International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

A Hypertuned Pipeline Vector Using Meta Classifier Technique for Feature Selection in Multi Disease Prediction

Manjula Rani Indupalli[1], G.Pradeepini[2]
[1]Research Scholar, Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation,Vaddeswaram
Andhra Pradesh, India
e-mail: indupalli.manjula@gmail.com
[2]Professor, Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Vaddeswaram
Andhra Pradesh, India
e-mail: pradeepini_cse@kluniversity.in

Abstract— Automation of health sector plays a very important role especially during this pandemic due to the side effects of either vaccination or attack of the COVID. Most of the researchers designed a system to predict whether a person suffers from a particular disease or not. Few researchers worked on prediction variants of a single disease based on symptoms but due to this COVID-19, different people are getting attacked with different diseases as a side effect. This proposed system aims to identify the multiple diseases that a person may suffer from based on the symptoms. In this paper, the dataset obtained from the open access repository "Kaggle" contains 17 symptoms combinations to identify the one of the 41 types of diseases as class label. All the symptoms may not be important for identification, so in this model, the important features are identified using the pipeline vector of different Machine Learning approaches are passed as base line classifier and decision tree classifier as meta line to the elimination function. The model has got "99.48%" accuracy for selecting the essential features using bagging and boosting algorithms.

Keywords- Baseline, Metaline, Pipeline Vector, K-Best Selection, Feature Rank,Wrapper Method, Hyper Tune, Cross Validation

I. INTRODUCTION

During this stressful working scenario, it is very important to protect the health of a person without being affected by the other external sources or due to the side effects [7]. Sometimes, people neglect small health issues due to various reasons, which may lead to the serious problems as the time passes. The society needs a single window system, which can predict the multiple diseases based on the symptoms posted by the user. The proposed system uses the machine learning techniques for identification of essential features [8]. Any model with high dimensionality gives less performance in executing the model. The model with lesser data helps the model to exhibit better visualization for analyzing each attribute in detail.Machine learning has three ways to identify the essential features namely, statistical approaches using the correlation, wrapper methods and embedded methods[9]. In general, statistical approaches are simple to implement but they doesn't fit for huge amount of data with multiple relations. The dataset in the proposed system has many symptoms to analyze. Since linear relations like correlation doesn't suit for identifying the connectivity symptoms, this model cannot apply this mathematical approach. In Wrapper methods, the selection of features is computed using the elimination mechanism. The popular mechanism is "Recursive Feature Elimination" [10], in which the system starts with either in the forward direction by considering the empty set or in the backward direction by considering all the features. The model computes the support for each attributes and if it is better than the threshold value then it retains the value otherwise it discards the elements from the sets.

Variance is computed for each attribute and its impact over the data is measured to compute attributes significance. The wrapper component takes two parameters; first one is estimator which uses the concept of entropy to construct the decision tree [11]. This decision tree helps to identify the root node at each level and eliminates the impure values. These impure values are considered as unimportant. The second parameter is a static value that represents the number of attributes to be selected. The major disadvantage of this approach is determination of number of essential features ahead. Any standard approach cannot determine the number of required features before-hand. So the proposed model solves

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

this problem by identifying the accuracy based on the number of features selected. The minimum value with good accuracy is considered as the result. The selection of criteria plays a vital role in determining the performance. [12] Instead of decision tree, the model has chosen hyper tuned algorithm for passing as the estimator. Instead of one classifier, the model has passed more number of tuned algorithms to take the decision based on the majority voting. During this process, the model has achieved success using the ensemble approaches.

The paper has discussed the entire work in five sections; the introduction section discusses about the need of dimensionality reduction along with the process implemented in the recursive feature elimination. This section also describes about the need of modifications in the existing approaches. The literature survey section discusses about the research work performed by the different researchers for predicting multiple diseases using machine learning and deep learning approaches. The proposed methodology section initially discusses the results obtained with the implementation of the traditional approaches then it discusses the working of tuned algorithms, which are part of pipeline vector. The Results & Discussion section presents the algorithm to be chosen from the pipeline vector to act as estimator to the RFE approach. Then it represents the best estimator that suits the data for implementation of tuned algorithm. It also presents the essential symptoms selected to predict the multi diseases. The conclusion discusses the advantages of tuning process along with the future scope for classification process.

II. LITERATURE SURVEY

[1] Karthikeyan et al, improved SVM by enhancing the nature of radial bias nature. This work majorly focused on the chronic diseases like heart, kidney, and other genetically inherited diseases. This model collects the data from different data sources, so it applies different pre-processing techniques by performing statistical analysis with the help of naive Bayesian. The feature extraction is performed by combining the forward and backward traditional elimination process. The novelty of this approachlies in integrating the radial bias function of linear and non-linear models. The model updates its coefficient vector of kernel function based on the iteration and learning rate. The selection of the kernel function depends on nature of the record acquired, the model runs the t-test to analyze the variance and populates the kernel function with that hypothesized values.

[2] Anil et al, proposed novel feature extraction and selection process using Genetic approach known as “Lion with Butterfly”. In the next step, it has customized the neural network to classify the disease based on symptoms. The model has acquired the different diseased patients records and performed the first order statistical evaluation to verify the integrity of the medical records history. The major focus of this paper is to encode the optimal features of the dataset, to make the Big Data distributed platforms to work with the medical records at a faster rate. The model identified the pitfalls of Lion and Butterfly algorithm and combined them as follows:

a. Butterflies in the genetic algorithm attract the opposite files for mating by updating their stimulus intensity range by emitting the fragrance. But at some point, the searching is limited where the global and local position becomes saturated.

b. Lion algorithm is popular for optimizing the estimators of neural network, Lion and Lioness performs mating using the crossover operation. The model suffers from producing more generations because of the wrong assumption of the termination condition

c. The proposed algorithm defines the fitness function as maximum accuracy with minimum number of neurons. The model initially assumes sensor modality and switch probability. The model first computes the fragrance of the butterfly and generates some random population. Later, for every butterfly, by comparing the random value and switch probability, the male mating is done by using butterfly and female mating is performed by using lioness.

