Journal of Theoretical and Applied Information Technology
15th November 2022. Vol.100. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# PERFORMANCE ANALYSIS OF DECISION TREE EXTENDED CLASSIFIERS (DTEC) ON CLINICAL DATASET OF COVID PREDICTION USING OPTIMAL FEATURE SELECTION

**DR. P.S.S. AKILASHRI[1], G NITHYA[1]**

[1]Head of the Department, Department of Computer Science, National College
[1]Research Scholar, Department of Computer Science, National College
Affiliated to Bharathidasan University, Tiruchirappalli, India

E-mail: [1]akilas27@gmail.com , [1] nithya.gna@gmail.com

**ABSTRACT**

In Computer Science Technology, Artificial Intelligence, Data Analytics and Machine Learning plays a predominant role in decision-making strategies. By choosing suitable method and algorithms from any of these fields, a good decision can be taken. In health care industry, at present, prediction of Covid 19, is still a very big challenge, since false positive, false negative metrics are occurring frequently through various covid test. Our main aim, in this proposed work is to increase the prediction of accuracy of Covid, by considering the optimum number of features alone, by taking into consideration the laboratory measurement, and also by clinical test understanding. When target variable is binary, the classification algorithm based on Tree Based Extended Classifiers like Random Forest, AdaBoost, XGBoost can be proposed with necessary features. The results are observed from the proposed algorithms, that gets trained using the training dataset using standard data repository and it is being tested with the testing dataset. By analyzing the performance metrics, the obtained results showed that the prediction accuracy is increased and also false positive and false negative are reduced. In the proposed work, the tree based extended classifiers of Random Forest and Extended Gradient Boosting produces maximum 92% accuracy with 11 features using Gini Index. Apart from accuracy, the metrics such as false positive and false negative are playing the important role. In this proposed work, the false negative is as low as 5 out of 14 by XGB, and false positive with the minimum value of 3 out of 106 using Random Forest. Thus, these methods of covid predictions are useful for health care community, if it is being utilized in an efficient manner.

**Keywords**: *Feature Selection, Binary Classification, Tree Based Classifier, Covid, Clinical data, Confusion Matrix*.

## 1. INTRODUCTION

Data Analytics is meant for predicting the future with the available historical data. In predictive analytics, data is playing the predominant role. To perform data analytic research, there are various domains available like Software, IT services, Healthcare, Corporate, Finance etc. The domain which affects both social and economical way of human being is Covid, (for the past 3 years) from December 2019. The domain which affects more or less, that everyday life has to be considered as the highest need of the hour in research. In the Month of July 2021, WHO [31] created a website link,) to collect clinical strategy of Covid 19. It's a proof that with the help of clinical characteristics, various analyses can be done and also the reports outcome can help the clinician to take good decision. Apart from SARS-CoV, there was a similar kind of Coronavirus called MERS-CoV [1] was wild spread across the World in the year 2012-2017 especially in the areas of Arabian, African, American and Asian Countries. It has given more ideas about this SARS-CoV like, how to handle, what are the precautionary steps to

undergo and what are the decisions that are to be taken between the countries during travel etc.

This [32] source gives more information about covid and its impact on various countries also in the form of chart, map etc. From this description and study, there was a motivation occurred to perform predictive analytics in covid 19. In search of data of Covid, it was identified that, in [2],[3], covid prediction based on clinical data can be considered. It has blood hematology report which contains various features like Lymphocytes, Neutrophils, Leukocytes, Protein c reactive (very important field) etc. It is not necessary to use all the features for prediction, meanwhile Optimum Feature Selection can be done with the available methods, and very limited features are made ready for consideration. Then predictive models were applied in a covid balanced dataset with those selected features and the laboratory findings were used to prediction the disease more precisely. In this, not only accuracy, but also false positive and false negative [9] values are being taken for further deep analysis. It is compulsorily needed to reduce these values by optimizing. Thus, the main contribution on prediction of covid with clinical dataset using predictive models has to be analyzed in order to improve the accuracy on SARS-CoV dataset. Many studies were based on RTPCR/ CT images etc. But the treatment based on clinical dataset [2] is essential for more accuracy. After data preparation, Pearson correlation method is used to understand about the dataset, by calculating association between the columns to predict the label and analyzed by using heat map. From this, it was identified that the data type was nonlinear and the feature selection was applied on this to select the optimum features among the variables by utilizing the ranking methods (Gini, Kendall, Entropy), so that redundant and irrelevant data were removed from the list. In binary classification, decision tree classifiers already played the important role in best prediction of covid. So, the decision tree extended classifiers like Random Forest, Adaboost and Extended gradient boosting thus selected, applied and produced the highest accuracy and also other performance metrics have satisfied the expectation of this proposed work. This proposed work is explicitly organized as follows: Section 2: related works, Section 3: methodology, Section 4: feature selection using statistical methods, Section 5: tree-based models and its performance metrics applying extended tree based binary classification models for predictions, Section 6: observation and analysis of performance metrics of models especially with accuracy, false positive, and false negative. Finally, conclusion, and future enhancements.

## 2. RELATED WORKS

Before going further deep into analysis, it is compulsorily needed to know how the research has been carried out in this domain and also in the other related healthcare domain. Several researches have been done in data analytics with detailed medical clinical datasets and also with same kind of datasets. Here the review was done in two different ways. One, on the basis of covid domain how research was carried out and also analyzed the classification algorithms that were performed on covid datasets. Next, on how researches in other medical domains like liver, heart, breast diseases [5], [6] etc. are handled, how the research was carried out and the performance metrics were used for analysis.

### 2.1 COVID DOMAIN RELATED WORKS

Under feature selection concept of Covid, Vishankumar Gupta [7] has detected covid 19 cases using some of the features related to common features with confirmed death and cured cases in different States of India. Among 5 features, only 3 features were selected after performing feature selection using Random Forest. For prediction, 5 different models (Decision Tree, Random Forest, Multinomial Logistic Regression, Support Vector Machine, Neural Network) were applied on this dataset, but proved that Random Forest performed the best accuracy of 83.54% in confirmed cases. In this analysis, 2342 cases were taken, split up was held with 70-30 for training and testing as well. The best paper that used was deep learning by Talha Burak alakus et al [2] in order to ease the health care community by improving the prediction of covid using laboratory clinical data. Various performance metrics were considered for valuation. The accuracy of prediction is increased to 86.66%. Totally 600 cases were considered with 19 laboratory findings as input variable. This dataset is stored in GitHub and available for researchers. Apart from prediction of covid, Juan Dominguez-olmedo [4] aimed to develop a model to predict the severity of covid infection and mortality. Among 32 blood related features, 4 features have considered as more relevant. Among classification model, XGB was considered as a best model for prediction and also

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

other performance metrics were also increased. Mostly clinical data are used to predict the risk factors and Brain W Allwood et al [8] were dealt with second wave of covid 19 named as Beta Variant. The clinical dataset with biochemical parameters used to identify the high-risk factors related to survival or death. Finally, D-Dimer and HCO3 standard level were considered as highly warranted in clinical approach and identified the high risk of patients and improve clinical decision making.

