International Journal of Computational Science and Engineering Research
ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 1, Issue. 1, 2024 , Pages: 5 - 9
Received: 07 May 2024 ; Accepted: 27 August 2024 ; Published: 21 October 2024.
Research Paper , https://doi.org/10.63328/IJCSER-V1RI1P2

Covid-19 Identification and Surveillance System using AI

Dekka Satish [1*], **K. Narasimha Raju** [2]

[1] Department of CSE, Lendi Institute of Engineering & Technology (A), Vizianagaram, JNTU GV, AP, India ; satishmsc4u@gmail.com
[2] Department of CSE, Gayatri Vidya Parishad College of Engineering (A) , Visakhapatnam, Andhra University, AP, India ; rj.vizagg@gmail.com

Abstract: Efficient screening for SARS-CoV-2 is crucial for the timely and accurate diagnosis of COVID-19, helping to alleviate the burden on healthcare systems. To assess infection risk, predictive models that incorporate multiple variables have been developed. These models support medical professionals in prioritizing patient care, especially in regions with limited healthcare resources. In this study, we developed a machine learning algorithm trained on data from 51,831 individuals who underwent COVID-19 testing, of which 4,769 cases were confirmed positive. The test dataset comprised individuals tested during the following week, including 3,624 confirmed cases. Our model relies on easily obtainable data, derived from basic demographic and health-related questions, using publicly available information released by the Israeli Ministry of Health. This approach can help prioritize COVID-19 testing when resources are constrained. Additionally, this project incorporates a Convolutional Neural Network (CNN) for detecting COVID-19 through chest X-ray images, alongside an XGBoost model designed to identify symptoms, enhancing the overall diagnostic process.

Keywords: CNN, Machine Learning, X-ray image, Gradient Boost Algorithm, Python.

1. Introduction

On January 30, 2020, the Emergency Committee under the International Health Regulations of the World Health Organization (WHO) officially declared the outbreak of the novel coronavirus disease, later named COVID-19, as a “Public Health Emergency of International Concern.” Caused by the virus SARS-CoV-2, COVID-19 rapidly spread across all inhabited continents, becoming an unprecedented global public health crisis. By October 2020, the Johns Hopkins University Coronavirus Resource Center reported over one million global fatalities, with infections continuing to surge.

The rapid transmission of SARS-CoV-2 is largely attributed to its ease of spread via respiratory droplets through coughing, sneezing, and close contact. Common symptoms include fever, cough, and shortness of breath, though the illness can escalate to pneumonia, multi-organ failure, and even death. Despite ongoing efforts, no definitive cure has been developed to date. Most global and national responses have focused on mitigating further transmission through public health measures such as travel restrictions, lockdowns, and social distancing. Given the high transmissibility of SARS-CoV-2 and uncertainties in its transmission dynamics, early responses heavily relied on broad, non-pharmaceutical interventions to save lives. The escalating number of infections underscored the need for a multidisciplinary approach, combining medical expertise and economic strategies to confront the crisis effectively. In this context, machine learning and artificial intelligence (AI) emerged as valuable tools to assist in various aspects of pandemic management.

AI technologies, particularly in healthcare, are being utilized to screen potential COVID-19 cases, track virus transmission, identify promising clinical trials, and aid in vaccine development. On March 16, 2020, the White House, in collaboration with research institutions and technology companies, launched a call to action to support AI-driven initiatives in response to COVID-19. In line with this, the Allen Institute for AI, in partnership with major research groups, released the COVID-19 Open Research Dataset (CORD-19), a publicly available and regularly updated repository of scholarly articles to accelerate scientific discovery. Hundreds of research teams have since collaborated globally, contributing to real-time data collection, predictive modeling, and recovery solutions. Machine learning has been especially useful in forecasting infection risks and supporting healthcare decision-making. This paper aims to highlight these efforts and examine the critical role of machine learning in addressing challenges

posed by SARS-CoV-2. Prominent among these efforts is the work of the European Laboratory for Learning and Intelligent Systems (ELLIS), which integrates top academic minds and industry experts to harness AI for societal and economic benefit. This paper discusses various research projects associated with the ELLIS network, along with contributions from the Council of Europe's Ad hoc Committee on Artificial Intelligence (CAHAI).

