JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 01
2024

Research Article
https://doi.org/10.56979/701/2024

Diagnosis of Pulmonary Tuberculosis by Posterior-Anterior Lung X-Ray

Riasat Ali[1], Nouman Arshid[2], Muhammad Ikramul Haq[1], Zaib Un Nisa[1], Syed Asad Ali Naqvi[1], Muhammad Waseem Iqbal[3] , and Khalid Hamid[4*]

[1]Department of Computer Science, Superior University, Lahore, 54000, Pakistan.
[2]Department of Information Technology, Superior University, Lahore, 54000, Pakistan.
[3]Department of Software Engineering,Superior University, Lahore , 54000, Pakistan.
[4]Department of Computer Science, NCBA & E University East Canal Campus, Lahore, Pakistan.
[*]Corresponding Author: Khalid Hamid. Email: khalid6140@gmail.com

Received: January 29, 2024 **Accepted:** May 03, 2024 **Published:** June 01, 2024

Abstract: Tuberculosis (TB) remains a pressing global health issue, with an estimated 10.6 million cases projected by 2021. In Pakistan, TB prevalence is notably high, comprising 61% of the WHO Eastern Mediterranean TB burden. TB, primarily caused by Mycobacterium bacteria, affects multiple organs, often presenting with subtle or asymptomatic symptoms. Despite the gravity of the disease, early detection methods are limited, typically relying on model lung segmentation techniques. This research aims to enhance TB detection using chest X-ray images through a novel, model-less segmentation approach. By extracting statistical, geometric, and Hog descriptor features from lung images, coupled with various classifiers such as Convolutional Neural Network (CNN), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), the study achieved promising results. The highest accuracy attained was 91.88% using self-extracted features and linear regression, while CNN demonstrated competitive performance with an accuracy of 89.58%. To bolster the findings, visualization techniques were employed, confirming CNN's superior ability to discern patterns from segmented lung areas, thereby contributing to higher detection accuracy. This innovative approach holds significant potential for expediting computer-assisted TB diagnosis, benefiting clinical practice and public health initiatives.

Keywords: Tuberculosis; Model-less Segmentation; Convolutional Neural Network (CNN); Support Vector Machine (SVM); K-Nearest Neighbor (KNN); Public Health Intervention.

1. Introduction

The introduction to the research paper provides a comprehensive overview of tuberculosis (TB), its historical context, current prevalence, symptoms, and the challenges associated with its diagnosis. It emphasizes the global burden of TB and highlights the need for improved detection methods, particularly in countries like Pakistan, where TB remains a significant public health concern [20].

The problem statement of the research is twofold: first, it addresses the limitations and gaps in existing TB detection research, such as the lack of preprocessing steps, unbalanced datasets, and reliance on model-based segmentation, which may lead to suboptimal accuracy and generalization. Second, it underscores the urgency of the TB problem, especially in high-burden countries like Pakistan, where a substantial portion of TB cases goes undetected or unreported, contributing to the disease's persistence and spread [14] [13].

The motivation for the proposed research lies in the dire need for a more effective and comprehensive TB detection framework that addresses the limitations mentioned earlier and leverages advanced machine learning and image processing techniques to improve accuracy and reliability. The research aims to develop an intelligent system capable of detecting TB from chest X-ray images by integrating preprocessing, segmentation, feature extraction, and classification steps [21] [7] .

The research scope encompasses publicly available datasets from reputable sources, such as Chinese and American hospital databases, to ensure data quality and diversity. The research will focus on preprocessing techniques for noise reduction and enhancement, segmentation without reliance on lung models, extraction of edge, shape, and texture features from regions of interest (ROIs), and evaluation of various classification models for performance measurement [3].

The proposed research is particularly relevant to Pakistan, where TB remains a significant public health challenge, ranking fifth in the world for TB cases. Despite the high burden of TB in Pakistan, a considerable number of cases go unreported or undetected, highlighting the urgent need for improved detection methods [4].

