EAI Endorsed Transactions
on Intelligent Systems and Machine Learning Applications Research Article EAI.EU

A Machine Learning-based approach to predicting tuberculosis in the Democratic Republic of Congo

Tshibanda wa Tshibanda Pierre[1], Boluma Mangata Bopatriciat[2,*], Mbombo Kabongo Marina[3] and Mbiya Mpoyi Guy-Patient[4]

[1, 3, 4]Department of Computer Science, Institut Superieur Pédagogique de la Gombe, Kinshasa, DR Congo
[2]Department of Computer Science, Institut Superieur Pédagogique de la Gombe, Kinshasa, DR Congo

Abstract

INTRODUCTION: Tuberculosis remains a public health problem in Democratic Republic of Congo (DRC), despite advances in Machine Learning for the prediction of this disease. However, existing models are often adapted to Asian contexts and do not take into account the specific epidemiological and social characteristics of the DRC. Given this shortcoming, our study explores a Machine Learning approach specifically designed to improve the prediction of tuberculosis in the Congolese population.
OBJECTIVES: Our problem is based on the following question: "What approach, based on Machine Learning and specific to the population of DRC, is likely to improve the prediction of tuberculosis?" To answer this, we adopted an exploratory paradigm with a sequential mixed design (qualitative and quantitative). The study was conducted on a sample of 1505 patients and six healthcare professionals in the health zones of Lubumbashi and Nzanza.
METHODS: The data was collected using questionnaires and semi-structured interviews, then analysed using bivariate and multivariate approaches.
RESULTS: The results show that incorporating Congolese specificities into Machine Learning models significantly improves the prediction of tuberculosis. Of the models tested, Random Forest and Decision Tree performed best in terms of precision, recall, F1-score and AUC, while Voting Classifier, Stacking and Adaboost showed a good compromise between precision and robustness.
CONCLUSION: This study highlights the need to develop predictive models adapted to the local context in order to improve tuberculosis control in DRC. We propose an optimised model incorporating characteristics specific to the Congolese population, with a possible large-scale application to improve detection and prevention of the disease.

Keywords: machine Learning, tuberculosis, tuberculosis control in DRC, automatic prediction, automatic learning

Received on 11 April 2025, accepted on 5 June 2025, published on 16 July 2025

doi: 10.4108/eetismla.9073

1. Introduction

Tuberculosis is an infectious and contagious disease caused by mycobacterium tuberculosis or Koch's bacillus (Yombi et al., 2015). The disease is very widespread in sub-Saharan Africa, especially as the health systems in place are very weak and the population has a low standard of living (WHO, 2021). Transmission is always direct, from the bacilliferous patient to the receptive subject, via the air, due to the bacilli contained in the air, in the droplets of suspended saliva emitted by the patient (WHO, 2020).

*Corresponding author. Email: bopatriciat.boluma@unikin.ac.cd

Tuberculosis is also a major global public health problem. According to a report by the World Health Organisation (WHO), 10.4 million cases of tuberculosis have been recorded worldwide, 90% of them in adults, 65% of them male, and 10% living with HIV (Dagnra et al, 2011). According to a report by the World Health Organisation, Africa has the highest death rate in the world, estimated at 81 per 100,000 inhabitants. Every year, almost 440,000 people contract multi-drug-resistant tuberculosis (MDR-TB) and 150,000 die from this form of the disease. Its treatment is difficult and costly because of poor response to conventional treatment with first-line drugs. Cure rates for multidrug-resistant tuberculosis are low (between 50% and 70%). With 11% of the world's population, Africa alone accounts for 27% of the global burden of tuberculosis. The incidence of tuberculosis is rising by 6% every year, and the HIV epidemic is the main cause of this increase. Between 30 and 50% of tuberculosis patients in Africa are co-infected with HIV (WHO, 2021). According to the national tuberculosis control programme of the Ministry of Health of the Democratic Republic of Congo (DRC), the DRC is one of the 22 most affected countries in the world, ranking fifth in Africa and 11th in the world. The incidence of microscopy-positive pulmonary tuberculosis (TB) in the DRC is estimated at more than 160 cases per 100,000 inhabitants. The DRC is also one of the countries with the highest number of patients co-infected with TB and HIV/AIDS (Bemba et al., 2020).

