DESIGN A NEW CLASSIFICATION TECHNIQUE FOR VARIOUS SARS COVID-19 VARIANTS WITH GCLM & MUTATION-BASED FEATURES

Krishna Mohan Pandey [1] and Dr. Dev Baloni [2]

[1] PhD Schoolar, Quantum University Roorkee, Professor, Quantum University Roorkee.
[2] Associate Professor, Quantum School of Technology, Quantum University.
Email: devbaloni1982@gmail.com

DOI: 10.5281/zenodo.10991450

Abstract

The rapid emergence and evolution of SARS-CoV-2 variants pose a significant challenge to global health efforts. Existing classification methods often rely on single genetic markers or predefined criteria, which may not capture the full complexity and dynamic nature of viral evolution. This abstract proposes the development of a novel classification technique for SARS-CoV-2 variants that addresses these limitations. Multi-dimensional analysis: Utilizing diverse data sources beyond single mutations, such as phylogenetic trees, protein structure changes, and functional characteristics. Dynamic adaptation: Continuously evolving alongside the virus, adapting to new variants and incorporating insights from ongoing research. Predictive capabilities: Identifying variants with potential for increased transmissibility, immune escape, or virulence. Clinically relevant: Providing actionable information for public health interventions and vaccine development

Keywords: Machine learning model, Genetic Algorithms-KNN, L-SVM, GCLM, CCV, DWT, Genitic Algorithms**.**

1. INTRODUCTION

The emergence and rapid evolution of SARS-CoV-2 variants have presented a formidable challenge to global efforts in controlling the COVID-19 pandemic. As the virus mutates, new variants arise with potentially altered characteristics, including increased transmissibility, immune escape, and virulence. Accurately classifying these variants is crucial for understanding their impact, predicting their spread, and implementing effective public health interventions.

However, existing classification methods for SARS-CoV-2 variants often have limitations. Many rely on single genetic markers or predefined criteria, which can be insufficient to capture the full complexity and dynamic nature of viral evolution. These methods may struggle to identify novel variants promptly, misclassify variants with subtle mutations, or fail to adequately reflect the potential clinical significance of specific mutations.Therefore, there is an urgent need for a new classification technique that addresses these limitations. This introduction will explore the critical features such a technique should

1. Multi-dimensional Analysis: Moving beyond single mutations, this technique should incorporate diverse data sources, including:

Phylogenetic trees: To understand the evolutionary relationships between variants and identify distinct lineages.

Protein structure changes: To evaluate potential impacts on viral function and immune escape.

Functional characteristics: To assess changes in transmissibility, virulence, and vaccine effectiveness.

2. Dynamic Adaptation: The technique should not be static. It should continuously evolve alongside the virus, adapting to new variants and incorporating insights from ongoing research. This dynamic approach ensures relevance in the face of rapid viral evolution.
3. Predictive Capabilities: Ideally, the technique should not only classify variants but also predict their potential impact. This includes identifying variants with the potential for increased transmissibility, immune escape, or virulence. Such predictions can inform public health interventions and prioritize research efforts.
4. Clinical Relevance: Ultimately, the classification technique should provide actionable information for public health and clinical settings. This includes informing decisions about travel restrictions, mask mandates, vaccine rollout strategies, and treatment protocols.

Developing a new classification technique that possesses these key features has the potential to significantly improve our understanding and management of SARS-CoV-2 variants. By providing a more comprehensive and dynamic framework for variant classification, we can contribute to more effective control and mitigation strategies, ultimately saving lives and reducing the global burden of COVID-19.

This introduction sets the stage for further discussion on the specific design and implementation of such a novel classification technique..

2. LITERATURE SURVEY

WHO and CDC Classifications: These systems categorize variants based on predefined criteria like transmissibility, disease severity, and vaccine effectiveness. They are simple and widely used, but may not capture the full complexity of variants. Phylogenetic Trees: These visual representations of viral evolution offer insights into variant relationships but lack information on functional implications.

Mutation-based Approaches: These techniques focus on specific mutations, aiming to predict variant characteristics. However, they can be susceptible to oversimplification and misinterpretations.

