HOG-Based Machine Learning Models for Classifying COVID-19 in Chest X-Ray Images

Greeshma K V[1], Dr. J. Viji Gripsy[2]

[1]Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India, Email: greeshmakv@gmail.com

[2]Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India, Email: gripsy@psgrkcw.ac.in

Received: 10.07.2023 Revised: 19.08.2023 Accepted: 08.09.2023

ABSTRACT

Chest X-ray analysis remains a pivotal tool for the initial screening of COVID-19, despite limitations in sensitivity and specificity. This study investigates the integration of Histogram of Oriented Gradients (HOG) features with various machine learning algorithms to classify chest X-ray images into COVID-19, Pneumonia, and Normal categories. HOG features provide a robust foundation for feature extraction, and their integration with machine learning models such as Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (KNN) has been comprehensively evaluated. The research highlights the strengths of SVM and Logistic Regression, which achieved an accuracy of 96% and an MCC of 0.92, showcasing their effectiveness for the task. In contrast, KNN and Random Forest exhibited moderate performance, while Decision Tree algorithms showed significant limitations. These findings underline the foundational role of HOG features and machine learning models in advancing automated diagnostic systems. While hand-crafted features like HOG laid the groundwork, the field is evolving rapidly with the advent of more sophisticated approaches, including deep learning. Future research should focus on optimizing algorithms for better accuracy, integrating deep learning-based methods, and enhancing model generalizability. This work underscores the importance of interdisciplinary collaboration and emerging technologies in developing reliable diagnostic tools to combat the ongoing global pandemic.

Keywords: COVID-19, X-ray, image classification, HOG, machine learning, medical imaging

1. INTRODUCTION

The COVID-19 pandemic has underscored the urgent need for innovative and efficient diagnostic solutions to address the challenges of timely identification and management of the virus. Among the various diagnostic modalities, chest X-ray imaging plays a pivotal role, offering rapid, accessible, and non-invasive assessment of lung abnormalities associated with COVID-19. Despite its utility, traditional chest X-ray analysis often suffers from subjective interpretation, limited sensitivity, and challenges in differentiating overlapping pulmonary conditions. These limitations necessitate the development of advanced computational tools to enhance diagnostic precision and reliability.

This study proposes a novel framework that integrates hand-crafted features, specifically the Histogram of Oriented Gradients (HOG), with machine learning algorithms for automated classification of COVID-19 in chest X-ray images. HOG features are well-known for their effectiveness in capturing edge and gradient structures, making them particularly suitable for identifying discriminative patterns in medical images. By leveraging this feature extraction technique, the framework aims to address the limitations of traditional diagnostic approaches while maintaining transparency in decision-making.

To evaluate the efficacy of this approach, machine learning algorithms such as Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors (kNN), and Decision Trees were employed. These algorithms were trained to classify chest X-ray images into three categories: COVID-19, Pneumonia, and Normal. The performance of each model was rigorously assessed using key metrics such as accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC).

This research not only advances the integration of artificial intelligence in medical diagnostics but also emphasizes the need for interpretable and explainable models in healthcare. The use of hand-crafted features and traditional machine learning models offers a transparent framework that allows clinicians to understand the rationale behind predictions, fostering trust and acceptance in clinical applications. By addressing the limitations of existing diagnostic tools and exploring the potential of HOG-based machine learning models, this research contributes to the broader efforts in combating the COVID-19 pandemic. It aims to provide a robust and accessible solution for healthcare systems, particularly in resource-constrained settings, ultimately enhancing diagnostic accuracy, efficiency, and trustworthiness.

2. LITERATURE REVIEW

This literature review explores the advancements in machine learning techniques applied to the detection of COVID-19 through chest X-ray and CT images. Several studies have demonstrated the potential of machine learning in enhancing diagnostic accuracy, offering promising solutions for early detection and diagnosis, especially in resource-constrained healthcare settings.