Key objectives include examining the progress of machine learning technologies in combating COVID-19, identifying current achievements and limitations, and analyzing how AI applications have influenced different sectors during the pandemic. Each case study discussed highlights the context of use, machine learning methods applied, and outcomes from a data science perspective. One notable advantage of AI-based diagnostic systems is their ability to enhance the accuracy of COVID-19 diagnosis, optimize resource allocation, and reduce healthcare worker exposure. Access to reliable predictive models is crucial for understanding the potential spread and impact of infectious diseases, and for guiding policy decisions and evaluating public health interventions [1].

COVID-19, like other coronaviruses such as MERS-CoV and SARS-CoV, demonstrates complex, nonlinear transmission patterns, making traditional epidemiological models less effective under certain conditions. The novel coronavirus, first identified in Wuhan, China, in late 2019, rapidly spread within 30 days to become a global threat. On February 11, 2020, the WHO officially named the disease COVID-19. Given its high person-to-person transmissibility, AI-powered tools and digital health technologies can play a vital role in curbing the virus's spread. The increasing availability of electronic health records has opened new opportunities for applying machine learning algorithms to enhance diagnosis and predict disease progression [2].

As of May 16, 2020, COVID-19 had affected over 213 countries and regions worldwide, with millions infected and a significant number of deaths. The rapid increase in cases overwhelmed medical facilities and highlighted the limitations of available healthcare resources. Delays in diagnostic testing and the high costs associated with large-scale testing created challenges for timely and effective treatment [3].

This project proposes the use of machine learning methods to predict the likelihood of COVID-19 infection in individuals. The goal is to develop an intelligent system that assists in prioritizing testing and treatment, particularly when resources are constrained, thereby supporting healthcare professionals in delivering timely and accurate care [4].

2. Literature Survey

Machine Learning (ML), a subset of Artificial Intelligence (AI), originated from the field of pattern recognition and is primarily used to structure and interpret data for meaningful analysis. In recent years, ML has been widely adopted across numerous sectors, including healthcare, finance, defense, and space exploration. As a rapidly advancing discipline, ML enables computers to learn from data and improve their performance without being explicitly programmed. It involves training models on existing data to identify patterns and make predictions about future outcomes. The core concept behind machine learning is the ability to build statistical models based on input data, allowing systems to make informed decisions without human intervention. By identifying trends and correlations within datasets, ML algorithms can generate accurate predictions and automate various analytical tasks. One commonly used ML technique is the Support Vector Machine (SVM), which seeks to identify the optimal hyperplane that separates data points belonging to different categories. The data points closest to the hyperplane, known as support vectors, play a critical role in defining this boundary [5].

Artificial Neural Networks (ANNs) are designed to emulate the functioning of the human brain. The basic building block of an ANN is the neuron, which processes inputs and produces outputs. Neurons are interconnected to form a network, which is then trained using data to minimize prediction errors. Optimization algorithms are applied during training to enhance the model's accuracy. Another robust algorithm is the Random Forest (RF), which leverages ensemble learning and random sampling to achieve high accuracy and generalization [6]. Random Forests consist of multiple decision trees, and as the number of uncorrelated trees increases, so does the accuracy of the predictions. RF models can also handle missing data effectively. Several relevant studies have contributed to the understanding and prediction of COVID-19 using AI techniques:

Nanshan Chen et al, conducted a retrospective study at Jinyintan Hospital in Wuhan, China, analyzing epidemiological characteristics, symptoms, lab findings, CT scans, and clinical outcomes of COVID-19 patients. While not focused on predictive modeling, this study offers valuable clinical insights [7].