The model constructs belief network by using bias function as the visible component of the hidden layers. The numbers of layers are optimized using the Lion algorithm.

[3] Rachel et al, developed annotation model for labeling the record with multiple diseases. The model gathers the radiology images related to different images and clusters are performed to identify the labels of the each image. This model applies transfer learning on the ResNet by replacing the residual components with skip connection, known as “CT-Net”. During the scanning process, if any image is stored as low quality or contains noisy information, then this model enhances the quality of the image by applying Weiner filter. The model also finds the abnormalities by transplantation technique by analyzing the nodules, opacity, and effusion to find the p-value of the DeLong. The evaluation of the model is computed using AUROC curve instead of accuracy, recall.

[4] Mathur et al, implemented a framework that can predict the 40 different types of skin related diseases. This model has modified the “DenseNet-161”by including the three components. The pre-processing of the images is performed using the augmentation techniques and color balancing techniques. The mobile application takes the scanned images of Lesion and analyzes their patternsof 5,104 patients. The validity of the model is compared with silico, cross validation,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

and AUC regions. The model combined 24 layers of 2D-CNN as a single batch and 32 layers of max pooling as another batch. These batches undergo focal and log loss to minimize the error rate. The novelty of this approach lies in designing the top layers of the CNN with cubic interpolation normalization, which helps to normalize the blur parts and to create augmented images.

[5] Yazeed Zoabi et al, focused on identification of COVID-19 disease based on symptoms using ML approaches. The model collected basic information like gender and age. It also collected regular symptoms associated with COVID. The model contains unbalanced dataset and unbiased features. The system performs SMOTE analysis for balancing the data in association with feature and class label. It experiments with SHAP algorithm to analyze the impact every feature on the class label [13]. The model developed at the level-0 constructed base line model with the help of decision tree variants. Rather than ranking mechanism, the impact analysis has given good essential features with increasing slope towards maximum orientation under region of characteristics.

Table 1: Analysis of Existing Approaches

S.No	Author	Algorithm	Merits	Demerits
1	Karthikeyan	Improved SVM	Since the model has heterogeneous data, instead of linear model the combination of linear and non-linear gives better results	The model implements decision tree rules to solve the ambiguous results. Analyzing huge amount of data using these rules involves a lot of complications because of variations in data
2	Anil	Lion with Butterfly	The mating process helps the module to generate limited population	The algorithm converges quickly and enters into local optima with equal number of male and female options.	problem
3	Rachel	CT-Net	The model has constructed a framework known as “SARLE”, which can identify 83 types of disease.	For few diseases, symptoms are also almost similar. The system need to develop hybrid rules, which is a biggest limitation in this system.
4	Mathur	Transfer Learning on DenseNet-161	The model utilized efficient techniques for analyzing different color and textures of the skin	The model doesn’t analyzeall the relevant information
5	Yazeed Zoabi	Meta Classification with Boosting	SHAP analysis improved all metrics related to model design	The model has limited its scope only to covid.

III. PROPOSED SYSTEM

In traditional approaches, researches are conducted on single disease like heart attack, diabetics, and recently COVID-19 but very fewer researches focused on developing a single window system that can identify the multiple diseases. This section describes the implementations of the traditional approaches as follows:

Many of the applications in prediction of important symptoms selection either implemented mathematical coefficients or KBestSelection [14]. The computation of correlation can be presented in three ways as shown in figure 1.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

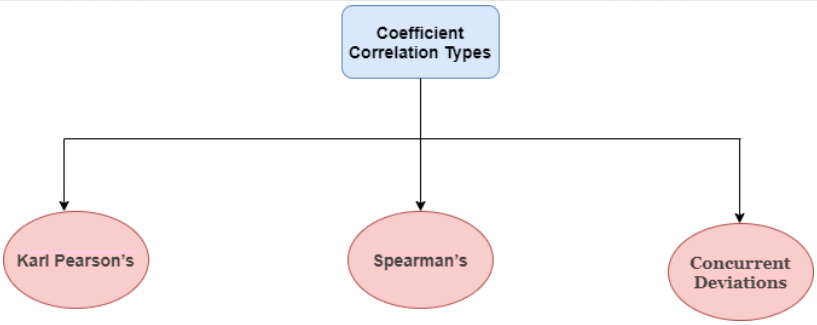

Figure 1: Mathematical Methods of Correlation Coefficient

Karl Pearson computes represents the linear numerical relation between two symptoms but to identify the accurate disease, the system need a coefficient that can analyze relation between all the symptoms [15]. The spearman's coefficient computes the linear relation between multiple symptoms. The mathematical computation is shown in equation (1)

$$Spearman\_Coefficient = 1 - \frac{6*\sum_{i=1}^{m} Distance^2}{m*(m^2-1)} \text{ - (1)}$$

Where,

Distance represents the sum of differences between all the symptoms

m represents number of records available in the dataset

The interpretation of the result is presented in table 2 to infer the conclusions on the dataset.

Table 2: Interpretation of Spearman's Coefficient

S.No	Coefficient	Interpretation on Correlation
1	1	Perfectly correlated
2	0.99 to 0.75	High degree of correlation
3	0.74 to 0.35	Medium degree of correlation
4	<=0.34	Low degree of correlation

The results obtained by the spearman's correlation is presented in the figure 2.

