Advancing Machine Learning in COVID-19 Diagnostics: Symptom-Based Classification and Ensemble Techniques

Jayendra S. Jadhav[1], Jyoti Deshmukh[2]

[1]PhD Research Scholar, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, Mumbai, India-400053, jayendra071985@gmail.com,
[2]Department of Computer Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, Mumbai, India-400053, Jyoti.Deshmukh@mctrgit.ac.in

KEYWORDS
COVID-19, Ensemble Learning, Disease Detection, Machine Learning

ABSTRACT
The COVID-19 pandemic necessitated the development of diagnostic methods that are not only rapid and accurate but also capable of distinguishing COVID-19 from similar respiratory diseases. The use of ensemble learning has proven effective in enhancing diagnostic accuracy through detailed symptom analysis. Previous studies have relied on traditional machine learning techniques like logistic regression and decision trees for early detection. These methods often struggle with symptom overlap, a challenge that ensemble learning addresses by combining predictions from various models to improve diagnostic precision. This study implements an ensemble learning framework that integrates diverse models to refine the accuracy of COVID-19 diagnoses based on patients' symptoms. The process includes data preprocessing, feature engineering, and optimizing ensemble methods such as random forests and gradient boosting. The Symptoms Based COVID-19 Classification Algorithm offers a straightforward diagnostic approach by assessing symptom proportions against a threshold. Conversely, the COVID-19 Detection Using an Ensemble Learning Model employs a sophisticated ensemble of models, enhancing diagnosis through weighted symptom analysis. The Symptoms Based Algorithm achieves 85% accuracy with some limitations in specificity, whereas the Ensemble Learning Model shows superior performance, achieving 90% accuracy and effectively minimizing false positives. Although the Symptoms Based Algorithm is useful for quick assessments, the Ensemble Learning Model's accuracy and comprehensive analysis make it more suitable for clinical application. Future efforts will focus on integrating broader data sources and validating these models in practical scenarios.

1. INTRODUCTION

The onset of the COVID-19 pandemic underscored the critical need for diagnostic methods that are rapid, accurate, and efficient, particularly for newly emerging infectious diseases. In its early stages, COVID-19 presented symptoms such as fever, dry cough, and fatigue—clinical indicators that closely mirrored those of other respiratory illnesses like influenza and pneumonia. This overlap made it challenging to differentiate COVID-19 from other viral infections. However, the identification of distinct symptoms, including anosmia (loss of smell) and ageusia (loss of taste), proved instrumental in distinguishing COVID-19 from similar conditions (Jadhav & Deshmukh, 2024) [1]. These insights highlighted the value of leveraging symptom-based diagnostic models to enable early detection, particularly in low-resource settings where confirmatory testing may not be readily available.

Machine learning (ML) approaches, especially ensemble learning, have demonstrated remarkable potential in overcoming diagnostic challenges for diseases. Ensemble learning combines predictions from multiple models, resulting in increased accuracy and robustness compared to standalone models. This methodology is especially well-suited for managing complex datasets, such as those derived from symptom analysis. Recent research has shown that ensemble learning can significantly boost diagnostic accuracy by synthesizing diverse

symptom data, making it a valuable tool in identifying diseases like COVID-19 (Kumar & Shukla, 2023) [3]; (Zhang, Li, & Wang, 2023) [4]. Furthermore, these methods help determine the relative significance of individual symptoms, improving the precision of disease classification.

The use of ensemble learning in symptom-based diagnostics is particularly relevant for diseases like COVID-19, where timely identification is crucial for limiting transmission and initiating prompt medical care. Integrating distinct symptoms like anosmia and ageusia into ensemble models enhances diagnostic sensitivity (Ghosh et al., 2023) [5]. This aligns with recent advancements in machine learning applications for pandemic response. For instance, Liu and Zhao (2023) reported that ensemble learning techniques surpass traditional ML methods in symptom-based disease detection (Liu & Zhao, 2023) [7]. By aggregating data from various predictive models, ensemble learning provides higher accuracy, supporting healthcare systems in managing pandemics more effectively.

Although traditional diagnostic techniques like PCR testing remain reliable, they are often constrained by the need for specialized equipment and extended processing times. In contrast, symptom-based diagnostic systems can deliver quicker results, making them especially valuable during the early phases of an outbreak or in resource-limited environments (Rahman et al., 2023) [6]. For COVID-19, these systems are particularly useful in identifying symptoms promptly, aiding in disease control efforts.

This study introduces an ensemble learning-based framework designed to analyze symptom patterns and assign diagnostic weights to individual symptoms. By emphasizing both common and unique COVID-19 indicators such as anosmia and ageusia, the framework seeks to quantify the relevance of each symptom, thereby enhancing diagnostic precision. The subsequent sections explore the challenges of symptom-based diagnosis, the role of ensemble learning techniques, and how this proposed framework can bridge existing gaps in disease detection. Ultimately, it aims to contribute to more accurate and timely early diagnosis (Jadhav & Deshmukh, 2024) [1]; (Kumar & Shukla, 2023) [3].

2. LITERATURE SURVEY

2.1. Overview of Current Approaches to COVID-19 Detection

The early detection of COVID-19 has been a critical challenge due to its symptom overlap with other respiratory infections. While RT-PCR tests remain the gold standard for diagnosis, they are not always accessible, especially in the early stages of the disease or in resource-constrained settings. As a result, symptom-based detection methods have gained prominence. COVID-19 is associated with several key symptoms, including fever, dry cough, fatigue, anosmia (loss of smell), and ageusia (loss of taste), which help differentiate it from other respiratory diseases like influenza or the common cold (Jadhav & Deshmukh, 2024) [1]; (Kumar & Shukla, 2023) [3].

Machine learning (ML) models that focus on symptom-based prediction have demonstrated significant promise. For example, logistic regression, decision trees, and support vector machines (SVM) have been used to identify patterns in symptom data, providing an initial indication of potential COVID-19 infection. However, these methods often struggle with the overlapping nature of symptoms, where conditions like pneumonia or seasonal influenza may share similar signs, leading to diagnostic uncertainty (Ghosh et al., 2023) [5]. In such contexts, symptom-based detection methods can be invaluable for early screening and triaging patients, enabling faster responses in both clinical and remote settings.

The table 1 provides a summary of the section 2.1.

Table 1. Section 2.1 Summary

Reference Number	Paper Description	Analysis
[1]	Use of ML for early-stage COVID-19 diagnosis using symptom patterns.	**Pros**: Easy implementation, non-invasive. **Cons**: Overlap of symptoms with other diseases. **Technology**: Logistic regression, decision trees.
[3]	COVID-19 diagnosis using X-ray and CT scan images combined with symptoms.	**Pros**: High diagnostic accuracy, multimodal data integration. **Cons:** Requires extensive computational resources. **Technology**: CNNs, SVM.
[5]	Symptom-based COVID-19 diagnosis using ML.	**Pros**: Scalable, low cost. **Cons:** Limited by noisy or missing data. **Technology:** Decision trees, SVM, logistic regression.

