Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.0651

2026, 45(2) 146-161

XAI-Driven Stacking Ensemble for Cross-Disease Feature Modification and Clinical Insights

Shakeel Ahmad [a], Sheikh Muhammad Saqib [b], Asif Hassan Syed [a], Ali Ahmed Abdelrahim [a], Nashwan Alromema [a]

[a] *Department of Computer Science, Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah, Saudi Arabia.*

[b] *Department of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan*

* Corresponding author: sarahmad@kau.edu.sa

Received: 08 December 2025, Revised: 19 February 2026, Accepted: 31 March 2026, Published: 01 April 2026

KEYWORDS

Stacking Classifier

Random Forest (RF)

Gradient Boosting (GB)

Support Vector Classifier (SVC)

Diverse Counterfactual Explanations (DiCE)

Explainable AI (XAI)

ABSTRACT

There is an increasing need to have timely and accurate clinical advice in management of diseases. Predictive models should be precise and in addition, clinically significant. The current literature is lacking in this aspect, and it does not satisfy this standard in many cases. This is filled by a complex also Stacking Classifier that attempts to predict in a multi-class manner five health conditions Bronchitis, Cold, Flu, Healthy and Pneumonia out of only structured symptom and vital sign data. The group incorporates the Random Forest (RF), Gradient Boosting (GB) and Support Vector Classifier (SVC) as the base learners, with a Logistic Regression (LR) meta-classifier condensing them. In contrast to the previous ones, which may rely only on high accuracy, our approach is based on the more developed interpretability framework, beginning with Explainable AI (XAI) through Local Interpretable Model-agnostic Explanations (LIME) to guarantee the transparency of predictions. The originality as well as its distinctiveness is the use of the Diverse Counterfactual Explanations (DiCE) framework that approximates the minimum changes in the independent features to convert a patient to a new disease and, thus, creates an individualized intervention approach. ADASYN pre-processing resulted in a good overall accuracy and Weighted F1-score of 94% in the model. Together with the actionability made available through DiCE, these overall results indicate that our proposed model is the most performant solution, which can provide excellent performance and unprecedented clinical value compared to current literature. This combined XAI framework extends past the conventional diagnostic tools to offer a proscriptive, action-based and trustworthy system of the personalized management of respiratory health. The performance rates of the proposed stacking model are also always higher compared to those presented in the literature of benchmarks on the case of multi-class respiratory disease prediction.

© Mehran University of Engineering and Technology 2026

1 Introduction

Precise and timely detection of infectious diseases has become a key need to reduce the risk of transmission and allow clinical decision-making through intelligent data-driven modeling with highly similar symptoms between respiratory diseases, including cold, flu, and pneumonia [1] The complexity of the matter and the lack of distinction between the symptoms of diseases like cold, flu, and pneumonia have been the main driving force behind the development of a set of intelligent, data-driven models with the capability to identify the disease in question precisely and at its early stages [2]. The use of modern technology, and particularly the use of Machine Learning (ML) has created a massive possibility of the system growth of the healthcare sector. The goal of these innovations is to enhance the effectiveness of diagnostic processes, the timeliness of the identification of new treatments, and a radical increase in the number of patients [3]. Recent studies have largely been devoted to the exploitation of ML and Deep Learning (DL) models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [4], to achieve high precision disease prediction. Multimodal Data Fusion (MDF) is a concept in which modern methods combine various data sources, such as Electronic Health Records (EHRs), wearable data, and imaging, to enhance predictive accuracy and improve timeliness of intervention [5]. This wave of innovation has developed strong systems of identification and treatment of different diseases, which in most cases have high performance prediction models [6, 7]

Although the literature to date has had incredible success in trying to model how highly accurate predictions can be made, it is mostly concerned with classification of a disease (what the patient has) and not with clinical advice on what to do (altering the result). One of the areas of research is hugely vacant and is the one that involves the use of predictive models not only to diagnose a disease, but also to learn the least, but the most physiological modifications involved in changing a diagnosis of a patient. Such a counter-factual approach provides practical wisdom to medical intervention and preventive healthcare planning. To close this gap, this study suggests a new framework, which is a high-performance ensemble classification model combined with advanced XAI analysis to make an accurate prediction and clinical actionable to the diagnosis of Bronchitis, Cold, Flu, Healthy, and Pneumonia. The design of the proposed system consists in the assessment of various Ensemble Architectures namely Hard Voting Classifier, Soft Voting Classifier, and Stacking Classifier that are based on a wide range of base models: Random Forest Classifier (RF), Gradient Boosting Classifier (GB), Support Vector Classifier (SVC), and Logistic Regression (LR). The last and most effective model is then capitalized on in two phases of interpretability:

- Local Transparency: The description of the specific feature contributions that make an individual patient receive a certain diagnosis using the LIME framework.

- Clinical Actionability: The use of the Diverse Counterfactual Explanations (DiCE) model to identify the smallest, realistic changes in the independent features of a patient (e.g., Body Temperature, Oxygen Saturation, Heart Rate) needed to transform the prediction of a model in one disease into a prediction in a different disease (cross-disease conversion).

The key purposes of this study are summarized as following:

- To develop, train, and statistically evaluate multiple machine learning models, including individual base classifiers (RF, GB, SVC, LR) and sophisticated ensemble architectures (Hard Voting, Soft Voting, Stacking).
- To identify the optimal prediction model based on comprehensive statistical metrics, including Accuracy, Mean Absolute Error (MAE), and class-specific F1-Scores.
- To provide local, transparent, and interpretable explanations for the model's predictions using the LIME analysis framework.
- To implement a DiCE-based counterfactual analysis to generate actionable clinical insights by quantifying the minimal feature modifications required for cross-disease prediction conversion.

2 Literature Review

The next discussion is a critical review of the current literature on disease prediction, which has two major limitations: it utilizes insufficient classes and lacks actionable counterfactual explanations. This comparison analysis shows that our multi-class Stacking model is novel.

Work [8] used the complex Convolutional Neural Network (CNN) models of VGG16, ResNet50V2, and DenseNet121 with Explainable AI (XAI) methods, with high-performance, which reached up to about 99%.

© Mehran University of Engineering and Technology 2026

Nonetheless, this accuracy was high on image-based inputs and was restricted to particular conditions such as pneumonia, COVID-19, or general lung opacities but did not deal with the issue of multi-disease classification at the same time. The proposed model currently overcomes this weakness by taking structured data on symptom and vital signs to a multi-class prediction task (Bronchitis, Cold, Flu, Healthy and Pneumonia) by an ensemble stacking process giving actionable insights with counterfactual explainability.

On the same note, in another study [9], a hybrid ensemble method was used which involved the use of CNNs and transfer learning alongside an XAI with the accuracy of 96.21. Although in this model the performance was strong in lung diseases, the coverage of multi-diseases and absence of structured symptom or vital sign data were still limitations that would limit its use. The unique feature of our work lies in providing multi-disease, symptom-based classification, which can be interpreted by LIME, and the benefit of counterfactuals is that it can provide direct clinical advice.

Ensemble CNNs (DenseNet169, MobileNetV2, ViT) are popular with [10] reporting 93.91] as the accuracy. Yet, their greatest flaw was the exclusive focus on a binary classification task (pneumonia versus normal) and the lack of the counterfactual analysis of any kind. The study is an extension to multi-class disease prediction based on the easily accessible vital and symptom data, incorporating a stacking ensemble with counterfactual explanations to provide usefulness.

Another ensemble method that used InceptionV3, VGG16, ResNet50, and ViT models with [11] only could obtain a 96 percent accuracy, but had only the ability to classify pneumonia. More importantly, it did not include the essential step of integrating vital/symptom data, and failed to include the counterfactual reasoning needed to make actionable recommendations. By comparison, our multi-class model offers practical insights based on DiCE and makes all the predictions interpretable based on LIME.

Accurate results were also highly reported by [12], with 99.98% accuracy using an ensemble of CNNs with the help of traditional ML classifiers (RF, SVM, Decision Tree). However, this study was limited to binary classification (pneumonia and normal) which restricted its clinical utility and disease scope in general. Our approach extends this to various specific respiratory diseases, is designed on symptom/vital sign data, and incorporates interpretability as well as counterfactuals as clinical action.

The research into non-image-based diagnostics showed that structure-based data, in the case of the ensemble model of XGBoost, CatBoost, LGBM, and MLP [13], reached the accuracy of about 90 percent (AUROC about 0.90). This system, however, was meant to do mortality prediction rather than multi-class disease classification, thus a completely different clinical outcome. The idea of multi-class disease prediction based on structured features is re-purposed in our work with an ensemble-based approach paying special attention to making the work interpretable and actionable in terms of clinical advice.

The other ML ensemble that included SHAP analysis [14] gave good results with 97.08% accuracy. Nevertheless, it was limited in its usage, as it only covered influenza and lacked disease coverage because of a small dataset. Our system is extended to a full multi-disease coverage, uses symptom/vital features, and uses counterfactual and LIME interpretability to provide actionable advice to single patients.