The summary sets the stage for the proposed research by providing a comprehensive overview of TB, highlighting its global significance, underscoring the limitations of existing detection methods, and outlining the research objectives, scope, and motivation. Through the proposed research, the aim is to develop a robust TB detection framework that addresses existing gaps and contributes to more accurate and timely diagnosis, particularly in high-burden countries like Pakistan [16].

2. Literature Review

The field of tuberculosis (TB) detection has seen significant advancements in recent years, driven by the integration of machine learning and image processing techniques with clinical data. TB, a centuries-old disease, continues to pose a formidable challenge to global health efforts. Despite historical progress in controlling TB in some regions, it remains a persistent threat, particularly in resource-limited settings where access to advanced medical diagnostics is limited. In response to this challenge, researchers have explored various approaches to improve TB detection accuracy and efficiency [6] [24].

Recent studies have highlighted the potential of deep convolutional neural networks (CNNs) in enhancing TB detection capabilities. CNNs leverage their ability to extract temporal and spatial features from images, making them well-suited for analyzing chest X-ray images, a common diagnostic tool in TB screening. Transfer learning, a technique that adapts pre-trained CNN models to new datasets, has emerged as a promising approach to overcome limitations associated with small datasets and lengthy training periods. By fine-tuning pre-trained models on TB-specific datasets, researchers have achieved notable improvements in detection accuracy [19].

Several frameworks and systems have been developed to automate TB detection and screening processes using machine learning algorithms. These systems typically involve preprocessing steps, such as lung region segmentation, followed by feature extraction and classification. Multiple instance learning (MIL) has emerged as an alternative pattern classification method for TB detection, offering advantages in terms of computational efficiency and scalability. Additionally, deep learning techniques, such as deep CNNs, have been applied to digital chest radiography for TB detection, demonstrating the feasibility of automated TB screening using advanced imaging technologies [2] [12].

Despite these advancements, challenges remain in optimizing TB detection systems for real-world clinical settings. The limited availability of large and diverse datasets poses a significant obstacle to training robust and generalizable models. Moreover, the performance of TB detection algorithms may vary depending on factors such as patient demographics and imaging modalities. Addressing these challenges requires collaborative efforts among researchers, clinicians, and policymakers to collect high-quality data and develop standardized evaluation metrics for assessing detection accuracy [22] [5].

In addition to technical challenges, ethical considerations must be addressed in the development and deployment of automated TB detection systems. Ensuring patient privacy and confidentiality, as well as mitigating biases in algorithmic decision-making, are critical considerations in the design and implementation of TB detection frameworks. Furthermore, efforts to integrate TB detection systems into existing healthcare infrastructure must take into account resource constraints and accessibility issues, particularly in low-resource settings [23].

Overall, the integration of machine learning and image processing techniques holds tremendous promise for improving TB detection and screening efforts worldwide. By harnessing the power of advanced computational methods, researchers can develop more accurate, efficient, and scalable TB detection systems that have the potential to save lives and reduce the burden of this devastating disease.

Continued research and innovation in this field are essential to realizing the full potential of automated TB detection in global health initiatives [18].

2.1. Machine Learning Classifiers

The machine learning classifiers used the self-extracted features for the classification of the datasets. Machine learning classifiers used the well-balanced training and the testing datasets for the classification. The machine learning classifiers that are mostly used for the classification in MATLAB are listed below [12] [10].

1. Support vector machine
2. K-nearest neighbors
3. Decision tree
4. Ensembles
5. Linear regression

2.1.1. Support Vector Machine

The study utilized Support Vector Machine (SVM) as a supervised learning model for dataset analysis and classification of self-extracted features. SVM's architecture involves training on datasets to classify new data, utilizing kernel methods for prediction across different classes [7] [17].