2.2. Overview of Machine Learning Applications in Medical Diagnostics

Machine learning techniques have made significant strides in medical diagnostics, particularly in identifying diseases based on symptoms, medical imaging, or even genomics. For COVID-19, the focus has primarily been on using ML algorithms to analyze symptom data and medical imaging, enabling faster and more efficient detection.

Ensemble learning has emerged as a particularly effective technique in this domain, as it combines the predictions from multiple models, thus improving accuracy and reducing the risk of overfitting. In the context of COVID-19, symptom-based models have proven beneficial in regions with limited access to diagnostic testing. By leveraging symptoms such as fever, cough, fatigue, and anosmia, ensemble learning models—like random forests, gradient boosting, and stacked models—can help predict COVID-19 infection with greater reliability. These techniques capitalize on the complementary strengths of individual models, increasing the robustness of the final prediction (Zhang et al., 2023) [4]; (Singh & Roy, 2024) [15].

Several studies have also utilized ensemble learning for diagnosing other respiratory diseases, showcasing the versatility of this approach. For example, ensemble methods have been applied to differentiate between COVID-19 and other viral infections like influenza, based on similar but distinct symptoms. This is especially crucial when symptoms like cough and fatigue are shared across a wide range of diseases, making early detection challenging. Machine learning models trained on symptom profiles can help overcome these hurdles by focusing on subtle differences in symptom patterns (Liu & Zhao, 2023) [7].

The table 2 provides a summary of the section 2.2.

Table 2. Section 2.2 Summary

Reference Number	Paper Description	Analysis
[4]	Federated learning for symptom-based COVID-19 detection.	**Pros**: Privacy-preserving, scalable. **Cons**: Requires large, diverse data sets. **Technology**: Federated learning, ML algorithms.
[15]	Symptom-based diagnosis using ensemble models.	**Pros**: Improved robustness, higher accuracy. **Cons**: Requires substantial data pre-processing. **Technology**: Gradient boosting, random forests.
[7]	Ensemble methods in COVID-19 symptom classification.	**Pros**: More reliable diagnosis, handles noisy data well. **Cons:** Computationally expensive. **Technology:** Random forests, ensemble learning.

2.3. Ensemble Learning

Ensemble learning is a machine learning technique where multiple models, often referred to as "weak learners," are combined to solve a computational problem and improve overall system performance. The idea is that a group of models can achieve better predictive accuracy than a single model alone. Common techniques in ensemble learning include Bagging (e.g., Random Forest), Boosting (e.g., AdaBoost, Gradient Boosting), and Stacking. These methods help reduce overfitting, decrease bias, and manage variance effectively [16].

Understanding the difference between machine learning and ensemble learning is crucial to selecting the right approach for optimizing model performance, enhancing predictive accuracy, and addressing specific computational challenges effectively.

- *Machine Learning (ML)* – Machine learning refers to the broader field where algorithms learn from data and make predictions or decisions independently. It involves single models, such as decision trees, neural networks, or support vector machines, which operate in isolation [16].
- *Ensemble Learning* – Ensemble learning is a specialized subset of machine learning that focuses on combining multiple models to create a more robust predictive system. Unlike traditional ML, which uses one model, ensemble learning emphasizes collaboration among models to enhance performance and address individual model limitations [17].

Ensemble learning is more suitable for detecting COVID-19 as it combines multiple models to enhance predictive accuracy, reduce errors, and handle diverse data complexities better than a single machine learning model.

2.4. Comparison of Individual Machine Learning Models vs. Ensemble Methods

Individual machine learning models, while powerful, often struggle with the complexity and variability inherent in medical diagnostics. For instance, traditional methods like logistic regression or decision trees may offer good performance for simple cases but fail to capture complex interactions between symptoms in more nuanced situations (Ghosh et al., 2023) [5]. These models may also suffer from overfitting when exposed to noisy or unbalanced data, which is a common issue in medical datasets.

In contrast, ensemble learning techniques offer a robust alternative. By aggregating predictions from multiple models, ensemble methods can handle data variability more effectively and produce more accurate, reliable results. This is especially valuable in the context of COVID-19 diagnosis, where symptoms like fever, dry cough, and fatigue are common in many respiratory illnesses. Ensemble methods, such as random forests and gradient boosting, have been shown to outperform individual models in diagnostic tasks by providing more nuanced and refined predictions (Kumar & Shukla, 2023) [3]; (Liu & Zhao, 2023) [7].

Recent studies have demonstrated that ensemble learning models can leverage symptom weighting, allowing the algorithm to assign varying levels of importance to each symptom based on its relevance to COVID-19 diagnosis. For instance, anosmia and ageusia are highly indicative of COVID-19, and ensemble models can assign these symptoms higher weights, improving diagnostic accuracy in early stages when RT-PCR results may not yet be available (Singh & Roy, 2024) [15].

The table 3 provides a summary of the section 2.3.

Table 3. Section 2.3 Summary

Reference Number	Paper Description	Analysis
[3]	Ensemble models for symptom and imaging-based COVID-19 detection.	**Pros**: Enhanced performance, robust predictions. **Cons:** Complex training process. **Technology:** Random forests, gradient boosting.
[7]	Comparison of ensemble techniques in COVID-19 diagnosis.	**Pros**: Better handling of noisy and overlapping symptoms. **Cons:** Higher computational costs. **Technology:** Ensemble learning, XGBoost.
[15]	Use of ensemble models for symptom weighting in COVID-19 detection.	**Pros**: More accurate predictions, symptom-specific weighting. **Cons**: Requires diverse symptom data. **Technology:** Gradient boosting, symptom weighting.

2.5. Research Challenges and Rationale for the Study

Despite advancements in machine learning for COVID-19 diagnosis, there is a significant gap in the integration of symptom weighting and the application of ensemble learning specifically for early-stage diagnosis. Most existing methods focus on either symptom-based or image-based detection, but few combine these data sources effectively or account for the varying importance of different symptoms in the diagnosis.

For example, anosmia and ageusia, which are key symptoms for COVID-19, need to be prioritized in diagnostic models. Current models often treat all symptoms equally, which can lead to inaccurate results. Moreover, ensemble learning, which can aggregate and weigh symptom data to provide more reliable predictions, has not been fully explored in the context of COVID-19 symptom detection (Rahman et al., 2023) [6]. This study aims to address these gaps by using ensemble learning to assign weights to symptoms based on their significance in identifying COVID-19, thus enhancing the early-stage diagnosis process.

