Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2024 (17 Jan);38.5. https://doi.org/10.47176/mjiri.38.5

Identification of Risk Factors Associated with Tuberculosis in Southwest Iran: A Machine Learning Method

Neda Amoori[1], Bahman Cheraghian[2], Payam Amini[3], Seyed Mohammad Alavi[1]*

Received: 12 Aug 2023 Published: 17 Jan 2024

Abstract

Background: Tuberculosis is a principal public health issue. Reducing and controlling tuberculosis did not result in the expected success despite implementing effective preventive and therapeutic programs, one of the reasons for which is the delay in definitive diagnosis. Therefore, creating a diagnostic aid system for tuberculosis screening can help in the early diagnosis of this disease. This research aims to use machine learning techniques to identify economic, social, and environmental factors affecting tuberculosis.

Methods: This case-control study included 80 individuals with TB and 172 participants as controls. During January-October 2021, information was collected from thirty-six health centers in Ahvaz, southwest Iran. Five different machine learning approaches were used to identify factors associated with TB, including BMI, sex, age , marital status, education, employment status, size of the family, monthly income, cigarette smoking, hookah smoking, history of chronic illness, history of imprisonment, history of hospital admission, first-class family, second-class family, third-class family, friend, co-worker, neighbor, market, store, hospital, health center, workplace, restaurant, park, mosque, Basij base, Hairdressers and school. The data was analyzed using the statistical programming R software version 4.1.1.

Results: According to the calculated evaluation criteria, the accuracy level of 5 SVM, RF, LSSVM, KNN, and NB models is 0.99, 0.72, 0.97,0.99, and 0.95, respectively, and except for RF, the other models had the highest accuracy. Among the 39 investigated variables, 16 factors including First-class family (20.83%), friend (17.01%), health center (41.67%), hospital (24.74%), store (18.49%), market (14.32%), workplace (9.46%), history of hospital admission (51.82%), BMI (43.75%), sex (40.36%), age (22.83%), educational status (60.59%), employment status (43.58%), monthly income (63.80%), addiction (44.10%), history of imprisonment (38.19%) were of the highest importance on tuberculosis.

Conclusion: The obtained results demonstrated that machine-learning techniques are effective in identifying economic, social, and environmental factors associated with tuberculosis. Identifying these different factors plays a significant role in preventing and performing appropriate and timely interventions to control this disease.

Keywords: Tuberculosis, Classification, Risk factor, Machine Learning

Conflicts of Interest: *None declared*
Funding: *The Vice-Chancellor for Research at Ahvaz Jundishapur University of Medical Sciences provided financial support.*

Cite this article as: Amoori N, Cheraghian B, Amini P, Alavi SM. Identification of Risk Factors Associated with Tuberculosis in Southwest Iran: A Machine Learning Method. *Med J Islam Repub Iran.* 2024 (17 Jan);38:5. https://doi.org/10.47176/mjiri.38.5

Introduction

Tuberculosis(TB) is a significant public health concern and is currently recognized as the most fatal treatable infectious disease in the world (1). As per the estimates of the World Health Organization (WHO), in 2020, 9.9 million individuals were infected with tuberculosis and this disease caused 1.3 million deaths (2). Worldwide, Eastern Mediterranean region, Iran,and Khuzestan province had an inci-

Corresponding author: *Dr Seyed Mohammad* Alavi, *alavi_sm@ajums.ac.ir*

1. Infectious and Tropical Diseases Research Center, Health Research Institute, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran
2. Department of Biostatistics and Epidemiology, School of Public Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran
3. School of Medicine, Keele University, Staffordshire, UK

↑What is "already known" in this topic:
Numerous studies have investigated the factors affecting tuberculosis based on socio-demographic variables, and each of these studies has identified a few factors.

→What this article adds:
This study used a machine learning method and investigated economic, social, and environmental factors. Tuberculosis is influenced by 16 main factors, including first-class family, friends, health centers, hospitals, stores, markets, workplaces, history of hospital admission, BMI, sex, age,education, employment status, income, addiction, and history of imprisonment.