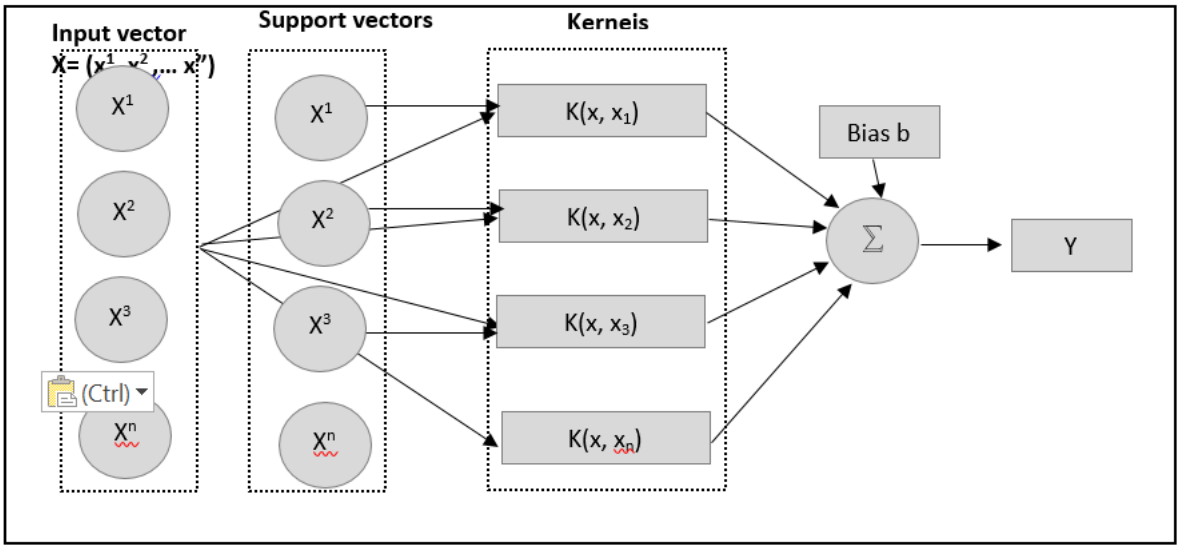

Figure 1. SVM Network

2.1.2. Nearest Neighbor

The k-nearest neighbor algorithm, or KNN classifier, is a simple non-parametric classification learning algorithm. It operates by utilizing a database of different classes to predict new datasets based on the trained dataset. KNN does not assume any underlying data distribution, making it effective for classification. Its architecture lacks generalization but provides accurate results on self-extracted features. It determines predictions by measuring distances between objects and selecting the nearest neighbor's class.

2.1.3. Ensembles

Ensembles are meta-algorithms that combine multiple machine-learning techniques to improve prediction accuracy by reducing variance and bias. They incorporate both supervised and clustering learning algorithms. The architecture involves two key steps: exploiting dependence between base learners through sequential methods to boost overall performance and exploiting independence between base learners through parallel methods. Ensembles consist of various classifiers using clustering-based and supervised-based classification. Bagging, a clustering learning algorithm, is used to decrease variance in training and prediction datasets while boosting converts weak learners into strong ones to enhance performance. The PeMS-8 dataset is in nps formatted so firstly I converted it into a CSV file using the PANDAS library. After the conversion of the file, I check the missing value from the dataset to achieve high accuracy during the model training. In the below table, I will show the five rows of the dataset.

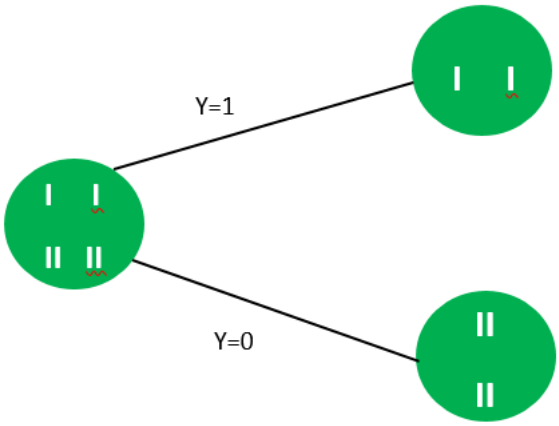

Figure 2. PeMS-8 Dataset

2.1.4. Linear Regression

Linear regression is a supervised learning algorithm that delivers classification results in continuous values due to its regression-based architecture. It utilizes mathematical techniques for optimal dataset prediction and relies on well-balanced training datasets for accurate classification.