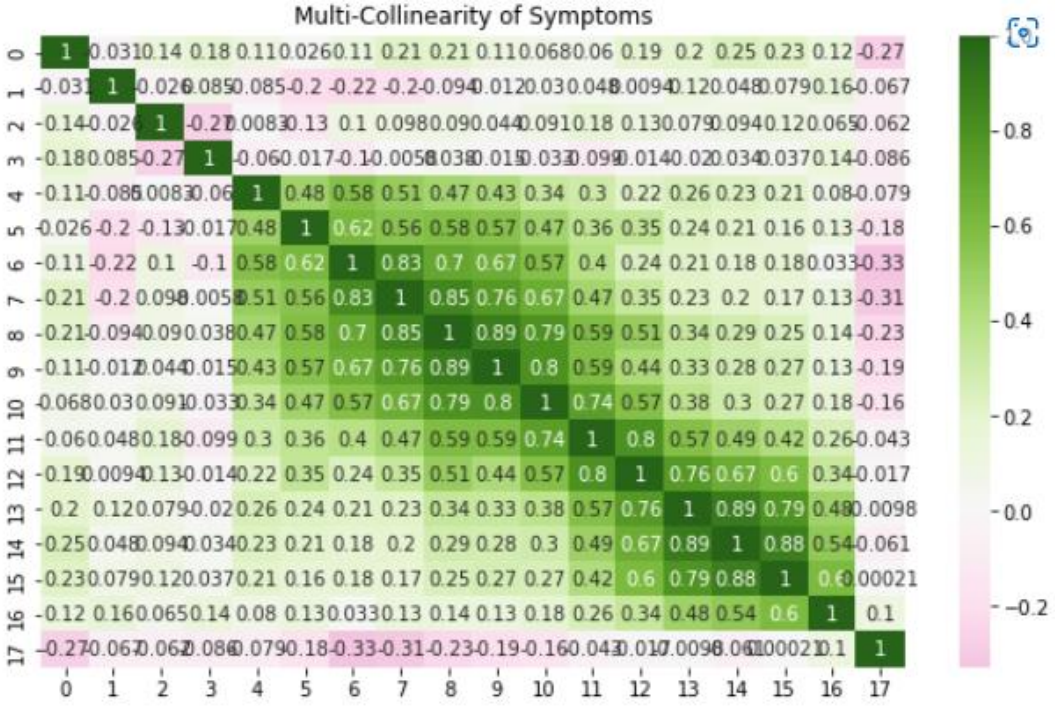

Figure 2: Spearman's Correlation for Multi Symptoms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

The high degree correlation symptoms are interpreted from figure 1 and table 1 are presented in table 3.

Table 3: Highly Correlated Symptoms

S.No	Symptom Number	Symptom Number	Coefficient Value
1	6	7	0.83
2	7	8	0.85
3	7	9	0.76
4	8	9	0.89
5	8	10	0.79
6	9	10	0.8
7	10	11	0.74
8	11	12	0.84
9	12	13	0.76
10	13	14	0.89
11	13	15	0.79
12	14	15	0.88

From the table 3, it is evident that 10 symptoms have high degree of correlation. The model concluded that remaining 7 symptoms are not correlated, so these are considered as essential symptoms but unfortunately the accuracy obtained with these symptoms is "67.13%". Another drawback associated with this model is it has picked symptoms which have more number of missing values [16,17]. The second traditional approach the model considered is "KBestSelection", which has obtained the essential features as shown in figure 3

	Specs	Score
6	Symptom_7	676.985874
7	Symptom_8	521.313740
0	Symptom_1	361.920958
8	Symptom_9	254.084605
5	Symptom_6	105.924513
9	Symptom_10	85.422658
10	Symptom_11	85.218838
16	Symptom_17	52.730522
14	Symptom_15	51.656757

Figure 3: Essential Features using K-Best

In general any Machine Learning algorithm assumes minimum half of the features are essential for the prediction of class label. So, this paper, the model wants to identify top 9 Best features using selection approach. In this paper, the model has chosen chi2 as a metric to rank the features [18]. It assumes two hypothesis for every features in which one hypothesis assumes that a particular feature is important for predicting the disease and another hypothesis assumes that a particular feature is not important for prediction. The reason for choosing this metric is both the features and class labels exists in categorical form. The mathematical notation for the chi-square is illustrated in equation (2)

$$Chi\_Square(Feature) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(Actual_{ij}-predicted_{ij})^2}{predicted_{ij}} \quad (2)$$

Where,

n denotes number of features

m denotes number of records

$Actual_{ij}$ denotes the original observations in the dataset

$Precited_{ij}$ denotes the expected observations from the inferences drawn

The accuracy with the best features is only "72.93%", which is better than the correlation but still the accuracy has to be improved. The proposed algorithm combines the wrapper and filter methods by using a pipeline vector. The pipeline vectors need some classifiers to eliminate the features[19]. The proposed model has chosen the below traditional algorithms as base classifiers:

1.1. Logistic Regression: The dataset contains multiple classes as labels, in general, any dataset that contains more number of discrete ordinal data then it is better to apply LR approach because it internally uses Softmax activation function so it distributes the features in such a way that sum of the probability distribution should be equal to 1[28]. It also minimizes the error rate by computing cross entropy instead of regular absolute and square errors. During the process of error minimization it maximizes the probability of likelihood between the two classes distribution[20]. The probability distribution for every class can be computed in two ways as shown in figure 4.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

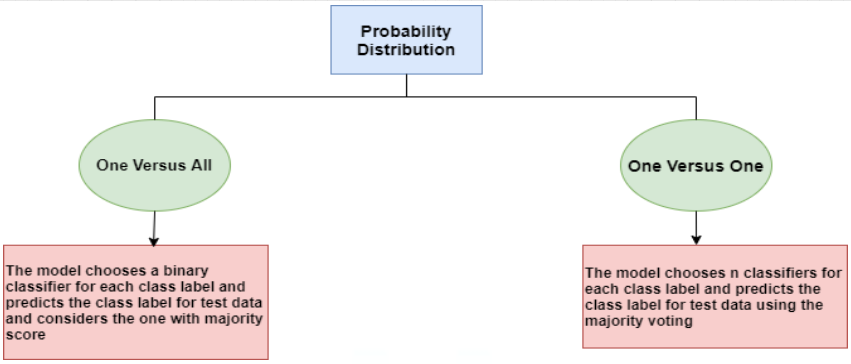

Figure 4: Categorization of Probability Distribution in LR Model

1.2. perceptron: The proposed system observed that there is some synaptic connection between the symptoms. The architecture of the neural networks is represented in figure 5. In the figure 5, the X- Input vector represents the symptoms of diseases and random weights are initialized to each feature [21]. This architecture computes the summation of dot product