[DOI: 10.47176/mjiri.38.5]

dence of 132, 114, 16, and 14.5 cases per 100,000 individuals, respectively, in 2020, which has had a relatively slow decline compared to previous years (3, 4).

Because the incidence of TB declines slowly, there is a renewed tendency to identify novel TB control measures. In addition to existing TB drugs, finding vaccines and designing shorter regimens have been some strategies. However, knowing the factors that lead to TB in some people and not in others (risk factors) can help them focus more on finding new TB control strategies in public health (5). Reported risk factors for tuberculosis include male gender, HIV infection, smoking, alcohol consumption, marital status,comorbidities such as diabetes, cancer, the use of immunosuppressive drugs, family history of tuberculosis, overcrowding In different environments and places, and poor socio-economic status. The current effort to find, treat, and cure any person with this disease is not enough (6, 7). There is a need to re-examine the socio-economic and environmental factors and understand the effective factors to adjust TB control policies. Recent studies have revealed that TB can be transmitted in crowded environments among households that are not related.These findings suggest that the transmission of tuberculosis may occur during exposure to infectious cases in social settings (8).

Identifying social sites associated with TB may help monitor public health, increase case findings, and identify areas that may reduce tuberculosis and the risk of transmission due to exposure to environmental interventions, such as UV radiation or improved air conditioning, exposure to the disease (9). The use of totally non-parametric machine learning models is growing in a variety of scientific disciplines. The principal purpose of this model is to determine the influential factors, the relationships between them, and prediction and estimation. This topic plays a substantial role in medicine and health data analysis because of the data type (10). This research aims to use machine learning techniques to identify and investigate economic, social, and environmental factors affecting tuberculosis.

Methods

Study population

Employing a case-control study, the risk factors associated with TB were evaluated. The subjects selected for the study were newly registered pulmonary TB patients with bacteriological confirmation who were over the age of 18 and presented at thirty-six health centers in Ahvaz, southwest Iran, between January and October 2021. The controls were those individuals with health issues other than TB who presented to the same health centers. The cases and controls were matched in age (within 5-year age bands).

The inclusion criteria for the case and control groups were as follows: TB patients over 18 years old with a positive culture of Mycobacterium tuberculosis who had TB treatment records in the health centers under auspicious.Controls were individuals over the age of 18 who did not suffer from tuberculosis. Both groups lived in areas covered by health centers in Ahvaz. The exclusion criteria for both the case and control groups were as follows: individuals under the age of 17 and those who suffered from mental disabilities or severe mobility limitations.

Sample size and sampling method

The sample size of 80 cases and 172 controls was estimated based on the following: the data available in previous studies, using the formula for determining the sample size in analytical studies and taking into account the level of confidence 95% and the power of the test 80%, with a 2:1 control to case ratio, using Stata 13.0 software. The sampling was conducted using an easy non-probability method for both groups. Since the beginning of the study, all the people who met the criteria for entering the study were selected as a sample. This procedure continued until the final size of the study sample was reached.

Data collection

Face-to-face interviews, standard forms, and checklists were the tools for data collection. This study examined factors associated with tuberculosis, including age, sex, BMI, marital status, education, employment status, size of the family, monthly Income, cigarette smoking, hookah smoking, history of chronic illness, history of imprisonment, history of hospital admission, first-class family, second-class family, third-class family, Friend, co-worker, neighbor, market, Store, hospital, health center, workplace, restaurant, park, mosque, basij base, hairdresser and school. The results of this study are included in the thesis, while the other part is published in a different journal (11).

Statistical analysis

We depicted the descriptive properties of the data through the frequency (percentage) and mean (standard error) for categorical and continuous variables, respectively. After checking the normality of data distribution and comparing the mean of observation across the categories of the response, an independent samples t-test was utilized.An independent chi-square was employed to evaluate the independence of categorical variables from the final result.

Classification algorithms are used in Machine Learning to predict the class label of a given data point. The main advantages of using Machine Learning in datasets with a relatively higher number of variables to the sample size are bias reduction due to robustness toward different sets of data, improving accuracy due to cross validation and iterative computations in background, the higher capability of determining complex patterns and correlations in the dataset among the variables, and the ability to make a prediction. It has been widely argued that the application of Machine Learning approaches in small sample data is associated with higher classification accuracy (12, 13).