Hussain et al. (2020) introduced a machine learning framework that focused on classifying texture features from portable chest X-rays. Their method achieved 100% accuracy in distinguishing COVID-19 from normal cases and demonstrated high sensitivity and specificity in differentiating between COVID-19, bacterial pneumonia, and non-COVID-19 viral pneumonia. This study highlighted the critical role of deep learning in improving diagnostic efficiency, particularly in environments where portable X-rays are widely used. Similarly, Johri et al. (2021) proposed an analytical framework that combined multiple machine learning algorithms, optimizing feature extraction processes to improve classification accuracy. Their model demonstrated high performance, emphasizing the ability of machine learning to assist healthcare professionals in making timely diagnoses.

In a similar vein, Khan (2021) developed an automated system for COVID-19 detection using image processing combined with machine learning algorithms. The system was designed for rapid diagnostics, reducing processing time while maintaining high accuracy, which is crucial in managing patient loads during the pandemic. Khan et al. (2020) also explored the classification of chest X-ray images, evaluating various machine learning algorithms to identify the most effective models. Their findings underscored the importance of careful model selection to ensure reliable and accurate results, contributing to the growing body of research supporting machine learning in infectious disease diagnostics.

Greeshma & Viji Gripsy (2021) review AI-based methods for biomedical image classification and retrieval, focusing on content-based image retrieval (CBIR) and techniques like Bag of Visual Words and deep convolutional neural networks. The study highlights challenges in handling heterogeneous medical images and emphasizes preprocessing and feature extraction as critical steps for efficient image retrieval systems. The study by Greeshma & Sreekumar (2019) highlights the effectiveness of the HOG feature descriptor combined with multiclass SVM for classifying the Fashion-MNIST dataset, demonstrating its potential in image recognition tasks. Their findings underscore the importance of selecting robust feature extraction techniques and classifiers to achieve high accuracy in classification challenges.

Manav et al. (2023) focused on optimal feature selection for COVID-19 detection from chest X-ray images. They found that selecting the right features significantly improved classification accuracy, reinforcing the importance of feature selection in the development of effective diagnostic models. Mijwil (2021) further explored machine learning techniques for classifying lung X-ray images to detect COVID-19, using algorithms like Support Vector Machines (SVM) and Random Forests. The study highlighted that choosing appropriate models was crucial for enhancing diagnostic capabilities. In a similar effort, Öztürk et al. (2021) proposed a classification framework that used shrunken features from both X-ray and CT images, reducing dimensionality while improving classification accuracy. This research emphasized the versatility of combining different imaging modalities, such as X-rays and CT scans, to enhance diagnostic performance.

Rehouma et al. (2021) reviewed machine learning applications for COVID-19 detection, highlighting the importance of deep learning in achieving high accuracy rates. They suggested that hybrid models incorporating both X-ray and CT scans could improve diagnostic performance further. Samsir et al. (2021) conducted a comparative analysis of machine learning algorithms for COVID-19 detection, finding that deep learning models outperformed traditional algorithms in terms of accuracy and sensitivity. This research reinforced the importance of selecting the right algorithm to develop reliable diagnostic tools.

Finally, Saygılı (2021) proposed a computer-aided detection system that integrated machine learning methods to identify COVID-19 from both CT and X-ray images. This system demonstrated the potential of machine learning to provide automated and reliable assessments, assisting healthcare professionals in making timely diagnoses. Collectively, the reviewed studies highlight the diverse and evolving role of machine learning in medical imaging, underscoring its potential to revolutionize diagnostic accuracy and efficiency in the detection of COVID-19.

These studies collectively illustrate significant advancements in the use of machine learning techniques for detecting and classifying COVID-19 from chest X-ray and CT images. The diverse methodologies, including deep learning, feature selection, and hybrid models, have demonstrated improved diagnostic accuracy and efficiency. These findings underscore the critical role of artificial intelligence in healthcare, particularly in managing pandemics like COVID-19. Future research will further refine these models, ensuring their integration into clinical workflows for broader applications in medical diagnostics.