In literature, when search was held, it was surprised that there was a disease similar to SARS-COV, and the research was taken as in MERS-COV [1] disease prediction that the researchers applied the algorithms for both binary and multi class classification. And the performance metrics were noted. The outcome of this work made by the classification algorithms SVM, DTree and KNN were noted. Among all these 3, SVM and DTree produced the best accuracy of 86.4 %.

*Table 1: Covid Recent Related Works*

| Year of publication and Ref number | Domain with Dataset | Objective of the Proposed work | Data analytics techniques | Software/tool used | Outcome |
|---|---|---|---|---|---|
| 2022, April [24] | Basic features related to patient for covid 19 , and the features were not related to clinical. | Early detection of covid 19 | Basic traditional methods and Hawks optimization method was applied | NA | After performing Hyper parameter tuning by HHO, XGB produced the better accuracy 92.3% when compared with other methods |
| 2020, April [25] | Blood hematochemical test result with about 280 records | Detection of Covid 19 using Machine Learning | Decision Tree, Logistic regression, K Nearest Neighbour | NA | With two different features (entire dataset and a dataset without Gender feature) and produced the accuracy of 82%. |
| June 2022 [26] | Radiographs images and clinical variables | Work was trying to identify any need of ICU for the patient | Deep learning with (CNN) and Random Forest analysis method | SciPy 1.7.1 in Python 3.8 and R Studio version 1.3.1056. | Image based analysis produced accuracy as 75% (approx.), and 81% in both clinical and combination of image and clinical |
| 2022 [27] | Data collected related to Procalcitonin variable were collected in the hospital | To identify how the procalcitonin is playing its role in covid 19 for its severity. | Statistical analysis made | SPSS | The result show that the procalcitonin level cannot distinguish the presence of COVID-19 pneumonia in RT-PCR positive patients. High procalcitonin levels may predict that the disease becomes more complicated |
| 2022 July [28] | Covid 19 patients personal health details and covid details | Identifying the important feature that influencing the prediction of Covid 19 | Logistic model tree, Decision Tree, C4.5,C5.0 decision tree, | NA | Among all the classification models, CART produce highest accuracy as 78.24 when compared with all other and also important |

| | | | | | |
|---|---|---|---|---|---|
| | | | CART models were applied | | features were selected from the list of features |
| Mar 2022 [29] | Covid 19 dataset | Severity of the patient is to be classified | Adaboost algorithm for ensemble method | Sklearn with Python libraries | Adaboost algorithm performed well with prediction rate in short with highest improved accuracy. |
| 2022 July [30] | Clinical dataset | Predicting severity and mortality of covid 19 | Other than traditional techniques, XGB was applied | NA | XGB produce highest accuracy of 92 % when compare with other traditional ML methods. And identified important features for mortality prediction. |

Authors are generally preferring performance metrics like accuracy, precision, recall and F1 score which are tabulated in Table 1. Recently, there is a huge hike in research on prediction of covid 19 which are not only states whether the patient is infected or not, and also it deals with the patient severity and mortality. But prediction and analysis in medical domain requires more consideration on false positives and false negatives especially in the case of covid 19 prediction. Such works were tabulated in the Table 1.

## 2.2 MEDICAL DOMAIN RELATED WORKS

Table 2 consist of various extensive details pertinent to the works that were handled by various authors related to data analytics techniques, especially on medical domain. In most of the works, decision tree and other extended algorithms like random forest and XGB produced the best accuracy.

*Table 2 Major Medical Domain Related Works*

| Year of publication and Ref number | Domain with Dataset | Objective of the Proposed work | Data analytics techniques | Software/tool used | Outcome |
|---|---|---|---|---|---|
| July 2022 [10] | ASD Autism (Adult, Adolescent, child) | To improve the highest accuracy of autism prediction with various level | Random Forest, Decision Tree classifier for prediction and SHAP method for feature selection | Colab Python tool in google platform | F beta, Recall, Precision scores are the other classifiers performance were calculated during prediction. |
| May 2022 [11] | Data1 In USA's center of disease controls around 95984 patients with 20 features. | Applied feature selection to improve the predictive performance and also to improve the | NB, MLP, KNN, RF, LR, Dt are the machine learning classifiers on the dataset after performing feature selection using Boruta, RR and RFE | NA | Identified that deep learning models are expecting more data for learning than other classifier models. It has achieved the |

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

| | | | | | |
|---|---|---|---|---|---|
| | Data2 Clinical dataset contains 600 patients with 20 features | execution tie of training set. So that test set execution can be improved. | feature selection method. | | accuracy of 91% using Random Forest. |
| Aug 2021 [12] | Liver dataset | To diagnostic liver disease using machine learning algorithms | LR, RF, XGB, SVM, ADB, KNN and Decision tree models on the dataset that contains 583 records with 10 features. | NA | Dataset split up into various sets from 50 :50 to 90: 10 in training. But only 80 :20 produced the best accuracy as 83.76% |
| 2021 [13] | UCI Parkinson's disease dataset | To construct the ML algorithm and can be used for e- health care system playing the expedite role in severity prediction. | LXGB, XGB, GB, ADABOOST were applied to diagnose the Parkinson disease | Jupyter, Anaconda tools were used in this proposed work | Prediction performance based on parameters accuracy, recall, F1score, AUC, Youden, Specificity etc. Here the model is improved by tuning the hyperparameter with Gridsearch CV |
| June 2021 [14] | Liver disease, UCI-MLR from California with 14 features | To identify blood donors and non-blood donors using clinical data | SVM for prediction and Pearson correlation to find the association between the data | NA | Proposed method using SVM achieved the accuracy of 98.23% |
| July 2020 [15] | Cleveland heart disease dataset | To improve the prediction of heart disease using hyper parameters optimization | XGBoost hyperparameter technique is applicable than Bayesian optimization. One hot encoding technology to improve accuracy | NA | Accuracy, specificity, sensitivity, F1score and AUC were used for performance evaluation and produced 91.8% accuracy. |
| July 2020 [16] | At Beijing author conducted survey and data was collected about the people with the age between middle to elderly people (around 380 people) | Risk of Type 2 diabetes prediction on the basis of the people habits | XGBoost model after performing 10-fold cross validation. To integrate many weak classifiers together to form a strong classifier. | Python | XGBoost had done its prediction work with outstanding learning capacity and earned 89.09% of accuracy when compared with SVM, RF, and KNN. and AUC was 91.82%. |

