



SECB3203-01(PROGRAMMING FOR BIOINFORMATIC)

Semester 01 2025/2026
Section 01

Project Proposal

Faculty of Computing

Dataset:

<https://www.kaggle.com/datasets/miadul/tuberculosis-x-ray-dataset-synthetic>

Project Title:

Tuberculosis Disease Classification Using Synthetic Chest X-Ray Images and Machine Learning Techniques

Student Name	Matric Number
NATIJAH BINTI HUDA	A23CS0142
NUR IMAN BINTI MOHD ZAHARI	A23CS0158
LIANA DARWISYAH BINTI AZMAN	A23CS0102

LECTURER'S NAME : DR. SEAH CHOON SEN

SUBMISSION DATE :

TABLE OF CONTENTS

1.0 Introduction.....	2
1.1 Problem Background.....	3
1.2 Problem Statement.....	3
1.3 Objectives.....	4
1.4 Scopes.....	4
1.4.1 Data Used.....	4
1.4.2 Techniques to Be Used.....	5
1.4.3 Methodology.....	5
1.4.4 Limitations of the Research.....	6
1.5 Conclusion.....	7
1.6 References.....	7

1.0 Introduction

Tuberculosis (TB) is a contagious infectious disease caused by *Mycobacterium tuberculosis* and primarily affects the lungs. Despite advancements in medical science, TB remains one of the leading causes of death worldwide, particularly in low- and middle-income countries. Early detection and accurate diagnosis are critical in preventing disease transmission and ensuring effective treatment. One of the most commonly used diagnostic tools for TB screening is chest X-ray imaging due to its cost-effectiveness and wide availability.

In recent years, bioinformatics and data-driven approaches have played an important role in healthcare research, especially in disease detection and classification. The integration of machine learning techniques with medical imaging allows the automation of image analysis, reducing dependency on manual interpretation by radiologists. This project leverages Python programming and bioinformatics analysis techniques to develop a classification model capable of distinguishing tuberculosis-infected chest X-ray images from normal images.

This project adopts a current bioinformatics trend which is disease classification using machine learning and image-based data analysis and applies computational techniques to a synthetic chest X-ray dataset. By utilizing synthetic data, the project avoids ethical and privacy issues associated with real clinical data while still providing a realistic environment for developing and evaluating predictive models.

1.1 Problem Background

Tuberculosis diagnosis traditionally relies on laboratory tests and expert interpretation of medical images, which may not always be accessible in rural or resource-limited regions. Chest X-ray screening is widely used; however, interpreting X-ray images requires specialized expertise and is subject to inter-observer variability. Inconsistent interpretations may lead to misdiagnosis or delayed treatment.

Furthermore, the development of automated TB detection systems faces challenges due to the limited availability of large-scale, well-annotated medical imaging datasets. Privacy regulations and ethical constraints often restrict access to real patient data, slowing down research and model development. Synthetic datasets provide a viable alternative by generating realistic medical images that support experimentation and learning.

By applying bioinformatics programming techniques and machine learning models to synthetic chest X-ray data, this project aims to explore automated TB classification and provide hands-on experience in data preprocessing, analysis, model development, and evaluation using Python.

1.2 Problem Statement

The key problems addressed in this project are as follows:

- Manual interpretation of chest X-ray images for tuberculosis diagnosis is time-consuming and dependent on expert radiologists.
- Human error and subjectivity in X-ray interpretation may lead to misclassification of TB cases.
- There is limited access to large, labeled real-world medical datasets due to privacy, ethical, and legal constraints.
- Existing datasets may suffer from class imbalance or insufficient sample size, reducing model reliability.
- Many developing regions lack automated diagnostic tools to support early TB screening.
- Machine learning models require systematic preprocessing and feature extraction to achieve acceptable performance, which is often overlooked.
- There is a need for practical exposure to bioinformatics programming techniques that integrate data science and healthcare applications.

1.3 Objectives

The objectives of this project are:

1. To collect and utilize a synthetic tuberculosis chest X-ray dataset for bioinformatics analysis.
2. To preprocess and clean the dataset using Python-based data wrangling techniques.
3. To perform exploratory data analysis (EDA) to understand data distribution and relationships.
4. To develop machine learning models for classifying chest X-ray images into tuberculosis and normal classes.
5. To evaluate the performance of the developed models using suitable evaluation metrics.
6. To document the analysis process, results, and findings through GitHub and a formal project report.

1.4 Scopes

The scope of this project is defined as follows:

1.4.1 Data Used

- A synthetic chest X-ray dataset for tuberculosis classification obtained from Kaggle.
- The dataset consists of labeled X-ray images categorized into Tuberculosis and Normal classes.
- Only image data and corresponding labels are used, no real patient or clinical metadata is included.

1.4.2 Techniques to Be Used

- Python programming for all data processing and analysis tasks.
- Data wrangling techniques using Pandas and NumPy, including:
 - Handling missing values
 - Data normalization and scaling
 - Data formatting and transformation
- Exploratory Data Analysis (EDA):
 - Descriptive statistics
 - Correlation analysis
 - Data visualization using Matplotlib
- Machine learning and modeling techniques, including:
 - Regression-based models
 - Classification models
 - Model evaluation and refinement
- Model evaluation methods such as:
 - Accuracy
 - Mean Squared Error (MSE)

- R-squared
- Cross-validation and grid search (if applicable)

1.4.3 Methodology

The project methodology includes the following stages:

1. Dataset collection and understanding
2. Data preprocessing and normalization
3. Exploratory data analysis
4. Feature extraction and model development
5. Model evaluation and optimization
6. Result interpretation and documentation
7. Version control and collaboration through GitHub

1.4.4 Limitations of the Research

- The dataset used is synthetic, which may not fully represent real-world clinical variations.
- The developed model is intended for academic and educational purposes only.
- Clinical validation using real patient data is outside the scope of this project.
- Computational resources are limited to student-level hardware and software environments.
- The model's performance may not generalize well to real hospital settings without further validation.

1.5 Conclusion

This project focuses on applying bioinformatics and machine learning techniques to classify tuberculosis using synthetic chest X-ray images. By integrating Python programming, data analysis, and model development, the project provides practical experience in disease classification and computational healthcare research. Although the study is limited to synthetic data and academic use, it demonstrates the potential of automated approaches in supporting tuberculosis screening and highlights the relevance of bioinformatics in modern healthcare applications.

1.6 References

Showkatian, E., Salehi, M., Ghaffari, H., Reiazi, R., & Sadighi, N. (2022).

Deep learning-based automatic detection of tuberculosis disease in chest X-ray images. *Polish Journal of Radiology*, 87(1), 118–124.

<https://doi.org/10.5114/pjr.2022.113435>

Wajgi, R., Yenurkar, G., Nyangaresi, V. O., Wanjari, B., Verma, S., Deshmukh, A., &

Mallewar, S. (2024). Optimized tuberculosis classification system for chest X-ray images: Fusing hyperparameter tuning with transfer learning approaches. *Engineering Reports*, 6(11). <https://doi.org/10.1002/eng2.12906>

Sharma, V., None Nillmani, Gupta, S., & Shukla, K. K. (2023). Deep learning models for

tuberculosis detection and infected region visualization in chest X-ray images.

Intelligent Medicine. <https://doi.org/10.1016/j.imed.2023.06.001>