Prognosis of Vitamin D Deficiency Severity using SMOTE optimized Machine LearningModels

B Padmaja[a], Battu Ramya Reddy[b], R Vikrant Sagar[c], Heetesh Kumar Pradhan[d], G Chandra Sekhar[e], E Krishna Rao Patro[f]

[a, e, f] *Assistant Professor, Department of CSE, Institute of Aeronautical Engineering, Hyderabad, India*

[b, c, d] *Student, Department of CSE, Institute of Aeronautical Engineering, Hyderabad, India*

Abstract: According to global health studies, Deficiency in vitamin D (VDD) is a major public health problem, and there is a strong need for developing a prediction method that can use non-invasive approaches. Invasive methods include the use of medical instruments and tests that can take a long time to predict the outcome of a VDD procedure. This paper proposes to use machine learning classification algorithms for predicting VDD. The machine learning algorithms include Random Forest (RF), Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), AdaBoost classifier (AB), Extra Trees classifier algorithm (ET), and Logistic Regression (LR). The article evaluates the output of different machine learning classification methods for estimating the severity of VDD in humans. The main goal is to find the most reliable model and incorporate it into the prediction system.

Keywords: Vitamin D Deficiency (VDD), SMOTE, Python, Classification algorithms, Visualization, Prediction model.

1. Introduction

Vitamin D helps your body absorb the essential calcium from food for your bone health. A deficiency in vitamin D occurs when the body does not receive enough vitamin D, either due to a lack of sunlight or a poor diet. Vitamin D is an essential nutrient that has a wide range of effects on the human body. Around the world, almost one billion people suffer from severe Vitamin D deficiency [12]. Vitamin D deficiency is more common in the elderly, people of African or Asian descent, and those who are obese. However, research is gradually revealing the role of vitamin D in the prevention of a wide range of health problems. Vitamin D level evaluation is presently very costly, and it is determined using pharmacological methods. In earlier studies, the results were contrasted between statistical models, and machine learning algorithms were not used to predict severity.

Traditional statistical models, such as LR [8], are used to measure the intensity of VDD, but their success is impaired by their predictive performance limit and numerous parameters. Because of its high performance and ease of application, machine learning is one of the innovations that are most rapidly evolving in many fields. Machine learning has increased in recent years in the deployment and use of medical applications. In identifying new trends in the health sector, machine learning results are helpful for the effective application of preventive public health initiatives. Machine learning aims mainly at learning from input data, also known as training data, and predicting the future using new data.

The different parameters considered are exposure to sunlight, age, milk intake, gender, exercise routine, BMI, bone mass, fat, height, and weight. The main aim of the study is to predict the seriousness of VDD using ML classification models including Random Forest (RF), Multilayer Perceptron (MLP), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), Stochastic Gradient Descent (SGD), AdaBoost classifier (AB), Extra Trees classifier algorithm (ET), and Logistic Regression (LR). The output of all machine learning algorithms is analysed and compared with various performance measures like support, F1-score, recall, and precision [11]. To remove the unbalanced data, the Synthetic Minority Oversampling Technique (SMOTE) is used. If the difference between positive and negative performance values is minimal in any analysis of a dataset, it is a balanced dataset. If the difference is large, the dataset is unbalanced. The SMOTE technique is used to balance the dataset.

2. Related Works

A literature survey of various papers is depicted in Table I. A support vector regression technique approach used for deficiency of vitamin D prediction by Shuyu Guo, Robyn M. LuCas1, AnneLouise, Ponsonby showed that the Radial Basis Function Support Vector Regression model provided a stronger correlation of 25(OH)D measured and predicted scores than the MLR (Multiple Linear Regression) models, by obtaining permission to use the Ausimmune dataset, their research demonstrated an accuracy of 95% [9]. Maryam Tayefi, Kayhan Gonoodi, Mohsen Nematy, M. Ghayour Mobarhan, Gordon A. Ferns, Saeid Eslami, Maryam Saberi Karimian, Alireza Amirabadi Zadeh, Alireza Moslem, Susan Darroudi, Zahra Abasalti and S. Kazem Farahmand evaluated vitamin D potential risks using the model decision tree with accuracy 85.3%. 14 variables were used in the algorithm.