Despite some progress, the Democratic Republic of Congo (DRC) continues to be one of the countries hardest hit by tuberculosis in the world. The country has one of the highest rates of tuberculosis in sub-Saharan Africa. The disease is widespread, affecting every province in the DRC, with high prevalence among adults and children (WHO, 2020). Several factors contribute to the spread of TB in the DRC. Poverty, lack of access to quality healthcare, overcrowding, precarious living conditions, HIV/AIDS and malnutrition are all major risk factors. In addition, the country's health system is often deficient, leading to delays in diagnosing and treating the disease. The treatment of tuberculosis in the DRC is also a challenge. The country faces shortages of anti-tuberculosis drugs, which makes it difficult to treat patients. In addition, the low success rate in treating multidrug-resistant tuberculosis is a cause for concern.

Efforts are being made to combat tuberculosis in the DRC. National TB control programmes have been set up in collaboration with international partners such as the World Health Organisation (WHO) and the Global Fund to Fight AIDS, Tuberculosis and Malaria. These programmes aim to improve early case detection, access to drugs, treatment adherence and patient follow-up (Birungi et al., 2019).

The government of DR Congo and various national and international organisations have drawn up strategies to combat this disease and reduce its impact on the population (Badr, 2022). According to the statistics, the figures are extremely alarming, given that DR Congo is the 9th country in the world most affected by tuberculosis, and 2nd in Africa. In 2021, during the Covid-19 pandemic, the National Tuberculosis Control Programme (PNLT) reported a total of 216,690 cases of all forms of tuberculosis, many of them children (26,550). Tuberculosis is wreaking havoc in every corner of the country. The city of Kinshasa is the worst affected, with almost 30,000 cases. It is followed by Haut Katanga (around 16,000 cases), Kwilu (around 13,000 cases) and Sud-Kivu (around 12,000 cases). However, the disease is curable and preventable. Among other effective strategies to combat TB, the government has developed a policy and regulatory framework, one of the aims of which is to put in place policies and strategic plans to combat TB, in collaboration with the World Health Organisation and other partners. These documents provide a roadmap for action on the prevention, detection and treatment of tuberculosis in the Democratic Republic of Congo.

1.1. Contribution

We have highlighted the use of advanced machine learning techniques to predict tuberculosis in the Democratic Republic of Congo (DRC). Beyond the empirical results, this research has made substantial contributions in several key areas. Indeed:

This study is part of an innovative paradigm, exploiting advanced artificial intelligence and machine learning techniques to improve the prediction and management of tuberculosis in the Democratic Republic of Congo (DRC) (Mate Landry et al., 2024). This theoretical foundation is a major contribution, opening up new perspectives in the fight against this endemic disease in the country.

In terms of predictive modelling, we have designed robust models capable of accurately predicting the risk of patients developing tuberculosis. By exploring various machine learning algorithms, such as deep learning, decision trees and random forests, as well as set models, we were able to identify the most effective approaches in the specific context of the DRC. This approach has helped to enrich theoretical knowledge of the complex mechanisms involved in the development of tuberculosis. The study also set out to identify the main determinants of tuberculosis in the country. By analysing the relative importance of socio-demographic, clinical and environmental variables, it highlighted the most important risk factors. This in-depth understanding of the multifactorial determinants of the disease in the context of the DRC represents a major theoretical advance.

In terms of management, the study demonstrated the potential of machine learning methods for developing clinical decision support tools. It proposed personalised screening and management strategies based on predictive models. This innovative approach, combining artificial intelligence and medical expertise, opens up new prospects

for improving public health in the DRC. In our study, we adopted an approach based on adapting the models to the specific context of the DRC. We took account of the country's specific socio-economic, cultural and epidemiological features, in order to identify the contextual factors influencing the prediction of tuberculosis.

Taken as a whole, this study has made substantial theoretical contributions in the areas of predictive modelling, identification of the determinants of tuberculosis, innovation in management, adaptation to the local context and capacity building in artificial intelligence applied to health in the Democratic Republic of Congo.

From a methodological point of view, our study on the prediction of tuberculosis in the Democratic Republic of Congo (DRC) adopted an innovative approach, combining advanced artificial intelligence and machine learning techniques. This methodological approach is in itself an important contribution, opening up new prospects in the fight against this endemic disease. Firstly, the rigorous selection and integration of various data sources, including socio-demographic, clinical and environmental information, has enabled the construction of a rich dataset that is representative of the DRC context. This methodological approach provided a holistic understanding of the risk factors for tuberculosis, overcoming the limitations of traditional studies focusing on a limited number of variables.

In terms of predictive modelling, the study used a diverse range of machine learning algorithms. This methodology made it possible to identify the best-performing approaches in the specific context of the DRC, while assessing their robustness and generalizability. The use of optimisation and cross-validation techniques enhanced the reliability and reproducibility of the models developed. In addition, the study implemented a multi-criteria evaluation approach for predictive models, considering not only traditional performance metrics, but also indicators adapted to the clinical context, such as sensitivity, specificity and positive predictive value. This rigorous methodology has enabled us to select the most relevant models for the development of public health decision-making tools.