3. Methodology

The study proposes a tuberculosis detection framework using machine learning and image processing techniques, comprising preprocessing, segmentation, feature extraction, and classification steps. It involves establishing two databases for classifying tuberculosis and segmenting lungs, followed by three large-scale experiments. Firstly, two U-Net models are compared for lung segmentation. Secondly, nine pre-trained networks classify original chest X-ray images for tuberculosis, and their reliability is evaluated using the Score-CAM technique. Thirdly, the same networks classify tuberculosis in segmented lung images, with performance assessment using Score-CAM.

The proposed methodology aims to provide accurate tuberculosis identification by employing self-extracted and classifier-based features. Figure illustrates the suggested technique flowchart, detailing each step of the diagnosis process. It begins with data acquisition, followed by image processing operations like segmentation, preprocessing, and feature extraction. The final stage involves classification using various machine-learning algorithms. Each step of the process is thoroughly explained within the study [7] [15].

3.1. Data Acquisition

The proposed methodology utilizes two internet datasets containing standard digital chest X-ray images for tuberculosis sourced from the National Library of Medicine. One dataset is from the No. 3 People's Hospital in Shenzhen, China, while the other is from Montgomery County, Maryland, USA. These datasets encompass images from both normal and abnormal chest X-rays, representing patients with and without tuberculosis. Notably, the USA dataset boasts a pixel spacing of 0.0875mm in both vertical and horizontal directions. Employing the region of interest (ROI) technique, lung sections are extracted from the images in both datasets, encompassing frontal and posterior views of the ribs. For further insights, Table 2 offers a detailed breakdown of both datasets across various parameters [8].

3.2. Pre-Processing

The next stage in the proposed methodology involves pre-processing the chest X-ray images, which comprises resizing and dimension adjustments, noise removal techniques, and contrast enhancement. Pre-processing is deemed critical before segmentation to mitigate errors and enhance performance. Different pre-processing techniques are applied to both datasets.

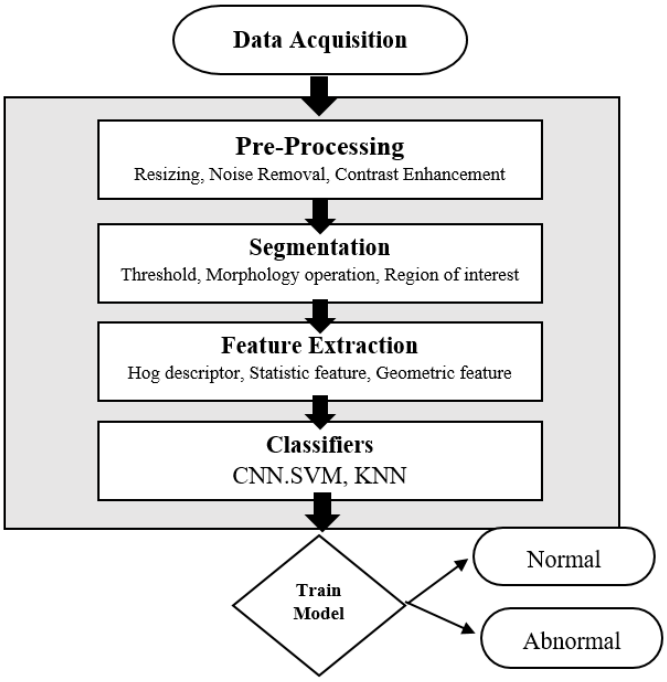

Figure 3. Preprocessing

The figure illustrates the step-by-step pre-processing procedure. Beginning with the input image, adjustments are made to its dimensions and size to accommodate various requirements for subsequent stages. Noise in the images is then eliminated to ensure clarity for further processing (Can AI Help in Screening Viral and COVID-19 Pneumonia? | IEEE Journals & Magazine | IEEE Xplore, n.d.). Image brightness is crucial for region extraction during the area of interest (ROI) phase, which involves three key steps detailed below.