3. Methodology

The methodology adopted in this study involves the implementation of machine learning models for the classification of chest X-ray images into three categories: COVID-19, Normal, and Pneumonia. The approach integrates the use of Histogram of Oriented Gradients (HOG) for feature extraction and evaluates the performance of various machine learning algorithms on the extracted features.

A. Dataset Collection and Preprocessing

The dataset consists of chest X-ray images classified into three classes: COVID-19, Normal, and Pneumonia. Images were organized into training and testing sets stored in structured directories. Each image was resized to a uniform dimension of 224×224 pixels to ensure consistency in feature extraction.

B. Feature Extraction and Data Preprocessing

HOG was employed as the feature extraction technique due to its robustness in capturing edge and texture features essential for image classification. Parameters for HOG were carefully selected, including 9 orientations, a cell size of 8×8 pixels, and a block size of 2×2 cells, normalized using the L2-Hys method. The resulting feature vectors were used as input for the classification models. The extracted HOG features were stored as arrays, and the labels were encoded into numerical format using a Label Encoder for compatibility with machine learning models.

C. Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) is a feature extraction technique widely used in image processing and computer vision. It captures the structural information of objects by calculating the gradient orientation and magnitude in localized regions of an image. The image is divided into small cells, and for each cell, a histogram of gradient directions is computed, weighted by the corresponding gradient magnitudes. These histograms are then normalized over overlapping blocks to enhance the feature's invariance to changes in illumination and contrast. HOG is particularly effective in detecting edges and shapes, making it suitable for object recognition and classification tasks. In this study, HOG features were extracted from chest X-ray images, enabling robust representation of the structural patterns critical for classifying COVID-19, Pneumonia, and Normal cases.

D.Model Selection

For the classification task, the following machine learning algorithms were chosen based on their characteristics and suitability for the extracted Histogram of Oriented Gradients (HOG) features. Each algorithm offers unique strengths, enabling a comprehensive comparison:

(i) Support Vector Machine (SVM)

SVM is a powerful supervised learning algorithm known for its effectiveness in handling high-dimensional data and its ability to find the optimal hyperplane that separates classes. For this study, the radial basis function (RBF) kernel was selected to map the HOG features into a higher-dimensional space, enabling better separation of non-linearly separable data. SVM's robustness against overfitting in small to medium-sized datasets and its well-defined mathematical foundation make it an ideal choice for medical image classification.

(ii) k-Nearest Neighbors (kNN)

The kNN algorithm is a simple yet effective distance-based classifier that predicts the class of a sample based on the majority class of its k nearest neighbors in the feature space. It is highly interpretable and requires minimal parameter tuning, making it a good baseline for evaluating other models. However, kNN's performance is heavily influenced by the choice of k and the distance metric, and it can be computationally expensive during inference, especially with large datasets.

(iii) Random Forest (RF)

Random Forest is an ensemble learning algorithm that constructs a multitude of decision trees during training and combines their outputs for improved accuracy and robustness. This method reduces the risk of overfitting associated with individual decision trees and provides reliable performance across various datasets. For this research, RF's ability to handle non-linear relationships and its inherent feature importance evaluation capabilities were key factors in its selection.

(iv) Decision Tree (DT)

The Decision Tree (DT) algorithm classifies data by recursively splitting it based on feature thresholds, forming a tree-like structure with decision rules at nodes and class labels at leaves. Its interpretability makes it valuable for understanding decision-making and feature importance. However, DT tends to overfit, limiting its effectiveness with complex datasets.

(v) Logistic Regression (LR)

Logistic Regression is a statistical machine learning algorithm commonly used for binary and multi-class classification tasks. It models the relationship between the dependent variable and one or more independent variables by estimating probabilities using the logistic function (sigmoid curve). Its simplicity, interpretability, and efficiency in handling linearly separable data make it a reliable baseline model for medical image classification. Despite its linear nature, Logistic Regression performed competitively in terms of accuracy and MCC, demonstrating its effectiveness for this dataset.