Equally high accuracies were achieved with ensemble DL models that used CT/X-ray [15], of 98-99%. This had been performed in the framework of binary classification, where the disease problem was the COVID-19 and the pneumonia, and most importantly, vital and symptom data had not been utilized. Our proposed stacking ensemble, which works on the symptom/vital data can provide multi-class prediction and can be interpreted through LIME, and counterfactual explanations are created to be clinically useful.

Finally, a hybrid model using ResNet-50 combined with texture features and an SVM classifier [16] reported 97% accuracy. Like many others, this system was image-based and limited to binary classification (pneumonia versus normal), lacking the essential counterfactual component. Our work addresses this critical gap by delivering multi-class prediction across respiratory diseases using structured feature data, ensuring the results are both interpretable and actionable through the integration of DiCE and LIME.

3 Proposed Methodology

The study is conducted in a multi-stage, systematic research process, with the initial stages of data preparation and the final stages of creating and

interpreting high-order models of ensembles. Fig. ***1*** is a schematic diagram of the entire course of work.

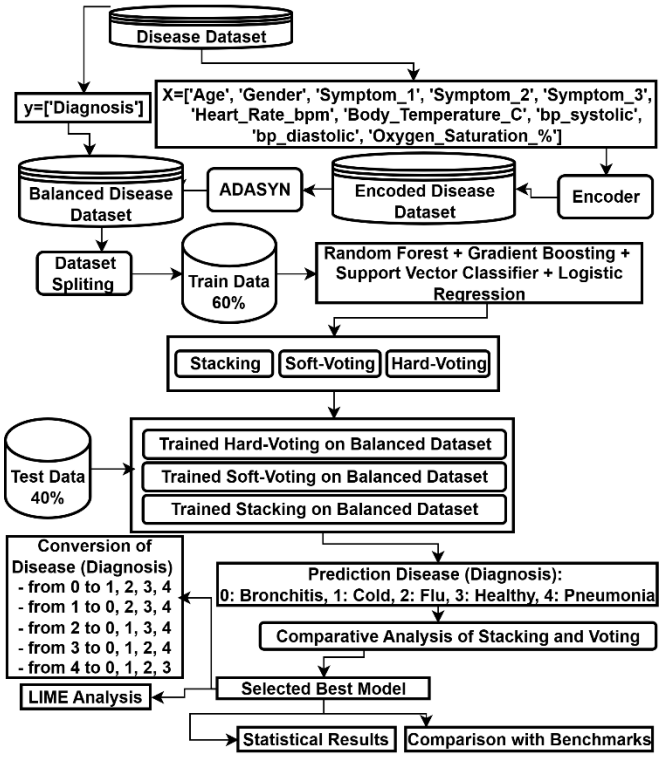

Fig. 1. Proposed Methodology for Disease Prediction with Analysis

3.1. Dataset Preparation

The structured respiratory symptom dataset used in this study is publicly available. It can be accessed at Kaggle [17] under the title "Holistic Health & Lifestyle Score Dataset".

The foundation of this study rests on a rich dataset comprising clinical and physiological measurements. The feature set includes Age, three categorical symptom variables (Symptom_1, Symptom_2, Symptom_3), and four continuous physiological parameters: Heart_Rate_bpm, Body_Temperature_C, Oxygen_Saturation_%, bp_systolic, and bp_diastolic. The objective is to predict the Diagnosis—a target variable encompassing five distinct health states: 0: Bronchitis, 1: Cold, 2: Flu, 3: Healthy, and 4: Pneumonia.

3.1.1 Addressing Class Imbalance

The distribution of classes in the dataset was deemed to be very skewed after conducting the first look. The initial sample sizes of the five classes were 334 (Bronchitis), 163 (Cold), 292 (Flu), 1167 (Healthy) and an insignificantly low 44 (Pneumonia). Such a drastic inequality is graphically displayed in Fig. ***2***. The data in Fig. ***2*** had a high class imbalance. Particularly, the most dominant sample was the so-called Healthy (Diagnosis=3) which comprised a majority of the observations. The Pneumonia type (Diagnosis=4) was, on the other hand, represented badly, with only 44 samples. It is directly on such skewed data that modeling would tend to produce a biased towards majority class classifier, which would have a poor generalization property and a high misclassification rate of the minor diseases.

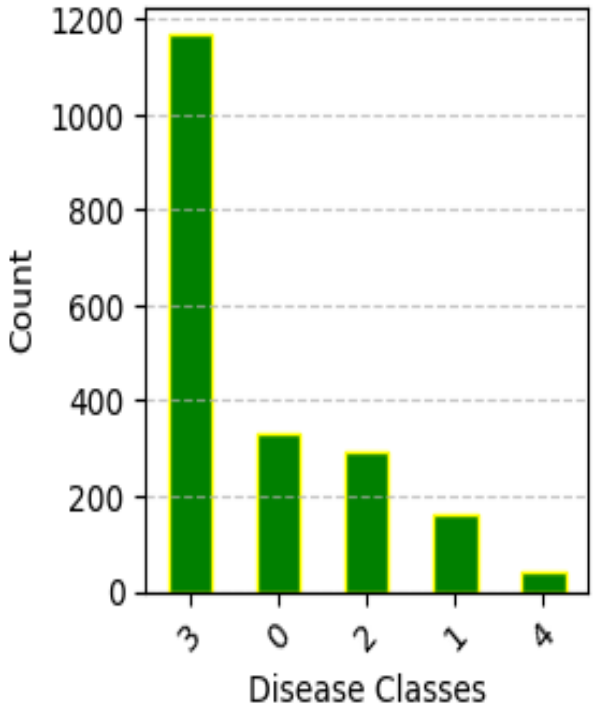

Fig. 2. Disease Class Distribution Before Balancing

3.1.2 Adasyn for Class Balancing

In order to diminish this problem and achieve the strong training of our classification models, we used the Adaptive Synthetic (ADASYN) sampling method on the training data. ADASYN creates synthetic data points along the minority classes according to the density distribution of the neighbors, in effect, driving the count of classes towards parity. This intervention made the distribution of the classes significantly balanced as it was as described below and as shown in Figure 3.

Figure 3 shows the effectiveness of the application of the ADASYN algorithm in the case of a successful result. The number of classes that represent all the five health states has now been strongly concentrated around the 1100 to 1200 range. This almost balanced distribution across the entire set of diagnostic categories is essential in training objective models that would be able to predict all the diseases accurately including the initially uncommon category of Pneumonia.

© Mehran University of Engineering and Technology 2026

Figure 3: Class Distribution after using ADASYN

3.1.3 Feature Encoding and Data Splitting

Prior to feeding the data into the machine learning models, all categorical features were converted into numerical representations using label encoding.

- Feature: 'Symptom_1', 'Symptom_2', 'Symptom_3' (8 unique categories):

0: Body ache, 1: Cough, 2: Fatigue, 3: Fever, 4: Headache, 5: Runny nose, 6: Shortness of breath, 7: Sore throat.

- Feature: 'Diagnosis' (5 unique categories):

0: Bronchitis, 1: Cold, 2: Flu, 3: Healthy, 4: Pneumonia.

The complete 5806 sample data was then comprehensively divided into training and test samples in the ratio of 60 and 40 respectively. This split ratio was selected to give a large enough training set to use complex models with and a large hold-out test set to test the performance of the complex models in an effective way.

Table 1 presents the results of the data partitioning process, where the training dataset had 3,483 samples and the testing dataset had 2,323 samples. In the case of the training set, the features matrix will be of shape (3483, 9), and the target vector will be of shape (3483,). Likewise, the 9-feature dimension is preserved in the features and target vectors of the testing and has 2,323 samples each. This proves that the nine independent features are maintained in both sets that will be used in the training and evaluation stages.

Sample records from the resulting training and test datasets, illustrating the encoded numerical features, are presented in Table 3 and Table 2, respectively.

Table 1

Shapes of Test and Train Data

Dataset	Features Shape	Target Shape	Total Samples
Train	(3483, 9)	(3483,)	3483
Test	(2323, 9)	(2323,)	2323

Table 2.

Sample from Training Data

Age	Symptom 1	Symptom 2	Symptom 3	Heart Rate bpm	Body Temperature C	Oxygen Saturation %	bp systolic	bp diastolic	Diagnosis
64	2	5	7	100	38.8	94	128	88	3
27	5	4	3	60	37.3	90	112	116	3
52	1	4	3	95	39.49271	90	113	111	2
31	1	3	4	85	36.9779	93	166	71	0
49	4	1	2	109	36.18224	98	105	112	1

Table 3.