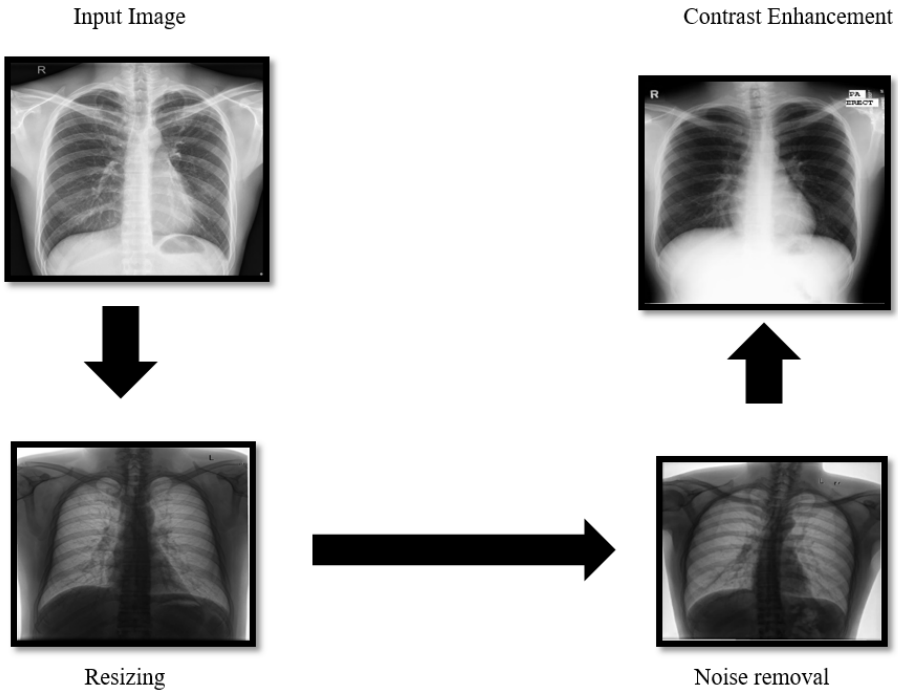

Figure 4. Chest X-ray Images

3.3. Segmentation

Segmentation is a crucial phase as it determines the accuracy of lung region extraction. The proposed methodology employs gray thresholding techniques along with morphological operations to exclude the

region of interest from preprocessed images. Various functions and techniques are applied during segmentation to ensure reliable tuberculosis detection. Figure illustrates the segmentation flow in a step-by-step manner. Enhancing the noise-free images through code comparison is essential as original images may lack the necessary brightness for easy lung area extraction. Histogram equalization is used to enhance the brightness of chest X-ray images for improved visualization and segmentation results. The pre-processing section of the proposed method is depicted through step-by-step images.