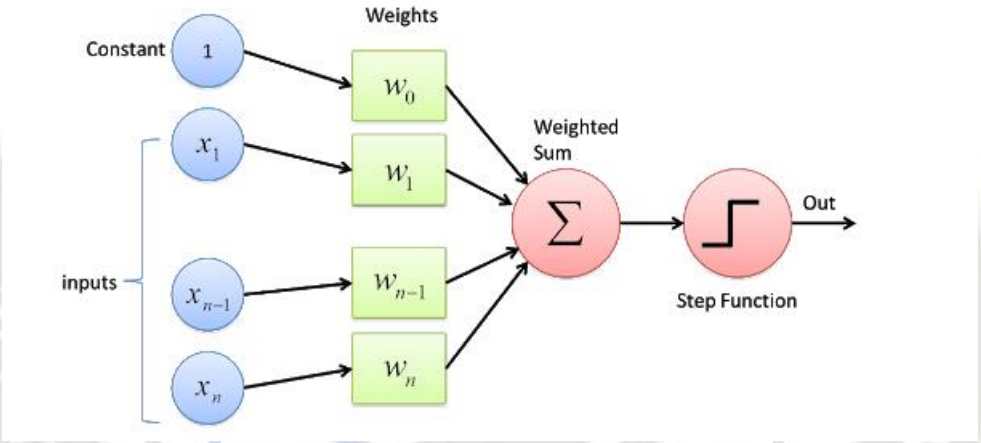

Figure 5: Architecture of perceptron

and enhances the property using the "sigmoid" activation function. The proposed system to predict the disease based on the symptoms, initially focused on selection minimum number of symptoms so that diagnosis to the patients can be found earlier. With the increase in the number of symptoms, doctors need more tests to confirm their presence because of their commonness in low and high pandemic diseases. The scenario can be explained as follows where "Cough" is a common symptom to common fever, Tuberculosis, Pneumonia, and COVID-19. Based on cough alone, doctor cannot claim that patient is suffering from "Covid", so a combination of symptoms has to be analyzed. The doctor suggests Chest X-ray, which takes time to get the report [23]. The selection of combinations plays a vital role in developing an intelligent autonomous diagnosis system. The dataset contains nearly 10,000 symptoms suppose a model using the original dataset need to predict class label need 17**1000 combinations, which is highly impossible [22]. So, the proposed system mainly focused on reducing the combination of symptoms which are efficient for prediction. The entire process of selection is illustrated in figure 6.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

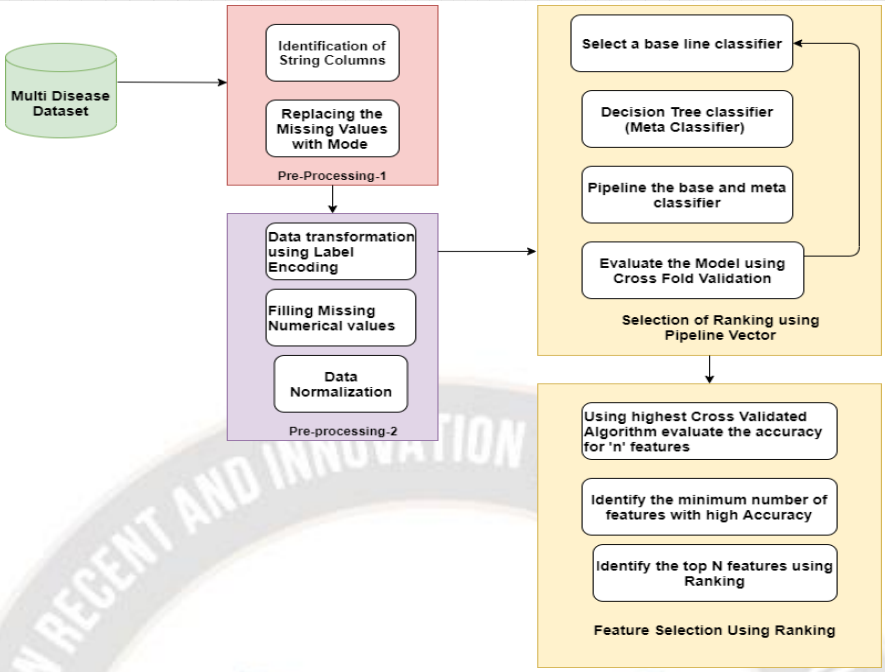

Figure 6: Feature Selection Process using Pipelined Vector of Random Forest

3.1. Data Pre-processing: In this proposed system, data pre-processing is a two folded approach because all the features in the dataset are categorical in nature. In the first fold, ML algorithms need numerical data to work and using SMOTE analysis, it is recognized that most of the symptoms are empty and the dataset is unbalanced. So, the model has applied central tendency nature "mode", to replace the missing fields of categorical data [25]. In the second fold, the model performed label encoding to transform the data from categorical to numerical as shown in figure 7. The figure shows sample column transformation with few records.