Five different machine learning approaches were used for classification, including Random Forest(RF), Support Vector Machine(SVM), Least Square Support Vector Machine(LSSVM), K-Nearest Neighborhood (KNN), and Naïve Bayes (NB). In RF, the dataset is subject to sampling to shape the trees through substitution, and random combinations of predictors are selected at the nodes. LSSVM is a theory of statistical learning that takes a linear function from least squares as a function of loss. The KNN classification generally focuses on the Euclidean distance the train samples defined and a test set and it was built based on the

[DOI: 10.47176/mjiri.38.5]

need for discriminating where accurate parametric estimates of likelihood densities became uncertain or difficult to evaluate. The well-known Bayes' theorem, which follows a clear, simple, and very fast classifier, is used to create the Naïve Bayes classification model (because of mutually independent attributes assumption) (14).

Numerous metrics, including specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), and overall accuracy, were utilized to evaluate the discriminative quality of the computational models. The dataset was divided into test sets (30% of individuals) and train sets (70% of individuals), each containing 176 and 76 individuals, respectively. The assessment criterion was listed as the average of the 500 iterations after we validated each model 500 times.

The statistical programming R software version 4.1.1 (http://www.R-project.org) packages, including randomForest, naïve Bayes, e1071, rpart, ipred, rminer, caret, adabag, magrittr, qwraps2, CORElearn, MASS, mda, klaR, and MASS were utilized to analyze the data. The type one error was assumed to be 0.05.

Results

The participants were 80 cases and 172 controls. The patients' mean age was 34.1 (±15.3) years old, while the controls' was 32.5 (±12.0) years old. Table 1 depicts that 77.5% of the cases and 51.7% of the controls were male. Primary education was recorded in 51 of the cases (63.8%) and 36 (20.9%) of the controls, and about 38 (47.5%) of the cases and 74 (43%) of the controls had a formal job. Five percent of cases had a monthly income of between $ 200-100, and 60 (34.9%) of the controls had a monthly income exceeding. Among patients with tuberculosis, 44(55%) and in the control group 22 (12.8%) had a history of hospitalization. The history of addiction and imprisonment was between 24(30%) and 21 (26.3%) in the cases, while in the control group, it was 2 (1.2%) and 1 (0.6%), respectively (Table 1).

In Table 2, five machine learning models, SVM, RF, LSSVM, KNN, and NB, are used to detect 39 factors affecting tuberculosis. As can be seen in this table, the researcher compared these five models in terms of accuracy(ACC), area under the curve(AUC), negative predictive value (NPV), positive predictive value (PPV), specificity, and sensitivity. Except for RF, in other models, the best value is highlighted. Accuracy has been used to compare these models. The four models, SVM, KNN, LSSVM, and NB, had the highest accuracy of 0.99, 0.99, 0.97, and 0.95, respectively.

Table 3 shows the significance of the variables that are most important for tuberculosis according to the Naïve Bayes model.This model showed that among social contacts, first-class family (20.83*%*) and friend (17.01*%*) and spatial contacts,health center (41.67*%*), hospital (24.74*%*), store (18.49*%*), market (14.32*%*), and workplace (9.46*%*)