By evaluating these algorithms, the study aims to identify the most effective machine learning model for classifying COVID-19 chest X-rays, while also exploring the trade-offs between complexity, interpretability, and performance.

4. RESULTS AND DISCUSSIONS

The classification performance of the selected machine learning algorithms—Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Random Forest (RF), and Decision Tree (DT)—was evaluated using the extracted Histogram of Oriented Gradients (HOG) features. The results were analyzed in terms of accuracy, Matthews Correlation Coefficient (MCC), precision, recall, and F1-score for three classes: COVID-19, Normal, and Pneumonia.

Table 1. Results of various machine learning algorithms

Algorithms	Accuracy	MCC
SVM	0.96	0.92
Logistic Regression	0.96	0.92
KNN	0.90	0.80
Random Forest	0.90	0.78
Decision Tree	0.79	0.56

The performance of various machine learning algorithms for classifying chest X-ray images into COVID-19, Pneumonia, and Normal classes was evaluated using HOG features, revealing significant differences in their effectiveness. Support Vector Machine (SVM) demonstrated the highest overall accuracy of 96% and an MCC of 0.92, with exceptional precision (0.97) and recall (0.96) for COVID-19, underscoring its robustness in handling high-dimensional features and effectively separating classes. Logistic Regression mirrored SVM's performance, achieving identical accuracy and MCC values, further validating its utility in such classification tasks.

Conversely, k-Nearest Neighbors (kNN) and Random Forest (RF) both achieved an accuracy of 90% but faced challenges with the imbalanced dataset, particularly in identifying COVID-19 cases, as reflected by their low recalls of 0.53 for this critical class. Despite robust recall values for Pneumonia (0.95 for kNN and 0.98 for RF), these algorithms struggled with minority class detection. The Decision Tree (DT) model demonstrated the lowest performance, with an accuracy of 79% and MCC of 0.56, providing limited reliability for COVID-19 classification due to overfitting and insufficient handling of complex datasets.

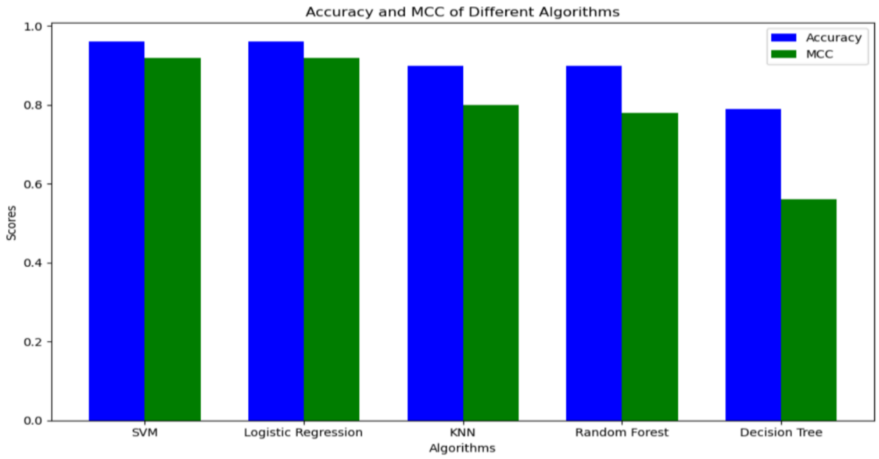

Fig. 1 Accuracy and MCC of Different Algorithms

The results highlight the strengths and weaknesses of each algorithm in the context of COVID-19 X-ray classification. SVM emerged as the most accurate and reliable model, benefiting from its ability to find optimal decision boundaries in high-dimensional feature spaces. Its performance demonstrates the utility of HOG features when combined with sophisticated classifiers for medical image analysis. In contrast, kNN and RF showed moderate performance but struggled with class imbalance, particularly in detecting COVID-19 cases. The limitations of kNN in managing imbalanced datasets and RF's dependence on sufficient feature representation were apparent.