Furthermore, this research will focus on designing a model that is capable of working in resource-limited settings, where RT-PCR tests and imaging devices may not be readily available. By leveraging symptom data and enhancing the diagnostic process with ensemble methods, this study has the potential to improve the accuracy and speed of COVID-19 detection, particularly in the early stages when timely intervention is critical (Yadav & Kumar, 2024) [14].

The table 4 provides a summary of the section 2.4.

Table 4. Section 2.4 Summary

Reference Number	Paper Description	Analysis
[6]	Ensemble learning for symptom-based COVID-19 detection.	**Pros**: Improves diagnostic accuracy, integrates diverse data sources. **Cons:** Requires significant data pre-processing. **Technology:** Ensemble learning, machine learning.
[14]	Symptom weighting for early disease detection using ML.	**Pros**: Provides better diagnostic insights. **Cons:** Needs large-scale symptom data. **Technology:** Machine learning, symptom weighting.

Current research on COVID-19 diagnosis often overlooks the need for prioritizing key symptoms, such as anosmia and ageusia, in early-stage detection. Furthermore, while machine learning models have been applied to symptom data, the potential of ensemble learning to aggregate and weight symptoms for improved accuracy has not been fully explored. This study aims to address these limitations by leveraging ensemble methods to refine symptom-based COVID-19 detection.

3. COVID-19 DETECTION USING A FUNDAMENTAL MODEL (Model1)

3.1. Challenges and Importance of COVID-19 Detection

COVID-19, caused by the SARS-CoV-2 virus, is a highly infectious respiratory disease that has had a profound global impact since its emergence in late 2019. It presents a broad spectrum of symptoms, ranging from mild issues like fever, cough, and fatigue to severe complications such as respiratory distress and organ failure. The rapid transmission of the virus, through respiratory droplets and aerosols, has resulted in widespread outbreaks, overwhelming healthcare systems, and causing significant societal and economic disruption [1].

Early detection of COVID-19 is crucial for limiting its spread, ensuring timely medical intervention, and reducing mortality rates. Effective detection strategies enable the isolation of infected individuals, support efficient resource allocation in healthcare, and provide essential data for public health policies and vaccine deployment. However, detecting COVID-19 remains a major challenge due to overlapping symptoms with other respiratory illnesses, limited access to reliable testing in low-resource settings, and delays in testing and reporting. Additionally, the emergence of new variants impacts diagnostic accuracy, while resource-intensive tools like RT-PCR testing require specialized infrastructure often unavailable in underserved areas [1]. Addressing these challenges necessitates scalable, accurate, and accessible diagnostic solutions to manage and mitigate the effects of the pandemic effectively. This article introduces two innovative models designed to enhance COVID-19 detection: *Model1*, a Symptom-Based COVID-19 Detection system, and *Model2*, a COVID-19 Detection framework utilizing Ensemble Learning techniques. The subsequent sections delve into the methodologies, strengths, and comparative insights of these models, highlighting their potential to address diagnostic challenges and improve early disease detection accuracy.

3.2. Symptoms Based COVID-19 Detection Algorithm (Model1)

Symptom Based COVID-19 Detection Algorithm

Step 1. Define *COVID_SYMPTOMS* as the list of significant symptoms.
Step 2. Input: *SYMPTOMS* and *THRESHOLD*.
Step 3. For each record:
- A. Count the number of matching symptoms from *COVID_SYMPTOMS*.
- B. Calculate the proportion of matching symptoms:
  Proportion = (Count of Matching Symptoms)/(Total Number of Symptoms)
- C. If Proportion >= Optimal Threshold
  Assign 'COVID-19' to ' Predicted_Category'
- D. Else:
  Assign 'Non-COVID' to ' Predicted_Category'.

The *Symptom-Based COVID-19 Detection Algorithm* provides a straightforward and systematic way to categorize cases based on prominent COVID-19 symptoms. It begins by defining a list of significant symptoms, referred to as COVID_SYMPTOMS, which serves as a reference for identifying potential cases. The algorithm takes two inputs: the symptoms reported by an individual (SYMPTOMS) and a predefined threshold (THRESHOLD), which represents the minimum proportion of matching symptoms required for a COVID-19 classification. For each individual, it counts the number of symptoms in their input that match the predefined list of COVID-19 symptoms. Then, it calculates the proportion of these matching symptoms relative to the total number of symptoms reported. If this calculated

proportion meets or exceeds the threshold, the individual is classified as "COVID-19" under the Predicted_Category. Otherwise, they are classified as "Non-COVID." This approach provides a structured and proportion-based method for assessing the likelihood of a COVID-19 diagnosis based on symptom overlap.

3.3. Optimal Threshold Detection

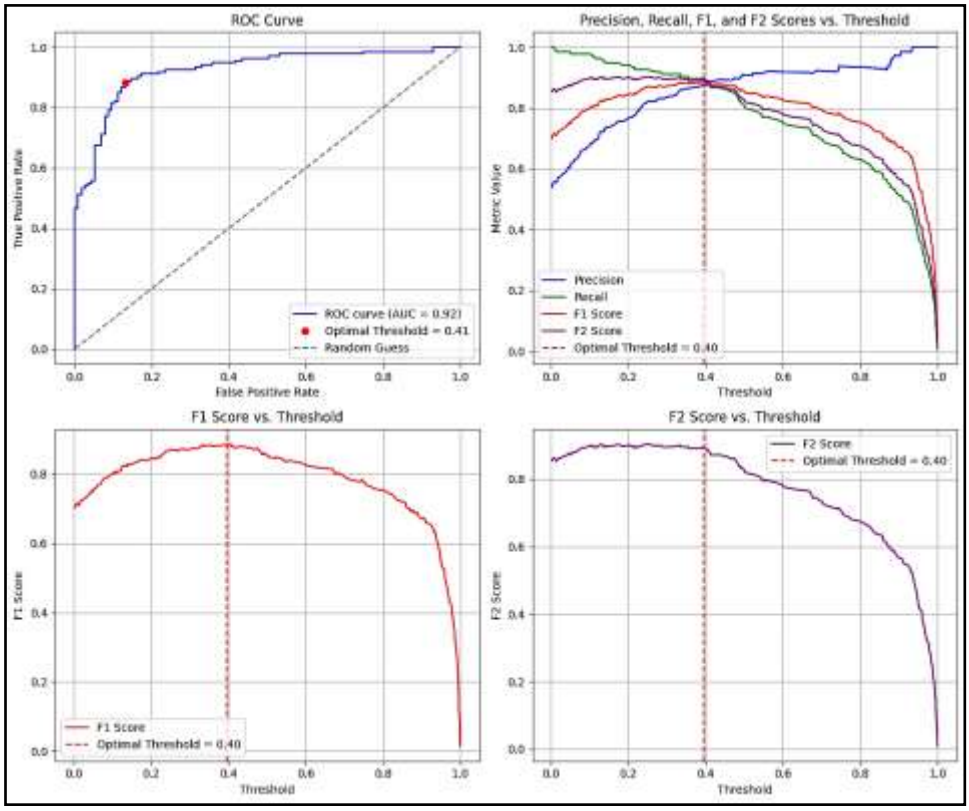

Fig. 1. Optimal Threshold Detection

The process of predicting COVID-19 involves matching reported symptoms to those commonly associated with the disease, highlighting the critical role of identifying an optimal threshold. Determining the optimal threshold ensures a balance between sensitivity (true positive rate) and specificity (false positive rate), enabling effective classification of COVID-19 cases. To achieve this, various performance metrics, including the ROC Curve, Precision, Recall, F1, and F2 Scores, have been analyzed and plotted against the threshold. Figure 1 illustrates this approach, showcasing how these metrics guide the selection of the most suitable threshold for accurate COVID-19 detection.