Table 1. Patients' characteristics in the case and control groups

Variable		Cases (N=80)	Controls (N=172)	*P*-value
Age, mean (SD)		34.1 (±15.3)	32.5 (±12.0)	0.080
Sex, n (%)	Male	62 (77.5%)	89 (51.7%)	<0.001
	Female	18 (22.5)	83 (48.3)	
BMI, n (%)	Underweight	22 (27.5)	9 (5.2)	<0.001
	Normal	45 (56.3)	114 (66.3)	
	Overweight	9 (11.3)	30 (17.4)	
	Obese	4 (5.0)	19 (11.0)	
Marital status, n (%)	Single	24 (30.0)	47 (27.3)	0.620
	Married	49 (61.3)	113 (65.7)	
	Divorced	3 (3.8)	10 (5.8)	
	Widowed	4 (5.0)	2 (1.2)	<0.001
Educational status, n (%)	Illiterate	16 (20)	2 (1.2)	
	Read and Write	13 (16.3)	3 (1.7)	
	Up to elementary	23 (28.7)	31 (18.0)	
	Secondary school	15 (18.7)	47 (27.3)	
	College or more	13 (16.3)	89 (51.7)	
Employment Status, n (%)	Employed	38 (47.5)	74 (43.0)	<0.001
	Unemployed	42 (52.5)	98 (57.0)	
Size of the family, n (%)	< 2	25 (31.2)	21 (12.2)	<0.001
	2-4	48 (60.0)	133 (77.3)	
	>4	7 (8.8)	18 (10.5)	
Monthly Income, n (%)	Less than $100	24 (30.0)	21 (12.2)	0.040
	$100-200	26 (32.5)	59 (34.4)	
	$200-300	18 (22.5)	32 (18.6)	
	More than $300	12 (15.0)	60 (34.9)	
Cigarette smoking, n (%)	Yes	34 (42.5)	31 (18.0)	<0.001
	No	46 (57.5)	141 (82.0)	
Duration of Cigarette smoking (year), n (%)	<1	7 (8.8)	12 (6.9)	<0.001
	1-3	10 (12.5)	10 (5.8)	
	>3	16 (20.0)	9 (5.2)	
Hookah smoking, n (%)	Yes	10 (12.5)	29 (16.9)	0.372
	No	70 (87.5)	143 (83.1)	
Duration of hookah smoking, n (%)	<1	1 (1.3)	5 (2.9)	0.721
	1-3	2 (2.5)	5 (2.9)	
	>3	8 (10.0)	19 (11.1)	

[DOI: 10.47176/mjiri.38.5]

Table 1. Continued

Variable		Cases (N=80)	Controls (N=172)	P-value
History of chronic illness, n (%)	Yes	50 (62.5)	54 (31.3)	<0.001
	No	30 (37.5)	118 (68.6)	
History of addiction, n (%)	Yes	24 (30.0)	2 (1.2)	<0.001
	NO	56 (70.0)	170 (98.8)	
History of imprisonment, n (%)	Yes	21 (26.3)	1 (0.6)	<0.001
	No	59 (73.7)	171 (99.4)	
History of hospital admission, n (%)	Yes	44 (55.0)	22 (12.8)	<0.001
	No	36 (45.0)	150 (87.2)	
Contact Type, n (%)	First-class family	421 (30.0)	667 (29.7)	<0.001
	Second-class family	245 (17.5)	502 (22.3)	
	Third-class family	85 (6.0)	221 (9.8)	
	Freind	380 (24.5)	410 (20)	
	Co-Worker	215 (14)	480 (24)	
	Neighbor	70 (2.5)	120 (6)	
Place of Contact, n (%)	Market	231 (22.8)	255 (16.6)	<0.001
	Store	178 (17.6)	254 (16.5)	
	Hospital	51 (5.0)	67 (4.3)	
	Health center	150 (14.8)	277 (17.9)	
	Workplace	201 (19.9)	529 (34.4)	
	Restaurant	30 (2.9)	67 (4.3)	
	Park	26 (2.6)	33 (2.1)	
	mosque	55 (5.4)	96 (6.2)	
	Basij base	12 (1.2)	21 (1.4)	
	Hairdresser's	62 (6.1)	140 (9.0)	
	school	13 (1.3)	54 (3.5)	

Table 2. A comparison of the Five applied Machine Learning techniques using the accuracy measures