The Decision Tree model, while interpretable, underperformed in terms of accuracy and recall, emphasizing the trade-offs between simplicity and effectiveness. Its inability to generalize well across the dataset highlights the need for more robust models in medical image classification tasks. These findings highlight SVM as the most reliable algorithm for COVID-19 detection, emphasizing the importance of selecting robust models capable of managing class imbalances in medical diagnostics.

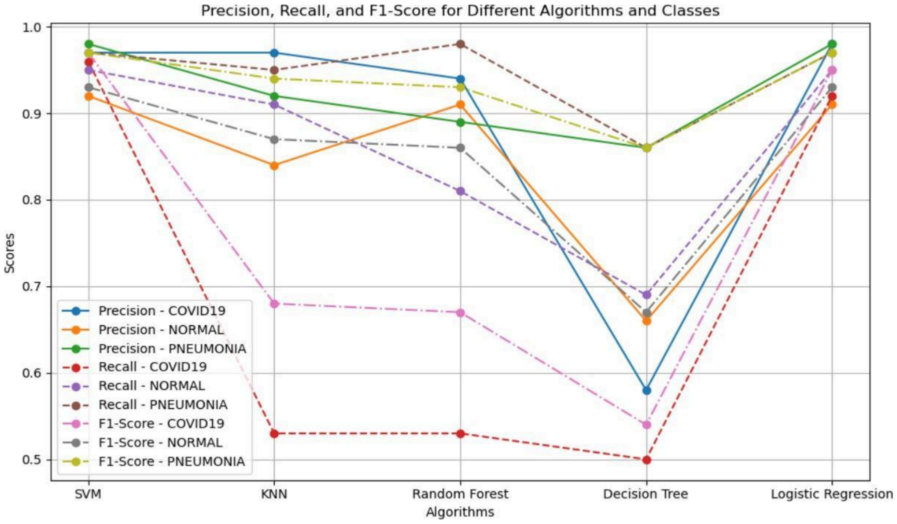

Fig. 2 Precision, Recall and F1-Score for Different Algorithms and Classes

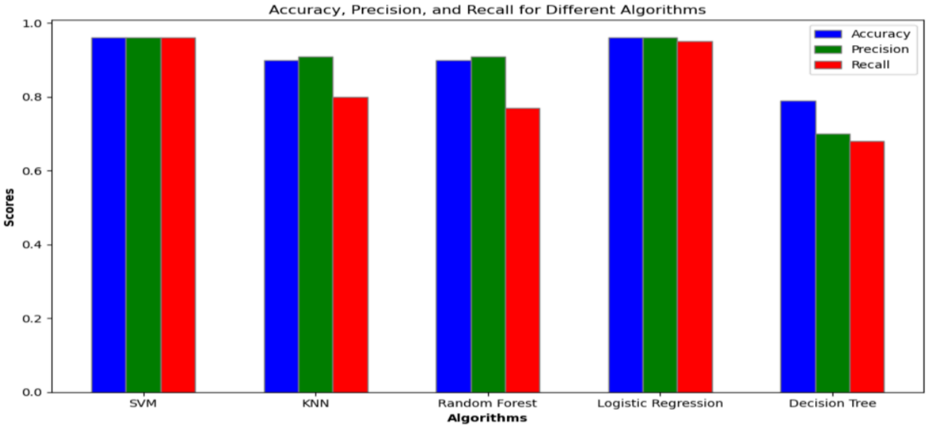

Fig. 3 Accuracy, Precision, Recall for Different Algorithms

5. Conclusion and Future Scope

This study explores the potential of combining Histogram of Oriented Gradients (HOG) features with machine learning algorithms to classify chest X-ray images into COVID-19, Normal, and Pneumonia categories. The analysis demonstrates that Support Vector Machine (SVM) is the most effective model among those tested, achieving an accuracy of 96% and an MCC of 0.92. SVM's ability to handle high-dimensional feature spaces and separate overlapping classes makes it particularly well-suited for this task.