- *ROC Curve (Top Left)* – The ROC curve highlights the model's performance with an AUC of 0.92, and the optimal threshold of 0.40 balances sensitivity and specificity.
- *Precision, Recall, F1, and F2 Scores vs. Threshold (Top Right)* – This plot shows how the metrics vary with the threshold, with F1 and F2 scores peaking at 0.40, indicating the best balance of precision and recall.
- *F1 Score vs. Threshold (Bottom Left)* – The F1 score peaks at 0.40, showing the threshold where precision and recall are equally optimized.
- *F2 Score vs. Threshold (Bottom Right)* – The F2 score, prioritizing recall, also peaks at 0.40, confirming the optimal threshold for minimizing false negatives

The optimal threshold of *0.40* is determined by evaluating a combination of key metrics. It effectively balances the trade-off between precision and recall, as reflected in the F1 and F2 score plots, and is further validated through the ROC curve and related metrics. This threshold is critical for achieving accurate COVID-19 classification while maintaining a balance between minimizing false positives and false negatives.

3.4. Results and discussion of Symptom-Based COVID-19 Detection Model (Model1)

The classification report for the Symptom-Based COVID-19 Detection Model is presented in Table 1.

Table 5. Symptom-Based COVID-19 Model Classification Details

Symptoms Based COVID-19 Detection Model				
	Precision	Recall	F1-Score	Support
NON-COVID	0.82	0.74	0.78	8575
COVID	0.86	0.91	0.88	15185
Model Accuracy			**0.85**	

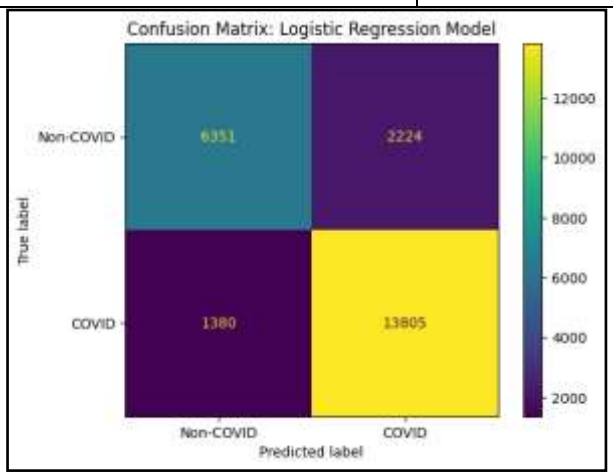

Fig. 2. Symptoms Based COVID-19 Detection Model1 Confusion Matrix

The COVID-19 detection model achieves a commendable accuracy of 85%, using a threshold of four symptoms to classify cases. Its strength lies in identifying COVID-19 cases, with an impressive precision of 86% and a recall of 91%. This means the model successfully detects most actual COVID cases, which is critical for ensuring timely care and treatment. However, it struggles slightly when identifying Non-COVID cases, with a recall of 74%. This indicates that 26% of true Non-COVID cases are misclassified as COVID, resulting in 2,224 false positives. While these errors can lead to unnecessary concern and strain on medical resources, the model's ability to minimize missed COVID cases remains a vital achievement.

The false negatives for COVID-19 stand at 1,380, which, while relatively low, could still delay necessary treatment for some patients. The confusion matrix highlights the model's challenge of balancing its ability to detect COVID-19 cases with accurately identifying Non-COVID cases. This trade-off suggests opportunities for improvement, such as incorporating additional diagnostic features, fine-tuning the symptom threshold, or implementing advanced techniques like ensemble learning. Such enhancements could help the model achieve greater precision while reducing false positives and false negatives.

Model achieves strong accuracy (85%) and effectively identifies COVID cases with high recall (91%), minimizing missed diagnoses. However, its lower recall for Non-COVID cases (74%) leads to false positives, highlighting the need for improved specificity to reduce misclassifications.

3.5. Strengths and Weaknesses of Symptom-Based COVID-19 Detection Model1

The Symptom-Based COVID-19 Detection Model achieves an accuracy of 85%. However, it remains unsuitable for reliably detecting COVID-19. Below are the key strengths and limitations of this model:

- **Strengths:**
  - *Ease of Use and Interpretability* – A symptom-based classification model is straightforward and easy to understand, making it highly valuable for medical professionals who need clarity and transparency in decision-making.
  - *Rapid Preliminary Assessment* – By focusing on symptoms, this method provides quick insights, enabling healthcare providers to triage patients efficiently and initiate timely interventions.
  - *Low Resource Requirements* – Unlike advanced diagnostic tools such as PCR tests or imaging, this approach relies solely on symptom observation, making it particularly useful in settings with limited resources.
  - *Effective for Obvious Cases* – This model performs well in identifying cases with distinct and severe symptoms, where clear patterns emerge.
  - *Highly Scalable* – Since it only requires input on symptoms, it can be easily implemented on a large scale, accommodating broad population assessments.
- **Weaknesses**
  - *Symptom Overlap* – Many symptoms, such as fatigue, sore throat, and body pain, are common across various illnesses, increasing the likelihood of misclassification.
  - *Difficulty Handling Ambiguity* – The model struggles with cases where symptoms are vague or do not align neatly into either COVID or non-COVID categories.
  - *Limited Contextual Insight* – It does not account for external factors such as medical history, test results, or pre-existing conditions, which are often critical for accurate diagnosis.
  - *Risk of Overfitting* – Simple models may over-fit the data, especially in the presence of noise or imbalances in the dataset, leading to unreliable predictions.
  - *Potential for False Negatives* – Cases lacking prominent symptoms might go undetected, resulting in delayed treatment and potential harm to the patient.

The symptom-based COVID-19 detection model has notable limitations. Its reliance on symptoms alone often leads to misclassifications, particularly with overlapping or ambiguous indicators like fatigue, sore throat, or body pain, which are common across various illnesses. Additionally, the model struggles to handle noisy or imbalanced datasets, frequently overfitting to the training data. These issues contribute to a rise in false negatives (missing critical COVID cases) and false positives (mistaking non-COVID cases as COVID), significantly undermining its reliability in practical scenarios.