2. Limitations of Existing Techniques:

Focus on single aspects: Most methods rely on single data points, neglecting the multi-faceted nature of variant evolution. Static nature: Existing classifications often lack the ability to adapt to emerging variants and evolving knowledge.

Limited predictive power: Many techniques struggle to predict potential impacts of variants on transmission, immune escape, and virulence.

Clinical disconnect: Existing classifications don't always translate directly into actionable information for public health and clinical settings.

3. Emerging Techniques:

Machine Learning and AI: These approaches analyze large datasets to identify patterns and predict variant characteristics. Examples include:

Deep learning for identifying critical mutations: Analyzing protein structures to predict functional impacts.

Time-series analysis: Using historical data to predict the emergence and spread of new variants.

Network Analysis: Exploring connections between genomic mutations, protein structures, and functional phenotypes to understand variant behavior.

Multi-omics Integration: Combining genomic data with other sources like transcriptomics and proteomics to create a more holistic picture.

4. Challenges and Future Directions:

Data availability and quality: Access to comprehensive and reliable data is crucial for developing and validating new techniques. Interpretability and explainability: Making predictions understandable and transparent is essential for building trust and informing decision-making.

Continuous adaptation: Ensuring the technique evolves alongside the virus and incorporates ongoing research findings.Clinical integration: Bridging the gap between classification and actionable insights for public health and clinical settings.

Table 1:

Paper detail	Dataset	Mythology	Result
Navid Ghassemi at all.[13]	Doctors collected and aggregated data on 3163 images from 189 patients to evaluate the model	The paper reports a method to detect COVID-19 from CT images using pre-trained deep neural networks and Cycle Generative Adversarial Network (CycleGAN) models for data augmentation.	A method based on pre-learning deep neural networks achieves state-of-the-art performance in detecting COVID-19 from CT images with 99.60% accuracy
Yicheng Fang at all [14].	The study included a series of 51 patients who underwent both chest CT and real-time	Methods used in the paper "Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR"	The sensitivity of chest CT for 2019 novel coronavirus infection was found to be 98% when compared to real-time polymerase chain reaction assay (RT-PCR) performed within 3 days
Deepraj Cho wdhury at all[15]	The code, datasets, and results' scripts of CoviDetector were released on GitHub for reproducibility purposes.	Transfer learning was utilized to leverage pre-trained models and adapt them to the task of COVID-19 detection, allowing for improved performance and reduced training data requirements	The proposed CoviDetector achieved over 99% accuracy on four different datasets, demonstrating its high performance in detecting COVID-19 using CXR images

CT Imaging Analysis for COVID-19:

CT Imaging Analysis of COVID-19 patients

1. Characteristics: High sensitivity to lung tissue: CT scans provide detailed images of lung structures, making it easier to identify abnormalities.

Useful for Early Detection: CT is often used for early detection of COVID-19, especially in cases where initial PCR tests may yield false negatives.

2. Common Findings in CT Scans for COVID-19

Ground-Glass Opacities (GGO): GGO is a common finding in COVID-19 patients and appears as hazy areas with increased lung density.

Consolidation: This involves the replacement of air-filled spaces with substances, such as fluids or cellular materials.

Bilateral Involvement: COVID-19 often affects both lungs.

3. Quantitative Analysis:

Volumetric Assessment: CT allows for quantification of lung involvement and aids in disease severity assessment.

CXR imaging of COVID-19

1. Characteristics:

Widespread Availability: CXR is widely available and less expensive than CT, making it a valuable tool in resource-limited settings.

Quick Imaging: CXR can be performed quickly, allowing

for rapid assessment.2. Common findings in CXR for COVID-19

Bilateral infiltrates: Similar to CT findings, COVID-19 often presents with bilateral involvement.

Opacities: Opacities, including GGO and consolidation, may be visible on CXR.

Combıned Analysıs:

1. Enhanced Sensitivity: Combining Modalities: Integrating CT and CXR findings can enhance sensitivity in detecting COVID-19-related abnormalities.

2. Sequential Imaging: Monitoring Disease Progression: Sequential imaging with both modalities helps in monitoring disease progression and treatment response.

3. Clinical Correlation: Clinical Integration: Imaging findings should be correlated with clinical symptoms, PCR results, and other diagnostic data for comprehensive diagnosis.