Sample from Test Data

Age	Symptom_1	Symptom_2	Symptom_3	Heart_Rate_bpm	Body_Temperature_C	Oxygen_Saturation_%	bp_systolic	bp_diastolic	Diagnosis
55	2	6	1	71	38.3	90	107	96	0
20	3	3	5	98	36.18025	96	104	87	1
52	4	3	0	88	39.22744	94	132	83	2
45	2	0	1	78	39.31392	95	174	71	2
59	1	3	0	106	39.46772	93	159	108	4

3.2. Ensemble Model Development

The primary contribution of this research involves the development and comparative analysis of three high-performance ensemble learning strategies. These models are built upon a foundation of carefully tuned, diverse base classifiers:

- Random Forest Classifier (RF): A highly stable tree-based model optimized with 100 decision trees to reduce variance and prevent overfitting.
- Gradient Boosting Classifier (GB): An adaptive model that builds trees sequentially, correcting the errors of the preceding trees, also using 100 estimators for strong predictive power.
- Support Vector Classifier (SVC): A powerful linear and non-linear model configured to output class probabilities for use in soft-voting.
- Logistic Regression (LR): Used primarily as the final meta-classifier in the stacking approach.

Using these base classifiers, three distinct ensemble architectures were constructed:

- Hard Voting Classifier: This model aggregates the class predictions from the base classifiers (RF, GB, SVC) and selects the class with the most votes as the final prediction. This method relies on simple majority rule.
- Soft Voting Classifier: Unlike hard voting, this approach averages the predicted probabilities of the base classifiers (RF, GB, SVC, requiring probability=True for SVC). The class with the highest average probability is chosen, making it generally more robust than hard voting.

© Mehran University of Engineering and Technology 2026

- Stacking Classifier: This advanced ensemble technique utilizes the predictions of the base classifiers (RF, GB, SVC) as input features for a final meta-classifier, which is a Logistic Regression model. This allows the meta-classifier to learn the optimal way to combine the strengths of the individual models.

The performance of these three models (Hard Voting, Soft Voting, and Stacking) was rigorously compared, and the single best-performing ensemble classifier was selected for all subsequent interpretability and counterfactual analyses.

3.3. Mathematical Model for Stacking

As the stacking ensemble was selected as the superior model based on empirical performance, its mathematical architecture is defined here. The stacking framework employs a two-tier architecture where the outputs of diverse base learners are utilized as input features for a meta-learner.

Proposed stacking model uses three base learners:

- Random Forest: RF
- Gradient Boosting: GB
- Support Vector Classifier: SVC

Let the input feature vector for a sample be in Eq-1:

$$\mathbf{x} = [x_1, x_2, \dots, x_p]^T \quad (1)$$

The base learners produce class probability predictions depicted in Eq-2:

$$\begin{aligned} \hat{y}^{(RF)} &= f_{RF}(\mathbf{x}) \\ \hat{y}^{(GB)} &= f_{GB}(\mathbf{x}) \\ \hat{y}^{(SVC)} &= f_{SVC}(\mathbf{x}) \end{aligned} \quad (2)$$

where f_{RF}, f_{GB}, f_{SVC}map input features to predicted probabilities (or decision function values in the case of SVC).

3.3.1. Meta-Learner (Logistic Regression)

The meta-learner takes the predictions of the base learners as input shown in Eq-3:

$$\mathbf{z} = [\hat{y}^{(RF)}, \hat{y}^{(GB)}, \hat{y}^{(SVC)}]^T \quad (3)$$

Then the final prediction is calculated in Eq-4:

$$\hat{y}^{(stack)} = f_{meta}(\mathbf{z}) = \sigma(w_0 + \sum_{i=1}^{3} w_i \hat{y}^{(i)}) \quad (4)$$

where:

- $\sigma(\cdot)$is the logistic sigmoid function: $\sigma(a) = \frac{1}{1+e^{-a}}$
- $w_0$is the bias term
- $w_i$are the weights learned by logistic regression
- $\hat{y}^{(i)}$are the base learner outputs

For multiclass classification, logistic regression generalizes to softmax as calculated in Eq-5:

$$\hat{\mathbf{y}}_c^{(stack)} = \frac{\exp\left(w_{0,c} + \sum_{i=1}^{3} w_{i,c}\, \hat{y}^{(i)}\right)}{\sum_{k=1}^{K} \exp\left(w_{0,k} + \sum_{i=1}^{3} w_{i,k}\, \hat{y}^{(i)}\right)} \quad (5)$$

where $c = 1, \dots, K$represents the class index.

3.3.2. Overall Stacking Function

Combining base and meta learners, the stacking model can be written as shown in Eq-6:

$$\hat{y}^{(stack)} = f_{stack}(\mathrm{x}) = f_{meta}(f_{RF}(\mathrm{x}), f_{GB}(\mathrm{x}), f_{SVC}(\mathrm{x})) \quad (6)$$

- This equation explicitly shows two-level mapping:
    1. Base learners f_{RF}, f_{GB}, f_{SVC}
    2. Meta-learner f_{meta}

3.3.3. Loss Function (for Training Meta-Learner)

For binary classification, logistic regression minimizes cross-entropy loss on meta-features calculated in Eq-7:

$$\mathcal{L} = -\frac{1}{N} \sum_{j=1}^{N} [y_j \log \hat{y}_j^{(stack)} + (1 - y_j) \log (1 - \hat{y}_j^{(stack)})] \quad (7)$$

For multiclass shown in Eq-8:

$$\mathcal{L} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{c=1}^{K} y_{j,c} \log \hat{y}_{j,c}^{(stack)} \quad (8)$$

where $y_{j,c} = 1$if sample $j$belongs to class c, else 0.

3.4. Statistical Results and Performance Metrics

The models were tested in terms of predictive performance that was assessed through a multi-class confusion matrix and the key metrics such as Accuracy, Precision, Recall, and F1-score. These actions give a

global evaluation of model performance in each of the five classes of disease.

3.5. LIME Analysis

In order to enhance transparency, the Local Interpretable Model-agnostic Explanations (LIME) framework was used to explain the results of the ensemble model. LIME describes individual predictions by determining the most effective features that contribute to every disease classification.

3.6. Modifying Features to Transform Each Disease into All Others

The last step uses counterfactual explanations through DiCE framework to find the least number of features changes needed to transform a predicted diagnosis to another disease category.

The counterfactual analysis investigates the changes that need to be made to shift the prediction in terms of each disease:

- From 0: Bronchitis to: 1: Cold, 2: Flu, 3: Healthy, 4: Pneumonia.
- From 1: Cold to: 0: Bronchitis, 2: Flu, 3: Healthy, 4: Pneumonia.
- From 2: Flu to: 0: Bronchitis, 1: Cold, 3: Healthy, 4: Pneumonia.
- From 3: Healthy to: 0: Bronchitis, 1: Cold, 2: Flu, 4: Pneumonia.
- From 4: Pneumonia to: 0: Bronchitis, 1: Cold, 2: Flu, 3: Healthy.

Age was rigorously limited in the counterfactual generation in order to keep bio-realism. The model gets the real age of the patient and puts it within [actual age, 120], such that the age is always constant or grows but never diminishes. Other continuous variables like Heart_Rate_bpm, Body_Temperature_C, Oxygen_Saturation_%, bp_systolic and bp_diastolic were left to be at liberty. Such adaptability facilitates a diagnosis to be changed by having minimum feature variations that can be identified in the model. The counterfactual analysis in such a way performs as a clinical decision-support mechanism as it indicates how the shift in symptoms or vital signs may actually alter the current prediction of health status of a patient. Algorithm is shown in Table 4.

Table 4.

Algorithm for Counterfactual Explanation Generation Using DiCE

Input:
• Trained black-box model **stacking_clf**
• Dataset **waterData**
• Input samples **inputdata**
• Continuous features ⟨Age, Symptom$_1$, Symptom$_2$, Symptom$_3$, Heart_Rate_bpm, Body_Temperature_C, Oxygen_Saturation_%, bp_systolic, bp_diastolic⟩
Output:
• Counterfactual explanations ⟨CF_1, CF_2, …, CF_n⟩
Algorithm Steps:

1. **Initialization**
   • Define continuous feature set
   • Create **DiCE data object** using **waterData**
   • Load trained model **stacking_clf**
   • Initialize **DiCE explainer** with random search
2. **DiCE Setup**
   • Data_dice ← (Dataset = waterData, Continuous Features, Outcome = Diagnosis)
   • Model_dice ← wrapper(stacking_clf)
   • Explainer ← Dice(Data_dice, Model_dice, method = "random")
3. **For each instance *i* in inputdata do**
   3.1 Extract instance → **first_row** ← **inputdata[i]**
   3.2 Get immutable feature → **actual_age** ← **first_row.Age**
   3.3 Define constraint → **Age ∈ [actual_age, 120]**
   3.4 Generate counterfactual CF_i ← GenerateCounterfactuals(first_row, Total_CFs = 1, Desired_Class = 0, Permitted_Range)
   3.5 Display **CF_i**
4. **End For**
5. **Return**
   • Counterfactual set ⟨CF_1, CF_2, CF_3, CF_4, CF_5⟩.

4 Experimental Results and Discussions

The initial step in our analysis involved comparing the predictive reliability of all deployed models. This comparison was first conducted by visually inspecting their respective confusion matrices, as shown in Fig. ***4***.