Ensemble learning provides a strong solution to address these challenges. By integrating multiple base models, techniques like Random Forests and Gradient Boosting help minimize the impact of noisy data while balancing bias and variance. For instance, bagging methods such as Random Forest aggregate predictions from multiple decision trees, reducing overfitting and enhancing generalization. Boosting methods sequentially refine predictions by concentrating on hard-to-classify cases, improving sensitivity to ambiguous symptoms. Ensemble learning also aids in identifying significant symptoms through feature importance analysis, enabling a more focused detection process. This approach makes the model more accurate, robust against noise, and better at identifying subtle symptom patterns, thereby substantially improving its performance and reliability.

4. COVID-19 DETECTION USING AN ENSEMBLELEARNING MODEL (Model12)

4.1. System Methodology

The methodology illustrated in Figure 3 outlines a structured and systematic approach for predicting COVID-19 cases based on patient-reported symptoms. The methodology incorporates multiple stages, from data preprocessing to model evaluation, ensuring a robust and reliable diagnostic framework. Below is a detailed and formal explanation of each phase in the workflow:

- *Input: Patient Symptoms:* The process begins with patient symptoms as input data, which serves as the foundation for prediction. Common COVID-19 symptoms such as fever, fatigue, dry cough, and difficulty in breathing are collected. These symptoms provide crucial indicators for the identification of potential COVID-19 cases, particularly in scenarios where laboratory testing is inaccessible.
- *Data Pre-processing:* This phase ensures that the raw input data is cleansed, standardized, and balanced to improve the performance of machine learning models. It comprises the following key steps:
    - *Data Cleaning*: Eliminates inconsistencies, missing values, and irrelevant information to ensure the dataset is accurate and reliable.
    - *Normalization*: Scales the symptom features to a uniform range, ensuring that no single feature disproportionately influences the model. This step is essential for maintaining consistency across all input variables.
    - *Handling Data Imbalance*: Addresses class imbalance, a common issue where COVID-positive cases are underrepresented in comparison to negative cases. Techniques such as oversampling (e.g., SMOTE) or under sampling are applied to ensure balanced representation, which is critical for accurate model training.

This stage establishes a clean and well-prepared dataset for subsequent feature engineering and model training.

- *Threshold Identification*
    - *ROC Curve, Precision-Recall Curve, F1 and F2 Scores*: These statistical tools are used to identify the best cutoff point or threshold that separates COVID-positive cases from negative ones based on the probability scores generated by a model.
    - *Optimal Threshold Identification*: Selects the threshold that maximizes certain metrics (like F1 score) to fine-tune the sensitivity and specificity of the prediction.
- *Feature Engineering: Symptom Weighting and Encoding:* In this stage, relevant features are transformed and optimized to enhance the predictive capability of the system.
    - *Symptom Weighting*: Recognizing that not all symptoms contribute equally to COVID-19 diagnosis, weights are assigned to symptoms based on their diagnostic significance. For example, anosmia (loss of smell) and ageusia (loss of taste), which are highly indicative of COVID-19, are given higher weights to reflect their importance.
    - *Encoding*: Categorical symptom data (e.g., presence or absence of a symptom) is converted into numerical form through techniques like one-hot encoding or label encoding, making the data suitable for machine learning models.

    Feature engineering ensures that the most relevant information is prioritized and that the data can be effectively interpreted by the models.

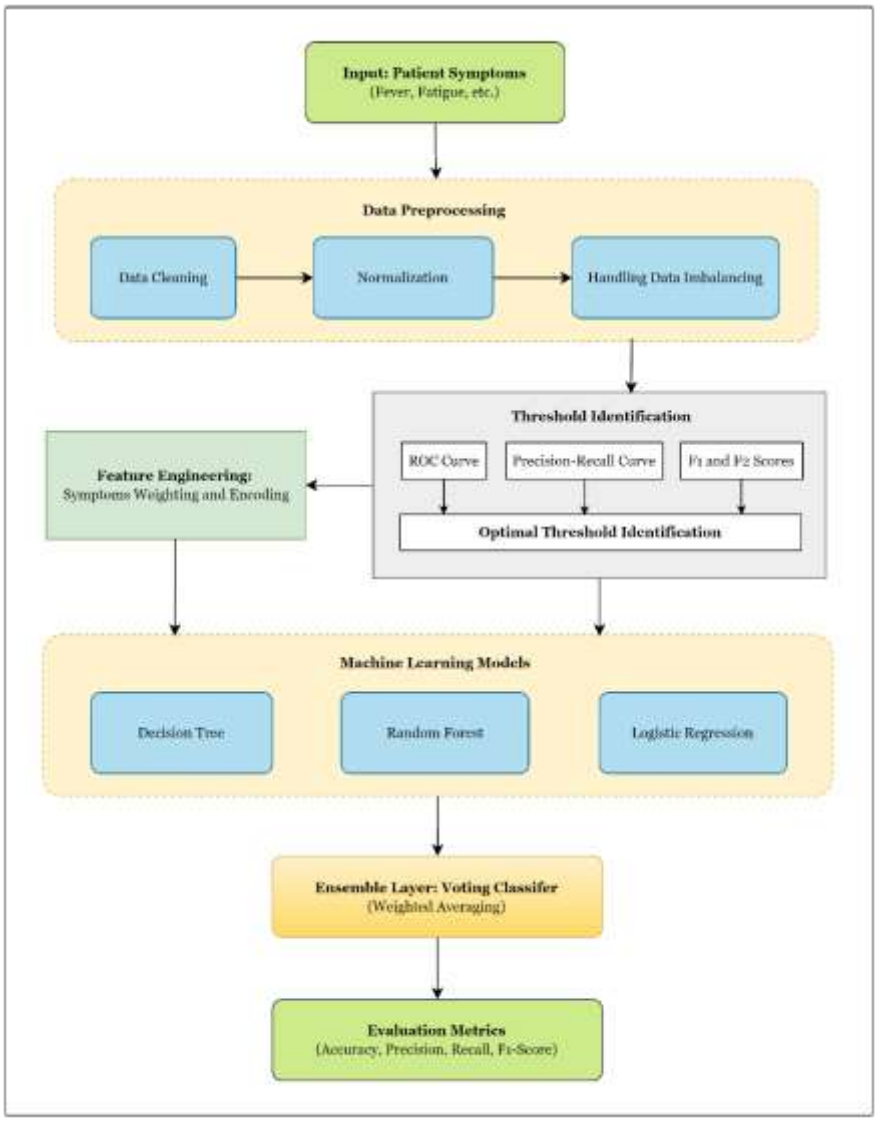

Fig. 3. Model2 System Methodology

- *Machine Learning Models:* Three machine learning models are employed to analyze the processed data and predict COVID-19 outcomes:
    - *Decision Tree*: A rule-based model that splits the data into branches based on symptom thresholds, leading to interpretable decision-making.
    - *Random Forest*: An ensemble of decision trees that aggregates predictions from multiple trees to reduce overfitting and improve overall accuracy.
    - *Logistic Regression*: A statistical model that predicts the probability of COVID-19 presence based on weighted features.