While k-Nearest Neighbors (kNN) and Random Forest (RF) performed moderately well, their limitations in handling imbalanced datasets were evident, particularly for detecting COVID-19 cases. Decision Tree (DT), despite its interpretability, underperformed, highlighting the trade-offs between simplicity and accuracy in machine learning models. The results underscore the efficacy of hand-crafted HOG features for medical image analysis, particularly when paired with advanced classifiers like SVM. However, the study also reveals areas where improvements can be made, especially in addressing class imbalance and enhancing the recall for minority classes such as COVID-19.

The findings of this study highlight several opportunities for future research and development in the field of medical image analysis. Addressing class imbalance remains a critical avenue, as improving the recall for minority classes such as COVID-19 is essential for reliable diagnostics. Techniques like Synthetic Minority Oversampling (SMOTE), cost-sensitive learning, or advanced ensemble methods can enhance model performance. Furthermore, integrating Histogram of Oriented Gradients (HOG) with modern deep learning frameworks, such as Convolutional Neural Networks (CNNs), offers an exciting direction. Feature fusion approaches that combine hand-crafted and deep features can leverage the strengths of both methodologies, improving classification accuracy and adaptability across diverse datasets.

Real-world validation and deployment are equally crucial for broader adoption. Future research should focus on expanding the dataset to include diverse imaging conditions, patient demographics, and multi-center data to ensure generalizability. Additionally, the proposed framework can be extended to classify other diseases observable in chest X-rays, such as tuberculosis or chronic obstructive pulmonary diseases (COPD). This research sets a foundation for developing robust, accessible, and scalable diagnostic tools, contributing to the global effort to enhance healthcare delivery and preparedness for future pandemics.

REFERENCES

[1] Absar, N., Mamur, B., Mahmud, A., Emran, T. B., Khandaker, M. U., Faruque, M. R. I., ...&Elkhader, B. A. (2022). Development of a computer-aided tool for detection of COVID-19 pneumonia from CXR images using machine learning algorithm. Journal of Radiation Research and Applied Sciences, 15(1), 32-43.

[2] Al-Jumaili, S., Al-Azzawi, A., Duru, A. D., & Ibrahim, A. A. (2021, October). Covid-19 X-ray image classification using SVM based on Local Binary Pattern. In 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 383-387). IEEE.

[3] Awotunde, J. B., Ajagbe, S. A., Oladipupo, M. A., Awokola, J. A., Afolabi, O. S., Mathew, T. O., &Oguns, Y. J. (2021, October). An improved machine learnings diagnosis technique for COVID-19 pandemic using chest X-ray images. In International Conference on Applied Informatics (pp. 319-330). Cham: Springer International Publishing.

[4] Cenggoro, T. W., &Pardamean, B. (2023). A systematic literature review of machine learning application in COVID-19 medical image classification. Procedia computer science, 216, 749-756.

[5] Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., &Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. Plos one, 15(6), e0235187.

[6] Eljamassi, D. F., &Maghari, A. Y. (2020, December). COVID-19 detection from chest X-ray scans using machine learning. In 2020 International Conference on Promising Electronic Technologies (ICPET) (pp. 1-4). IEEE.

[7] Erdaw, Y., &Tachbele, E. (2021). Machine learning model applied on chest X-ray images enables automatic detection of COVID-19 cases with high accuracy. International Journal of General Medicine, 4923-4931.

[8] Garlapati, K., Kota, N., Mondreti, Y. S., Gutha, P., & Nair, A. K. (2021, June). Detection of COVID-19 using X-ray image classification. In 2021 5th international conference on trends in electronics and informatics (ICOEI) (pp. 745-750). IEEE.

[9] Goyal, S., & Singh, R. (2023). Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques. Journal of Ambient Intelligence and Humanized Computing, 14(4), 3239-3259.

[10] Greeshma, K. V., & Viji Gripsy, J. (2021). A review on classification and retrieval of biomedical images using artificial intelligence. The Fusion of Internet of Things, Artificial Intelligence, and Cloud Computing in Health Care, 47-66.