Tools	Set	Methods Tool (95% confidence interval)					
		Sensitivity	Specificity	Positive predictive value (PPV)	Negative predictive value (NPV)	The area under the curve (AUC)	Accuracy (ACC)
RF	Train	0.80 (0.79-0.80)	0.64 (0.63-0.64)	0.95 (0.94-0.95)	0.28 (0.27-0.29)	0.78 (0.78-0.78)	0.72 (0.71-0.72)
RF	Test	0.76 (0.75-0.76)	0.66 (0.59-0.74)	0.99 (0.99-0.99)	0.04 (0.03-0.05)	0.75 (0.75-0.76)	0.71 (0.67-0.75)
SVM	Train	1.00 (1.00-1.00)	0.98 (0.98-0.99)	0.99 (0.99-0.99)	0.94 (0.92-0.96)	1.00 (0.99-1.00)	0.99 (0.99-0.99)
SVM	Test	1.00 (1.00-1.00)	0.99 (0.99-1.00)	1.00 (0.99-1.00)	0.94 (0.92-0.96)	1.00 (1.00-1.00)	0.99 (0.99-0.99)
LSSVM	Train	0.98 (0.96-0.99)	0.97 (0.96-0.98)	0.99 (0.98-0.99)	0.95 (0.92-0.98)	0.98 (0.96-0.99)	0.97 (0.96-0.99)
LSSVM	Test	0.98 (0.96-0.99)	0.97 (0.96-0.98)	0.99 (0.98-0.99)	0.95 (0.92-0.98)	0.97 (0.96-0.98)	0.97 (0.96-0.98)
NB	Train	0.99 (0.99-1.00)	0.98 (0.98-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)	0.99 (0.99-0.99)
NB	Test	0.97 (0.96-0.97)	0.94 (0.92-0.96)	0.97 (0.96-0.98)	0.93 (0.91-0.94)	0.96 (0.95-0.97)	0.95 (0.94-0.97)
KNN	Train	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)	1.00 (1.00-1.00)
KNN	Test	1.00 (1.00-1.00)	0.94 (0.92-0.96)	0.99 (0.99-1.00)	1.00 (1.00-1.00)	1.00 (0.99-1.00)	0.99 (0.99-1.00)

were of the highest importance. Among the demographic and other variables related to tuberculosis, history of hospital admission (51.82%), BMI (43.75%), sex (40.36%), age (22.83%), education (60.59%), employment status (43.58%), income (63.80%), addiction (44.10%), history of imprisonment (38.19%) had the greatest impact on tuberculosis.

Discussion

This study aimed to investigate the risk factors affecting tuberculosis using machine learning techniques.In the statistical analysis of our paper, five classification methods of RF, SVM, LS-SVM, NB, and KNN were compared using Accuracy criteria. Except for RF, other classification methods performed well. In this study, factors such as first-class family, friend, health center, hospital, store, market, workplace, history of hospital admission, BMI, sex, age, educational status, employment status, income, addiction, and history of imprisonment were of the highest importance in relation to tuberculosis.

The high incidence of tuberculosis in many countries can be ascribed to the country's socio-demographic, economic, and environmental characteristics, such as poverty, lack of knowledge, attitudes, and practices, overcrowding in various places, malnutrition, comorbidities, etc (15). A comprehensive understanding of epidemiological factors is crucial in developing national policy and directing health resources to control TB transmission and guarantee improved patient management. Since there is no single factor that can be fully ascribed to the occurrence of tuberculosis and there is a gap in information about the factors affecting the incidence of tuberculosis, it has been tried to examine various demographic, social, economic, and environmental factors in this study. Socio-demographic characteristics of the study participants showed that the mean age of patients was 34 years old (range 19-49) and also for the control group was 32 (range 20-44). In general, the findings are in accordance with other studies that have documented a rapid

[DOI: 10.47176/mjiri.38.5]