In the evaluation of model validation, the ROC curve was used. This research offers simple regulations of classification for categorizing deficiency of vitamin D potential risks, which may help improve vitamin D deficiency standard operating procedures [2]. [8] Wisconsin breast cancer dataset is taken to test the scalability of breast cancer prediction by ML algorithms in the big data context by S. Alghunaim and H. H. Al- Baity used different machine learning techniques for the breast cancer result in patients. They compared the effectiveness, efficiency & performance of the nine predictive models on the two platforms (SPARK and WEKA) in terms of specificity, accuracy, area under the curve (ROC), precision, and recall determining the best classification accuracy.

The scaled SVM classifier in the Spark framework outperformed the other classifiers with a performance measure of 97.13% and a low level of fault with the dataset, according to the experimental results. In the estimation of deficiency in vitamin D3 risk factors in patients, Freshtech Osmani and Masood Ziaeeb used a decision tree algorithm and found the result with an accuracy of 85.3%. In 2019, Patients were randomly selected in the study which was conducted in Khorasan Jonoobi province. R statistical software, version (3.4.1) is used for the analysis. The datasets were subjected to a variety of decision tree learning techniques. [10] Jun Ye, Ishir Bhan, Marcello Tonelli, Sherri-Ann M. Burnett Bowie, and Ravi Thadani studied to see whether vitamin D deficiency in dialysis patients could be classified based on routinely assessed clinical and demographic parameters.

There have been 980 dialysis patients. The logistic regression modelling, neural networks, and decision trees with vitamin D defect as the dependent variable have been developed using predictive models. Vitamin D deficiency occurred in 79 percent of the population. [3] Mohammad T. Al Hariri and Lubna I.Al Asoom studied a total of young Saudi women with an age range between 18 to 32 years. Data used here is Anthropometric data. 82.5% accuracy is obtained by using multiple linear regression models. The Pearson correlation determined that every model included risk factors that were significantly linked to the dependent variable. [13] E. Sohl, T. Merlijn, C. J. Netelenbos, M. A. Swart, N. M. Van Schoor, M. W. Heymans, P. Lips, and P. J. M. Elders performed a validation model where serum vitamin D insufficiency status in old is predicted. Researchers used backward selection multivariable logistic regression models and obtained 95 percent accuracy to identify predictors for inadequate serum 25(OH)D status. Questionnaires are used here. Questionnaires with statistical model classifiers have traditionally been used to predict the magnitude of VDD. The disadvantage is that the expense of evaluating the questionnaires is high and time consuming.

The use of super vector regression to predict vitamin D deficiency has been studied. The fact that they only used one validation dataset for their study is a limitation. As a result, this model was unable to accurately determine the risk of vitamin D deficiency to provide a more realistic foundation for vitamin D research. A study was carried out to develop a safe UVB LED indoor general lighting to aid in vitamin D synthesis in the human body. It did not, however, specify how much differential UVB irradiation dose should be given based on individual age and exposure.

Table 1. Literature Survey

Ref No.	Author Names	Classifiers Used	Dataset	Accuracy
[1]	Shuyu Guo Robyn M. LuCas AnneLouise Ponsonby	MLR Model SVR Model	Ausimmune dataset	95%
[2]	Kayhan Gonoodi Maryam Tayef	Decision tree model	Multi-dimensional data	85.3%
[3]	Lubna I. Al Asoom Mohammad T. Al Hariri	Multiple linear regression	Anthropometric data	82.5%
[4]	T. Merlijn P. J. M.	Multivariable logistic regression models	Questionnaire data	95%

[5]	S. Alghunaim H. H. Al-Baity	Support Vector Machine (SVM) Decision Tree(C4.5) Naive Bayes (NB) k Nearest Neighbors (k-NN)	Wisconsin Breast Cancer (original datasets	97.13%
[6]	J.-J. Beunza E. Puertas E. García-Ovejero G. Villalba E. Condes G. Koleva C. Hurtado M.F. Landecho	Decision tree Random Forest Support Vector Machines Neural Networks Logistic Regression	Open database of the Framingham Heart Study	85%
[7]	A. Jorge V. M. Castro A. Barnado V. Gainer C. Hong T. Cai R. Carroll J. C. Denny L. Crofford K.H. Costenbader K. P. Liao E. W. Karlson C. H. Feldman	SLE (Systemic lupus erythematosu) EHR algorithms	A centralized longitudinal EHR database of the Research Patient Data Repository (RPDR)	90%
[8]	H. Tamune J. Ukita Y. Hamamoto H. Tanaka K. Narushima N. Yamamoto	K-nearest Neighbors Logistic Regression Support Vector Machine Random Forest	The datasets utilized in the current study are available from the corresponding author upon reasonable request	95%
[9]	Souad Bechrouri Abdelilah Monir Hamid Mraoui El houcine Sebbar	Linear Regression Random Forest Multivariable Adaptive Regression Spline Support Vector Machine	The database is composed of biochemical data, blood tests were performed at the biochemistry laboratory of the Mohamed 6 UHC	80%