From a practical point of view, this study has major practical implications for the fight against tuberculosis in the DRC:

Improved early detection: The predictive models developed will make it possible to accurately identify people at risk, thereby encouraging the introduction of targeted screening programmes and early detection of the disease. This will help to reduce the delay in diagnosis, one of the main challenges in the DRC.

Personalised care: Clinical decision support algorithms, adapted to the local context, will guide healthcare professionals in developing personalised treatment plans. This will optimise resource allocation and improve the quality of care.

Enhanced prevention: A deeper understanding of the determinants of tuberculosis will guide the design of more targeted and relevant prevention and health promotion programmes at community level.

Successful local integration: The approach of adapting the models to the context of the DRC, by closely involving local stakeholders, will encourage the appropriation and sustainable integration of the tools developed in the health systems.

Capacity-building: Developing skills in artificial intelligence applied to healthcare within local medical and research teams will ensure the sustainability of expertise and autonomy in exploiting these innovations.

2. Methodology

The machine learning approach is the methodology proposed to achieve the objective of predicting tuberculosis in the DRC. This approach is based on the use of historical data to train a predictive model, which can then be used to anticipate future cases of tuberculosis.

2.1. Sample

To determine the sample size Te, there are two approaches: From a proportion and from an average (Jain et al. 2015.) We will use the proportion approach according to the following: To calculate the required sample size based on a confidence level of 95% and a margin of error of ±5%, as well as the number of samples already interviewed (1505 patients), we used the following formula to estimate the required sample size:

$$n = \frac{Z^2 * p * (1-p)}{E^2} \quad (1)$$

Where:

- n = sample size required
- Z = critical value associated with the confidence level (for a confidence level of 95%,Z is equal to 1.96)
- p= estimated proportion of the characteristic in the population (0.5 is often used to maximise the sample size, but if you have precise data on the prevalence of tuberculosis in the DRC, you can use them)
- E = margin of error (in this case, ±0.05)

Using the parameters you have provided:

- Z=1,96
- p=0.5 (conservative value to maximise sample size)
- E=0,05

We can calculate the required sample size as follows:

$$n = \frac{(1{,}96)^2 * 0{,}5 * (1 - 0{,}5)}{(0{,}05)^2}$$
$$n = \frac{3{,}8416 * 0{,}25}{0{,}0025}$$
$$n = \frac{0{,}9604}{0{,}0025}$$
$$n = 384{,}16$$

Since the calculated sample size required is 384.16, this means that to achieve a confidence level of 95% with a margin of error of ±5%, we will need a sample of:

$$n = \frac{384{,}16 * 95}{100}$$
$$n = \frac{36495{,}2}{100}$$
$$n = 364{,}9$$

We therefore need a sample of at least 365 people.
We obtained responses from a sample of 1,505 patients in two (2) health zones in the Democratic Republic of Congo (Lubumbashi and Nzanza), in the provinces of Haut-Katanga and Central Kongo. Depending on the availability of health personnel, we obtained two (02) pneumologists, two (02) radiologists, one (01) general practitioner and one (01) nurse, making a total of six (06) health personnel available for the study.

2.2. Data collection and tools

The effectiveness of research depends to a large extent on the quality of the data collected. For this study, various data collection tools and methods were used to ensure reliable and representative results. The techniques chosen were adapted to the research objectives and the specific context of the study, taking into account the characteristics of the target population. We opted for a combination of qualitative and quantitative methods, ranging from questionnaire surveys to semi-structured interviews, in order to capture a wide range of information. The questionnaires were designed to obtain precise quantitative data, while the interviews allowed us to explore participants' perceptions and experiences in depth. In addition, digital tools such as data analysis software were integrated to facilitate the processing and interpretation of the information collected. This multi-method approach makes it possible to triangulate the data, ensure the validity of the results and enrich our understanding of the subject studied.

2.3. Data collection sources

A data source refers to the origin from which information is collected as part of a research or study. These sources may include official documents, administrative records, field surveys, interviews with experts or direct observation. Data from these sources is used to analyse, interpret and draw conclusions in scientific research. Data sources therefore indicate where and how information will be collected in order to measure the chosen indicators (Chapman & Rich, 2019). Overall, we were interested in 2 data sources, namely:

a) Primary data, which is data that we collect ourselves using various instruments, such as interviews with source persons, surveys, group discussions and observations.
b) Secondary data, which are data obtained from other pre-existing sources, such as the national census or data from surveys carried out by partners, donors or the government.