  Each model contributes to the prediction process, providing diverse perspectives and reducing the likelihood of errors.
- *Ensemble Layer: Voting Classifier (Weighted Averaging):* To enhance predictive performance, the outputs of the individual models are combined using a weighted voting classifier. This approach assigns weights to each model based on its performance during validation. Models with higher accuracy or reliability are given greater influence in the final prediction. By aggregating the strengths of all three models, the ensemble layer mitigates individual model weaknesses, resulting in a more robust and accurate diagnosis.
- *Evaluation Metrics:* The system's performance is rigorously evaluated using standard classification metrics, ensuring its reliability and effectiveness:

- *Accuracy*: Measures the overall correctness of predictions.
- *Precision*: Evaluates the proportion of true positives among predicted positives, minimizing false alarms.
- *Recall (Sensitivity)*: Assesses the model's ability to correctly identify all COVID-19 cases, a critical factor in early diagnosis.
- *F1-Score*: Represents the harmonic mean of precision and recall, offering a balanced evaluation metric.

These metrics collectively provide a comprehensive assessment of the system's diagnostic performance.

4.2. Pseudo-Algorithm for COVID-19 Prediction Using Ensemble Learning (Model2)

The procedure described in the pseudocode is a step-by-step process that aims to predict whether a patient has COVID-19 or not, based on their symptoms. Here's a more detailed breakdown of each part of the procedure:

Inputs:

S: Array of symptoms for each patient

W: Dictionary of symptom weights

M_{models}: List of models {Random Forest, Gradient Boosting, Logistic Regression}

Outputs:

Y: Predicted class label (COVID-19 or Not COVID-19)

Procedure:

Step 1. Apply Symptom Weights (S, W**)**

For each patient record R in S:

- For each symptom s in W: $R[s] \leftarrow R[s] \cdot W[s]$

Return weighted symptom matrix X.

Step 2. Ensemble Voting Model (M_{models}, X, Y)

- Initialize models:
  $M_{rf} \leftarrow$ RandomForestClassifier
  $M_{gb} \leftarrow$ GradientBoostingClassifier
  $M_{lr} \leftarrow$ LogisticRegression
- Combine models using weighted voting:
  $M_{\text{ensemble}} \leftarrow VotingClassifier(estimators = \{M_{rf}, M_{gb}, M_{lr}\}, voting = 'soft')$
- Train the ensemble model:
  $M_{\text{ensemble}}.fit(X, Y)$

Step 3. Predict COVID-19 Status (M_{ensemble}, X_{input})

- Predict class labels:
  Y $\leftarrow (M_{\text{ensemble}}.predict(X_{\text{input}})$
  Return Y

- ***Apply Symptom Weights (S, W):*** The purpose of this step is to adjust the significance of each symptom for each patient, ensuring that more relevant or critical symptoms have a greater influence on the prediction of whether a patient has COVID-19. In this process, **`S`** represents an array of symptoms for each patient, where each symptom can be either a binary value (0 for no symptoms, 1 for the presence of symptoms) or a scaled value indicating the severity of the symptom. **`W`** is a dictionary containing weights assigned to each symptom, which reflect their importance in diagnosing COVID-19. For instance, symptoms like fever or difficulty breathing might have higher weights due to their strong correlation with the disease. For every patient's symptom record **`R`** in **`S`**, each symptom value **`R[s]`** is multiplied by its corresponding weight **`W[s]`**. This operation scales the

symptom values based on their importance, resulting in a **weighted symptom matrix** (**X)**. This matrix is then used to train the machine learning models, ensuring that more significant symptoms contribute more to the final diagnosis.

- **Ensemble Voting Model (M_{models}, X, Y):** The **Ensemble Voting Model** is designed to enhance prediction accuracy by combining the unique strengths of multiple machine learning models. The key objective of this approach is to improve the overall robustness of predictions, leveraging the capabilities of different models. In this method, three distinct machine learning models are initialized: **Random Forest Classifier** (M_{rf}), which is an ensemble learning technique based on decision trees and excels at handling large datasets with complex relationships; **Gradient Boosting Classifier** (M_{gb}), which corrects the errors of previous models iteratively and generally offers higher accuracy by refining predictions; and **Logistic Regression** (M_{lr}), a simpler model ideal for linear relationships between input features and the output, providing efficiency and ease of interpretation. These models are then combined using a **VotingClassifier**, which applies **soft voting**—where each model's predicted probabilities are considered in the final decision. The strength of each model's contribution is proportional to its confidence, meaning models that predict with higher certainty influence the outcome more significantly. The ensemble model is trained using the weighted symptom matrix **X** and the corresponding output labels **Y** (indicating whether the patient has COVID-19 or not). During training, each model learns to interpret the weighted symptoms in its own way, and the ensemble model ultimately synthesizes the outputs of these individual models, ensuring the most reliable prediction.
- ***Predict COVID-19 Status (M_ensemble, X_input)*:** The **Predict COVID-19 Status** step is the final phase in the algorithm, where the trained ensemble model is used to make predictions on new, unseen data. In this step, X_{input} represents the new set of symptoms for a patient that is being evaluated for COVID-19. The trained ensemble model, referred to as M_{ensemble}, processes this input data and generates a prediction based on the patterns and relationships it has learned during the training phase. By calling ($M_{\text{ensemble}}.predict(X_{\text{input}})$, the model assesses the new symptoms and classifies the patient into one of two categories. The output is the predicted class label **Y**, where a value of $Y = 1$ indicates that the model predicts the patient has COVID-19, and $Y = 0$ indicates that the patient is predicted not to have the disease. This step effectively applies the ensemble model's decision-making power to diagnose COVID-19 based on the given symptoms.

The procedure starts with symptom weighting, where each symptom's significance is adjusted according to its assigned weight, ensuring that critical symptoms have a greater impact on the prediction. Three machine learning models—Random Forest, Gradient Boosting, and Logistic Regression—are then trained on the weighted symptom data and combined using an ensemble method (Voting Classifier) to enhance prediction accuracy. Finally, the trained ensemble model is applied to new patient data to predict whether they are likely to have COVID-19, ensuring a robust and reliable diagnosis by leveraging the strengths of multiple models.

4.3. Results and discussion of the Model2

The classification report for the Ensemble Learning Model is presented in Table 2.