	Symptom_1	Symptom_2
0	itching	skin_rash
1	skin_rash	nodal_skin_eruptions
2	itching	nodal_skin_eruptions
3	itching	skin_rash
4	itching	skin_rash
...	...	...
4915	vomiting	headache
4916	skin_rash	pus_filled_pimples
4917	burning_micturition	bladder_discomfort
4918	skin_rash	joint_pain
4919	skin_rash	high_fever

Label Encoding

	Symptom_1	Symptom_2
0	33	35
1	24	27
2	33	27
3	33	35
4	33	35
...	...	...
4915	28	17
4916	24	31
4917	4	4
4918	24	20
4919	24	18

Figure 7: Label Encoding Process on Categorical Features

In this process, the model sorts all the symptoms in alphabetical order and start assigning unique values from 0 to n-1 with respect to that specific column [26]. The numerical data varies with different ranges so it needs to be standardized by reducing the distance between them. Sine different symptoms different notations to measure, instead of normalization then model chooses standardization where the distribution of data takes place in such way that mean of the entire column becomes 0 and standard deviation becomes 1. The result obtained after standardization is presented in figure 8.

```
[[ 1.46868498 -0.36516744  1.7907384  ...  0.18910916  0.1426824
   0.        ]
 [ 1.46868498  1.07638417 -0.483793   ...  0.18910916  0.1426824
   0.        ]
 [-0.76643989  1.29261691 -1.12350496 ...  0.18910916  0.1426824
   0.        ]
 ...
 [-1.07122964  0.42768595  1.57750108 ...  0.18910916  0.1426824
   0.        ]
 [ 0.96070206  1.43677207 -0.27055568 ...  0.18910916  0.1426824
   0.        ]
 [-1.07122964 -1.30217598 -0.483793   ... -5.69495136 -8.05582766
   0.        ]]
```

Figure 8: Standardization of Symptoms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

3.2. Pipeline Vector with Baseline and Metaline Classifiers: The traditional Recursive Feature Elimination (RFE) mechanism, which is a wrapper method is utilized for classification and regression process [27]. But in this process, the hyper tuning process of RFE helps in determining the number of essential features required because it computes different combinations possible. In traditional approaches, feature elimination is done using the decision tree algorithm which computes the entropy as purning metric. But due to stochastic nature of data, other ensemble algorithms may have good purning nature. So in this proposed system 4 Machine Learning algorithms and 1 DL algorithm are passed as baseline classifiers but the second level i.e., meta classifier is decision tree only. Every intelligent wrapping should be evaluated by considering the pipeline vector as input. In this approach, the model has applied cross validation using RepeatedStratifiedKFold because the model gets trained with all possible combinations of records [28]. The Hyper-tuned Wrapped REF pipeline vector algorithm is discussed in below section:

Algorithm for HWREF pipeline vector:

Input: Load the symptoms and disease dataset, SymData

Output: Accuracy of each wrapped model

Begin:

1. Define a dictionary with a set of Machine Learning algorithms, dict_alg

2. Initialize num_features←5

3. for i in dict_alg:

a. rfe[i]←RFE(estimator=dict_alg[i], num_features)

b. meta_model←DecisionTreeClassifier()

c. rfe_model[i]←pipeline(rfe[i],meta_model)

4. for j in rfe_model:

a. rfe_cv←RepeatedStratifiedKFold(n_splits=10)

b.rfe_score[j]←cross_val_score(rfe_model[j], X_train,y_train,scoring='accuracy')

c. print rfe_score[j]

The model has selected popular traditional approaches from each sector like logistic regression, CART trees, Random Forest from bagging, Gradient from boosting, perceptron from neural network. Out of these, algorithms, ensemble algorithms have achieved equal and highest accuracies. Since, the dataset contains nearly 5000 records, working with boosting algorithm with boosting combinations has taken lot of time to compute. So, the proposed model has chosen Random Forest (RF), ensemble algorithm for further processing.

3.3. Ranking the Features using RF Algorithm:

The major focus of the proposed system to reduce the number of symptoms, so it utilized the RF algorithm to compute the accuracy for "N" number of features. The original dataset contains first 5 symptoms without any missing values but remaining features as few missing values it doesn't mean that remaining symptoms are not important. The output for each evaluation is presented in table 4.

Table 4: Accuracies of little iteration

Number of Symptoms	Accuracy
1	62.4
2	90.8
3	97
4	99.6
5	99.8
6	99.9
7	99.9
8	100

The above observations are clearing stating with the increase in the number of features, the accuracy level increased. But with the number of symptoms as "8", the model has achieved 100% accuracy. So the minimum number of symptoms required to design a diagnosis system is obtained but out of the 17 symptoms, the crucial 8 symptoms are identified using the ranking mechanism. For every feature support is computed and if it is more than threshold value then it is marked as True otherwise is marked as False and other 1 some rank is assigned to the feature.

IV. EXPERIMENTAL RESULTS

a. Dataset Description: Table 5 represents the description about diseases in the dataset. The diseases mentioned here range from low level common diseases to high level diseases including genetically disorders. In the dataset, set of attributes are composed in a certain way to class label known as "diseases".

Table 5: List of Diseases in Dataset

S.No	Disease	S.No	Disease	S.No	Disease
1	Acne	15	Fungal infection	29	Jaundice
2	AIDS	16	Gastroenteritis	30	Malaria
3	Alcoholic hepatitis	17	GERD	31	Migraine
4	Allergy	18	Heart attack	32	Osteoarthristis
5	Arthritis	19	hepatitis A	33	Paralysis

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

					(brain hemorrhage)
6	Bronchial Asthma	20	Hepatitis B	34	Paroymsal Positional Vertigo
7	Cervical spondylosis	21	Hepatitis C	35	Peptic ulcer diseae
8	Chicken pox	22	Hepatitis D	36	Pneumonia
9	Chronic cholestasis	23	Hepatitis E	37	Psoriasis
10	Common Cold	24	Hypertension	38	Tuberculosi

					s
11	Dengue	25	Hyperthyroidism	39	Typhoid
12	Diabetes	26	Hypoglycemia	40	Urinary tract infection
13	Dimorphic hemorrhoids(piles)	27	Hypothyroidism	41	Varicose veins
14	Drug Reaction	28	Impetigo		

Table 6 represents the list of fewpopular symptoms collected for every disease to annotate the records with class labels.