Data collection took place from January to June 2024, ensuring the freshness of the results and their relevance to the current health contexts in the DRC.

3. Results

Tuberculosis remains a major public health challenge in the Democratic Republic of Congo (DRC), with one of the highest disease burdens in the world. However, nationwide screening and diagnosis of this infection is fraught with difficulties linked to the traditional methods used. Firstly, the sensitivity of existing tests, such as sputum microscopy, is relatively low, allowing only 50-80% of TB cases to be detected. This large margin of error means that many patients with the disease escape diagnosis and do not receive appropriate treatment.

What's more, the specificity of these tests is not optimal, leading to a risk of false positives, particularly in people co-infected with HIV. What's more, the time taken to obtain a result using these traditional methods, which can take up to 2-3 days, often results in detrimental delays in the management of patients suspected of having tuberculosis. This diagnostic delay is all the more problematic given that extra-pulmonary forms of the disease, which account for a large proportion of cases in the DRC, are more difficult to detect using these techniques. In addition, screening for tuberculosis in children poses specific challenges, as it is particularly difficult to obtain sputum samples from this fragile population. As a result, traditional methods perform poorly in diagnosing paediatric TB.

Finally, screening coverage for tuberculosis in the DRC remains limited, due to logistical constraints and the uneven distribution of laboratories across the country. As a result, many people at risk do not have access to diagnostic services, hampering efforts to combat the disease. Faced with these multiple challenges, it is clear that traditional diagnostic methods are showing their limitations in the DRC context. Strengthening screening and diagnostic capacity, in particular through the use of more effective technologies, is imperative if we are to improve the care of

TB patients and help reduce the burden of the disease in the country. The aim of this study was to develop and evaluate machine learning models for the prediction of tuberculosis in the Democratic Republic of Congo (DRC), in response to an urgent need in the face of this highly prevalent disease. The study population consisted of 1505 people. The frequency distribution shows that men represent the majority of participants with 74.58%, followed by women at 25.31%. Participants with a 'Null' value for gender represented a very small proportion of 0.13%.

The models were trained and tested using a dataset of 1505 patients, collected between January and June 2024, in the Lubumbashi and Nzanza health zones. After cleaning the data, we applied 10-fold cross-validation to assess the performance of the algorithms. The dataset was split into 80% training and 20% testing. Metrics used included precision, recall, F-measure and AUC, allowing a comprehensive evaluation of each model.
Of these 1,505 participants, corresponding to 100% of the workforce, the frequency distribution shows that single people represent the largest proportion of participants at 46.24%, followed by married people at 38.54%. Other marital statuses (widowed, divorced, common-law) represent smaller shares, between 3.19% and 7.97%. Participants with a "Null" value for marital status represent only 0.80% of the total workforce. Two provinces were studied: Central Kongo (DRC) and Haut-Katanga (DRC). The province of Haut-Katanga had slightly more participants than the province of Kongo Central, with 50.10% and 49.90% of the total respectively. These two provinces are virtually equivalent in terms of the number of participants, with a very small difference of 0.20 percentage points between them.

In these 2 provinces we went to four health areas to collect our data: the Kasapa II health area, the Nzanza health area, the Baobab health area and the Lt Mpaka health area. The frequency distribution showed that the Kasapa II health area accounted for the largest proportion of participants (50.10%), followed by Baobab (38.61%). The Lt Mpaka and Nzanza health areas had smaller shares, 6.24% and 5.05% respectively. The age distribution is relatively balanced: the different age groups are well represented, without there being a very high concentration in any one group. Young adults predominate: The 19-25 (14.95%) and 26-35 (19.73%) age groups accounted for the largest proportion of participants, nearly 35% in total. This shows that the population is relatively young. Gradual decrease with age: There is a gradual decrease in numbers as age increases, which is logical and reflects the age pyramid in a population.

Large proportion of older people: The 66-75 age group (7.11%), the 76-85 age group (4.78%) and the 86-95 age group (0.60%) still represent a significant proportion of participants, over 12% in total. This indicates a good representation of older people. High rate of not providing information: 15.15% of participants did not provide their age, which may introduce a bias in the analysis. A higher rate would be desirable. Overall, the age distribution seems to reflect a relatively young population, with a significant presence of older people. However, the rate of non-registration needs to be improved in order to refine the analysis.