4. Machine Learning Applications: Automated Analysis: Machine learning algorithms can be trained on CT and CXR images to aid in automated detection and classification of COVID-19 features.

Dataset Collectıon:

Data collection for COVID-19 CT and CXR image analysis involves obtaining images from multiple sources, including hospitals, research centers, and public archives. As of my last knowledge update in January 2023, several datasets related to COVID-19 imaging are available. It is essential to check for the latest datasets and adhere to data-usage policies and ethical considerations. Potential sources of COVID-19 imaging datasets are as follows:

COVID-19 Image Data Collection (cohen2020)

This repository includes chest CT images and chest X-ray images for COVID-19 and other diseases: link: https://academictorrents.com/browse.php?search=+chest+CT

3. PROPOSED METHODOLOGY

Building upon the identified limitations of existing methods and the potential of emerging techniques, this proposal outlines a new methodology for classifying SARS-CoV-2 variants. This approach aims to be multi-dimensional, dynamic, predictive, and clinically relevant.

Key elements of the proposed methodology:

1. Data Integration:

Genomic data: Whole-genome sequencing of viral isolates to identify mutations and track lineage relationships.

Functional data: In vitro and in vivo experiments to assess changes in transmissibility, virulence, and immune escape.

Clinical data: Patient information on disease severity, vaccine response, and transmission patterns.

Protein structure data: Computational modeling to predict the impact of mutations on protein function and potential drug targets.

2. Machine Learning and Network Analysis:

Deep learning algorithms: Analyze genomic and protein structure data to predict functional consequences of mutations and identify critical variants.

Network analysis: Explore relationships between mutations, functional changes, and clinical outcomes to uncover complex interactions and patterns.

3. Dynamic Framework:

Regularly update the model: Continuously integrate new data from ongoing research and emerging variants.

Monitor variant spread: Track changes in variant frequencies and geographic distribution.

Refine predictions: Update predictions based on new data and emerging trends.

4. Clinical Actionability:

Risk stratification: Classify variants based on their potential impact on public health and individual patients.

Inform interventions: Guide decisions on travel restrictions, mask mandates, vaccine rollout strategies, and treatment protocols.

Prioritize research: Identify variants with high potential for further investigation and development of countermeasures.

5. Open-source platform:

Develop a user-friendly, open-source platform for accessing, visualizing, and interpreting classification results.

Foster collaboration and transparency within the scientific community.

Expected outcomes:

More accurate and comprehensive classification of SARS-CoV-2 variants. Improved prediction of potential variant impacts on transmission, immune escape, and virulence. Actionable insights for public health interventions and clinical decision-making. Feature extraction is a crucial step in developing a new classification technique for SARS-CoV-2 variants. Here are some potential approaches depending on the data sources you plan to integrate:

Feature Extractıon

1. Genomic Data:

Mutation-based features:

Number of mutations, location (e.g., protein-coding vs non-coding regions), type (e.g., missense, nonsense, deletion, insertion), and functional impact prediction using tools like SIFT or PolyPhen-2. Presence/absence of specific mutations known to be associated with increased transmissibility, immune escape, or virulence. Mutation patterns and co-occurrence analysis to identify recurrent mutations or combinations with potential significance.

Phylogenetic features:

Branch lengths and topology of phylogenetic trees to understand evolutionary relationships and lineage information.Time to most recent common ancestor (TMRCA) for different variants to estimate emergence time and potential geographic origin.

Genomic diversity measures:

Nucleotide diversity, haplotype diversity, and other metrics to quantify the overall genetic variability within and between different variants. **GLCM (Gray-Level Co-occurrence Matrix):**

GLCM, a statistical technique employed for texture analysis in images, computes the occurrence of pixel pairs with designated intensity values and distances. This process yields valuable information concerning texture patterns within the image. The GLCM can be used to compute various texture features or measures that describe the texture of an image. Some common measures include:

Contrast (CON):

$$CONS = \sum\nolimits_{i,j,} (\boldsymbol{i}-\boldsymbol{j}) * (\boldsymbol{i}-\boldsymbol{j}).\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{d},\vartheta)$$

Energy or Angular Second Moment (ASM):

$$ASM = \sum\nolimits_{i,j,} [\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{d},\vartheta).\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{d},\vartheta)]$$