Table 3. The importance of variables based on Machine Learning methods

Variable	RF Mean Decrease Accuracy	RF Mean Decrease Gini	Naïve Bayes	LSSVM (Standardized Importance)	KNN (Attribute Evaluation)	SVM
Health center	0.006	2.094	41.67	1.82	-0.001	7.47
Hospital	0.000	0.472	24.74	1.82	-0.002	0.12
Store	0.001	0.970	18.49	0.00	0.005	8.26
Market	-0.001	0.643	14.32	0.00	0.009	6.99
Hairdresser's	0.000	0.482	7.03	0.00	0.002	1.05
Restaurant	0.000	0.143	0.61	0.00	0.002	7.01
Park	0.000	0.007	2.95	0.00	0.000	4.88
Workplace	0.000	0.192	9.46	0.00	0.001	0.88
Basij base	0.000	0.050	2.86	0.00	0.000	1.52
mosque	0.000	0.257	6.94	0.00	0.000	0.50
school	0.000	0.032	1.74	0.00	0.000	0.47
First-class family	0.000	0.296	20.83	0.00	0.000	1.29
Second class family	0.000	0.384	1.13	0.00	0.001	1.50
Third class family	0.001	0.540	0.17	0.00	-0.003	0.22
Freind	0.001	0.480	17.01	0.00	0.001	7.52
Co-Worker	0.001	0.458	8.25	0.00	0.004	0.09
Neighbor	0.000	0.297	8.33	0.00	0.000	5.53
BMI	0.002	1.077	43.75	0.00	0.001	0.30
sex	0.001	0.272	40.36	0.00	0.000	1.10
age	0.002	1.559	22.83	0.00	0.003	10.81
Marital status	0.000	0.218	4.51	0.00	0.000	0.43
Educational status	0.008	2.820	60.59	0.00	0.014	1.23
Employment Status	0.006	2.102	43.58	0.00	0.011	0.65
Size of the family	0.002	0.629	6.42	0.00	0.002	0.34
Monthly Income	0.001	0.645	28.13	0.00	0.001	0.55
Cigarette smoking	0.000	0.432	63.80	0.00	-0.002	0.79
Duration of Cigarette smoking (year)	0.000	0.361	3.13	0.00	0.000	0.09
heart attack	0.000	0.070	5.90	0.00	0.000	0.13
stroke	0.000	0.024	2.95	0.00	0.000	0.03
Diabetes	0.000	0.165	6.34	0.00	0.003	0.12
Kidney failure	0.000	0.102	4.77	0.00	0.001	0.48
Chronic Pulmonary	0.000	0.023	1.13	0.00	0.000	0.07
cancer	0.000	0.071	5.90	0.00	-0.001	0.07
HIV	0.000	0.000	4.77	0.00	0.000	0.35
Hepatitis C	0.000	0.014	7.73	0.00	0.000	0.20
History of addiction	0.008	1.917	44.10	0.00	0.011	1.45
History of imprisonment	0.003	0.945	38.19	0.00	0.007	1.80
History of hospital admission	0.009	2.297	51.82	0.00	0.006	1.46

rise in morbidity and mortality due to TB in this young adult population, mainly between 18 and 51 years old . The elevated probability of infection in this age group is related to the increased number of social contacts in the community during adolescence. This study showed that the majority (55.5%) were male and (45.5%) were female. A similar finding was observed in other studies in which 60.5% and 57.5% of the participants were male and female (16). In this study, a history of addiction and imprisonment were more reported among patients than in the control group.Other studies have reported similar results. In a study by Gabriel et al. in Malaysia, the results showed that the imprisonment history was 88% among patients and 36% among healthy individuals (17). In this study, monthly household income for cases was lower than in the control group, which was consistent with a study by Wondemagegn et al. (16). In developing countries, the majority of impoverished families are confronted with financial constraints that result in poverty, malnutrition, poor health, overcrowding, reduced attitudes toward health care, and the cycle of an agent-host environment that is vicious and increases the risk of infectious diseases such as tuberculosis. Overcrowding in different places can be a strong risk factor for tuberculosis (18).

Patients with a history of previous hospitalization were more than four times more likely to develop tuberculosis than patients without a history of previous hospitalization. This indicates that visiting health centers is a risk factor for tuberculosis, which suggests the necessity of establishing a robust infection control strategy in health centers. As per a longitudinal study, the hospital-acquired infection incidence rate was 28.15% (95%C.I:24.40, 32.30) per 1000 patient days, whereas the overall prevalence was 19.41% (95%C.I:16.97–21.85), Furthermore, pneumonia and other respiratory tract infections were ranked among the top ten diseases (19). Therefore, the transmission of the infection is common in the hospital, and people with a history of hospitalization have a higher chance of being infected during their stay. People who have been admitted for a longer period or more frequently are at a higher risk (20).