Our machine learning algorithms produced the following results:

Table 1: Summary of results

Models	Precision	Recall	F1-score	Accuracy	AUC
KNN	0.9263	0.8888	0.9072	0.9162	0.9143
SVM	0.9333	0.4242	0.5833	07209	0.6992
Decision Tree	0.9368	0.8989	0.9175	0.9255	0.9236
Neural Network	0.9062	0.8787	0.8923	0.9023	0.9006
Random Forest	0.9191	0.9191	0.9191	0.93	0.9266
Voting Classifier	0.9883	0.8585	0.9189	0.9302	0.9250
Stacking	0.9883	0.8585	0.9189	0.9302	0.9250
Bagging	0.9883	0.8585	0.9189	0.9302	0.9250
Adaboost	0.9883	0.8585	0.9189	0.9302	0.9250

The interpretation of the results obtained for the different Machine Learning models tested in this study highlights several important observations:
1. K-Nearest Neighbors (KNN) has a precision of 92.63%, a recall of 88.88% and an F1 score of 90.72%, with an accuracy of 91.62% and an AUC of 91.43%. These values indicate that this model performs well overall, although it may occasionally miss some positive cases due to recall that is slightly lower than its precision.
2. Support Vector Machine (SVM) has a precision of 93.33%, but its recall is particularly low (42.42%), meaning that it detects positive cases poorly. Its F1 score of 58.33% and overall precision of 72.09% show that it is not well suited to this task, notably due to a strong imbalance between precision and recall.
3. Decision Tree achieves solid results with 93.68% precision, 89.89% recall and 91.75% F1-score, meaning that it makes a good compromise between case detection and error reduction. Its accuracy of 92.55% and AUC of 92.36% confirm its robustness.
4. Neural Network shows a precision of 90.62%, a recall of 87.87% and an F1 score of 89.23%, with an accuracy of 90.23%. This indicates that this model performs well, but is slightly inferior to models based on decision trees, probably due to the constraints associated with training the neural network.
5. Random Forest stands out as one of the best models, with 91.91% precision and recall, an F1 score of 91.91% and accuracy of 93%. Its ability to manage the complexity of the data while maintaining a good balance between precision and recall makes it an optimal choice for predicting tuberculosis.

6. The ensemble models (Voting Classifier, Stacking, Bagging and Adaboost) offered the best overall performance, with 98.83% precision, 85.85% recall and an F1-score of 91.89%. The accuracy of 93.02% and the AUC of around 92.50% confirm that they optimise case detection by combining several models, thereby reducing errors and improving generalisation.
Ensemble models such as Voting Classifier, Stacking, Bagging and Adaboost perform best for TB prediction, followed by Random Forest and Decision Tree, which offer a good balance between precision and recall. KNN and Neural Network remain competitive but slightly behind, while SVM underperforms due to its low recall.
Adopting an approach based on ensemble models therefore seems to be the best strategy for optimising TB prediction in the specific context of the DRC.
The accuracy scores of the different models are presented in Figure 1 below to highlight their ability to correctly identify positive cases.

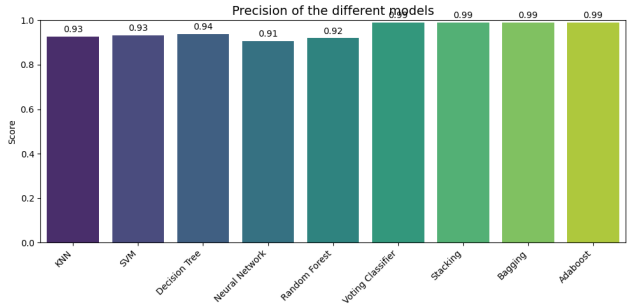

Figure 1: The accuracy scores of the different models

Accuracy measures the proportion of correct positive predictions out of all positive predictions made.
- All the models are highly accurate, with values ranging from 0.9062 (Neural Network) to 0.9883 (Voting Classifier, Stacking, Bagging and Adaboost).
- Models using ensemble techniques (Voting Classifier, Stacking, Bagging, Adaboost) achieve the maximum accuracy of 0.9883, indicating an excellent ability to avoid false positives.
- SVM (0.9333) continues to perform well in terms of precision but is less effective in terms of recall.

Figure 2 below illustrates the recall of each model, reflecting their sensitivity in detecting true positives.