| | | | | | |
|---|---|---|---|---|---|
| Feb 2020 [17] | --- | Prediction of fibrosis stages marked from F1 to F4 using feature selection and classification algorithms | For feature selection correlation as filter and Genetic algorithm as wrapper were applied. and also, RF, NB, LR and SVM were applied for prediction of stages. Here out of 28 features 21 features were selected. | NA | Combination of genetic algorithm with Decision tree combination produced the accuracy of 85%. |
| Nov 2019 [18] | Physical observation related to Hypertension | To predict Hypertension using physical examination clinical value | RFE and Cross validation for optimal feature selection. SVM, C4.5, DT, RF and XGB are the classification models used | NA | XGBoost produced best prediction performance with accuracy 94.36%, F1score – 87.5, AUC – 92.7 XGBoost follows divide and conquer method. |
| 2018 [19] | Breast cancer with UCI ML REPOSITORY | Performance with highest criterion in the domain should be achieved | SVM, KNN, Naïve, J48, RF and MLP models were applied. FS with Entropy, Gain and Gain ratio to split the tree | NA | Accuracy, sensitivity and specificity, ROC and confusion matrices are increased. But finally Random Forest was produced the high accuracy of 98.77% |

## 3. METHODOLOGY

In figure 1, this proposed work is clearly explained and the steps are followed in this work.

### 3.1 Dataset Description

From [3] the covid, clinical dataset was taken for processing and analyzing. It was taken by this author from [2] and it was balanced, as clinical laboratory findings which consist of 600 cases. It may represent either covid positive or negative cases details. Around 520 have negative cases, and the remaining 80 was positive cases. For data analytics using classification model, the dataset is necessarily divided into training and testing phase datasets. Around 480 cases were considered as training phase dataset and 120 as testing phase dataset. Among 120, 16 cases were covid positive and the remaining was negative. To access the dataset the link is given and also all the records were stored in the comma separated value (csv) format, as it is easy to access the data using Colab. [33]

Journal of Theoretical and Applied Information Technology
15th November 2022. Vol.100. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

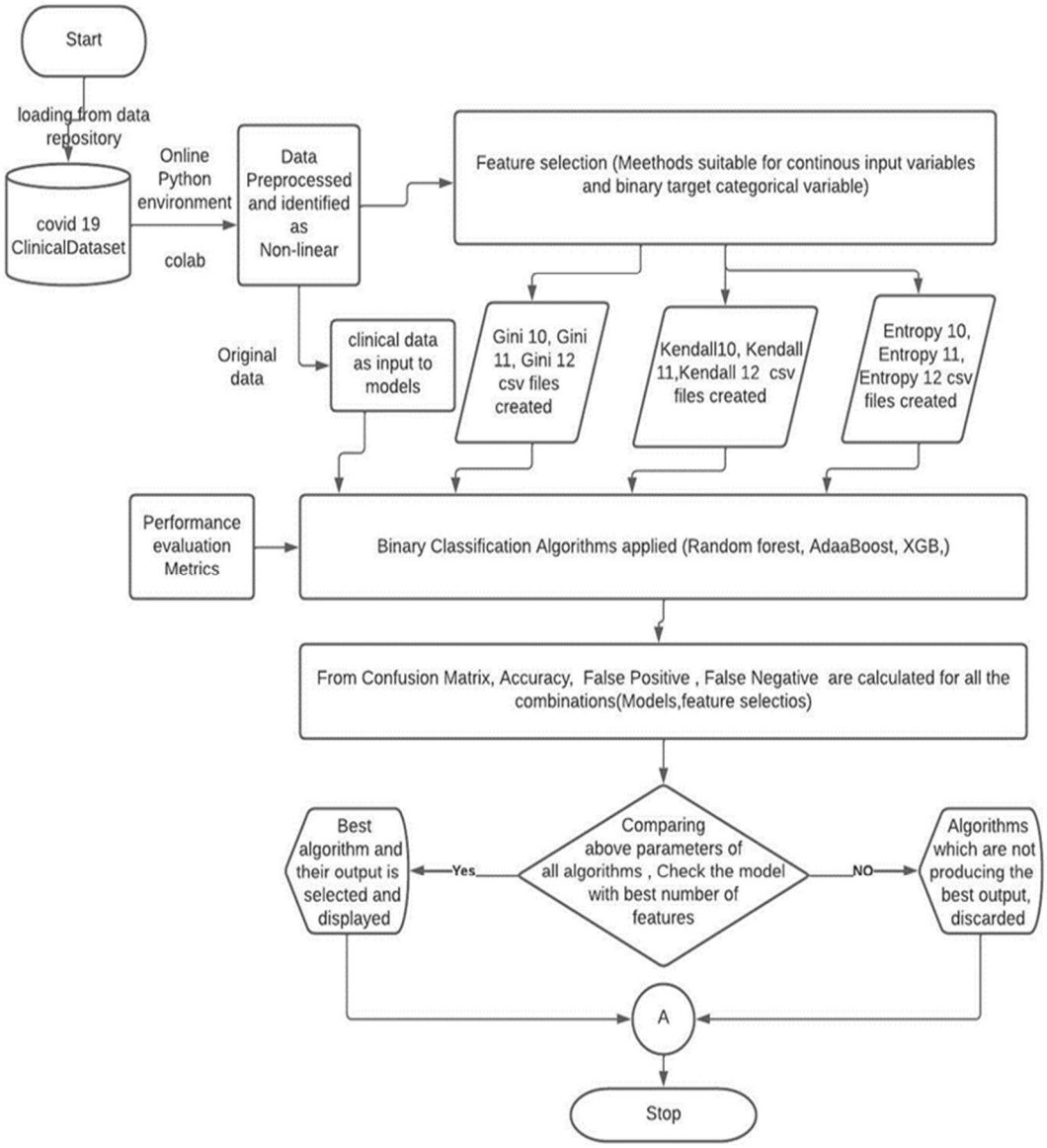


*Figure 1 Detailed Explanation About The Methodology Of This Proposed Work*

### 3.2 Data Preprocessing

Since the dataset is already ready for processing, as balancing is already performed, and there are no missing values present in the dataset. So next step is feature selection. When higher the number of columns present, it is necessary to identify any occurrence of redundant, irrelevant or any kind of unwanted data in the dataset for prediction. It can be removed by using certain statistical methods. By observing the data type, suitable method has to be chosen for features selection. In the dataset, the input variables are continuous and the target variable is categorical. For further analysis, association between the data can be established using Pearson's Correlation [14] Method, and identified that the values were non-linear and also weakly correlated. By considering all these factors, Gini index (Gini importance), Entropy and Kendall were selected for feature selection on

this Covid 19 dataset. There is another advantage that by reducing unwanted features, cost, time and space for classification model were also likely to be reduced. Using the literature, it was able to identify that for binary classification prediction, suitable models were tree-based or extended tree-based models since, many models are well suitable for this kind of dataset. To prove the model performance and the proposed work, the results are analyzed using performance metrics, so that the best model can be declared for clinical dataset with selected features, and the health care community can use this work for their applications. Remaining steps will be followed in further sections.