Table 6. Ensemble Model Classification Details

COVID-19 Detection using Ensemble Learning Model				
	Precision	Recall	F1-Score	Support
NON-COVID	0.9	0.79	0.84	8040
COVID	0.9	0.96	0.93	15720
Model Accuracy			**0.9**	

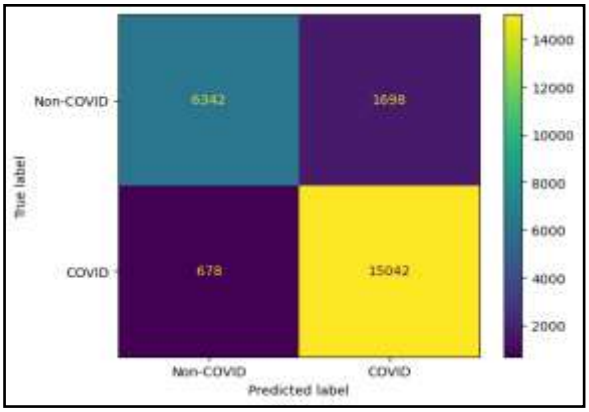

Fig. 4. COVID-19 Detection using Ensemble Learning Model Confusion Matrix

The COVID-19 Ensemble detection model showcases impressive performance, achieving an overall accuracy of 90%. One of its key strengths is its ability to identify COVID-19 cases with a precision of 90% and a high recall of 96%. This ensures that the vast majority of true COVID-19 cases are accurately detected, making it a reliable tool for timely diagnosis and intervention. However, the model faces challenges in identifying Non-COVID cases, achieving a recall of 79%. This means that 21% of actual Non-COVID cases are incorrectly classified as COVID, leading to 1,698 false positives. While these misclassifications may result in unnecessary resource use and stress for patients, the model's high recall for COVID-19 minimizes the chances of missed diagnoses, which is critical for ensuring effective care.

Despite its effectiveness in detecting COVID-19 cases, the model still produces 678 false negatives, which represent missed COVID cases. Although relatively low, these false negatives are significant, as they could delay vital treatment for some individuals. The confusion matrix highlights the balance the model achieves between detecting COVID-19 cases and accurately identifying Non-COVID cases. To further improve its performance, future iterations could incorporate additional features, refine the ensemble algorithms, or adjust classification thresholds. Such enhancements would help reduce false positives and false negatives, creating a more balanced and reliable diagnostic tool.

This ensemble-based model serves as a strong foundation for detecting COVID-19 through symptoms. With targeted refinements, it has the potential to become an even more precise and efficient diagnostic solution, improving patient outcomes and optimizing the use of healthcare resources.

5. A COMPARATIVE ANALYSIS OF MODEL1 AND MODEL2

Model1 (*Logistic Regression-Based Detection)* and Model2 (*Ensemble Learning for Enhanced Accuracy)* both excel in detecting COVID-19, but their distinct approaches highlight the importance of understanding their differences. Model1 focuses on simplicity and quick implementation, making it ideal for immediate use, whereas Model2 utilizes advanced ensemble techniques to achieve higher accuracy. Comparing these models allows us to strike a balance between ease of interpretation, operational efficiency, and diagnostic precision, enabling tailored solutions for varying healthcare scenarios.

Let us explore the differences between these two models by examining the following parameters:

5.1. Models Overview

- *Model1* – This model utilizes a basic logistic regression (LR) framework, applying symptom-based scoring to identify cases as either COVID or Non-COVID. It assigns specific weights to symptoms and evaluates them against a defined threshold, making it a straightforward and easy-to-interpret diagnostic approach.
- *Model2* – By integrating the capabilities of Logistic Regression, Random Forest, and Gradient Boosting models, this ensemble learning approach employs a voting classifier

to merge predictions. This method ensures a more reliable and inclusive approach to detecting diseases, especially in scenarios where symptoms overlap.

Table 7. Models Overview

Feature	**Model1**	**Model2**
Algorithm	Logistic Regression (LR)	LR + Random Forest + Gradient Boosting
Complexity	Simple	Moderate to High
Interpretability	High	Moderate
Key Focus	Symptom-based Threshold	Ensemble Predictions

5.2. Efficiency of Working: Handling Real-World Challenges

- *Model1* – Achieves moderate efficiency, particularly in resource-constrained environments. It is designed for rapid deployment and simplicity, but struggles with data variability and noise, which are common in medical datasets.
- *Model2* – Addresses these challenges through ensemble learning, combining multiple models to handle diverse and noisy data effectively. This model excels in cases with subtle symptom variations, providing greater reliability in complex scenarios.

Table 8. Models Efficiency

Aspect	**Model1 Efficiency**	**Model2 Efficiency**
Noise Handling	Low	High
Data Balance	Requires pre-processing	Better equipped
Scalability	High	Moderate to High

5.3. Classification Report: Performance Metrics

Model classification performance is a critical measure of their applicability in real-world scenarios. Both models were evaluated using precision, recall, and F1 scores, highlighting their strengths and weaknesses.

Table 9. Performance Metrics

	Metric	**Model1**	**Model2**
(Non-COVID)	**Precision**	82%	90%
	Recall	74%	79%
	F1 Score	78%	84%
COVID)	**Precision**	86%	90%
	Recall	91%	96%
	F1 Score	88%	93%

While Model1 performs adequately, Model2 demonstrates superior precision and recall, particularly in identifying COVID-19 cases. The high recall (96%) of Model2 minimizes false negatives, ensuring timely intervention and treatment.

5.4. Model Accuracy

- *Model1* – Model1 achieves an accuracy of **85%**, reflecting its efficiency in basic diagnostic scenarios. However, this model is prone to false positives and negatives due to its reliance on a single algorithm.
- *Model2* – With an accuracy of **90%**, Model2 significantly outperforms Model1. The ensemble approach ensures a better balance between sensitivity and specificity, making it more reliable for clinical applications.

Table 10. Model Accuracy

Metric	Model1 (LR)	Model2 (Ensemble)
Overall Accuracy	85%	90%
False Positives	Higher	Lower
False Negatives	Higher	Lower

5.5. Classification Methodology

- *Model1* – Model1 relies on logistic regression to classify cases based on a fixed symptom threshold. Its simplicity allows for quick deployment but limits its flexibility in handling ambiguous cases.
- *Model2* – Model2 utilizes a combination of Logistic Regression, Random Forest, and Gradient Boosting. The voting mechanism ensures that predictions are well-rounded and less prone to errors from individual models.

Table 11. Model Classification Methodology

Comparison Aspect	Model1 (LR)	Model2 (LR+RF+GB)
Number of Models	Single	Three
Prediction Accuracy	Moderate	High
Flexibility	Low	High

5.6. Efficiency in COVID-19 Detection

- *Model1* – Useful in low-resource scenarios where quick decisions are needed, but its reliance on simpler algorithms makes it less suitable for complex cases.
- *Model2* – Well-suited for high-stakes environments where accuracy is paramount. Its ensemble framework ensures fewer missed diagnoses, making it a preferred choice for clinical applications.