Other results of this study showed that different places such as health centers, hospitals, workplaces, markets, and shops had the highest association with tuberculosis and may be potential sites of transmission. The categorical analysis showed that healthcare locations are the first junctions and are undoubtedly important locations for infection control. Our findings are substantiated by previous studies that have identified healthcare settings as high-risk environments for the transmission of TB. These findings highlight

[DOI: 10.47176/mjiri.38.5]

opportunities for health ministries to decrease community exposure to transmission risk by implementing environmental improvements, such as enhanced ventilation and UV irradiation at these venues (21). Reducing transmissibility in TB hotspots by targeting environmental interventions in areas where the risk of exposure to the disease is high may have significant community advantages. The workplace has been reported as another important contact point for the community,and possible causes may be poor ventilation, long-term and close exposure to infections. The results of our study are similar to the studies reviewed by Chamie et al. (22). According to the findings, implementing environmental interventions throughout the visited sites and avoiding focus on areas with low risk of TB exposure is necessary (23).

Among social contacts between individuals, family history among first-class relatives and close friends was of the most importance.This is similar to a study by Shimeles et al., the results of which showed that among the sick people who were exposed to patients with active TB, 18.1% were exposed to a family member and 7.3% to friends who were infected with TB (24).

Strength of the study

First, standard checklists are used, and interviews are carefully conducted by the researcher. Second, this study used appropriate statistical tools and techniques to find the relationship and impact of selected demographic, socio-economic, and environmental factors.

Limitations of this study

First, proxy interviews were required for some participants because they were either very ill or poorly educated. Second, we obtained information from participants about one month before the start of the study. Thus, recall bias is inevitable and potentially affects the results.

Conclusion

This study employed machine learning methods to identify the effective economic, social, and environmental factors in tuberculosis. The researcher used five models, RF, SVM, LS-SVM, NB, and KNN, and except for RF, the rest of the models had the highest accuracy. In the final analysis, the study identified 16 key risk factors for tuberculosis, including BMI, sex, age, educational status, employment status, income, addiction, history of imprisonment, history of hospital admission, as well as contact with first-class family and friends and various places including a health center, hospital, store, market, workplace. Hence, it is essential that TB control efforts incorporate a strategy that prioritizes socio-economic concerns such as overcrowding and poverty. Furthermore, preventing the transmission of TB requires significant measures to control infection at the level of healthcare facilities and other places.

Acknowledgments

We express our gratitude to all the observers who have provided invaluable assistance in the execution of this project.

Ethics approval and Consent to participate

The Ethics Committee of Ahvaz Jundishapure University of MedSciencesences (IR.AJUMS.REC.1400.2 40) approved the present study that was part of a PhD thesis. The relevant guidelines and regulations of the Helsinki Declarations were followed by all methods. All participants were informed and provided with written consent/assent.They were assured that the information gathered would be confidential. Unique study identifiers were used to anonymize all collected data.

Conflict of Interests

The authors declare that they have no competing interests.