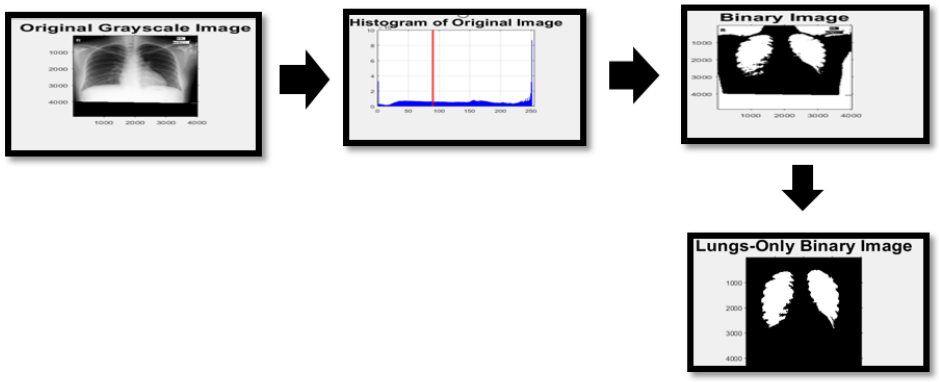

Figure 5. Segmentation Flow

3.4. Feature Extraction

In the feature extraction phase of the proposed methodology, various features are extracted from the lung region (ROI) to facilitate tuberculosis detection. This includes statistical, geometric, and Hog descriptor features obtained through different techniques. The diagram illustrates the list of different features extracted for tuberculosis detection. Segmented images are utilized for this phase, as described in the diagram. All features are extracted for both datasets, with further details provided in a step-by-step manner.

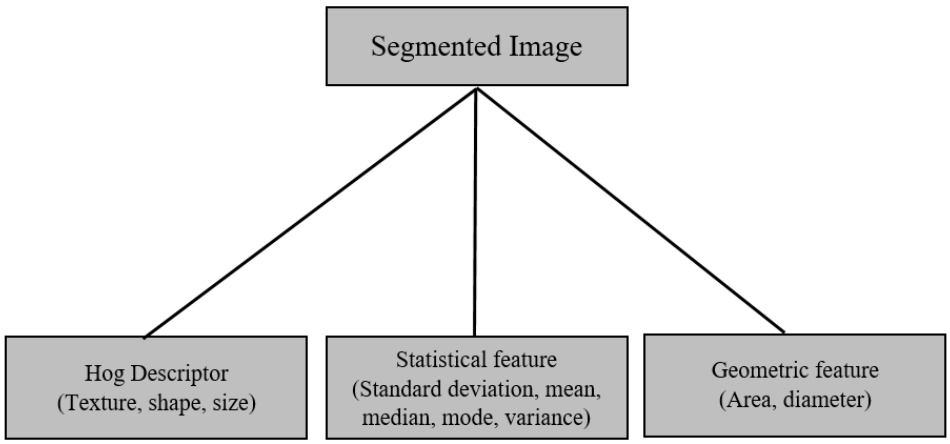

Figure 6. Feature Extracted

4. Results and Discussions

The proposed methodology evaluates tuberculosis (TB) detection effectiveness using parameters such as accuracy, error rate, sensitivity, and specificity. Two experiments are conducted: one utilizing Convolutional Neural Network (CNN) models and the other using self-extracted features with machine learning classifiers.

For the CNN experiment, various CNN models like AlexNet and GoogleNet are employed. Results show that GoogleNet achieves the highest accuracy of 90.45% for the Shenzhen, China dataset and 89.58% for the combined dataset. Graphical representations demonstrate the performance of CNN models, indicating GoogleNet's superiority.

In the experiment with self-extracted features, classifiers like Linear Discriminant, SVM, KNN, Decision Tree, Ensembles, and Linear Regression are used. Linear Regression achieves the highest accuracy of 86.86% for the Montgomery, USA dataset and 85.86% for the Shenzhen, China dataset.

Combining both datasets yields a higher accuracy of 90.00% with Linear Regression. Comparative analysis with existing research highlights the proposed methodology's advantages in preprocessing techniques, segmentation, feature extraction, and classification, achieving an Area under the Curve (AUC) of 90.45% with CNN and 86.86% with Linear Regression.

5. Conclusion and Future Work

The proposed methodology aims to enhance tuberculosis detection through early diagnosis, crucial for combating its widespread impact, particularly in impoverished regions. It employs gray threshold segmentation and morphological operations to isolate lung areas, followed by feature extraction and classification using various classifiers. Notably, GoogleNet achieves a top accuracy of 90.45%, while linear regression reaches 97.00%, utilizing statistical and texture features. The Montgomery US dataset achieves 86.86% accuracy, and the Shenzhen China dataset achieves 85.86% with linear regression. The study also explores the effectiveness of deep convolutional neural networks, with DenseNet201 outperforming other models on segmented lung images. This method holds promise for aiding TB diagnosis, particularly in high-burden regions like Pakistan, where collaboration with local experts and datasets is planned for implementation.