It can be seen that in Fig. ***4***, there are distinct differences in model performance as a result of a comparison of the six confusion matrices. The base classifiers especially the Support Vector machine (SVM) and the Logistic Regression (LR) have strong off-diagonal values which means that the misclassification rates are greater. The ensemble models, Hard Voting, Soft Voting, and Stacking, on the

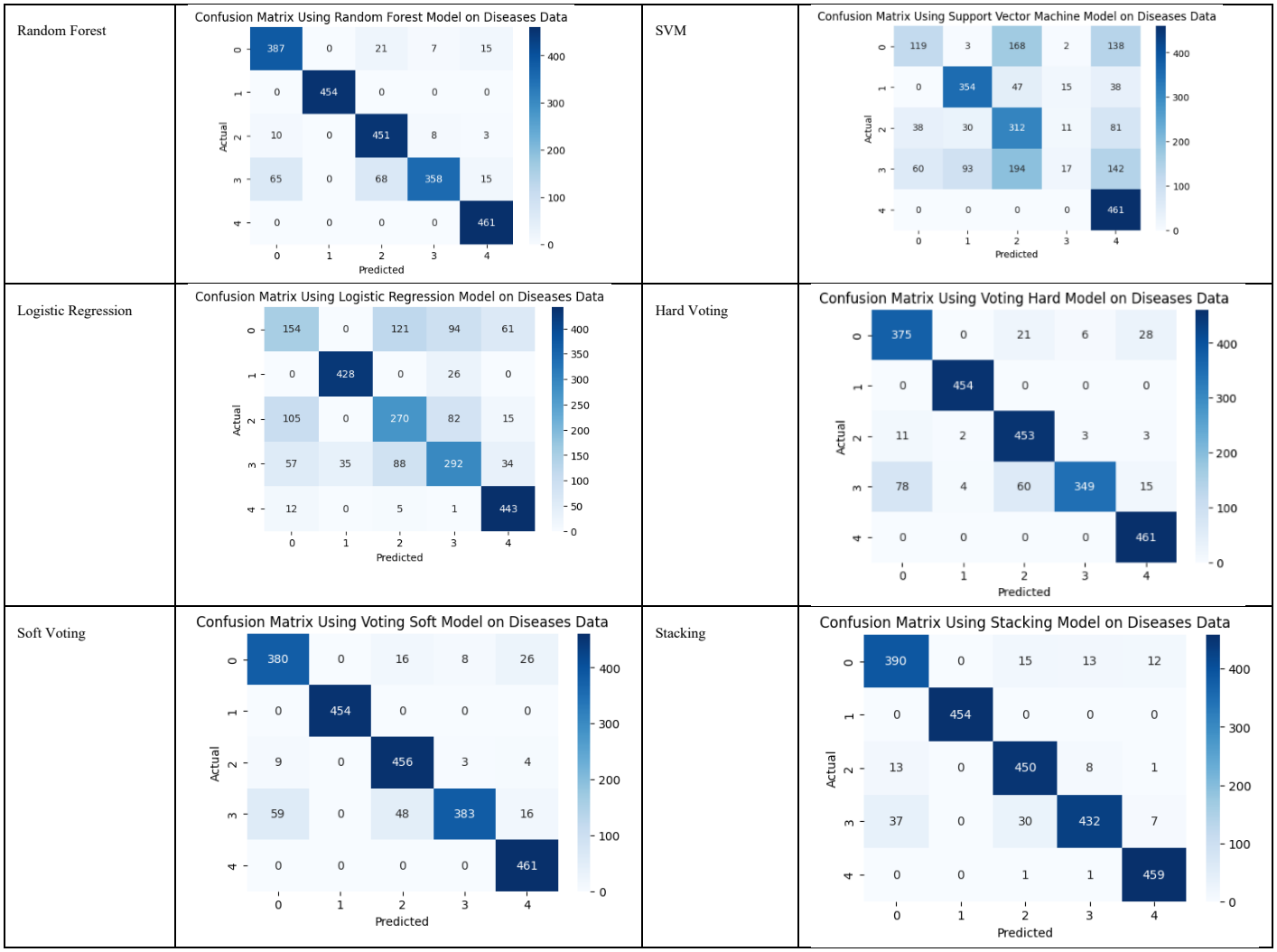

Fig. 4. Confusion Matrices for all Models

other hand, have a stronger degree of diagonal patterns indicating better prediction accuracy. Of these, Stacking model is the least misclassified, which means that it has better generalization abilities.

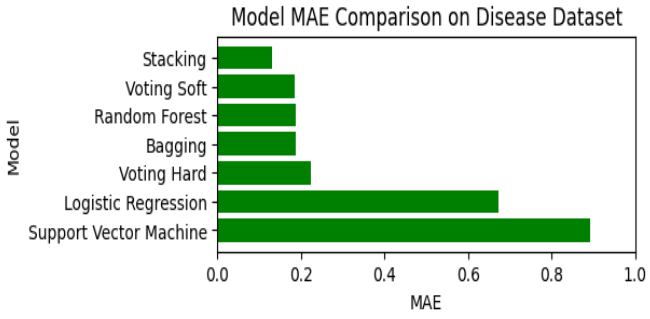

Fig. 5. Model MAE Comparison on Disease Dataset

Mean Absolute Error (MAE) and Accuracy were employed to numerically estimate the models. Fig. ***5*** and Fig. ***6*** bar charts reveal that Stacking model has lowest MAE and highest accuracy indicating that it can better predict the population compared to individual classifiers.

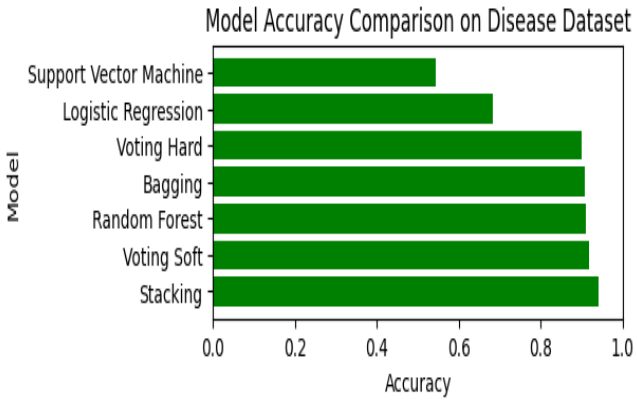

Fig. 6. Model Accuracy Comparison on Disease Dataset

4.1. Statistical Results for Stacking Model

The statistical performance of the Stacking model was studied considering that it had the best performance. The weighted F1-score in Table 5 is 0.94, which is in

© Mehran University of Engineering and Technology 2026

line with the attained accuracy. The proposed model works well in the Cold class (Precision, Recall, F1-score = 1.00) and exhibits a high recall (1.00) in the Pneumonia previously underrepresented class, which serves as a good indicator of the efficiency of the ADASYN balancing method. The minimum recall is found in the Healthy class which means that there are a few conservative predictions in healthy people.

Table 5.

Confusion Matrix Measures for Stacking Model for Disease Prediction

Class / Metric	Precision	Recall	F1-Score	Support
Bronchitis (0)	0.89	0.91	0.90	430
Cold (1)	1.00	1.00	1.00	454
Flu (2)	0.91	0.95	0.93	472
Healthy (3)	0.95	0.85	0.90	506
Pneumonia (4)	0.96	1.00	0.98	461
Weighted Avg	0.94	0.94	0.94	2323
Accuracy	0.94			

Fig. *7* shows the confusion matrix which gives the results of the classification by the Stacking model. The predictions are concentrated mostly along the diagonal which means that there was a high rate of classification of all the diseases. There is only a small group of misclassifications in which some Healthy cases are forecasted to be Bronchitis.

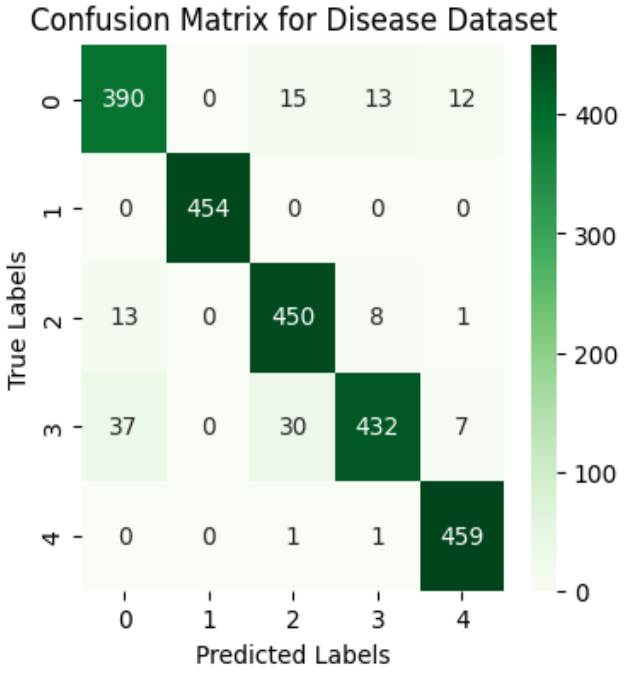

Fig. 7: Heat Map on Stacking Model for Disease Prediction

4.2. Cumulative Gain Curve on Stacking Model for Disease Prediction

Fig. ***8*** The cumulative gain curves of the five disease classes are shown in figure 8. Each of the curves is very high in comparison to the random baseline, which is a good sign of a well-functioning model. Most of the true positives are found in a small percentage of samples in the model and this implies good prediction efficiency.