3. Methodology

The main aim of our proposed system is to find the most accurate machine learning classification models which can predict vitamin D deficiency. The machine learning algorithms include Gradient Boosting Classifier, Extra Trees Classifier, Gaussian NB Classifier, SGD Classifier, MLP ANN, Logistic Regression, K Neighbours Classifier, Decision Tree Classifier, and Random Forest Classifier.

Imbalanced data is particularly difficult for a predictive modelling job because of the uneven classification of data. If the difference between positive and negative performance values is minimal in any analysis of a dataset, it is a balanced dataset. If the difference is large, the dataset is unbalanced. To remove the unbalanced data, the Synthetic Minority Oversampling Technique (SMOTE) is used. The inputs which are taken by a dashboard for a new patient are gender, bone mass index, age, waist, exercise, body fat, height, milk consumption, sunlight exposure, BMI, and weight.4 levels are taken into consideration.

Merits of proposed system:

SMOTE Technique is used for balancing the data. Extra Tree Classifier outperformed other classifiers with 73.3%. By using the most accurate machine learning model and creating a dashboard using python flask framework which predicts the level of deficiency of vitamin D in a new patient.

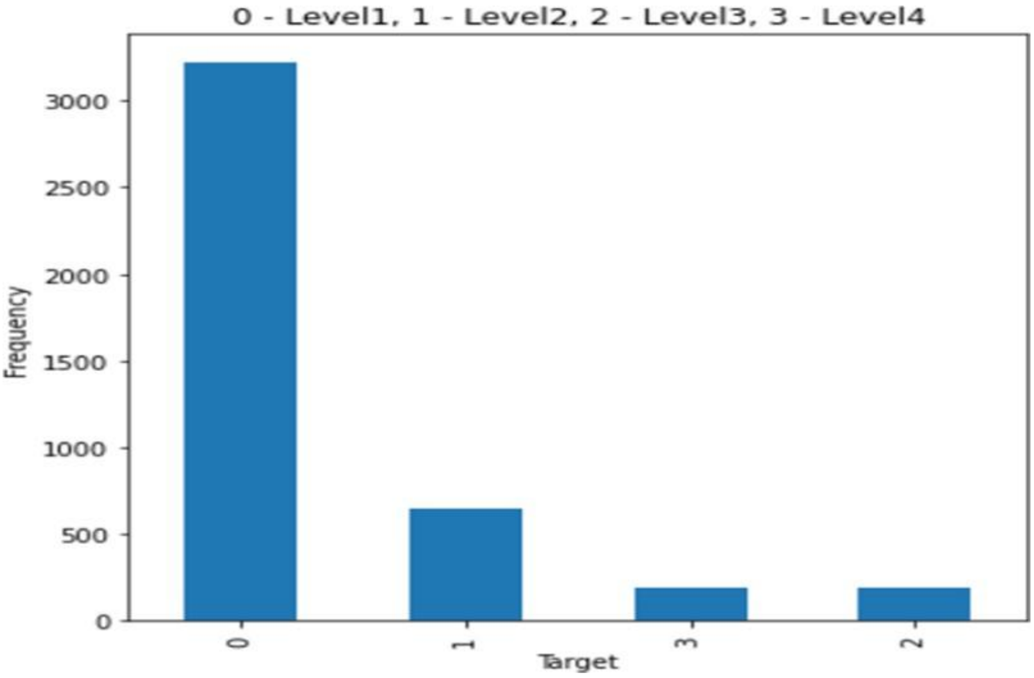


Figure 1. Targeting different levels for different frequency

In figure 1, the numbers 0, 1, 2, 3 on the x-axis represent the severity of level 0-sufficiency, level 1-insufficiency, level 2- deficiency, and level 3-severe deficiency, respectively. The frequency of the number of people is indicated on the Y-axis. According to the statistics, level 0 has over 3000 number of people, level 1 has 700 number of people, level 2 has around 250 number of people, and level 3 has around 250 number of people.