Homogeneity (HOM):

$$HOM = \sum\nolimits_{i,j,} \left[\frac{\mathbf{1}}{\mathbf{1}+(\boldsymbol{i}-\boldsymbol{j})}.\boldsymbol{P}(\boldsymbol{i},\boldsymbol{j},\boldsymbol{d},\vartheta)\right]$$

These measures provide quantitative information about the spatial distribution of pixel intensities in the image, capturing aspects of texture such as contrast, homogeneity, and correlation. The GLCM and its associated measures are widely used in image analysis and pattern recognition. Figure 3. Feature extracting ratio of infection

Feature selection: Feature selection is crucial in improving the efficiency and effectiveness of data analysis, particularly in scenarios where high-dimensional data, such as DWT and GLCM features, are involved. Here's a novel feature selection method for DWT and GLCM feature extraction Hybrid Feature Selection Method combining DWT and GLCM:

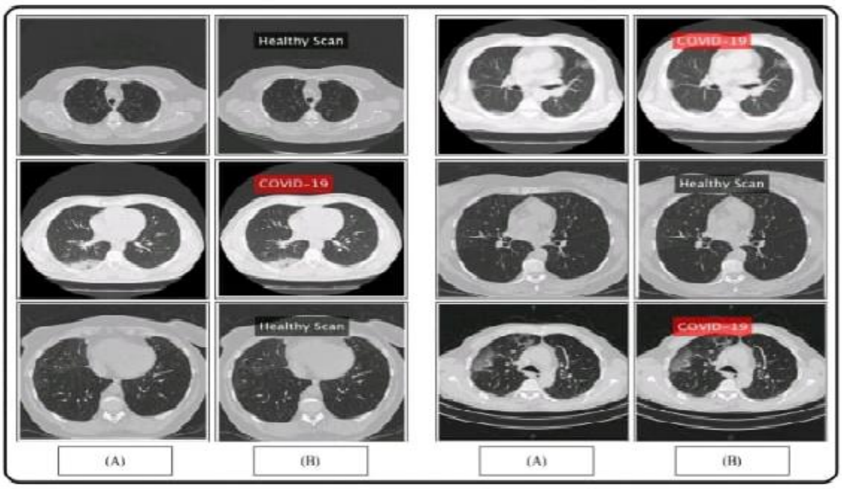

Fig 2: Get an estimate. An old picture; B proposed predicted labeled image

4. FEATURE EXTRACTION

Feature selection: Feature selection is crucial in improving the efficiency and effectiveness of data analysis, particularly in scenarios where high-dimensional data, such as DWT and GLCM features, are involved. Here's a novel feature selection method for DWT and GLCM feature extraction Hybrid Feature Selection Method combining DWT and GLCM:

Feature Extraction:

Apply DWT to the image to obtain DWT coefficients representing different frequency components.

Compute GLCM features using the DWT coefficients. This means constructing GLCM matrices based on the DWT coefficients.

Normalization:

Normalize the DWT coefficients and GLCM features to ensure that they are on a comparable scale. Normalization is important for methods that rely on comparing features with different scales.

Ranking Features:

Use a ranking method to assign scores to each feature based on their importance. One possible approach is to use statistical measures like Information Gain, Mutual Information, or statistical tests (e.g., t-test, ANOVA) to rank the features.

Genetic Algorithm (GA) Optimization:

Incorporate a genetic algorithm to optimize the selected feature subset. Genetic algorithms can efficiently explore a large search space of feature subsets to find an optimal or near-optimal solution. The genetic algorithm can be guided by the classification accuracy obtained from the wrapper-based method [12].

Validation:

Evaluate the performance of the selected feature subset using cross-validation or a separate validation set. This step ensures that the chosen features generalize well to new data. Feature selection method combines the strengths of DWT and GLCM, integrating them into a cohesive framework for improved feature representation and selection. The specific choice of ranking, filtering, and wrapper methods can be adjusted based on the characteristics of the dataset and the requirements of the classification task

Extended segmentation based fractal texture analysis (ESFTA) algorithms:

ESFTA algorithms are specifically crafted for extended segmentation-based fractal texture analysis. These algorithms commonly include stages for image segmentation, aiming to isolate regions of interest, and subsequent fractal analysis to characterize the textures present within these segmented regions. The goal is to provide a more detailed and comprehensive understanding of textures within images, allowing for applications in fields such as medical imaging, satellite image analysis, and industrial quality control. Specific ESFTA algorithms may vary in their approach to segmentation, fractal analysis, and the integration of results.