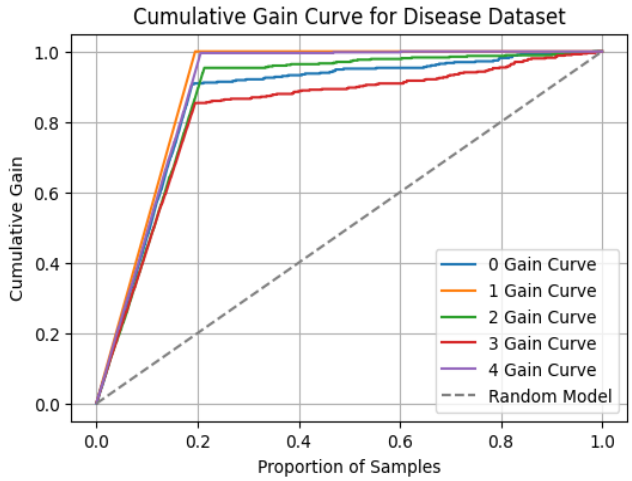

Fig. 8. Cumulative Gain Curve for Disease Dataset

4.3. Lift Curve on Stacking Model for Disease Prediction

Fig. ***9*** compares the lift curve that gauges the effectiveness of the model to random selection. The curve starts at the highest lift value which implies that the model is almost twice as effective in detecting positive cases. Such a high start-up lift proves the high predictive ability of the stacking classifier.

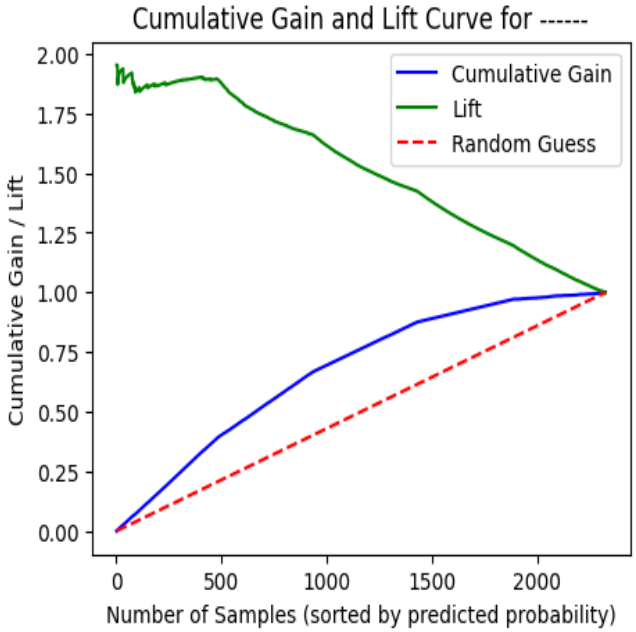

Fig. 9. Lift Curve for Disease Dataset

© Mehran University of Engineering and Technology 2026

4.4. ROC Curve on Stacking Model for Disease Prediction

The multiclass ROC curves were displayed in Figure 10 with One-vs-Rest strategy. The AUC values of all the classes are very high (above 0.98), which means high discriminative capability. The curves nearest to the top-left corner verify high sensitivity at low false positive rate.

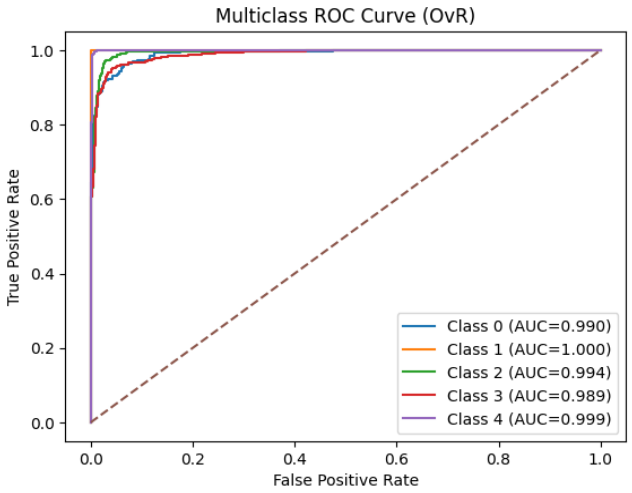

Figure 10: ROC Curve for Disease Dataset

4.5. Precision-Recall Curve on Stacking Model for Disease Prediction

Fig shows the Precision-Recall curves of all classes of diseases. The large mean values of precision show that it is highly performing with skewed data. The stable curves show that the stacking model is accurate and stable in its predictions.

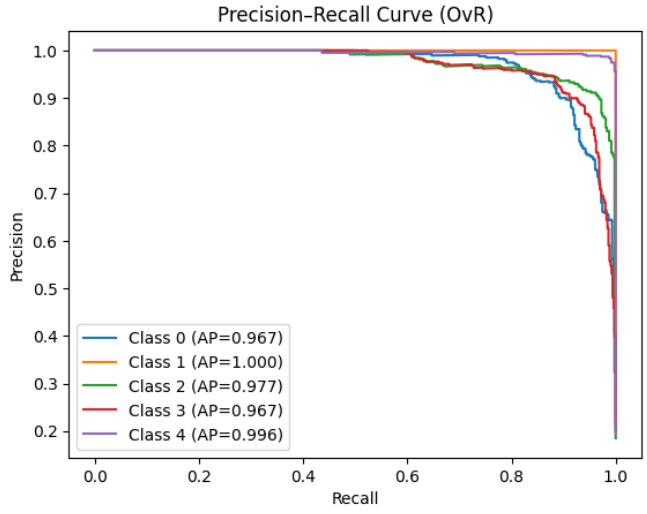

Fig. 11. Precision-Recall Curve for Disease Dataset

4.6. XAI Analysis as LIME

The purpose of using the Local Interpretable Model-The Stacking model is interpreted by use of LIM to justify why this particular prediction was made. It brings out the features that impacted positively or negatively on the diagnosis of individual patient records.

In Fig. ***11*** (Record-501, Age 55), the model successfully forecasted Bronchitis which was dominated by high values of Symptom 2 that suggests respiratory symptoms. Lower Saturation of oxygen and low Symptom_1 slightly steered the prediction away but was unbalanced to larger positive factors.

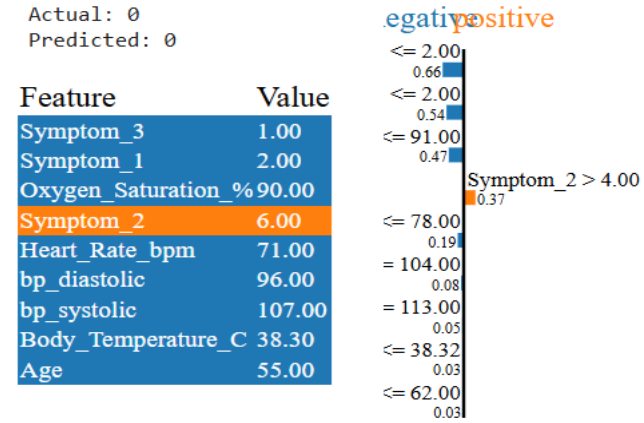

Fig. 11. LIME Analysis for Record-501 with Actual and Predicted as Bronchitis (0)

In Fig. ***12***, 'cold' was also well predicted by the model and strongly dependent on the very low body temperature and moderate oxygen saturation. The low negative values show that there is great confidence in the prediction.

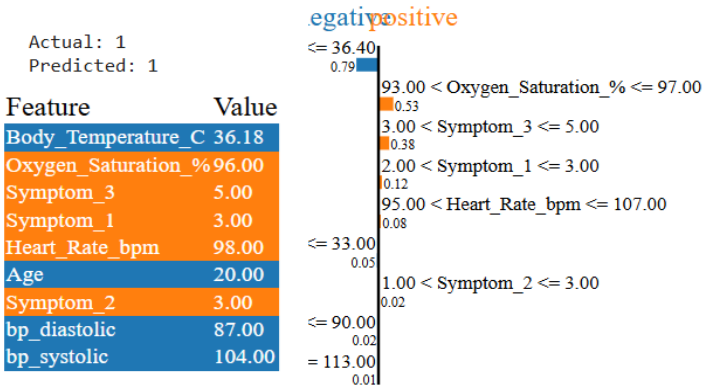

Fig. 12. LIME Analysis for Record-3162 with Actual and Predicted as Cold (1)

In Fig. ***13*** for Record-3782, very high body temperature and moderate saturation levels of oxygen are the primary factors influencing the Flu prediction.