Shuai Wang et al, applied deep learning techniques to CT scans to extract visual features for identifying COVID-19-related abnormalities. Their work highlights the effectiveness of AI in diagnostic imaging, distinguishing

COVID-19 cases from other types of pneumonia.

Dawei Wang et al, examined clinical data from Zhongnan Hospital in Wuhan, including demographic, lab, and radiological information, providing insights that could enhance prediction models for COVID-19.

Halgurd S. Maghdid et al, proposed a novel AI-based framework utilizing smartphone sensors and uploaded CT images to detect COVID-19 and assess pneumonia severity. The multi-sensor approach focuses on symptom-based prediction.

Ali Narin et al, developed an automated diagnostic tool using three CNN-based models—ResNet50, InceptionV3, and Inception-ResNetV2—to identify COVID-19 cases via chest X-rays. The study evaluates classification accuracy across these models [8].

Another study proposed a model using three clinical indicators to predict COVID-19 mortality risk, employing the XGBoost algorithm. Unlike our approach, their focus was on mortality prediction rather than initial diagnosis.

Comparative analyses in various studies have explored ML models for forecasting COVID-19 outbreaks in different countries. These studies emphasize the potential of machine learning in public health surveillance. Further benchmarking of machine learning and deep learning methods, as well as ICU scoring systems, has been performed using publicly available clinical datasets to evaluate performance on diverse clinical prediction tasks. In contrast to the above-mentioned research, our study focuses specifically on predicting COVID-19 infection using clinical data rather than imaging or sensor inputs. While many existing models rely on CT scans, symptoms, or national case data for prediction, there is limited research utilizing purely clinical records for accurate diagnosis. This thesis aims to fill that gap by applying machine learning algorithms to clinical features of confirmed COVID-19 patients. Additionally, our study seeks to identify the most influential clinical attributes that impact the performance of the predictive model [9].

3. Methodology

Data Pre-processing

Data pre-processing plays a vital role in building effective machine learning models. Raw data often contains inconsistencies such as missing values or outliers, which can negatively affect the accuracy and reliability of the model. Handling Missing Values: To address gaps in the dataset, we applied a simple imputation technique using the `Simple Imputer` from the `scikit-learn` Python library. The imputation was done using the mean of each feature, replacing any missing entries accordingly.

Encoding Categorical Features: Categorical variables were transformed using One-Hot Encoding, implemented through the `OneHotEncoder` module in Python. This technique converts categorical attributes into a binary matrix format, allowing the model to interpret them numerically without implying any ordinal relationship.

Implementation

The experiment was carried out using Python IDLE, which serves as the default integrated development and learning environment for Python. The implementation was structured in multiple stages, as outlined below:

Data Segmentation: After collecting the dataset, the patient records were divided into multiple subsets containing 100, 150, 200, 250, 300, and 355 entries respectively.

Cross-Validation: A 5-fold cross-validation method was employed to ensure the robustness and reliability of the results. This technique was used to shuffle and partition the data for testing each machine learning model.

Model Evaluation: Each machine learning algorithm was tested using 5-fold cross-validation across the different record sets. The prediction accuracy of each model at varying dataset sizes was then compared to identify the most suitable algorithm for this specific dataset.

4. Existing System

A common method to ensure data confidentiality when outsourcing is to encrypt the data before storing it externally.

Searchable encryption techniques allow users to store encrypted data in the cloud while still enabling keyword-based searches over the encrypted content. Numerous approaches have been developed under various threat models, offering functionalities such as single keyword search, similarity search, Boolean search with multiple keywords, ranked search, and ranked search with multiple keywords. Among these, multi-keyword ranked search has gained considerable interest due to its practical applicability [10]. More recently, dynamic searchable encryption schemes have been introduced, enabling data owners to insert and delete documents in the encrypted dataset stored on the cloud. This is especially important, as data updates are a frequent requirement in real-world applications.