Table 6: List of symptoms in the dataset for defining the class label

S. No	Symptom	S.No	Symptom	S.No	Symptom
1	Abdominal_Pain	21	Dark_Urine	41	Muscle_Weakness
2	Acidity	22	Dehydration	42	Nausea
3	Altered_Sensorium	23	Diarrhea	43	Neck_Pain
4	Anxiety	24	Dischromic Patches	44	Nodal_Skin_Eruptions
5	Back_Pain	25	Dizziness	45	Obesity
6	Blackheads	26	Extra_Marital_Contacts	46	Pain_During_Bowel_Movements
7	Bladder_Discomfort	27	Fatigue	47	Pain_In_Anal_Region
8	Blister	28	Foul_Smell_Of Urine	48	Patches_In_Throat
9	Bloody_Stool	29	Headache	49	Patches_In_Throat
10	Blurred_And_Distorted_Vision	30	High_Fever	50	Pus_Filled_Pimples
11	Breathlessness	31	Hip_Joint_Pain	51	Red_Sore_Around_Nose
12	Bruising	32	Indigestion	52	Restlessness
13	Burning_Micturition	33	Joint_Pain	53	Scurring
14	Chest_Pain	34	Knee_Pain	54	Shivering
15	Chills	35	Lethargy	55	Silver_Like_Dusting
16	Cold_Hands_And_Feets	36	Loss_Of_Appetite	56	Skin_Peeling
17	Constipation	37	Loss_Of_Balance	57	Skin_Rash
18	Continuous_Feel_Of_Urin	38	Mood_Swings	58	Spinning_Movements
19	Continuous_Sneezing	39	Sunken_Eyes	59	Stiff_Neck
20	Cramps	40	Movement_Stiffnes	60	Stomach_Pain

b. Results Obtained:

Figure 9 shows the results obtained by the different traditional and ensemble ML algorithms by combining the DT as meta classifier to the elimination. Out of these different algorithms, bagging and boosting has got equal and highest accuracy. Among the executed algorithms, the logistic regression algorithm (lr) has achieved second accuracy. The perceptron (per) doesn't suit for this dataset because the model has got less accuracy, which means it suffers from underfit problem.

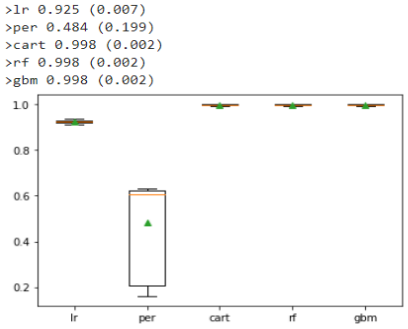

Figure 9: Selection of Algorithm for Feature Selection

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

Figure 10 shows the number of symptoms minimum needed to get good accuracy out of 17 symptoms using the bagging algorithm know as "Random Forest". In every iteration, the model prints the accuracy for number of features=1 to number of features= 17, but from 8 number of features, the accuracy has reached 100%, since it is minimum number, the model found that it 8 features are enough to design a good system.

```
>1 0.624 (0.029)
>2 0.908 (0.030)
>3 0.970 (0.014)
>4 0.996 (0.003)
>5 0.998 (0.002)
>6 0.999 (0.002)
>7 0.999 (0.002)
>8 1.000 (0.000)
>9 1.000 (0.000)
>10 1.000 (0.000)
>11 1.000 (0.000)
>12 1.000 (0.000)
>13 1.000 (0.000)
>14 1.000 (0.000)
>15 1.000 (0.000)
>16 1.000 (0.000)
```

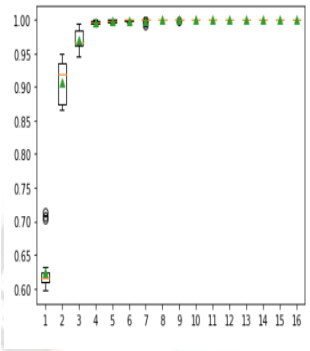

Figure 10: Numerical & graphical representation for selecting 'n' features

Figure 11 represents the 8 selected features which are treated as important by the model to predict a disease by assigning the rank to each feature using the Random Forest algorithm. The columns which are marked as True with rank 1 are considered as the essential symptoms. When compared with the other traditional approaches the proposed algorithm has got the optimal symptoms to decide the nature of disease.

```
Column: 0, Selected True, Rank: 1.000
Column: 1, Selected True, Rank: 1.000
Column: 2, Selected True, Rank: 1.000
Column: 3, Selected True, Rank: 1.000
Column: 4, Selected True, Rank: 1.000
Column: 5, Selected True, Rank: 1.000
Column: 6, Selected False, Rank: 3.000
Column: 7, Selected False, Rank: 2.000
Column: 8, Selected True, Rank: 1.000
Column: 9, Selected False, Rank: 4.000
Column: 10, Selected True, Rank: 1.000
Column: 11, Selected False, Rank: 6.000
Column: 12, Selected False, Rank: 5.000
Column: 13, Selected False, Rank: 9.000
Column: 14, Selected False, Rank: 8.000
Column: 15, Selected False, Rank: 7.000
Column: 16, Selected False, Rank: 10.000
```

Figure 11: Feature Ranking using RF Algorithm

V. CONCLUSION

Intelligent diagnosis system to predict and that can recommend precautions is needed to save a life. A multi diagnosis system helps the persons to get the suggestions at an early stage using a single application. The model has identified 8 essential symptoms using Random Forest that can predict the disease that a person is suffering from. Using the traditional "KBestSelect", the model has selected 9 symptoms with 72.93% accuracy. The random forest algorithm starts with all the all symptoms and starts constructing tree based on information gain, gini index and computes impurity of each tree, the tree with less impurity is identified and its root node is considered, if its average as root node cross over more than half of the computations then it is further considered in the list. The support () method computes the average score of every tree that has same root node if the normalize sum of these trees is 1 then the model assigns Boolean value "True" for that root node as a mark of important feature. In the future work, the selected pipeline vector can be hyper tuned by optimizing the important estimators with genetic algorithms.