References

1. Can AI Help in Screening Viral and COVID-19 Pneumonia? | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Retrieved May 11, 2024, from https://ieeexplore.ieee.org/abstract/document/9144185.
2. Chalie, M., & Mossie, Z. (2023). Pulmonary Disease Identification and Classification Using Deep Learning Approach. Ethiopian International Journal of Engineering and Technology, 1(2), Article 2. https://doi.org/10.59122/144CFC16.
3. Hamid, K., Aslam, Z., Delshadi, A. M., Ibrar, M. I., Mahmood, Y., & Iqbal, M. W. (2024). Empowerments of Anti-Cancer Medicinal Structures by Modern Topological Invariants.
4. Hamid, K., Ibrar, M., Delshadi, A. M., Hussain, M., Iqbal, M. W., Hameed, A., & Noor, M. (2024). ML-based Meta-Model Usability Evaluation of Mobile Medical Apps. International Journal of Advanced Computer Science and Applications (IJACSA), 15(1), Article 1. https://doi.org/10.14569/IJACSA.2024.0150104.
5. Hamid, K., & Iqbal, M. waseem. (2022). Topological Evaluation of Certain Computer Networks by Contraharmonic-Quadratic Indices. Computers, Materials and Continua, 74, 3795–3810. https://doi.org/10.32604/cmc.2023.033976.
6. Hamid, K., Iqbal, M. waseem, Aqeel, M., Rana, T., & Arif, M. (2023). Cyber Security: Analysis for Detection and Removal of Zero-Day Attacks (ZDA) (pp. 172–196). https://doi.org/10.1201/9781003190301-10.
7. Hamid, K., Iqbal, M. waseem, Fuzail, Z., Muhammad, H., Basit, M., Nazir, Z., & Ghafoor, Z. (2022). Detection of Brain Tumor from Brain MRI Images with the Help of Machine Learning & Deep Learning. https://doi.org/10.22937/IJCSNS.2022.22.5.98.
8. Hamid, K., Iqbal, M. waseem, Muhammad, H., Basit, M., Fuzail, Z., † Z., & Ahmad, S. (2022). Usability Evaluation of Mobile Banking Applications in Digital Business as Emerging Economy. 250. https://doi.org/10.22937/IJCSNS.2022.22.2.32.
9. Hamid, K., Iqbal, M. waseem, Muhammad, H., Fuzail, Z., & Nazir, Z. (2022). ANOVA BASED USABILITY EVALUATION OF KID'S MOBILE APPS EMPOWERED LEARNING PROCESS. Qingdao Daxue Xuebao(Gongcheng Jishuban)/Journal of Qingdao University (Engineering and Technology Edition), 41, 142–169. https://doi.org/10.17605/OSF.IO/7FNZG.
10. Hamid, K., Iqbal, M. waseem, & Niazi, Q. (2022). Discovering Irregularities from Computer Networks by Topological Mapping. Applied Sciences, 12, 1–16. https://doi.org/10.3390/app122312051.
11. Hamid, K., Muhammad, H., Basit, M., Hamza, M., Bhatti, S., Bukhari, S., & Hassan. (2022). EXTENDABLE BANHATTI SOMBOR INDICES FOR MODELING CERTAIN COMPUTER NETWORKS MUHAMMAD WASEEM IQBAL M AMEER HAMZA. Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition), 41, 11–2022.
12. Hamid, K., Muhammad, H., Iqbal, M. waseem, Nazir, A., shazab, & Moneeza, H. (2023). ML-BASED META MODEL EVALUATION OF MOBILE APPS EMPOWERED USABILITY OF DISABLES. Tianjin Daxue Xuebao (Ziran Kexue Yu Gongcheng Jishu Ban)/Journal of Tianjin University Science and Technology, 56, 50–68.
13. Hamid, K., Muhammad, H., Iqbal, M. waseem, Nazir, Z., Irfan, D., & Rashed, R. (2022). EMPOWERMENTS OF CHEMICAL STRUCTURES USED FOR CURING LUNGS INFECTIONS BY MODERN INVARIANTS. Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition), 41, 439–458. https://doi.org/10.17605/OSF.IO/SY6KH.
14. Hopewell, P. C., Pai, M., Maher, D., Uplekar, M., & Raviglione, M. C. (2006). International Standards for Tuberculosis Care. The Lancet Infectious Diseases, 6(11), 710–725. https://doi.org/10.1016/S1473-3099 (06)70628-4.
15. Iqbal, M. W., Hamid, K., Ibrar, M., & Delshadi, A. (2024). Meta-Analysis and Investigation of Usability Attributes for Evaluating Operating Systems. Migration Letters, 21, 1363–1380.
16. Matsuoka, S., Uchiyama, K., Shima, H., Suzuki, K., Shimura, A., Sasaki, Y., & Yamagishi, F. (2004). Relationship between CT findings of pulmonary tuberculosis and the number of acid-fast bacilli on sputum smears. Clinical Imaging, 28(2), 119–123. https://doi.org/10.1016/S0899-7071 (03)00148-7.
17. Memon, A., Nazir, A., Hamid, K., & Iqbal, M. waseem. (2023). AN EFFICIENT APPROACH FOR DATA TRANSMISSION USING THE ENCOUNTER PREDICTION M. ASHRAF NAZIR KHALID HAMID MUHAMMAD WASEEM IQBAL. Tianjin Daxue Xuebao (Ziran Kexue Yu Gongcheng Jishu Ban)/Journal of Tianjin University Science and Technology, 56, 92–109. https://doi.org/10.17605/OSF.IO/RM3UJ.
18. Murphy, K., Habib, S. S., Zaidi, S. M. A., Khowaja, S., Khan, A., Melendez, J., Scholten, E. T., Amad, F., Schalekamp, S., Verhagen, M., Philipsen, R. H. H. M., Meijers, A., & van Ginneken, B. (2020). Computer aided detection of tuberculosis on chest radiographs: An evaluation of the CAD4TB v6 system. Scientific Reports, 10(1), 5492. https://doi.org/10.1038/s41598-020-62148-y.
19. Nazir, Z., Iqbal, M. waseem, Hamid, K., Muhammad, H., Nazir, A., Hussain, N., & ann, Q. (2023). VOICE ASSISTED REAL-TIME OBJECT DETECTION USING YOLO V4- TINY ALGORITHM FOR VISUAL CHALLENGED. 56, 2023. https://doi.org/10.17605/OSF.IO/APQYH.