Limitations of the Existing System

Reduced Data Usability: Traditional keyword-based retrieval methods, which work effectively on unencrypted

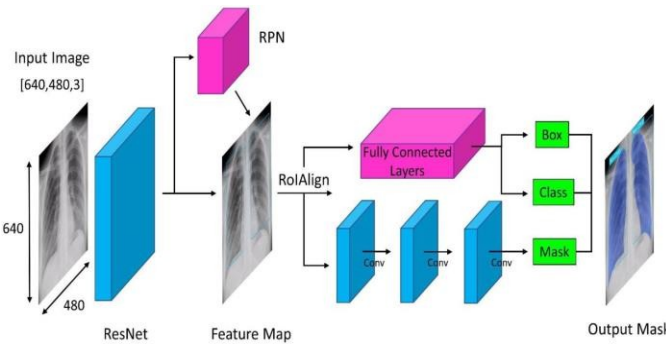
Vol. 1, Issue 1, 2024 https://doi.org/10.63328/IJCSER-V1RI1P2

data, cannot be directly applied to encrypted datasets. Downloading the entire dataset for decryption and local processing is not feasible due to bandwidth and storage limitations. High Computational Overhead: Existing methods often involve significant processing costs for both cloud servers and end users, making them less practical for large-scale or real-time applications.

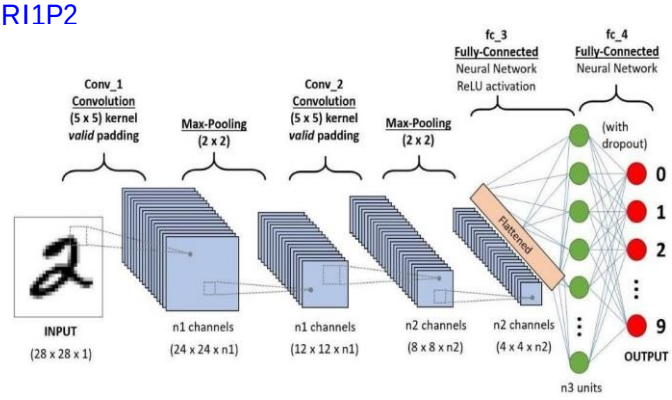
Proposed System

This paper introduces a secure tree-based search framework for encrypted cloud data that supports both multi-keyword ranked search and dynamic document operations. The approach integrates the vector space model with the widely-used Term Frequency–Inverse Document Frequency (TF-IDF) scheme for both index construction and query formulation. This combination facilitates efficient and accurate multi-keyword ranked searches. To enhance search performance, a tree-structured index is developed, and a custom "Greedy Depth-First Search" algorithm is proposed to navigate this structure effectively [11].

For the machine learning component, a multilayer Convolutional Neural Network (CNN) architecture is utilized. The model includes a single hidden layer with two sigmoid-activated neurons and an output layer consisting of two SoftMax neurons. The hidden layer is visually represented by a red dashed rectangle, while the output layer is shown as an orange rectangle. The network outputs [12] a probability distribution across two classes, corresponding to survival and mortality predictions. The architecture of the neural network is further illustrated through synaptic connections, depicted as blue lines, each carrying a distinct weight. A comprehensive analysis of the model's features is included in Appendix B, where 21 attributes are prioritized based on their relevance. Additionally, the proposed machine learning algorithm is applied for detecting symptoms, enhancing the overall diagnostic capability of the system.



The secure k-Nearest Neighbors (KNN) algorithm is utilized to encrypt both index and query vectors, while still ensuring accurate computation of relevance scores between the encrypted vectors.



To address various attack scenarios across different threat models, we propose two secure search mechanisms:

Basic Dynamic Multi-Keyword Ranked Search (BDMRS) designed for environments where only ciphertext is known. And Enhanced Dynamic Multi-Keyword Ranked Search (EDMRS)– tailored for situations where an adversary may possess additional background knowledge.