© Mehran University of Engineering and Technology 2026

These characteristics make Flu obviously different in comparison with other respiratory diseases.

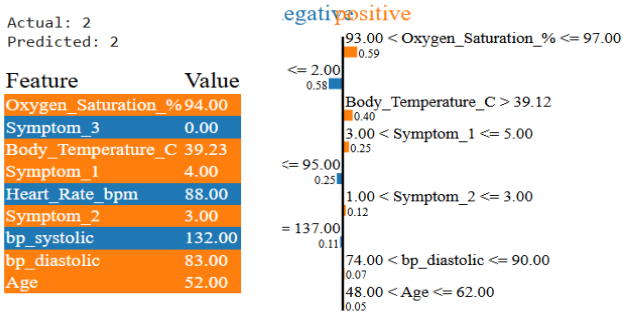

Fig. 13. LIME Analysis for Record-3782 with Actual and Predicted as Flu (2)

In Fig. ***14*** (Record-1322, Age 42), the model forecasted Healthy because the vital signs including normal temperature and moderate oxygen saturation were stable. Even though there were some weak symptoms, they were not so intense to alter the projection towards a type of disease.

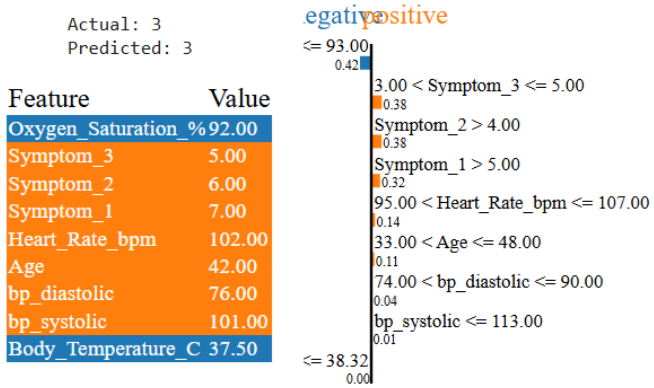

Fig. 14. LIME Analysis for Record-1322 with Actual and Predicted as Healthy (3)

For Record-5023 (Age 28) in Fig. ***15***, Powerful symptoms and high blood pressure have an effect on the Pneumonia prediction. This combination of factors had an overriding negative impact on the ultimate decision due to oxygen saturation.

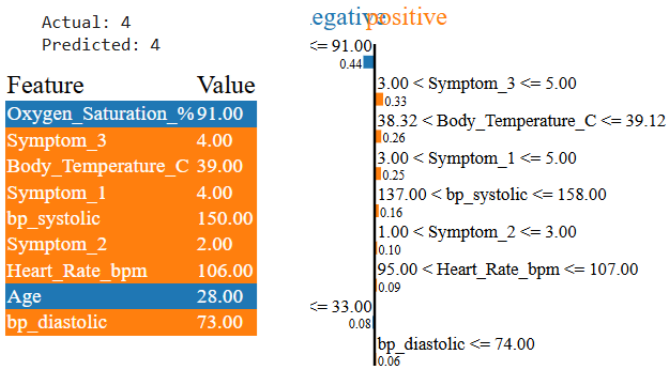

Fig. 15. LIME Analysis for Record-5023 with Actual and Predicted as Pneumonia (4)

4.7. Modified Features for Cross Diseases

The final analysis uses the DiCE counterfactual framework to identify minimal changes in patient health metrics needed to shift the Stacking model's diagnosis to another class. Vital features could vary freely, while Age was restricted to remain constant or increase to maintain physiological realism.

4.7.1. Cold, Flu, Healthy, Pneumonia to Bronchitis

In Table 6, two primary levers are thrown to bring the predictions to Bronchitis: reduce Oxygen Saturation (Cold) or reduce Body Temperature (fever) (the effect of Flu and Pneumonia). In the Healthy patient, a shift in Symptom26 of Shortness of breath with 6 to 1 (Cough) with an intermediate rebound in Heart Rate (102 bpm to 107 bpm) is sufficient to activate the diagnosis of Bronchitis.

Table 6

Modified Features for Cold, Flu, Healthy, Pneumonia to Bronchitis

Type	Age	S1	S2	S3	Heart_Rate	Body_Temp	O_2_Sat	bp_sys	bp_dia	Diagnosis
Actual	20	3	3	5	98	36.18	96	104	87	Cold
Modified	20	3	3	3	98	36.18	92	104	87	Bronchitis
Actual	52	4	3	0	88	39.23	94	132	83	Flu
Modified	52	4	3	0	88	37.10	94	132	83	Bronchitis
Actual	42	7	6	5	102	37.50	92	101	76	Healthy
Modified	42	7	1	5	107	37.50	92	101	76	Bronchitis
Actual	59	1	3	0	106	39.47	93	159	108	Pneumonia
Modified	59	1	5	0	106	38.20	93	159	108	Bronchitis

4.7.2. Bronchitis, Flu, Healthy to Pneumonia to Cold

In Table 7, for nearly all illnesses (Bronchitis, Flu, Pneumonia), the change to Cold is overwhelmingly driven by the normalization of vitals. Specifically, a large drop in Body Temperature (from a feverish state to a normalized or slightly low temperature) coupled with an increase in Oxygen Saturation is the minimal counterfactual needed, aligning with the expected profile of a non-severe Cold diagnosis.

4.7.3 Bronchitis, Cold, Healthy, Pneumonia to Flu

In Table 8, the shift to Flu is primarily driven by the introduction or maintenance of a fever (Body_Temp increase from Cold and Healthy) and changes in Heart Rate. Notably, the Pneumonia case only required a dramatic drop in Heart Rate (from 106 bpm to 61 bpm) to shift the high-fever, high-bp case into a Flu diagnosis, showing that the model uses relative heart stability as a key differentiator for severe respiratory illness.

© Mehran University of Engineering and Technology 2026

Table 7.

Modified Features for Bronchitis, Flu, Healthy to Pneumonia to Cold

Type	Age	S1	S2	S3	Heart_Rate	Body_Temp	O_2_Sat	bp_sys	bp_dia	Diagnosis
Actual	55	2	6	1	71	38.30	90	107	96	Bronchitis
Modified	55	4	6	1	71	35.60	99	107	96	Cold
Actual	52	4	3	0	88	39.23	94	132	83	Flu
Modified	52	4	3	0	88	36.20	97	132	83	Cold
Actual	42	7	6	5	102	37.50	92	101	76	Healthy
Modified	42	7	6	5	102	35.50	96	101	76	Cold
Actual	59	1	3	0	106	39.47	93	159	108	Pneumonia
Modified	59	1	3	0	106	35.70	98	159	108	Cold

Table 8.

Modified Features for Bronchitis, Cold, Healthy, Pneumonia to Flu

Type	Age	S1	S2	S3	Heart_Rate	Body_Temp	O_2_Sat	bp_sys	bp_dia	Diagnosis
Actual	55	2	6	1	71	38.30	90	107	96	Bronchitis
Modified	55	2	6	1	71	38.30	96	93	96	Flu
Actual	20	3	3	5	98	36.18	96	104	87	Cold
Modified	20	3	3	5	98	39.70	96	104	74	Flu
Actual	42	7	6	5	102	37.50	92	101	76	Healthy
Modified	42	7	6	3	82	38.40	92	101	76	Flu
Actual	59	1	3	0	106	39.47	93	159	108	Pneumonia
Modified	59	1	3	0	61	39.47	93	159	108	Flu

4.7.4. Bronchitis, Cold, Flu, Pneumonia to Healthy

Based on Table 9, it is common that the Healthy diagnosis involves normalizing blood pressure (increase in bp_diastolic with Bronchitis and Flu) or, counter-intuitive with Cold, a decrease of O_2_Sat (96% to 90 %) to achieve a Healthy diagnosis. In the case of the severe Pneumonia the least amount of change that needed to be made included increasing the Age (59 to 94) and a change in Symptom_1 (Cough to Sore throat) which is showing that in certain areas of the data the model correlates the characteristics to less severe results despite the severity of the vitals.

Table 9.

Modified Features for Bronchitis, Cold, Flu, Pneumonia to Healthy

Type	Age	S1	S2	S3	Heart_Rate	Body_Temp	O_2_Sat	bp_sys	bp_dia	Diagnosis
Actual	55	2	6	1	71	38.30	90	107	96	Bronchitis
Modified	55	2	6	6	71	38.30	90	107	101	Healthy
Actual	20	3	3	5	98	36.18	96	104	87	Cold
Modified	20	3	3	5	98	36.18	90	104	87	Healthy
Actual	52	4	3	0	88	39.23	94	132	83	Flu
Modified	52	4	7	0	88	39.23	94	132	95	Healthy
Actual	59	1	3	0	106	39.47	93	159	108	Pneumonia
Modified	94	7	3	0	106	39.47	93	159	108	Healthy

4.7.5. Bronchitis, Cold, Flu, Healthy to Pneumonia

Based on Table 10, the transformation into Pneumonia is realistically focused on the joint deterioration of a variety of health measures. To the Cold patient the only three changes needed were an increase in the Heart Rate (98 bpm to 110 bpm), an increase in the Body Temperature (36.18°C to 39.50°C) and a decrease in the Oxygen Saturation (96% to 91%). In the case of Bronchitis, it was only necessary to change Symptom_3 (Cough) to Symptom4 (Headache) and this demonstrates the sensitivity of the model to the occurrence of certain combinations of symptoms despite the vitals remaining the same.