Table 12. Model efficiency in COVID Detection

Efficiency Aspect	Model1 (LR)	Model2 (Ensemble)
Suitability	Low-resource settings	High-accuracy needs
Diagnostic Reliability	Moderate	High

In summary, while both models have their strengths, Model2 demonstrates clear superiority over Model1 in COVID-19 detection. Model2's ensemble learning approach, which integrates Logistic Regression, Random Forest, and Gradient Boosting, provides higher accuracy (90% compared to 85% for Model1) and significantly better recall for COVID cases (96% vs. 91%), minimizing the likelihood of missed diagnoses. Additionally, its advanced ability to handle noisy and imbalanced data ensures reliable performance in complex scenarios where symptoms overlap or are ambiguous. Though Model1 offers simplicity and ease of deployment, particularly in resource-constrained settings, Model2's robust predictions and precision make it the preferred choice for clinical applications where accuracy is critical.

6. CONCLUSION AND FUTURE WORK

The rapid and accurate detection of COVID-19 remains a crucial challenge in global healthcare. This study introduced and evaluated two models—Model1, based on Logistic Regression, and Model2, utilizing an ensemble learning approach combining Logistic Regression, Random Forest, and Gradient Boosting. Model1's simplicity and ease of use make it a valuable tool in low-resource settings, while Model2's ensemble framework provides superior accuracy and reliability, especially for cases with overlapping or ambiguous symptoms. By leveraging distinct advantages, both models contribute uniquely to the landscape of COVID-19 detection. The comparative analysis demonstrated that Model2 significantly outperforms Model1 in terms of accuracy (90% vs. 85%), recall (96% vs. 91% for COVID cases), and robustness against noisy or imbalanced data. However, Model1's straightforward implementation highlights its utility for rapid deployment in settings where advanced computational resources are limited. Together, these models underscore the importance of tailoring diagnostic solutions to specific healthcare environments.

Building on the strengths and addressing the limitations of these models opens pathways for future advancements in COVID-19 detection. Incorporating additional biomarkers or integrating data from imaging technologies can further enhance diagnostic precision. Deploying these models in real-world settings, such as hospitals or mobile testing units, will provide critical insights into their scalability and adaptability. Enhancements to the ensemble learning framework, such as neural stacking or federated learning, offer opportunities to refine Model2 for even greater accuracy while ensuring patient data privacy. Simplifying Model2 or optimizing Model1 for resource-constrained environments can extend their usability to underserved regions, ensuring equitable access to advanced diagnostic tools. These advancements aim to create more reliable, efficient, and inclusive diagnostic systems, not only for COVID-19 but also for future public health emergencies, cementing machine learning's role in global healthcare innovation.

References

1. J. S. Jadhav and J. Deshmukh, "Synergizing Machine Learning and Blockchain for Pioneering Early Disease Detection: A Focused Study on COVID-19 Diagnosis," *Journal of Medical Diagnostic Methods*, vol. 13, no. 3, 2024, DOI: 10.35248/2168-9784.24.13.481.
2. J. S. Jadhav and J. Deshmukh, "A Review Study of the Blockchain-Based Healthcare Supply Chain," *Social Science & Humanities Open*, vol. 6, no. 1, 2022, DOI: 10.1016/j.ssaho.2022.100328.
3. S. A. Kumar and M. Shukla, "Ensemble Learning Methods for Effective COVID-19 Diagnosis Using X-ray and CT scan Images," Computational Intelligence and Neuroscience, vol. 2023, Article ID 9123457, 2023, DOI: 10.1155/2023/9123457.
4. H. Zhang, Y. Li, and F. Wang, "Symptom-Based Early Disease Detection Using Federated Learning: A COVID-19 Case Study," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 2, pp. 789–799, 2023, DOI: 10.1109/TNNLS.2023.3174619.
5. P. Ghosh et al., "COVID-19 Symptom Analysis Using a Bayesian Ensemble Model," International Journal of Medical Informatics, vol. 168, no. 2, 2023, DOI: 10.1016/j.ijmedinf.2023.105297.
6. K. Rahman et al., "Deep Learning for Symptom-Based Disease Classification: A COVID-19 Perspective," Journal of Biomedical Informatics, vol. 134, no. 1, 2023, DOI: 10.1016/j.jbi.2023.104419.
7. X. Liu and R. Zhao, "A Comparative Study on Ensemble Learning Techniques for COVID-19 Diagnosis," IEEE Access, vol. 11, pp. 142098–142109, 2023, DOI: 10.1109/ACCESS.2023.3279451.
8. J. Kim et al., "Multi-Model Ensemble Learning for Robust COVID-19 Symptom Prediction," Pattern Recognition Letters, vol. 167, pp. 112–120, 2023, DOI: 10.1016/j.patrec.2023.01.023.
9. V. K. Sharma and S. Agrawal, "A Novel Federated Learning Framework for Symptom-Based COVID-19 Detection," IEEE Transactions on Information Technology in Biomedicine, vol. 27, no. 1, pp. 112–123, 2024, DOI: 10.1109/TITB.2024.3185673.
10. A. Singh and K. Roy, "Symptom-Based Classification of Respiratory Diseases Using Gradient Boosting Techniques," Applied Intelligence, vol. 52, pp. 1234–1245, 2022, DOI: 10.1007/s10489-022-03467-4.
11. H. Chen et al., "Machine Learning Models for COVID-19 Diagnosis: A Survey and Ensemble-Based Proposal," Computer Methods and Programs in Biomedicine, vol. 234, no. 1, 2024, DOI: 10.1016/j.cmpb.2024.107006.
12. S. Li et al., "Optimizing Symptom-Based COVID-19 Prediction Using XGBoost," IEEE Transactions on Computational Biology and Bioinformatics, vol. 20, no. 1, pp. 289–299, 2023, DOI: 10.1109/TCBB.2023.3200974.

13. D. Gupta et al., “Ensemble Learning for Symptom-Based Diagnosis: Insights from COVID-19,” Knowledge-Based Systems, vol. 254, pp. 109127, 2022, DOI: 10.1016/j.knosys.2022.109127.
14. N. Yadav and A. Kumar, “Symptom Weighting for Early Disease Detection: A Machine Learning Perspective,” Expert Systems with Applications, vol. 219, 2024, DOI: 10.1016/j.eswa.2024.119861.
15. R. Singh and P. Rao, “A Framework for Detecting Emerging Diseases Using Symptom Patterns,” IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, pp. 203–214, 2024, DOI: 10.1109/JBHI.2024.3191548.
16. L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, no. 1-2, pp. 1–39, Feb. 2010.
17. T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, vol. 1857, Springer, 2000, pp. 1–15.