Key Advantages of the Proposed System

Efficient and Flexible Search Performance: Thanks to the uniquely structured tree-based index, the search process operates with sub-linear time complexity. It also supports seamless insertion and deletion of documents, making it suitable for dynamic data environments.

Comprehensive Searchable Encryption: The proposed system enables precise multi-keyword ranked search while also accommodating dynamic document operations such as updates, additions, and removals [13].

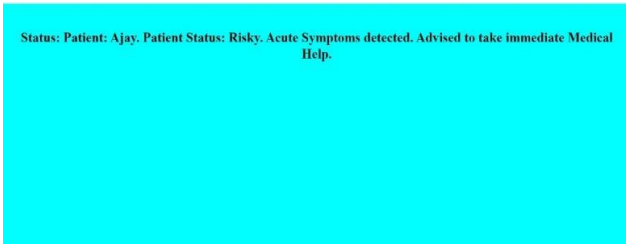
Optimized Search Complexity: The design of the index inherently maintains logarithmic search complexity. In practical scenarios, the use of the customized “Greedy Depth-First Search” algorithm significantly improves search efficiency. Additionally, parallel search capabilities can be employed to further reduce the overall time required for processing queries.

5. Proposed System

The proposed system presents a secure and efficient searchable encryption framework for cloud-based data storage, with support for multi-keyword ranked search and dynamic operations on encrypted document collections. The system addresses data privacy concerns while ensuring usability through sub-linear search time and accurate query results [14].

Vol. 1, Issue 1, 2024 https://doi.org/10.63328/IJCSER-V1RI1P2

This system combines tree-based index structures, secure kNN encryption, and a custom Greedy Depth-First Search (GDFS) algorithm to achieve efficient and secure search functionality over encrypted data. Additionally, it incorporates a machine learning component for predictive analysis based on user symptoms using a multilayer Convolutional Neural Network (CNN).



6. Conclusion and Future Scope

In this study, a systematic literature review was carried out to identify the most appropriate machine learning algorithm for predicting COVID-19 in patients. The review did not find definitive evidence to establish a single algorithm as the optimal classification method for accurate prediction. As a result, multiple algorithms were selected and trained using clinical data from patients.

To evaluate the accuracy of these machine learning models, each algorithm was trained with datasets of varying sizes. Performance was assessed using accuracy metrics, and the models were further analyzed to determine which clinical features had the most significant impact on COVID-19 prediction outcomes.

Machine learning holds vast potential in the healthcare sector. For future work, it is recommended to explore calibrated and ensemble methods, which may offer faster and more reliable results compared to individual algorithms. Additionally, AI-powered applications integrating data from various sensors and features could be developed to enhance early disease detection and diagnosis.

Given the critical role of predictive analytics in healthcare, there is also potential to design a forecasting system that could assess the risk of future disease outbreaks. Such a system would consider socio-economic and cultural factors, helping to prepare for and potentially prevent threats to global health.

References

[1]. Countries where Coronavirus has spread - Worldometer. Library Catalog: www.worldometers.info.
[2]. COVID-19 situation reports. Library Catalog: www.who.int.
[3]. Diagnosis of covid-19 and its clinical spectrum dataset. url=https://kaggle.com/einsteindata4u/covid19.
[4]. WHO Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020. Library Catalog: www.who.int.
[5]. WHO EMRO | Questions and answers | COVID-19 | Health topics.
[6]. Support Vector Machine Machine learning algorithm with example and code, January 2019. Library Catalog: www.codershood.info Section: Machine learn- ing.
[7]. Ali Al-Hazmi. Challenges presented by MERS corona virus, and SARS corona virus to global health. Saudi journal of biological sciences, 23(4):507–511, 2016.Publisher: Elsevier.
[8]. Sina F Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M Atkinson. Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188, 2020.
[9]. Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064– 1069, 2016.
[10]. Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, pages 19–48, 2010.
[11]. Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, pages 19–48, 2010. Publisher: InTech.
[12]. David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high- risk and high-cost patients. Health Affairs, 33(7):1123–1131, 2014.