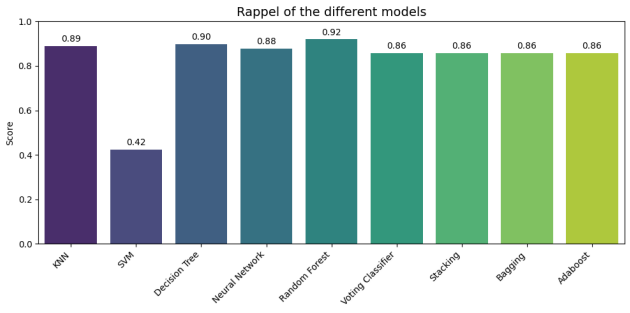

Figure 2: The recall of each model

Recall represents the model's ability to correctly identify all positive instances.
- Random Forest (0.9191) and Decision Tree (0.8989) show a good balance between precision and recall.
- SVM (0.4242) has the lowest recall, indicating that it misses a large proportion of positive cases.
- The ensemble models, although performing well in terms of precision, have a slightly lower recall (0.8585). This suggests that they may miss some positive cases.

The F1-scores shown in Figure 3 below give a balanced view of each model's performance, combining precision and recall.

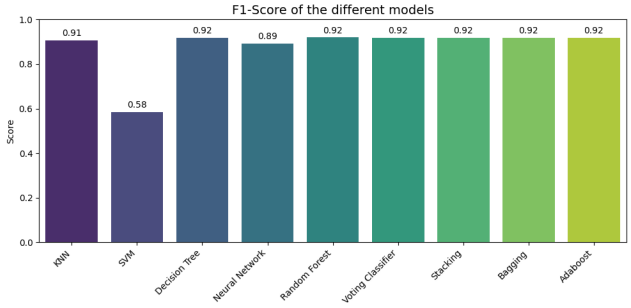

Figure 3: The F1-scores of each model's performance

The F1 score is a harmonic mean between precision and recall, providing a good indicator of the overall balance of a model.
- Random Forest (0.9191) and Decision Tree (0.9175) obtain the best F1 scores among the individual models, indicating a good balance between precision and recall.
- Ensemble models such as Voting Classifier, Stacking, Bagging and Adaboost performed strongly at 0.9189, suggesting that they are effective for global classification.

The accuracy values presented here in figure 4 compare the overall accuracy of the predictions made by the different models.

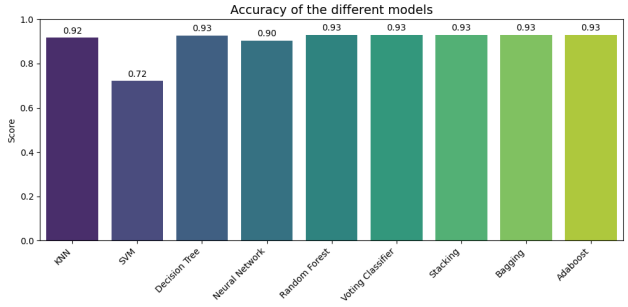

Figure 4: The accuracy of different models

Accuracy represents the total proportion of correct predictions (positive and negative).
- Random Forest (0.93) is the most accurate individual model, closely followed by Decision Tree (0.9255).
- All the ensemble models (Voting Classifier, Stacking, Bagging and Adaboost) achieve an accuracy of 0.9302, proving their robustness.
- SVM (0.7209) has the lowest accuracy, which is consistent with its low recall.

The AUC scores in Figure 5 below assess the ability of the models to distinguish between classes at different decision thresholds.

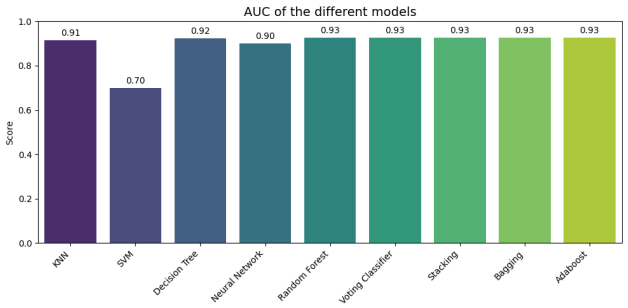

Figure 5: The AUC scores of the models

The AUC measures the model's ability to classify samples between positive and negative classes.
- Random Forest (0.9266) and Decision Tree (0.9236) show the best performance for a single model.
- The ensemble models have very similar performances (0.9250), indicating a strong ability to distinguish between classes.
- SVM (0.6992) performs significantly worse on this metric, which confirms the shortcomings observed in recall.

Generally speaking, the ensemble models (Voting Classifier, Stacking, Bagging, Adaboost) perform best overall, particularly in terms of precision and accuracy. Random Forest and Decision Tree offer a good compromise between precision and recall, which makes them interesting for applications requiring a balance between these two metrics. SVM performs less well, mainly because of its low recall and less competitive AUC. It can be improved by adjusting its hyperparameters or changing its kernel.