## 4. Feature Extraction using Statistical Methods

While considering the previous work, the maximum accuracy obtained from the same dataset is displayed in the figure. By applying various deep learning models [2] (ANN, CNN, LSTM and RNN), CNN produced the highest accuracy of 88. But with Boruta feature selection in [3], various conventional classification algorithms (NB, MLP, KNN, LR, DT and RF) were applied and the highest accuracy was attained with Random Forest as 91%. The Figure 2 represents the previous work accuracy comparison.

After review made with the literature for feature extraction [4], Gini, Entropy and Kendall tau coefficient [20] are the statistical filter methods that are suitable for this dataset (input variables are continuous and target variable is categorical) and the ranking is assigned to all the features by all these methods with respect to label.

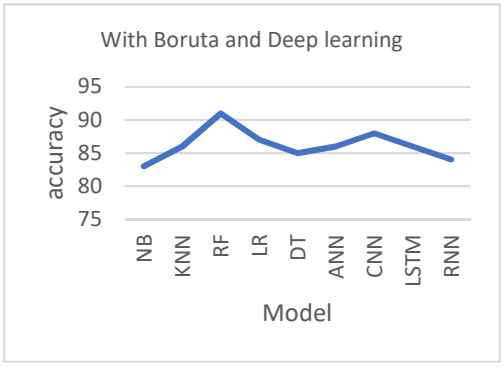


Figure 2 *Previous work Comparison*

### 4.1 Gini index on Feature Selection

In Gini index [21], it finds the difference between a node's impurity. It is necessary to take the decision of the node on which the forest is further splitter. Such valued feature is decided through Gini index through the following formula $\text{Gini index} = 1 - \sum_{i=1}^{c}(P_i)^2$ ,
where Pi is the probability of class i in a node.
The entire procedure is working through Random-forest classifier. Here the tree is constructed as forest and selectkbest() is the function used in scikit learning to identify the node and also the feature importance value (rank) which is assigned to every feature.

### 4.2 Entropy on Feature Selection

The most common feature selection technique for nonlinear, classification problem is Entropy method and it is very well suitable, as it finds the isolated correlation between them. In Entropy, when a node has less Entropy value then that node of that feature may get split in that branch. Else no split up in the branch. The formula for Entropy calculation for every feature is $\text{Entropy} = \sum_{i=1}^{c} -P_i log_2(P_i)$.
Using inbuilt class feature importance [22], the following steps are applied to select the features as
1. Importing the required libraries,
2. Loading and cleaning the data,
3. Building the Extra Trees Forest and computing the individual feature importance,
4. Visualize and compare the result.

### 4.3 Kendall on Feature selection

Kendall's rank coefficient is also suitable for nonlinear dataset. In Kendall, the advantage is its applicability to non-linear and the data does not require to be sorted before giving it as parameter / input. Kendall method assigns the tau value to estimate the strength of dependence between each feature which decides whether the feature is to be involved for prediction. This method is applied on non-parametric values to rank the given data. Ranking can be assigned with the range from -1 to 0 or 0 to 1. Perhaps, it can assume negative value to positive value as 0 to 1 since it doesn't mean much difference in this method. At the same time where there are 0 values between two features, then there is no relation between those features, and also if it is 1 then it has perfect relationship. All other remaining values are showing the actual relationships between them.

Finally, after ranking as in Table 1.1, optimum feature selection was decided, from all the three methods were made from the Table 1.2. As mentioned in the table the features having threshold value of count of occurrence (count >=2), were selected for further processing. Here only 11 features were selected, since after applying classification models, based on the performance metrics optimal features were identified from all the three feature selection methods**.** The following table clearly explains why only 11 features were selected for further processing.

## 5. TREE BASED MODELS AND ITS PERFORMANCE METRICS

In this proposed work, Feature selection methods were verified by applying various predictive models. From machine learning algorithm, as presented in the literature review in the section 2 of this paper, it was identified that many of the decision tree models were used to produce the best accuracy in many of the medical domain with same or related type of datasets. At the same time, other than accuracy metric, false negative and false positive has to be reduced. Because no positive patients and no negative patients can be marked wrongly, because it may lead to wrong guidance, such that, the chosen models were based on Tree-based and tree based extended classifiers that trained with optimal feature datasets selected by Gini, Entropy and Kendall. Here 3 Tree based models were considered as Random Forest, Adaboost and XGBoost. Initially, the model was trained with all the 19 features and the observations were made and represented in the Table 1.3. Models (Random Forest (RF), Extended Gradient Boosting (XGB), AdaBoosting(ADB), were applied on the dataset without any feature selection and the performance metrics were tabulated with Precision, Recall and F1 score of class label 1 and class label 0. And, also the work was handled with the same dataset for prediction using Deep learning models.

Later, the same models were applied on different combination of Gin index (15,12,11,10 features), Entropy (15,12,11,10 features) and Kendall (15,12,11,10 features) with all the 3 chosen models. In this proposed work, the evaluation has to be done with these metrics for both class 0 (covid negative patient) and class 1 (covid positive patient).

Since Covid is a most dangerous virus, it can cause any kind of negative effect, randomly by spreading over to the human being, which is even uncertain. Otherwise, if it is false negative, then spreading of virus may occur, and if it is false positive, then they may get a chance to admit with true positive patients and may get affected even after vaccination.

### 4.1 Random Forest:

Random Forest model is an advanced level of decision tree [13] model, where the dataset is taken as input and entire dataset is divided into various subsets and taken as input to each decision tree. Every time, each decision tree produces its own output. Then it is compared with the new data, and the majority of results are considered, and returned as performance metrics value for that dataset.

### 4.2 AdaBoost:

AdaBoost model is an ensemble learning algorithm [6], uses iterative procedures where this algorithm always learns from the mistakes of weak classification algorithms, so that this Algorithm can be turned as strong classifier.

### 4.3 Extended Gradient Boosting (XGB):

XGB is a method where decision trees are built in parallel [13], so that the level-wise strategy is applied with gradient value. The evaluation is made at the quality of every split and the majority results are returned as its output of the model.

## 5. RESULTS AND DISCUSSION

Based on the methodology described in that section, feature selections were made. After performing the optimum feature selection using Gini, Kendall and Entropy methods, the three different tree based extended classifiers were applied as per above and the accuracies were obtained (Figure 4) on the SARS-COV dataset and the results were shown in Table 1.4. All the tree-based models did not produce the same kind of output.