[11] Greeshma, K. V., & Sreekumar, K. (2019). Fashion-MNIST classification based on HOG feature descriptor using SVM. Int. J. Innov. Technol. Explor. Eng, 8(5), 960-962.

[12] Hasoon, J. N., Fadel, A. H., Hameed, R. S., Mostafa, S. A., Khalaf, B. A., Mohammed, M. A., &Nedoma, J. (2021). COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images. Results in Physics, 31, 105045.

[13] Hussain, L., Nguyen, T., Li, H., Abbasi, A. A., Lone, K. J., Zhao, Z., ...& Duong, T. Q. (2020). Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection. BioMedical Engineering OnLine, 19, 1-18.

[14] Johri, S., Goyal, M., Jain, S., Baranwal, M., Kumar, V., &Upadhyay, R. (2021). A novel machine learning-based analytical framework for automatic detection of COVID-19 using chest X-ray images. International Journal of Imaging Systems and Technology, 31(3), 1105-1119.

[15] Khan, M. A. (2021). An automated and fast system to identify COVID-19 from X-ray radiograph of the chest using image processing and machine learning. International journal of imaging systems and technology, 31(2), 499-508.

[16] Khan, N., Ullah, F., Hassan, M. A., &Hussain, A. (2020). COVID-19 classification based on Chest X-Ray images using machine learning techniques. Journal of Computer Science and Technology Studies, 2(2), 01-11.

[17] Manav, M., Goyal, M., & Kumar, A. (2023). Role of Optimal Features Selection with Machine Learning Algorithms for Chest X-ray Image Analysis. Journal of Medical Physics, 48(2), 195-203.

[18] Mijwil, M. M. (2021). Implementation of Machine Learning Techniques for the Classification of Lung X-Ray Images Used to Detect COVID-19 in Humans. Iraqi Journal of Science, 2099-2109.

[19] Öztürk, Ş., Özkaya, U., &Barstuğan, M. (2021). Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features. International journal of imaging systems and technology, 31(1), 5-15.

[20] Rehouma, R., Buchert, M., & Chen, Y. P. P. (2021). Machine learning for medical imaging-based COVID-19 detection and diagnosis. International Journal of Intelligent Systems, 36(9), 5085-5115.

[21] Samsir, S., Sitorus, J. H. P., Ritonga, Z., Nasution, F. A., &Watrianthos, R. (2021, June). Comparison of machine learning algorithms for chest X-ray image COVID-19 classification. In Journal of Physics: Conference Series (Vol. 1933, No. 1, p. 012040). IOP Publishing.

[22] Saygılı, A. (2021). A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods. Applied Soft Computing, 105, 107323.

Author Biography

Greeshma K V is currently the Assistant Professor, Department of Forensic Science, University of Calicut, Kerala Police Academy, Thrissur, Kerala, India. She has 8 years of teaching experience. She is also a research scholar at PSGR Krishnammal College for Women, Coimbatore. She received a bachelor's degree in computer applications from the M.G. University in 2009, and a Master's degree in computer applications from the IGNOU in 2014. She completed her Master of Philosophy in Computer Science from the University Amrita Vishwa Vidyapeetham, Coimbatore. She has published over 10 research works in National and International journals and books. Her research interests include image processing, deep learning, computer vision, and cyber forensics.

Dr. J. Viji Gripsy, has 16 years of teaching experience. Currently, she is working as an Associate Professor in Computer Science at PSGR Krishnammal College for Women, Coimbatore. Her area of Specialization is Mobile Ad-hoc and Wireless Sensor Networks. She has published over 25 research works in National and International journals and books. She is currently guiding 4 research scholars in the field of Mobile Networks and Data Mining. She is a life member of the International Association of Computer Science and Information Technology (IACSIT), the International Association of Engineers (IAENG), and the Indian Society for Technical Education (ISTE).