REFERENCES

[1] Harimoorthy, K., Thangavelu, M. RETRACTED ARTICLE: Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. J Ambient Intell Human Comput 12, 3715–3723 (2021). https://doi.org/10.1007/s12652-019-01652-0

[2] Dubey, A.K. Optimized hybrid learning for multi disease prediction enabled by lion with butterfly optimization algorithm. Sādhanā 46, 63 (2021). https://doi.org/10.1007/s12046-021-01574-8

[3] Draelos, R. L., Dov, D., Mazurowski, M. A., Lo, J. Y., Henao, R., Rubin, G. D., & Carin, L. (2020). Machine-Learning-Based Multiple Abnormality Prediction with Large-Scale Chest Computed Tomography Volumes. Medical Image Analysis, 101857. doi:10.1016/j.media.2020.101857

[4] Pangti, R., Mathur, J., Chouhan, V., Kumar, S., Rajput, L., Shah, S., … Gupta, S. (2020). A machine learning-based, decision support, mobile phone application for diagnosis of common dermatological diseases. Journal of the European Academy of Dermatology and Venereology. doi:10.1111/jdv.16967

[5] Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit. Med. 4, 3 (2021). https://doi.org/10.1038/s41746-020-00372-6

[6] Pangti, R., Mathur, J., Chouhan, V., Kumar, S., Rajput, L., Shah, S., Gupta, A., Dixit, A., Dholakia, D., Gupta, S., Gupta, S., George, M., Sharma, V. K., & Gupta, S. (2020). A machine learning-based, decision support, mobile phone application for diagnosis of common dermatological diseases. In Journal of the European Academy of Dermatology and Venereology (Vol. 35, Issue 2, pp. 536–545). Wiley. https://doi.org/10.1111/jdv.16967

[7] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[8] Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K.-S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. In Information Fusion (Vol. 63, pp. 208–222). Elsevier BV. https://doi.org/10.1016/j.inffus.2020.06.008

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

[9] Chen, J., Dai, X., Yuan, Q., Lu, C., & Huang, H. (2020). Towards Interpretable Clinical Diagnosis with Bayesian Network Ensembles Stacked on Entity-Aware CNNs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.286

[10] Akram Baig, M. M. . (2023). An Evaluation of Major Fault Tolerance Techniques Used on High Performance Computing (HPC) Applications. International Journal of Intelligent Systems and Applications in Engineering, 11(3s), 320–328. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2696

[11] Muthu, B., Sivaparthipan, C.B., Manogaran, G. et al. IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. Peer-to-Peer Netw. Appl. 13, 2123–2134 (2020). https://doi.org/10.1007/s12083-019-00823-2

[12] Mishra S, Tripathy HK, Mallick PK, Bhoi AK, Barsocchi P. EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis. Sensors. 2020; 20(14):4036. https://doi.org/10.3390/s20144036

[13] Eunhee Chang, Hyun Taek Kim & Byounghyun Yoo (2020) Virtual Reality Sickness: A Review of Causes and Measurements, International Journal of Human–Computer Interaction, 36:17, 1658-1682, DOI: 10.1080/10447318.2020.1778351

[14] Koo, M. M., Swann, R., McPhail, S., Abel, G. A., Elliss-Brookes, L., Rubin, G. P., & Lyratzopoulos, G. (2020). Presenting symptoms of cancer and stage at diagnosis: evidence from a cross-sectional, population-based study. In The Lancet Oncology (Vol. 21, Issue 1, pp. 73–79). Elsevier BV. https://doi.org/10.1016/s1470-2045(19)30595-9

[15] Zeeshan Ahmed, Khalid Mohamed, Saman Zeeshan, XinQi Dong, Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine, Database, Volume 2020, 2020, baaa010, https://doi.org/10.1093/database/baaa010

[16] Halu, A., De Domenico, M., Arenas, A. et al. The multiplex network of human diseases. npj Syst Biol Appl 5, 15 (2019). https://doi.org/10.1038/s41540-019-0092-5

[17] Halu, A., De Domenico, M., Arenas, A. et al. The multiplex network of human diseases. npj Syst Biol Appl 5, 15 (2019). https://doi.org/10.1038/s41540-019-0092-5

[18] Deepthi, Y., Kalyan, K.P., Vyas, M., Radhika, K., Babu, D.K., Krishna Rao, N.V. (2020). Disease Prediction Based on Symptoms Using Machine Learning. In: Sikander, A., Acharjee, D., Chanda, C., Mondal, P., Verma, P. (eds) Energy Systems, Drives and Automations. Lecture Notes in Electrical Engineering, vol 664. Springer, Singapore. https://doi.org/10.1007/978-981-15-5089-8_55

[19] V. Muthu Ganesh, Janakiraman Nithiyanantham. (2022) Heuristic-based channel selection with enhanced deep learning for heart disease prediction under WBAN. Computer Methods in Biomechanics and Biomedical Engineering 0:0, pages 1-20.

[20] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2021). Multiple disease prediction using Machine learning algorithms. In Materials Today: Proceedings. Elsevier BV. https://doi.org/10.1016/j.matpr.2021.07.361

[21] Z. Sun, H. Yin, H. Chen, T. Chen, L. Cui and F. Yang, "Disease Prediction via Graph Neural Networks," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 3, pp. 818-826, March 2021, doi: 10.1109/JBHI.2020.3004143.