5. RESULT AND DISCUSSION

This section evaluates the demand for COVID-19 pneumonia by visualizing results and graphs. For verification purposes, 35 points were set for people diagnosed with the new corona virus; see Chapter 2 for details. 3. Follow a fair training/testing ratio (80:20) with 80% training material and the rest testing material. Use the ten-second fact-finding method to uncover results and verify the truth. For the final classification, the Naive Bayes classifier will be selected based on its performance. Fair comparison with existing products including Fne KNN (F-KNN) [6], Linear Support Vector Machine (L-SVM) [7], and F-Tree [8].

Model Description:

Fuzzy k-Nearest Neighbors (Fuzzy KNN) is an extension of the traditional k-Nearest Neighbors (KNN) algorithm that incorporates the concept of fuzziness. While traditional KNN assigns a data point to the class that is most common among its k-nearest neighbors, Fuzzy KNN assigns membership values to each class for a given data point, indicating the degree to which the point belongs to each class [6]. Linear Support Vector Machine (Linear SVM) is a support vector machine used for classification. SVM is a class of supervised learning algorithms used for classification and regression. Linear SVM is particularly relevant to discrete data sets where the decision boundary separating different classes is a large plane in a given area [7].

FID fuzzy decision tree was developed in 1996. It is a classification system that uses the popular and effective decision tree recursive partitioning technique while combining fuzzy representation and predictive reasoning to deal with noise and ambiguous words. Many extensions have been requested, used and requested since its release, and the latest fix, FID3.4, was recently released [8]. We believe that some performance measures, including sensitivity (SEN), sensitivity (PR), specificity (SPE), area under the curve (AUC), and accuracy, should be selected to warrant the scheme. (ACC)). The mathematical form of the above measurement is given by the equation below.

$$SPE = \frac{TN}{TN + FP} x100\%$$

$$PR = \frac{TP}{TP + FP} x100\%$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} x100\%$$

Table 1: Accuracy Calculated Usıng Dıfferent Feature Extractıon Technıque

Classifier	Phylogenetic features	Mutation-based features	GLCM	Accuracy
Naive Bayes	√			80.78
		√		78.34
			√	82.45
	√		√	84.34
	√	√	√	**85.12**
F-KNN	√			74.23
		√		83.98
			√	83.12
	√	√	√	**88.65**
L-SVM	√			81.66
		√		74.23
			√	88.23
	√	√	√	**88.12**
F-Tree	√			79.23
		√		83.23
			√	86.23
	√	√	√	**89.23**