Table 10.

Modified Features for Bronchitis, Cold, Flu, Healthy to Pneumonia

Type	Age	S1	S2	S3	Heart_Rate	Body_Temp	O_2_Sat	bp_sys	bp_dia	Diagnosis
Actual	55	2	6	1	71	38.30	90	107	96	Bronchitis
Modified	55	2	6	4	71	38.30	90	107	96	Pneumonia
Actual	20	3	3	5	98	36.18	96	104	87	Cold
Modified	20	3	3	5	110	39.50	91	104	87	Pneumonia
Actual	52	4	3	0	88	39.23	94	132	83	Flu
Modified	52	4	3	6	110	39.23	94	132	83	Pneumonia
Actual	42	7	6	5	102	37.50	92	101	76	Healthy
Modified	42	2	6	5	102	38.50	92	101	76	Pneumonia

4.8. Cross Validation for Stacking Model

The Stacking Classifier was tested on a five-fold Stratified Cross-Validation in order to test the strength and generalization capability of the model. All the performance statistics are summarized in **Error! Reference source not found.**, where the model had a mean accuracy of 0.9507, F1 Macro of 0.9501 and ROC AUC of 0.9961 with very low standard deviations indicating high stability in the model. The class-prediction ROC analysis also indicated that it had high predictive power with all disease classes having AUC values near to 1.0. The learning curves in Fig. ***16*** illustrate that the training and validation scores begin to stabilize with rise in the training data hence good balance between bias and variance. Moreover, all the calibration curves in Fig. ***17*** indicate that the majority of predicted probabilities are relatively close to the ideal diagonal line, which proves that probability is estimated fairly well and the entire model is robust.

© Mehran University of Engineering and Technology 2026

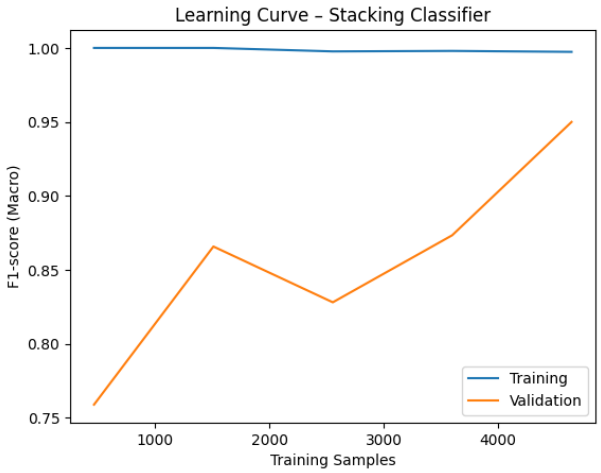

Fig. 16: Learning Curve from Cross Validation

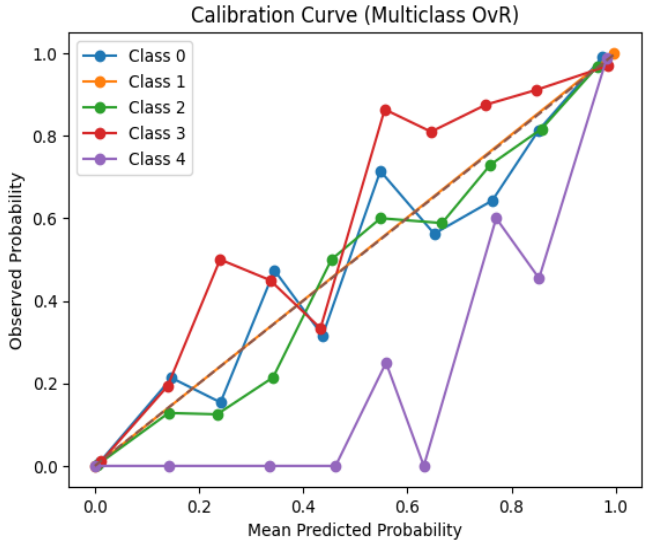

Fig. 17. Calibration Curve from Cross Validation

4.7. Comparison with Existing Works

Critical overview of the relevant studies presented in Table 11 allows quantitatively summarising the current state of the art and clearly defining the novelty of our work. Not only the reported performance metrics are compared but also the degree of interpretability and actionability obtained. Even high-accuracy models often do not provide Local Interpretable Model-agnostic Explanations (LIME) and, most importantly, counterfactual explanations, which explains the gap filled by our proposed Stacking model. Fig. ***18*** clearly demonstrates clear comparative visual representation.

4.8 Performance of Stacking with Respect to Other Models

The performance analysis of the implemented models is described by a set of classification reports and analysis of error metrics Table 12, Table 13 and Table 14 show the individual results of the base classifiers. Random Forest (Table 12) has a strong ability to diagnose being able to achieve an accuracy of 91% which is much higher than Support Vector Machine (Table 13) and Logistic Regression (Table 14) which produced lower accuracies of 55% and 68% respectively

Table 11.

Comparison of Proposed Work with Existing Works

Used Model(s) / Approach	Reported Performance (Accuracy / Key Metric)	LIME	Counterfactual
CNNs (VGG16, ResNet50V2, DenseNet121, etc.) + XAI [8]	99%	Yes	No
CNN + transfer learning + hybrid/ensemble + XAI [9]	96%	Yes	No
Ensemble CNNs (DenseNet169, MobileNetV2, ViT) [10]	93%	No	No
Ensemble InceptionV3, VGG16, ResNet50, ViT [11]	96%	No	No
Ensemble CNNs + ML classifiers (RF, SVM, Decision Tree) [12]	99%	No	No
Ensemble XGBoost, CatBoost, LGBM, MLP [13]	90%	No	No
ML ensemble + SHAP [14]	97%	No	No
Ensemble DL on CT/X-ray [15]	98%	No	No
Hybrid using ResNet-50 + texture features + SVM [16]	97%	No	No
Proposed Model Stacking (RF, GB, SVC meta-classifier LR)	94%	Yes	Yes

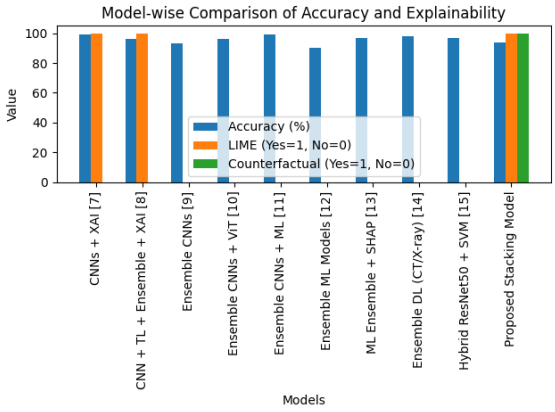

Fig. 18. Comparative Representation of Proposed Model with Benchmarks

© Mehran University of Engineering and Technology 2026

because they cannot model complex non-linear relationships of symptoms.

Table 12:

Random Forest Classification Performance

Class	Precision	Recall	F1-score
0	0.85	0.90	0.87
1	1.00	1.00	1.00
2	0.84	0.96	0.89
3	0.95	0.72	0.82
4	0.94	1.00	0.97
Weighted Avg	0.92	0.91	0.91
Accuracy	0.91		

Table 13.

Support Vector Machine Classification Performance

Class	Precision	Recall	F1-score
0	0.58	0.28	0.37
1	0.75	0.81	0.78
2	0.44	0.68	0.53
3	0.41	0.03	0.05
4	0.53	1.00	0.70
Weighted Avg	0.54	0.55	0.48
Accuracy	0.55		

Table 14.

Logistic Regression Classification Performance

Class	Precision	Recall	F1-score
0	0.43	0.34	0.38
1	0.91	0.99	0.95
2	0.54	0.55	0.54
3	0.60	0.55	0.57
4	0.81	0.96	0.88
Weighted Avg	0.66	0.68	0.67
Accuracy	0.68		

Table 15,

Table ***16*** and Table 17 compare ensemble methods as Voting (Hard) and Bagging have a consistent 90% accuracy whereas Voting (Soft) raises it by a minor margin to 92%. The highest overall performance with the maximum accuracy and weighted average F1-score of 94% is reached with the Proposed Stacking Model, which is described in Table 18. Class 1 (Flu) was perfectly reliable, and Class 4 (Pneumonia) was highly sensitive in this model.

Table ***19*** synthesizes these findings, contrasting the accuracy against the Mean Absolute Error (MAE), where the Stacking Model registers the lowest error rate of 0.1180.

Table 15.