20. Organization, W. H. (2013a). Global Tuberculosis Report 2013. World Health Organization.
21. Organization, W. H. (2013b). Systematic Screening for Active Tuberculosis: Principles and Recommendations. World Health Organization.
22. Ors, F., Deniz, O., Bozlar, U., Gumus, S., Tasar, M., Tozkoparan, E., Tayfun, C., Bilgic, H., & Grant, B. J. B. (2007). High-resolution CT Findings in Patients With Pulmonary Tuberculosis: Correlation With the Degree of Smear Positivity. Journal of Thoracic Imaging, 22(2), 154. https://doi.org/10.1097/01.rti.0000213590.29472.ce.
23. Peralta, G., Barry, P., & Pascopella, L. (2016). Use of Nucleic Acid Amplification Tests in Tuberculosis Patients in California, 2010–2013. Open Forum Infectious Diseases, 3(4), ofw230. https://doi.org/10.1093/ofid/ofw230.
24. Syed, W., Ahmed, A., Zubair, S., Iqbal, M. waseem, Arif, S., & Hamid, K. (2023). FUZZY-BASED EXPERT SYSTEM FOR TEST CASE GENERATION ON WEB GRAPHICAL USER INTERFACE FOR USABILITY TEST IMPROVEMENT. 42, 549–565. https://doi.org/10.17605/OSF.IO/H5ER9.