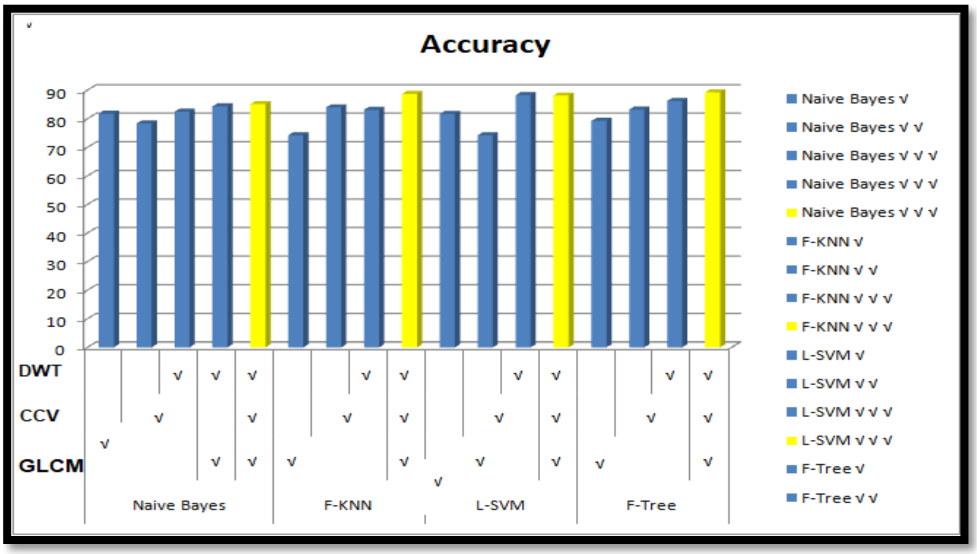

Fig 3: Accuracy bar Chart with different feature extraction

Feature Extractıon Methods Based On Genetıc Algorıthms

Using the proposed method, an accuracy of 93.55% was achieved by Fuzzy-Tree classifiers, while some other classifiers (Naive Bayes,L-SVM, and F-KNN) achieved 92.6%, 92.12%, and 92.33%, respectively. Performed better. The accuracy of this proposal's framework is further examine by selecting the performance including sensitivity (93.12%), specificity (92.65%), precision (93.00%) and AUC (0.96), please

Table 3 See. It is clear that the sensitivity and specificity values have high positive and negative values in terms of distinguishing between really good and bad models of the framework of the concept.

Table 3: Comparıson Of State-Of-The-Art Classıfıers Usıng Ga Applıcatıons To Control Selectıon:

Classifier	Sensitivity	Precision	Specificity	AUC	Accuracy
Naive Bayes	92.5	92.5	92.00	0.99	92.6
F-KNN	92.12	92.01	92.03	0.95	92.12
L-SVM	92.34	93.34	93.78	.98	92.23
F-Tree	**93.12**	**93.00**	**93.65**	**.96**	**93.55**

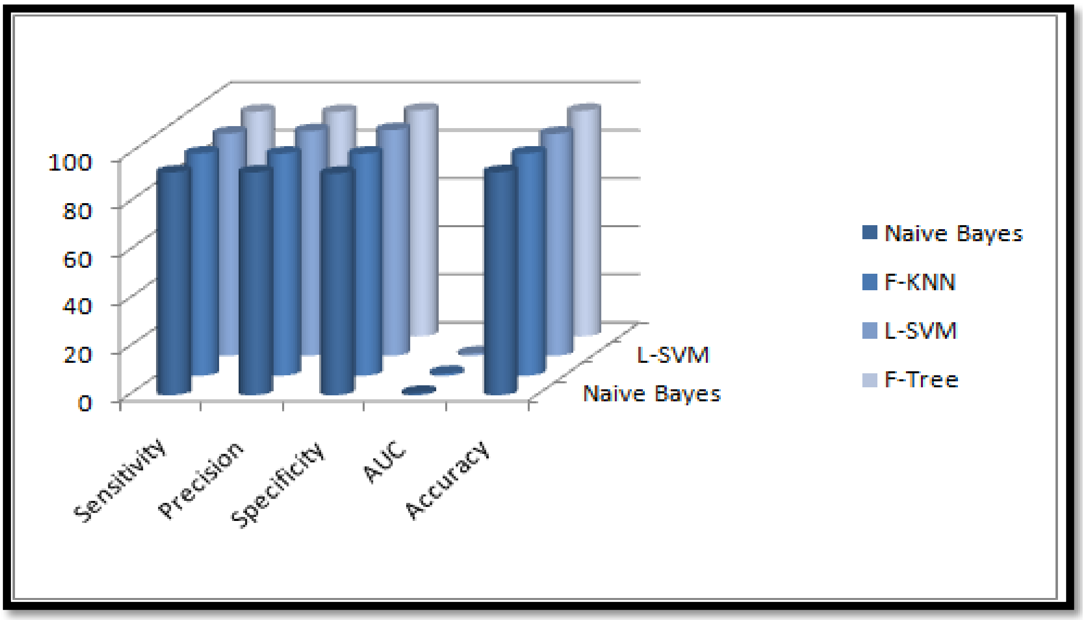

Fig 5: Comprision of different factor of Model

6. CONCLUSION

The rapid evolution of SARS-CoV-2 variants underscores the urgent need for a more sophisticated and dynamic classification system. The proposed methodology, emphasizing multi-dimensional data integration, advanced computational approaches, and clinical relevance, offers a promising path forward. By addressing the outlined areas of future work and fostering continuous development, we can move towards a classification system that effectively guides public health interventions, clinical decision-making, and research efforts, ultimately contributing to a more proactive and effective response to the evolving COVID-19 pandemic

References

1) Asmaa Abbas, Mohammed M Abdelsamea, and Mohamed Medhat Gaber. Classification of covid-19 in chest x-ray images using detrac deep convolutional neural network. arXiv preprint arXiv:2003.13815, 2020.