Voting (Hard) Classification Performance

Class	Precision	Recall	F1-score
0	0.83	0.87	0.85
1	0.99	1.00	1.00
2	0.84	0.97	0.90
3	0.97	0.69	0.81
4	0.91	1.00	0.95
Weighted Avg	0.91	0.90	0.90
Accuracy	0.90		

Table 16.

Voting (Soft) Classification Performance

Class	Precision	Recall	F1-score
0	0.86	0.86	0.86
1	1.00	1.00	1.00
2	0.87	0.97	0.92
3	0.96	0.76	0.85
4	0.91	1.00	0.95
Weighted Avg	0.92	0.92	0.91
Accuracy	0.92		

Table 17.

Bagging Classification Performance

Class	Precision	Recall	F1-score
0	0.83	0.84	0.84
1	1.00	1.00	1.00
2	0.83	0.95	0.88
3	0.91	0.71	0.80
4	0.93	1.00	0.96
Weighted Avg	0.90	0.90	0.90
Accuracy	0.90		

Table 18.

Proposed Stacking Model Classification Performance

Class	Precision	Recall	F1-score
0	0.92	0.89	0.90
1	1.00	1.00	1.00
2	0.91	0.96	0.93
3	0.94	0.88	0.91
4	0.95	1.00	0.97
Weighted Avg	0.94	0.94	0.94
Accuracy	0.94		

© Mehran University of Engineering and Technology 2026

Table 19.

Model Accuracy and Mean Absolute Error

Model	Accuracy	Mean Absolute Error (MAE)
Random Forest	0.91	0.1791
Support Vector Machine	0.55	0.8709
Logistic Regression	0.68	0.6870
Voting (Hard)	0.90	0.2135
Voting (Soft)	0.92	0.1898
Bagging	0.90	0.2157
Proposed Stacking Model	0.94	0.1180

5 Practical Applications and Domain Relevance

The clinical and domain applicability of this paper is exemplified with a number of real-life applications. First, being a Clinical Decision Support System (CDSS), the model can help doctors and nursing personnel to work in crowded clinical settings, as it offers more evidence-based suggestions on the diagnosis faster. Second, the model can be adopted well to Remote Healthcare and Telemedicine; it can be incorporated in mobile health applications and this enables a patient to check the symptoms at home and be notified promptly in case he or she needs the services of a professional medical team. Third, the system Improves Emergency Triage, the system can quickly sort the patients according to their severity; that is, distinguishing between life-threatening and non-life threatening conditions such as Pneumonia and common cold to maximize the allocation of bed and resources in the hospital. Lastly, Actionable Intervention (DiCE) is integrated to make sure that the system does not just give a label but it gives the specific vital sign(s) (like oxygen saturation or heart rate) that primarily drive a health state change and provide clinicians with a clear roadmap of what to do to a particular patient.

6 Conclusion

This paper designed and tested a number of ensemble machine learning models to predict across multiple classes, the respiratory health state, including Bronchitis, Cold, Flu, Healthy, and Pneumonia. The most successful of the tested methods was the Stacking Classifier that used the combination of Random Forest, Gradient Boosting, and Support Vector Classifier with Logistic Regression as the meta-classifier; the overall accuracy, as well as weighted F1-score of the developed approach, were 94%. The use of the ADASYN technique of data balancing greatly enhanced the model with respect to identifying the minority Pneumonia class correctly with a very high degree of recall.

In addition to predictive performance, another important shortcoming of current research is the inclusion of an Explainable AI (XAI) framework in this work. Local Interpretable Model-agnostic Explanations (LIME) offered clear patient-level explanations of predictions, whereas the Diverse Counterfactual Explanations (DiCE) allowed taking practical actions by determining the least feature modifications necessary to yield changes in disease predictions. This method favors interpretability as well as clinical decision-making.

Further studies will be aimed at enhancing model robustness by using other balancing techniques, increased and more varied data and other validation measures. The application of wearable sensor data, electronic health records and real-time clinical deployment may further increase the effectiveness of the framework towards a practical healthcare application.

7 References

[1] E. L. Ray and N. G. Reich, "Prediction of infectious disease epidemics via weighted density ensembles," *PLoS Comput. Biol.*, vol. 14, no. 2, 2018, doi: 10.1371/journal.pcbi.1005910.

[2] C. Viboud, L. Simonsen, and G. Chowell, "A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks," *Epidemics*, vol. 15, pp. 27–37, 2016, doi: 10.1016/j.epidem.2016.01.002.

[3] A. Mobeen, M. Shafiq, M. H. Aziz, and M. J. Mohsin, "Impact of workflow interruptions on baseline activities of the doctors working in the emergency department," *BMJ Open Qual.*, vol. 11, no. 3, 2022, doi: 10.1136/bmjoq-2022-001813.

[4] T. A. Kakani, J. Vedula, M. Mohammed, R. Gupta, K. Hudani, and N. Yuvaraj, "Developing Predictive Models for Disease Diagnosis using Machine Learning and Deep Learning Techniques," *Proc. 6th Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2025*, pp. 158–163, 2025, doi: 10.1109/ICICV64824.2025.11085894.

[5] D. M. Sunil and P. V. Joshi, "Early Prediction of Chronic Diseases Through Multimodal Data Fusion in Machine Learning: A Survey," *Proc. 7th Int. Conf. Intell. Sustain. Syst. ICISS 2025*, pp. 1108–1115, 2025, doi:

10.1109/ICISS63372.2025.11076438.

[6] Anjali Yadav, Shruti Dwivedi, Anubhav Dwivedi, Ujjwal Thakur, and Dr. Nikhat Akhtar, "Intelligent Disease Diagnosis: A Multi-Disease Prediction Approach Using Machine Learning," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 98–109, 2025, doi: 10.32628/ijsrset251235.

[7] H. Sadr *et al.*, "Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches," *Eur. J. Med. Res.*, vol. 30, no. 1, 2025, doi: 10.1186/s40001-025-02680-7.

[8] S. Sarker, S. R. Refat, F. F. Preotee, T. R. Shawon, and R. Tanvir, "Comprehensive Lung Disease Detection Using Deep Learning Models and Hybrid Chest X-ray Data with Explainable AI," *2024 27th Int. Conf. Comput. Inf. Technol. ICCIT 2024 - Proc.*, pp. 2279–2284, 2024, doi: 10.1109/ICCIT64611.2024.11022605.

[9] T. T. Ifty, S. A. Shafin, S. M. Shahriar, and T. Towhid, "Explainable Lung Disease Classification from Chest X-Ray Images Utilizing Deep Learning and XAI," *2024 IEEE 3rd Int. Conf. Comput. Mach. Intell. ICMI 2024 - Proc.*, 2024, doi: 10.1109/ICMI60790.2024.10586202.

[10] A. Mabrouk, R. P. D. Redondo, A. Dahou, M. A. Elaziz, and M. Kayed, "Pneumonia Detection on Chest X-ray Images Using Ensemble of Deep Convolutional Neural Networks," *Appl. Sci.*, vol. 12, no. 13, 2022, doi: 10.3390/app12136448.

[11] E. Yanar, F. Hardalaç, and K. Ayturan, "PELM: A Deep Learning Model for Early Detection of Pneumonia in Chest Radiography," *Appl. Sci.*, vol. 15, no. 12, 2025, doi: 10.3390/app15126487.

[12] R. Vyas and D. R. Khadatkar, "Ensemble of Deep Learning Architectures with Machine Learning for Pneumonia Classification Using Chest X-rays," *J. Imaging Informatics Med.*, vol. 38, no. 2, pp. 727–746, 2025, doi: 10.1007/s10278-024-01201-y.

[13] S. M. Baik, K. S. Hong, J. M. Lee, and D. J. Park, "Integrating ensemble and machine learning models for early prediction of pneumonia mortality using laboratory tests," *Heliyon*, vol. 10, no. 14, 2024, doi: 10.1016/j.heliyon.2024.e34525.

[14] M. S. I. Sumon, M. S. A. Hossain, H. Al-Sulaiti, H. M. Yassine, and M. E. H. Chowdhury, "Enhancing Influenza Detection through Integrative Machine Learning and Nasopharyngeal Metabolomic Profiling: A Comprehensive Study," *Diagnostics*, vol. 14, no. 19, 2024, doi: 10.3390/diagnostics14192214.

[15] X. Xue *et al.*, "Design and Analysis of a Deep Learning Ensemble Framework Model for the Detection of COVID-19 and Pneumonia Using Large-Scale CT Scan and X-ray Image Datasets," *Bioengineering*, vol. 10, no. 3, 2023, doi: 10.3390/bioengineering10030363.

[16] K. Hashmi, M.F., Husain, A.A., "Hybrid deep learning and handcrafted feature fusion for pneumonia detection in chest X-rays," *J. Electr. Syst. Inf Technol*, vol. 19, no. 94, 2025, doi: https://doi.org/10.1186/s43067-025-00284-z.

[17] "https://www.kaggle.com/datasets/miadul/holistic-health-and-lifestyle-score-dataset".

© Mehran University of Engineering and Technology 2026