Journal of Theoretical and Applied Information Technology
15th November 2022. Vol.100. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table 1.2 Number Of First 11 Features Occurred In The Respective Feature Selection Methods*

| Gini Importance | Entropy | Kendall | Count of Occurrence |
|---|---|---|---|
| Leukocytes | Leukocytes | Leukocytes | 3 |
| Platelets | Platelets | Platelets | 3 |
| Eosinophils | Eosinophils | Eosinophils | 3 |
| Monocytes | Monocytes | Monocytes | 3 |
| Patient age quantile | Patient age quartile | Patient age quartile | 3 |
| Proteina C R mg/dL | Proteina C R mg/dL | Proteina C R mg/dL | 3 |
| Red blood Cells | Red blood cells | Red blood cells | 3 |
| Hematocrit | Hematocrit | Hematocrit | 3 |
| Hemoglobin | Hemoglobin | Hemoglobin | 3 |
| Lymphocytes | Lymphocytes | | 2 |
| Neutrophils | Neutrophils | | 2 |
| | | Creatinine | 1 |
| | | Urea | 1 |

*Table 1.3 Performance Of Models On Classifier Work Vs Existing Deep Learning Work*

| Model | ACC | ROC | Pre_1 | Rec_1 | f1_1 | Pre_0 | Rec_0 | F1_0 |
|---|---|---|---|---|---|---|---|---|
| RF | 90 | 70 | 83 | 43 | 57 | 93 | 99 | 96 |
| XGB | 90 | 78 | 63 | 64 | 64 | 95 | 95 | 95 |
| ADB | 86 | 69 | 41 | 50 | 45 | 94 | 95 | 95 |
| (Deep learning) | 86.6 | 62.5 | - | - | - | 92 | 93.68 | 93 |

Every combination (feature selection method with classifier model) produced the accuracy which was analyzed and finalized that helps to understand which combination produces the best and highest accuracy. In the discussion given below, the detailed description about these combinations & also the chart is neatly explained based on the results.

### 5.1 Binary Classification Using Entropy

Random Forest, XGB and ADABoost, were applied on highest ranks features as 15, 12, 11 and 10, then observations were made. The highest accuracy, false negative and false positive produced by Entropy with every model. The only model that produced the highest accuracy as 92% in Random Forest with Entropy 11 features.

Journal of Theoretical and Applied Information Technology
15th November 2022. Vol.100. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

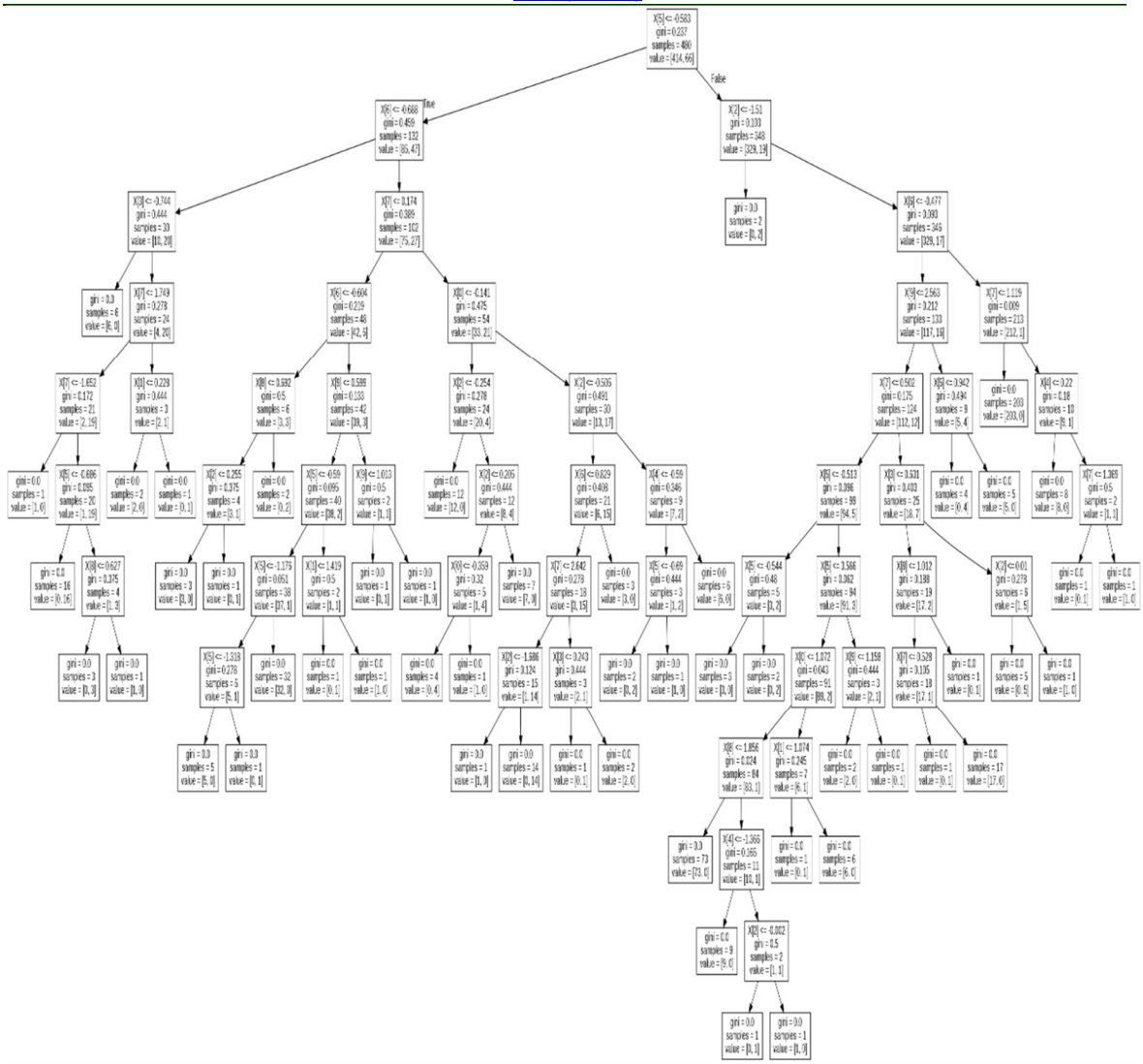

Figure 3 *A Sample Decision Tree To Perform Binary Classification Algorithm For The Given Dataset*

Supported value of class 0 (Negative Patients) and Class 1 (Positive Patients) are marked for false negative as 7, 6, 7 and 5 in the mentioned order itself. Whereas false positive values are 2,1,1 and 3. So to conclude, this model produces better result for false positive, since it reduces it to 2. In Figure 3, it explains easily how basic decision tree works on the dataset to perform binary classification to predict class 0 and class 1.