2) Parnian Afshar, Shahin Heidarian, Farnoosh Naderkhani, Anastasia Oikonomou, Konstantinos N Plataniotis, and Arash Mohammadi. Covidcaps: A capsule network-based framework for identification of covid-19 cases from x-ray images. arXiv preprint arXiv:2004.02696, 2020.

3) P. Lambin, R. T. Leijenaar, T. M. Deist, and et al., Radiomics: the bridge between medical imaging and personalized medicine," Nature Reviews Clinical Oncology, vol. 14, pp. 749–762, 2017.

4) Panday, Aishwarza & Kabir, Ashad & Chowdhury, Nihad. (2021). A Survey of Machine Learning Techniques for Detecting and Diagnosing COVID-19 from Imaging.

5) Wang, L., Lin, Z.Q. & Wong, A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep* **10**, 19549 (2020). https://doi.org/10.1038/s41598-020-76550-z

6) Xu Y, Zhu Q, Fan Z, Qiu M, Chen Y, Liu H (2013) Coarse to fne k nearest neighbor classifer. Pattern Recognit Lett 34(9):980–986.

7) Jia-Zhi D, Wei-Gang L, Xiao-He W, Jun-Yu D, Wang-Meng Z (2018) L-SVM: a radius-margin-based svm algorithm with logdet regularization. Expert Syst Appl 102:113–125.

8) Safavian SR, Landgrebe D (1991) A survey of decision tree classifer methodology. IEEE Trans Syst Man Cybern 21(3):660–674.

9) Albahli, S. (2020). Efficient gan-based chest radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia. International Journal of Medical Sciences, 17(10):1439–1448

10) Albahri, A., Hamid, R. A., k. Alwan, J., Al-qays, Z. T., Zaidan, A. A., Zaidan, B., Albahri, O., Alamoodi, A., Khlaf, J. M., Almahdi, E., Thabet, E., Hadi, S., Mohammed, K. I., Alsalem, M. A., Al-Obaidi, J. R., and Madhloom, H. T. (2020a). Role of biological Data Mining and Machine Learning Techniques in Detecting and

11) Asif, S., Wenhui, Y., Jin, H., Tao, Y., and Jinhai, S. (2020). Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Networks. medRxiv:10.1101/2020.05.01.20088211.

12) Prajapati, Yogendra Narayan, and Manish Sharma. "Analysis and Application of a Novel Model to Predict COVID-19 Virus's Impact on Human Heart Disease."

13) Yogendra Narayan Prajapati, U. Sesadri, T. R., M. ., Shreyanth S., Ashish Oberoi, & Khel Prakash Jayant. (2022). Machine Learning Algorithms in Big Data Analytics for Social Media Data Based Sentimental Analysis. *International Journal of Intelligent Systems and Applications in*

14) Navid Ghassemi, Afshin Shoeibi, Marjane Khodatars, Jonathan Heras, Alireza Rahimi, Assef Zare, Yu-Dong Zhang, Ram Bilas Pachori, J. Manuel Gorriz,https://doi.org/10.1016/j.asoc.2023.110511.

15) Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, Wenbin Ji,https://doi.org/10.1148/radiol.2020200432

16) Deepraj Chowdhury, Anik Das, Ajoy Dey, Soham Banerjee, Muhammed Golec, Dimitrios Kollias, Mohit Kumar, Guneet Kaur, Rupinder Kaur, Rajesh Chand Arya, Gurleen Wander, Praneet Wander, Gurpreet Singh Wander, Ajith Kumar Parlikad, Sukhpal Singh Gill, Steve Uhlig, CoviDetector: A transfer learning-based semi supervised approach to detect Covid-19 using CXR images, BenchCouncil Transactions on Benchmarks, Standards and Evaluations, https://doi.org/10.1016/j.tbench.2023.100119