References

1. Khan A, Marks S, Katz D, Morris SB, Lambert L, Magee E, et al. Changes in Tuberculosis Disparities at a Time of Decreasing Tuberculosis Incidence in the United States, 1994–2016. Am J Public Health. 2018;108(S4):S321-S6.
2. World Health Organization (2020): Global Tuberculosis Report. https://www.who.int/publications/i/item/9789240037021.
3. Shaikh MA, Malik NA. Spatial cluster analysis of new and relapsed cases of pulmonary tuberculosis by district: pakistan 2015. J Ayub Med Coll Abbottabad. 2019;31(2):293-5.
4. Tavakoli A. Incidence and prevalence of tuberculosis in Iran and neighboring countries. Zahedan. J. Res. Med. Sci.. 2017;19(7).
5. Arsang-Jang S, Mansourian M, Amani F, Jafari-Koshki T. Epidemiologic Trend of Smear-Positive, Smear-Negative, Extra Pulmonary and Relapse of Tuberculosis in Iran (2001-2015); A Repeated CrossSectional Study. Res J Health Sci. 2017;17(2):380.
6. Hagiya H, Koyama T, Zamami Y, Minato Y, Tatebe Y, Mikami N, et al. Trends in incidence and mortality of tuberculosis in Japan: a population-based study, 1997–2016. Epidemiol Infect.2019;147.
7. Ohene SA, Bakker MI, Ojo J, Toonstra A, Awudi D, Klatser P. Extra-pulmonary tuberculosis: A retrospective study of patients in Accra, Ghana. PloS One. 2019;14(1):e0209650.
8. Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, et al. The transmission of Mycobacterium tuberculosis in high burden settings. Lancet Infect Dis. 2016;16(2):227-38.
9. Yates T, Tanser F, Abubakar I. Plan Beta for tuberculosis: it's time to think seriously about poorly ventilated congregate settings. Int J Tuberc Lung Dis. 2016;20(1):5-10.
10. Balogun OS, Olaleye SA, Mohsin M, Toivanen P. Investigating machine learning methods for tuberculosis risk factors prediction: a comparative analysis and evaluation. Proceedings of the 37th Int Bus Manag (IBIMA). 2021.
11. Amoori N, Amini P, Cheraghian B, Alavi SM. Investigating the intensity of social contacts associated with tuberculosis: a weighted networks model. BMC Pulm Med. 2023;23(1):1-8.
12. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer. Behav Ther. 2020;51(5):675-87.
13. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PloS one. 2019;14(11):e0224365.
14. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann; 2016.
15. Organization WH. WHO global lists of high burden countries for tuberculosis (TB), TB/HIV and multidrug/rifampicin-resistant TB (MDR/RR-TB), 2021–2025: background document. 2021.
16. Mulu W, Mekkonnen D, Yimer M, Admassu A, Abera B. Risk factors for multidrug resistant tuberculosis patients in Amhara National Regional State. Afr Health Sci. 2015;15(2):368-77.
17. Culbert GJ, Pillai V, Bick J, Al-Darraji HA, Wickersham JA, Wegman MP, et al. Confronting the HIV, tuberculosis, addiction, and incarceration syndemic in Southeast Asia: lessons learned from Malaysia. J NeuroImmune Pharmacol. 2016;11(3):446-55.
18. Srivastava K, Kant S, Verma A. Role of environmental factors in transmission of tuberculosis. Dynamics of Human Health. 2015;2(4):12.
19. Ali S, Birhane M, Bekele S, Kibru G, Teshager L, Yilma Y, et al. Healthcare associated infection and its risk factors among patients

[DOI: 10.47176/mjiri.38.5]

admitted to a tertiary hospital in Ethiopia: longitudinal study. Antimicrob Resist Infect Control. 2018;7(1):1-9.
20. Yallew WW, Kumie A, Yehuala FM. Point prevalence of hospital-acquired infections in two teaching hospitals of Amhara region in Ethiopia. Drug Healthc Patient Saf. 2016;8:71.
21. Zelner JL, Murray MB, Becerra MC, Galea J, Lecca L, Calderon R, et al. Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. J Infect Dis. 2016;213(2):287-94.
22. Chamie G, Wandera B, Marquez C, Kato-Maeda M, Kamya MR, Havlir DV, et al. Identifying locations of recent TB transmission in rural Uganda: a multidisciplinary approach.Trop Med Int Health. 2015;20(4):537-45.
23. Chamie G, Kato-Maeda M, Emperador DM, Wandera B, Mugagga O, Crandall J, et al. Spatial overlap links seemingly unconnected genotype-matched TB cases in rural Uganda. PloS one. 2018;13(2):e0192666.
24. Yan J, Fan JG, Jing P, Ke W, Zhang PY, Wang HQ, et al. Risk of active pulmonary tuberculosis among patients with coal workers' pneumoconiosis: a case-control study in China. Biomed Environ Sci. 2018;31(6):448-53.

[DOI: 10.47176/mjiri.38.5]