.1186/s12916-018-1178-4.

[12] S. Sathitratanacheewin, P. Sunanta, and K. Pongpirul, "Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability," *Heliyon*, vol. 6, no. 8, p. e04614, Aug. 2020, doi: 10.1016/j.heliyon.2020.e04614.

[13] S. Hansun, A. Argha, S.-T. Liaw, B. G. Celler, and G. B. Marks, "Machine and Deep Learning for Tuberculosis Detection on Chest X-Rays: Systematic Literature Review," *J Med Internet Res*, vol. 25, p. e43154, Jul. 2023, doi: 10.2196/43154.

[14] D. R. Fakhma, "Diagnosis of TBC Disease Using SVM and Feedforward Backpopagation" Journal of Advances in Information Systems and Technology, vol 4, no 1, 2022

[15] T. P. Dao *et al.*, "A geospatial platform to support visualization, analysis, and prediction of tuberculosis notification in space and time," *Front. Public Health*, vol. 10, p. 973362, Sep. 2022, doi: 10.3389/fpubh.2022.973362.

[16] N. Tiwari, C. Adhikari, A. Tewari, and V. Kandpal, "Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic," *Int J Health Geogr*, vol. 5, no. 1, p. 33, 2006, doi: 10.1186/1476-072X-5-33.

[17] M. Barman, M. Panja, N. Mishra, and T. Chakraborty, "Epidemic-guided deep learning for spatiotemporal forecasting of Tuberculosis outbreak," Feb. 15, 2025, *arXiv*: arXiv:2502.10786. doi: 10.48550/arXiv.2502.10786.

[18] A. S. Cahyaningrum and N. A. Setiyadi, "Geo-Spatial Cluster Tuberculosis Of 2021-2023: Study In A District, Indonesia," Indonesian Journal of Global Health Research, vol. 6, no. 5, 2024.

[19] "TBC2024." Accessed: Apr. 12, 2025. [Online]. Available: https://ee-miftahuljnnah25.projects.earthengine.app/view/tbc2024

[20] S. García, J. Luengo, and F. Herrera, "Introduction," in *Data Preprocessing in Data Mining*, S. García, J. Luengo, and F. Herrera, Eds., Cham: Springer International Publishing, 2015, pp. 1–17. doi: 10.1007/978-3-319-10247-4_1.