[22] Bhanuteja, T., Kumar, K. V. N., Poornachand, K. S., Ashish, C., & Anudeep, P. (2021). Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach. In International Journal of Innovative Technology and Exploring Engineering (Vol. 10, Issue 9, pp. 67–72). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijitee.i9364.0710921

[23] Prof. Barry Wiling. (2018). Identification of Mouth Cancer laceration Using Machine Learning Approach. International Journal of New Practices in Management and Engineering, 7(03), 01 - 07. https://doi.org/10.17762/ijnpme.v7i03.66

[24] Men, L., Ilk, N., Tang, X., & Liu, Y. (2021). Multi-disease prediction using LSTM recurrent neural networks. In Expert Systems with Applications (Vol. 177, p. 114905). Elsevier BV. https://doi.org/10.1016/j.eswa.2021.114905

[25] Swarajya Lakshmi V Papineni, Snigdha Yarlagadda, Harita Akkineni, A. Mallikarjuna Reddy. Big Data Analytics Applying the Fusion Approach of Multicriteria Decision Making with Deep Learning Algorithms International Journal of Engineering Trends and Technology, 69(1), 24-28, doi: 10.14445/22315381/IJETT-V69I1P204.

[26] Mallikarjuna Reddy, A., Rupa Kinnera, G., Chandrasekhara Reddy, T., Vishnu Murthy, G., et al., (2019), “Generating cancelable fingerprint template using triangular structures”, Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, pp. 1951-1955(5), doi: https://doi.org/ 10.11 66 5/jctn.2019.7830.[22] P. S. Silpa et al., "Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), 2022, pp. 759-767, doi: 10.1109/ICOSEC54921.2022.9951883

[27] Wanjiku , M., Levi, S., Silva, C., Ji-hoon, P., & Yamamoto, T. Exploring Feature Selection Methods in Support Vector Machines. Kuwait Journal of Machine Learning, 1(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/131

[28] A. Mallikarjuna Reddy, K. S. Reddy, M. Jayaram, N. Venkata Maha Lakshmi, Rajanikanth Aluvalu, T. R. Mahesh, V. Vinoth Kumar, D. Stalin Alex, "An Efficient Multilevel Thresholding Scheme for Heart Image Segmentation Using a Hybrid Generalized Adversarial Network", Journal of Sensors, vol. 2022, Article ID 4093658, 11 pages, 2022. https://doi.org/10.1155/2022/4093658.9

[29] Uthayopas, K., de Sá, A. G. C., Alavi, A., Pires, D. E. V., & Ascher, D. B. (2021). TSMDA: Target and symptom-based computational model for miRNA-disease-association

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9s
DOI: https://doi.org/10.17762/ijritcc.v11i9s.7470
***Article Received**: 24 May 2023 **Revised:** 20 July 2023 **Accepted:** 02 August 2023*

prediction. In Molecular Therapy - Nucleic Acids (Vol. 26, pp. 536–546). Elsevier BV. https://doi.org/10.1016/j.omtn.2021.08.016

[30] Faris, H., Habib, M., Faris, M., Elayan, H., & Alomari, A. (2021). An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine. In Informatics in Medicine Unlocked (Vol. 23, p. 100513). Elsevier BV. https://doi.org/10.1016/j.imu.2021.100513

[31] Tseng, V.WS., Sano, A., Ben-Zeev, D. et al. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. Sci Rep 10, 15100 (2020). https://doi.org/10.1038/s41598-020-71689-1

[32] KUMAR, P., PRADEEPINI, G. and KAMAKSHI, P., 2019. Feature selection effects on gradient descent logistic regression for medical dataclassification. International Journal of Intelligent Engineering and Systems, 12(5), pp. 278-286.

[33] BANGARE, S.L., PRADEEPINI, G. and PATIL, S.T., 2018. Regenerative pixel mode and tumour locus algorithm development for brain tumour analysis: A new computational technique for precise medical imaging. International Journal of Biomedical Engineering and Technology, 27(1-2), pp. 76-85.

[34] Sri Silpa Padmanabhuni and Pradeepini Gera, "Synthetic Data Augmentation of Tomato Plant Leaf using Meta Intelligent Generative Adversarial Network: Milgan" International Journal of Advanced Computer Science and Applications(IJACSA), 13(6), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130628

[35] Zhou Y, He Y, Yang H, Yu H, Wang T, Chen Z, et al. (2020) Development and validation a nomogram for predicting the risk of severe COVID-19: A multi-center study in Sichuan, China. PLoS ONE 15(5): e0233328. https://doi.org/10.1371/journal.pone.0233328

[36] Abdar, M., Książek, W., Acharya, U. R., Tan, R.-S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. In Computer Methods and Programs in Biomedicine (Vol. 179, p. 104992). Elsevier BV. https://doi.org/10.1016/j.cmpb.2019.104992.

[37] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, "Leveraging bio-maximum inverse rank method for iris and palm recognition", International Journal of Biometrics, 2022 Vol.14 No.3/4, pp.421 - 438, DOI: 10.1504/IJBM.2022.10048978.

[38] Sudeepthi Govathoti, A Mallikarjuna Reddy, Deepthi Kamidi, G BalaKrishna, Sri Silpa Padmanabhuni and Pradeepini Gera, "Data Augmentation Techniques on Chilly Plants to Classify Healthy and Bacterial Blight Disease Leaves" International Journal of Advanced Computer Science and Applications(IJACSA), 13(6), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130618.

[39] V. NavyaSree, Y. Surarchitha, A. M. Reddy, B. Devi Sree, A. Anuhya and H. Jabeen, "Predicting the Risk Factor of Kidney Disease using Meta Classifiers," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972392.

[40] A. Mallikarjuna Reddy, V. Venkata Krishna, L. Sumalatha," Face recognition based on stable uniform patterns" International Journal of Engineering & Technology, Vol.7 ,No.(2),pp.626-634, 2018,doi: 10.14419/ijet.v7i2.9922 .