### 5.2 Binary Classification Using Kendall

Similar to Entropy Method, Random Forest, XGB, ADABoost, models were applied with highest ranks of features (15, 12, 11 and 10) shortlisted by Kendal. It is observed that with Random Forest Model and with first 11 features produced the best accuracy of 92% when compared with the other combinations.

### 5.3 Binary Classification Using Gini Importance

By using Gini importance (Gini index) for feature selections ranking, and random forest with first 11 features of Gini value and XGBoost with first 11 features of Gini importance produced the highest accuracy of 92%. All the other accuracy values were noticed as lesser than these combinations.

While considering the other models, some of them produced the accuracy of 91. In the existing work the prediction accuracy was 86.6 with original

clinical dataset. But by applying these 3 algorithms on covid optimized feature set, confusion matrix is created.

In confusion matrix, (Figure 5) there are 4 different values representing true negative, false positive, false positive and true positive. And also, while using confusion matrix, accuracy, precision, recall of class 1 and class 0 are remarkable. AUROC is the metric used to identify how much the model learned from the training dataset. But in medical domain, it's not enough to consider these metrics.

Specifically, it is necessary to consider false positive and false negative in prediction vs actual using testing dataset. Every false negative covid 19 patients create a big problem to the society. Sometimes, false positive leads to anxiety to such persons. So, both should be reduced at the maximum.

Among 120 patients of test dataset, only 14 patients hold positive. If all the 14 patients were classified correctly, then true positive = 14. If it is not, then false negative shows that covid patients were not classified properly. At the same time, among 106 values of Covid negative if all were correctly classified properly then True negative =106, else false positive occurred. So, some of the negative persons are wrongly classified as covid positive, and they may get anxiety unwantedly. Using RTPCR test, these results may occur frequently due to various reasons [23], but using Clinical Hematology test, these wrong predictions can be reduced with detailed analysis.

In [2], the covid 19 prediction based on various models' performance were analyzed using precision, recall and F1 score. Using CNNLSTM, the highest Precision was 92.35, Recall 93.68 and F1 Score as 93.00. In [3], the Covid 19 prediction after applying Boruta feature selection, the Random Forest model produced the performance values as Precision 94, Recall 95, F1 Score 94. Apart from these metrics, F1 Score (Macro) as 64. In both the analysis, there was no discussion about Covid 19 positive patient prediction analysis. In this proposed work, the evaluation criteria have given important critiques not only to accuracy, but also to false positive and false negative as mentioned in the figure 6 and figure 7. Such that for the society and health care community, the real threats of false positive and false negative have taken as important consideration to reduce. (i.e.) F1 score macro was produced as 77%, but it was only 64% in previous work [3].

| | PREDICTED | |
|---|---|---|
| MODEL NAME<br><br>Total value | NEGATIVE | POSITIVE |
| TRUE/FALSE | TRUE NEGATIVE | FALSE POSITIVE |
| FALSE/TRUE | FALSE NEGATIVE | TRUE POSITIVE |

ACTUAL

Figure 5 *Confusion Matrix (Actual vs prediction)*

Journal of Theoretical and Applied Information Technology
15th November 2022. Vol.100. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table 1.4 Feature Selection Method With Classifier Model Prediction Metrics*

| TO FIND THE BEST No. OF FEATURES FROM 3 DIFFERENT FEATURE SELECTION METHODS with COMBINATION OF CLASSIFIER MODEL BASED ON PERFORMANCE METRICS | | | | | | | |
|---|---|---|---|---|---|---|---|
| Performance Metrics | FEATURE SELECTION METHOD | MODEL WITH BEST ACCURACY USING NO. OF FEATURES | | | | | Best count |
| ACCURACY | GINI INDEX | Random Forest | 11 | XGBoost | 11 | 92 | 11 |
| | ENTROPY | Random Forest | 11 | | | 92 | |
| | KENDALL | Random Forest | 11 | - | | 91 | |
| FALSE NEGATIVE | GINI INDEX | XGBoost | 10 | - | | 3 | 10 |
| | ENTROPY | XGBoost | 11 | - | | 4 | |
| | KENDALL | AdaBoost | 10 | AdaBoost | 12 | 3 | |
| FALSE POSITIVE | GINI INDEX | Random Forest | 12 | Random Forest | 11 | 2 | 11 |
| | ENTROPY | Random Forest | 11 | Random Forest | 12 | 1 | |
| | KENDALL | Random Forest | 15 | - | | 1 | |
| | | | 11 | | | | 11 |

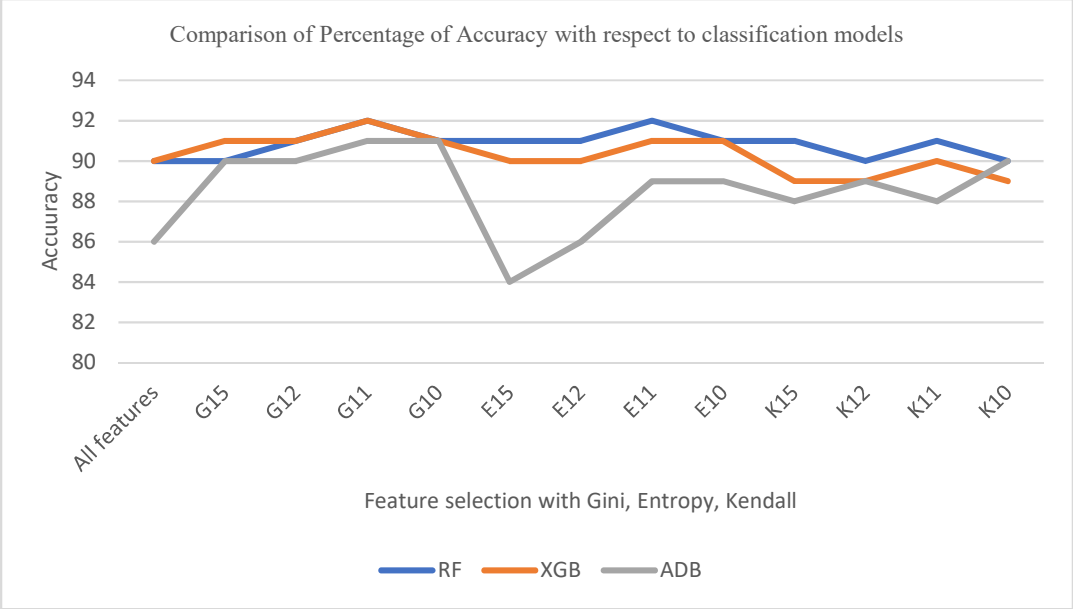


*Figure 4 Comparison Of Model Performance With Respect To Model And Accuracy (%).*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In the existing work, false positive and false negative were not considered for performance analysis but it is required. In this proposed work, the two false negative and false positive values were represented with figure 6 and figure 7, and it is able to observe that very minimum False negative occurred as 3 out of 14 using Kendall 10 and 12 features using Adaboosting. At the same time, the model XGB produced 4 out of 14 using Gini index of 10 features. But in the case of false positive, using Random Forest only 1 out of 106 were classified as False positive in Entropy 11, Entropy 12 and in Kendall 15 sub features which were remarked as better result. For false negative, Adaboosting played the best role, whereas for false positive random forest model produced the best result. It can be observed that if more trained values are available, then the result can be still optimized.