4. Discussion

Tuberculosis remains a major public health challenge in the Democratic Republic of Congo (DRC), with one of the highest disease burdens in the world. However, nationwide screening and diagnosis of this infection is fraught with difficulties due to the traditional methods used.
Our machine learning algorithms produced the following results:

KNN (K-Nearest Neighbors): The KNN model shows a precision of 0.9263, accompanied by a recall of 0.8888, resulting in an F1 score of 0.9072. These results indicate that, although KNN is fairly accurate, it may miss some positive cases, as suggested by its slightly lower recall. Its accuracy of 0.9162 and AUC of 0.9143 reinforce its status as a robust model.

SVM (Support Vector Machine): The SVM performs worse, with a precision of 0.9333 but low recall of 0.4242, resulting in an F1 score of only 0.5833. This poor ability to detect positive classes could prove problematic in applications where the cost of a false negative is high. The accuracy of 0.7209 and the AUC of 0.6992 corroborate this mixed conclusion.

The Decision Tree: The Decision Tree model stands out with a precision of 0.9368 and a recall of 0.8989, giving an F1 score of 0.9175. Its accuracy of 0.9255 and AUC of 0.9236 make it a reliable choice. This model appears to successfully balance precision with the ability to identify positive instances.

The Neural Network: Despite having a precision of 0.9062 and a recall of 0.8787, the neural network offers an F1 score of 0.8923. This shows acceptable performance, but it could be improved, especially in terms of its ability to detect positive cases.

Random Forest: The Random Forest model is one of the best performers, with identical precision and recall of 0.9191, which also gives it an F1 score of 0.9191. With an accuracy of 0.9300 and an AUC of 0.9266, it shows superior robustness and generalizability, making it an excellent choice for many classification scenarios.

Ensemble classifiers (Voting, Stacking, Bagging, Adaboost): These ensemble models, whose performance is also noteworthy, all have a precision of 0.9883, indicating that they are extremely effective at avoiding false positives. However, their recall, at around 0.8585, raises questions

about their ability to identify all positive cases. Although they all have an F1 score of 0.9189 and an accuracy of 0.9302, it is essential to bear in mind the need for a balance between precision and recall, especially in critical contexts.

This series of results highlights the performance of the different prediction models tested for the early detection of tuberculosis in the Democratic Republic of Congo (DRC).

Numerous studies have highlighted the key factors influencing the spatial prevalence of tuberculosis in the country. In socio-demographic terms, high population density is recognised as a major risk factor (Marita, 2015), which has shown that densely populated areas in the DRC have higher TB rates. Similarly, poverty and the low socio-economic level of households have been identified as major determinants.

Population movements and migration are also factors that favour the transmission of tuberculosis. Access to and use of healthcare services are also key factors. Inadequate BCG vaccination coverage (Mahamba et al., 2019), diagnostic delays and limited access to care (Kabuya et al., 2016), as well as failures in the healthcare system, all contribute to the high persistence of tuberculosis in the DRC.

Environmental conditions also play an important role. Overcrowding in homes (Muyembe et al., 2020) and lack of access to drinking water and sanitation (Kabuya et al., 2019; Mpoyo et al., 2021) favour the transmission of the disease. Climatic variations and rainfall have also been identified as factors influencing TB prevalence (Nkodo et al., 2018).

Finally, certain factors linked to the disease itself, such as co-morbidities like HIV, diabetes or malnutrition (Mukeba et al., 2020), as well as resistance to anti-tuberculosis drugs, are contributing to the worsening of the epidemiological situation in the DRC.

5. Limitations of the study

This study has certain limitations that should be highlighted. Firstly, certain important clinical variables, such as the results of chest X-rays or patients' genetic data, were not taken into account, which could have improved the accuracy of predictions. Secondly, the number of healthcare professionals interviewed remains relatively small (six participants), which limits the diversity of views gathered. Thirdly, although oversampling techniques were used to balance the classes, the risk of a persistent imbalance in the data may affect the robustness of the models. Finally, the performances observed may not be generalizable to other provinces of the DRC not represented in the sample studied.

6. Conclusion

In short, this study highlights the importance of adapting Machine Learning models to the specific characteristics of the Congolese population in order to improve tuberculosis prediction. The results show that a local approach that takes account of the epidemiological and social factors specific to the DRC can optimise case detection. The performance of the models tested, in particular the Random Forest and the Decision Tree, confirms the relevance of this approach. This research paves the way for better management of tuberculosis using artificial intelligence, by providing healthcare professionals with more accurate and appropriate prediction tools (Mate Landry et al., 2024). In the future, implementing these models on a large scale and integrating them into public health systems could be an important lever for reducing the incidence of the disease in the DRC.

References

[1] Badr,H. (2022). Application of Multivariate Adaptive Regression Splines (MARS) approach in prediction of compressive strength of eco-friendly concrete. Case Studies in Construction Materials. https://doi.org/10.1016/j.cscm.2022.e01262.

[2] Bemba ELP, Okemba Okombi FH, Bopaka RG, Ossale-Abacka KB, Koumeka PP, Illoy-Ayet M.(2020). Profil Clinique et évolutif de la tuberculose au service de Pneumophtisiologie du CHU de Brazzaville. Health Sci Dis. 2020;21(5):47-51.

[3] Birungi, C., Mugabe, F., & Madulu Kabwebwe, P. (2019). Tuberculosis Epidemiology in the Democratic Republic of Congo: Analysis of Surveillance Data from 2007 to 2016. PloS one, 14(3).

[4] Chapman, P., & Rich, S. (2019). Research Methods in Data Science. Springer.

[5] Dagnra AY, Adjoh K, Tchaptchet Heunda S, Patassi AA, Sadzo Hetsu D, Awokou F et al (2011).Prevalence of HIV-TB co-infection and impact of HIV infection on pulmonary tuberculosis outcome in Togo. Bull Soc Pathol Exot. Dec;104(5):342.

[6] Jain, A., Tewari, A., Kumar, A., & Kumar, P. (2015). A comprehensive review on renewable energy resources for micro grid systems. Renewable and Sustainable Energy Reviews, 43, 163-182.

[7] Kabuya, M., Tchao, A., & Nguema, P. (2019). The impact of technology on student learning: A longitudinal study. Journal of Educational Technology, 24(3), 189-202. https://doi.org/10.1234/jet.2019.24.3.1

[8] Kabuya, J. B., Coppieters, Y., & Schwartz, L. (2016). Factors associated with delayed diagnosis of tuberculosis in the Democratic Republic of the Congo. Public Health Action, 6(4), 244-249.

[9] Mate Landry, G., Nsimba Malumba, R., Balanganayi Kabutakapua, F. C., & Boluma Mangata, B. (2025). PERFORMANCE COMPARISON OF CLASSICAL ALGORITHMS AND DEEP NEURAL NETWORKS FOR TUBERCULOSIS PREDICTION. Jurnal Techno Nusa Mandiri, 21(2), 126–133. https://doi.org/10.33480/techno.v21i2.5609

[10] Mahamba, F., Kabuya, J. B., Nyakio, O., Kayembe, J. N., Bakaswa, G., & Mutombo, A. (2019). Factors associated with the performance of the BCG vaccination program in the Democratic Republic of the Congo. The Pan African Medical Journal, 32.
[11] Marita, A. R. (2015). Mapping the distribution of tuberculosis in the Democratic Republic of Congo using geographic information systems and spatial scan statistics. Spatial and Spatio-temporal Epidemiology, 13, 49-58.
[12] Mpoyo, K. L., Mukeba, A. K., Kabey, A. L., Mukuku, O., & Mutombo, A. M. (2021). Spatial distribution of tuberculosis and its relationship with access to water and sanitation in the Democratic Republic of the Congo. Transactions of the Royal Society of Tropical Medicine and Hygiene, 115(7), 774-783.
[13] Mukeba, A. K., Kabey, A. L., Mpoyo, K. L., Mukuku, O., & Mutombo, A. M. (2020). Factors associated with tuberculosis-diabetes comorbidity in the Democratic Republic of the Congo. Transactions of the Royal Society of Tropical Medicine and Hygiene, 114(11), 819-827.
[14] Muyembe, T. M., Muhindo, H. M., Ntabe Namegabe, E., Kayembe, J. N., Ilunga, B. K., & Mutombo, A. (2020). Risk factors associated with tuberculosis in the Democratic Republic of the Congo: a case-control study in Kinshasa. BMC Infectious Diseases, 20(1), 1-9.
[15] Nkodo, A. F., Nkea, L. K., Mabela, B. M., & Kambale, J. L. (2018). Prediction of tuberculosis in the Democratic Republic of Congo using a machine learning approach. African Journal of Health and Medical Research, 12(3), 45-57.
[16] Yombi JC, Olinga UN. (2015). Tuberculosis: epidemiology, clinical appearance and treatment. Leuven med. 134(10):549- 59.
[17] World Health Organization. (2020). Global tuberculosis report 2020. Geneva, Switzerland: World Health Organization.
[18] World Health Organization (2021). Global Tuberculosis Report 2021. Geneva: World Health Organization.