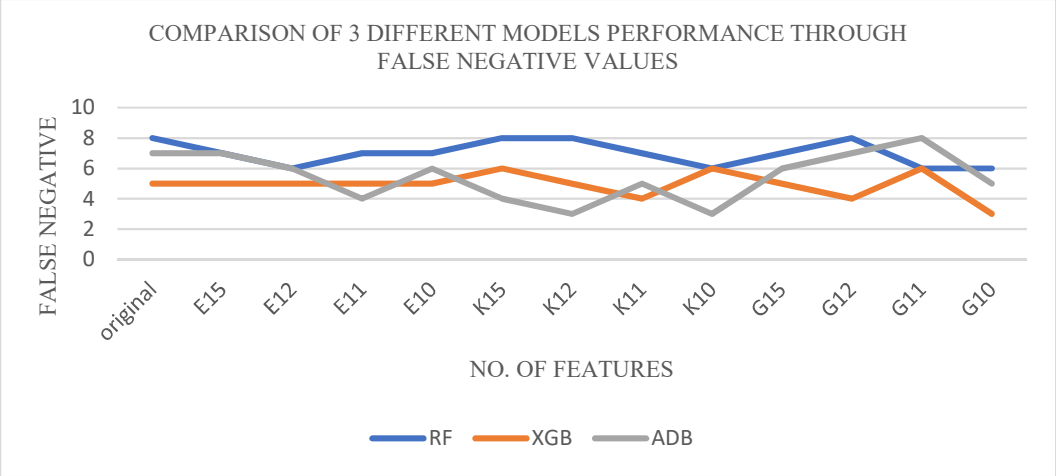


*Figure 6 False Negative Comparison*

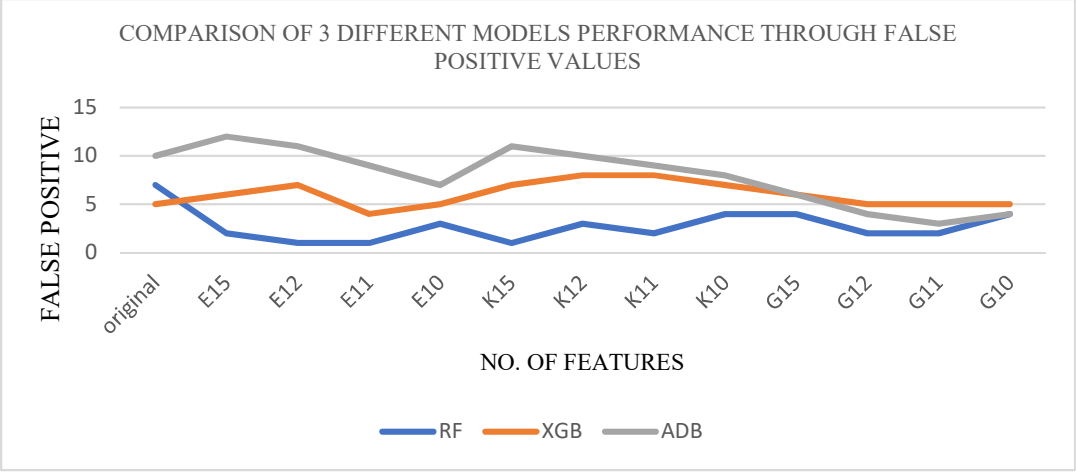


*Figure 7 False Positive Comparison*

## 6. CONCLUSION

To conclude the proposed work, by selecting various set of features and applying various models on those sub features, the highest accuracy is remarkably noted as 92%. While analyzing, these observations clearly stated that decision tree classifier or extended version of it, tried to create all possibilities of branches using ensemble or boosting method, so that the models

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

were able to understand using the training data, for further prediction. In such a way, Random Forest and XGB produced the best result in accuracy. This result shows that applying feature selection, and taking minimum number of features produced the highest accuracy than the original dataset. The proposed model provides the highest importance to the covid false positive and false negative. The model ensures that it is well trained to minimize the error in classification compared to previous work. The feature selection is optimized by comparing the features selected from 3 different methods to bring out the best possible accuracy. The boosting algorithm XGB, and Random Forest works well, due to its capacity to compare maximum number of combinations of nodes, flexibility, parallelization and distributed computing. Addition of data with variability ensures no overfitting. The parameters can be tuned or optimized. The features can further be given importance with weightage which improvises the accuracy of model. These will be considered as a future work for research. Thus, in further research work, it is necessary to identify where the model lacks not only in accuracy but also in false negative values, so that the health care community will be able to use these models with clinical dataset for their prediction with high accuracy. And the spreading of covid may reduce as early as possible.

## REFERENCES

[1] AlMoammar, A., AlHenaki, L., Kurdi, H. (2019). Selecting Accurate Classifier Models for a MERS-CoV Dataset. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 868. Springer, Cham. https://doi.org/10.1007/978-3-030-01054-6_74

[2] Alakus, T.B., Turkoglu, I.: Comparison of deep learning approaches to predict covid-19 infection. Chaos, Solitons & Fractals 140, 110120 (2020).

[3] Ali, S., Zhou, Y. and Patterson, M., 2022. Efficient analysis of covid-19 clinical data using machine learning models. *Medical & Biological Engineering & Computing*, pp.1-16.

[4] Domínguez-Olmedo JL, Gragera-Martínez Á, Mata J, Pachón Álvarez V. Machine Learning Applied to Clinical Laboratory Data in Spain for COVID-19 Outcome Prediction: Model Development and Validation. J Med Internet Res. 2021 Apr 14;23(4):e26211. doi: 10.2196/26211. PMID: 33793407; PMCID: PMC8048712

[5] Nahúm Cueto López , María Teresa García-Ordás , Facundo Vitelli-Storelli , Pablo Fernández-Navarro , Camilo Palazuelos and Rocío Alaiz-Rodríguez ; Exploring the performance of feature selection method using breast cancer dataset; January 2022, Indonesian Journal of Electrical Engineering and Computer Science 25(1):232, DOI:10.11591/ijeecs.v25.i1.pp232-23

[6] M. Ghosh, M. Mohsin Sarker Raihan, M. Raihan, L. Akter, A. Kumar Bairagi *et al.*, "A comparative analysis of machine learning algorithms to predict liver disease," *Intelligent Automation & Soft Computing*, vol. 30, no.3, pp. 917–928, 2021.

[7] Gupta, Vishan. (2021). Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model. Big Data Mining and Analytics. 4. 116-123. 10.26599/BDMA.2020.9020016.

[8] E. Tuba, I. Strumberger, T. Bezdan, N. Bacan in, M. Tuba Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine, Procedia Comput. Sci., 162 (Nov. 2019), pp. 307-315

[9] Valanis BG, Perlman CS. Home pregnancy testing kits: prevalence of use, false-negative rates, and compliance with instructions. Am J Public Health. 1982 Sep;72(9):1034-6. Doi: 10.2105/ajph.72.9.1034. PMID: 7102853; PMCID: PMC1650088.

[10] M. Masum, I. Nur, M. J. Hossain Faruk, M. Adnan, and H. Shahriar "A comparative study of machine learning -based Autism spectrum disorder detection with feature importance analysis",03 2022.[online]

[11] Ali, S., Zhou, Y. & Patterson, M. Efficient analysis of COVID-19 clinical data using machine learning models. *Med Biol Eng Comput* **60,** 1881–1896 (2022). https://doi.org/10.1007/s11517-022-02570-8

[12] Ghosh, M., Raihan, M. M. S., Raihan, M., Akter, L., Bairagi, A. K., Alshamrani, S. S., & Masud, M. (2021). A comparative analysis of

machine learning algorithms to predict liver disease. *Intelligent Automation and Soft Computing*, *30*(3), 917-928.

[13] M. M. Nishat, T. Hasan, S. M. Nasrullah, F. Faisal, M. A. -A. -R. Asif and M. A. Hoque, "Detection of Parkinson's Disease by Employing Boosting Algorithms," *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 2021, pp. 1-7, doi: 10.1109/ICIEVicIVPR52578.2021.9564108.

[14] Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data,FahadB. Mostafa, Easin Hasan,medRxiv 2021.04.26.21256121;doi: https://doi.org/10.1101/2021.04.26.2125612

[15] Budholiya K., Shrivastava S.K., Sharma V.A n optimized XGBoost based diagnostic system for effective prediction of heart disease J. King Saud Univ.-Comput. Inf. Sci. (2020) Google Scholar

[16] Wang L, Wang X, Chen A, Jin X, Che H. Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. Healthcare. 2020; 8(3):247. https://doi.org/10.3390/healthcare8030247

[17] A Survey on Predicting Advanced Liver Fibrosis Using Different Machine Learning Algorithms ,Krishnendu K B, Deepa S S, Published 29 February 2020, Computer Science, International journal of scientific research in science, engineering and technology

[18] Chang W, Liu Y, Xiao Y, et al. A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics (Basel)*. 2019;9(4):178. Published 2019 Nov 7. doi:10.3390/diagnostics9040178

[19] A. Saygılı ," Classification and diagnostic prediction of breast cancers via different classifiers", International Scientific and Vocational Studies Journal 2(2018) 56.

[20] Kaushalya Dissanayake, Md Gapar Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms", Applied Computational Intelligence and Soft Computing, vol. 2021, Article ID 5581806, 17 pages, 2021. https://doi.org/10.1155/2021/5581806

[21] Medina Garcia, V.H., Rodriguez Rodriguez, J., Ospina Usaquén, M.A. (2018). A Comparative Study Between Feature Selection Algorithms. In: Tan, Y., Shi, Y., Tang, Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science(), vol 10943. Springer, Cham. https://doi.org/10.1007/978-3-319-93803-5_7

[22] Krakovska O, Christie G, Sixsmith A, Ester M, Moreno S (2019) Performance comparison of linear and non-linear feature selection methods for the analysis of large survey datasets. PLoS ONE 14(3): e0213584. https://doi.org/10.1371/journal.pone.0213584

[23] Braunstein GD, Schwartz L, Hymel P, Fielding J. False Positive Results With SARS-CoV-2 RT-PCR Tests and How to Evaluate a RT-PCR-Positive Test for the Possibility of a False Positive Result. J Occup Environ Med. 2021 Mar 1;63(3):e159-e162. doi: 10.1097/JOM.0000000000002138. PMID: 33405498; PMCID: PMC7934325.

[24] Debjit, Kumar & Islam, Md & Rahman, Md & Pinki, Farhana & Nath, Rajan Dev & Al-Ahmadi, Saad & Hossain, Md & Mirazul Mumenin, Khondoker & Awal, Md.abdul. (2022). An Improved Machine-Learning Approach for COVID-19 Prediction Using Harris Hawks Optimization and Feature Analysis Using SHAP. Diagnostics. 12. 1023. 10.3390/diagnostics12051023.

[25] Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. J Med Syst. 2020 Jul 1;44(8):135. doi: 10.1007/s10916-020-01597-4. PMID: 32607737; PMCID: PMC7326624

[26] Munera N, Garcia-Gallo E, Gonzalez Á, Zea J, Fuentes YV, Serrano C, Ruiz-Cuartas A, Rodriguez A, Reyes LF. A novel model to predict severe COVID-19 and mortality using an artificial intelligence algorithm to interpret chest radiographs and clinical variables. ERJ Open Res. 2022 Jun 27;8(2):00010-2022. doi: 10.1183/23120541.00010-2022 PMID: 35765299; PMCID: PMC9059131

[27] Demir E., Giden R., Demir Z. The Importance of Procalcitonin Levels in COVID-19 Pneumonia. The Medical Journal of Mustafa Kemal University. 2022; 13(45): 1-5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[28] Moslehi, S., Rabiei, N., Soltanian, A.R. *et al.* Application of machine learning models based on decision trees in classifying the factors affecting mortality of COVID-19 patients in Hamadan, Iran. *BMC Med Inform Decis Mak* 22, 192 (2022). https://doi.org/10.1186/s12911-022-01939-x

[29] Sevinç E. An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. Comput Ind Eng. 2022 Mar;165:107912. doi: 10.1016/j.cie.2021.107912. Epub 2022 Jan 5. PMID: 35013637; PMCID: PMC8730510.

[30] Ramón A, Torres AM, Milara J*, et al* eXtreme Gradient Boosting-based method to classify patients with COVID-19 *Journal of Investigative Medicine* 2022;70:1472-1480

[31] https://www.who.int/publications/i/item/WHO-2019-nCoV-Clinical-Analytic-plan-2021.1

[32] https://ourworldindata.org/grapher/full list-cumulative-total-tests-per-thousand-bar-chart

[33] https://github.com/burakalakuss/COVID-19-Clinical .[6].