The proposed method's approach was divided into 6 phases: data exploration, pre-processing, visualization, handling imbalanced data, training of the model, and prediction, with the analysis emphasizing the handling of imbalanced datasets using SMOTE.

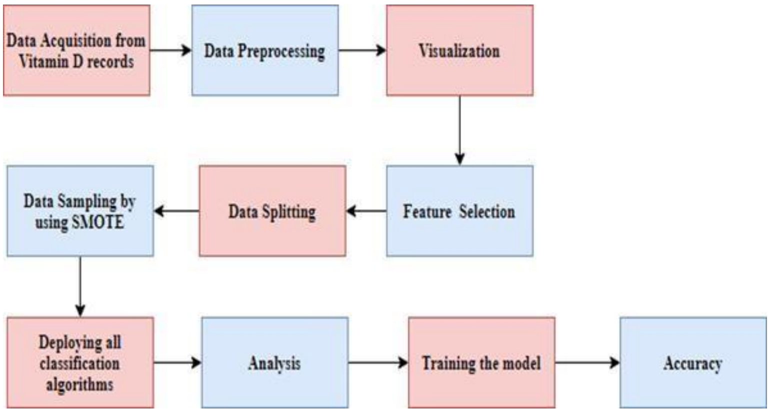

Data Acquisition from Vitamin D records
Data Preprocessing
Visualization
Feature Selection
Data Splitting
Data Sampling by using SMOTE
Deploying all classification algorithms
Analysis
Training the model
Accuracy

Figure 2. The process sequence depicting the prediction of VDD

Above figure 2 depicts the process sequence of predicting Vitamin D Deficiency. Data is primarily collected with parameters like age, height, weight, Body Mass Index (BMI), bone mass, body fat, sunlight exposure, exercise, gender, waist circumference, and milk consumption. First the pre-processing of data occurs where the data gets ready to be used in a machine learning model and it is considered as an important step. Then the processed data gets viewed by data visualization. Data can be analysed by viewing different parameters of the data. Then Feature Selection of the data occurs, and the data is split. The split data is then sampled by using SMOTE technique. After applying a smote technique, deployment of all classification algorithms occurs. Analysis is done on all the machine learning classification algorithms and the model with high accuracy is trained and incorporated in the dashboard.

Data Acquisition

People of both male and female are considered. Milk Consumption varying from 0 to 600. Sunlight Exposure varying from 5.0 to 30 hours. Whether exercise is involved or not in one's routine. Body Fat varies from 21.60% to 41.20%. Bone mass varies from 2.00-3.60. Waist Circumference in between 58cm to 92cm. Height varies from 0 to 300cm. Weight ranging from 61kgs to 91kgs.Body Mass Index (BMI)varying from 25.94 kg/m2 to 34.81 kg/m2. These all are different 11 parameters with their ranges that were taken in dataset. 4 levels are framed from the data as Level 0 - Sufficiency Severity, Level 1 - Insufficiency Severity, Level 2 - Deficiency Severity, and Level 3 - Severe Deficiency.

Pre-processing of data

It is the first and most important step in building a machine learning model. Data pre-processing is the procedure for preparing data for use in a machine learning model. Real data, which is in the unspecified format typically contains noises, missing values. This kind of data should not be used directly. The main purpose of the pre-processing task is cleaning and formatting the data and making it suitable for the model by increasing accuracy and efficiency. Getting the data and putting in comma-separated files (CSV) and then importing specific predefined python libraries NumPy, matplotlib, and pandas which are used in the pre-processing step. Extraction of dependent and independent variables is done after importing the dataset. Missing data is then handled, and categorical data is encoded. And finally, the performance of the model gets enhanced by forming a training set and test set from splitting the data.

Visualization of data

Visualization is crucial for data analysis. Presentation of data in pictorial format can be done by using data centric python packages like pandas, seaborn, and matplotlib. Seaborn is a fantastic Python visualization library that lets us create statistical graphics plots. The frequency of occurrence of phenomena that fall within a specific range of values and are arranged in consecutive and fixed intervals is represented by the histogram. The figure shows the analysis of the outcome which is done by using the seaborn python library. Matplotlib is an incredible Python visualization library for 2D array plots. Boxplot is a type of graph that shows the quartiles of a set of numerical data. It's an easy way to see how our data is organized. Figure 4 represents the data visualization of different terms like gender, age, waist, exercise, Sunlight Exposure, Milk Consumption, Height, Body Fat.

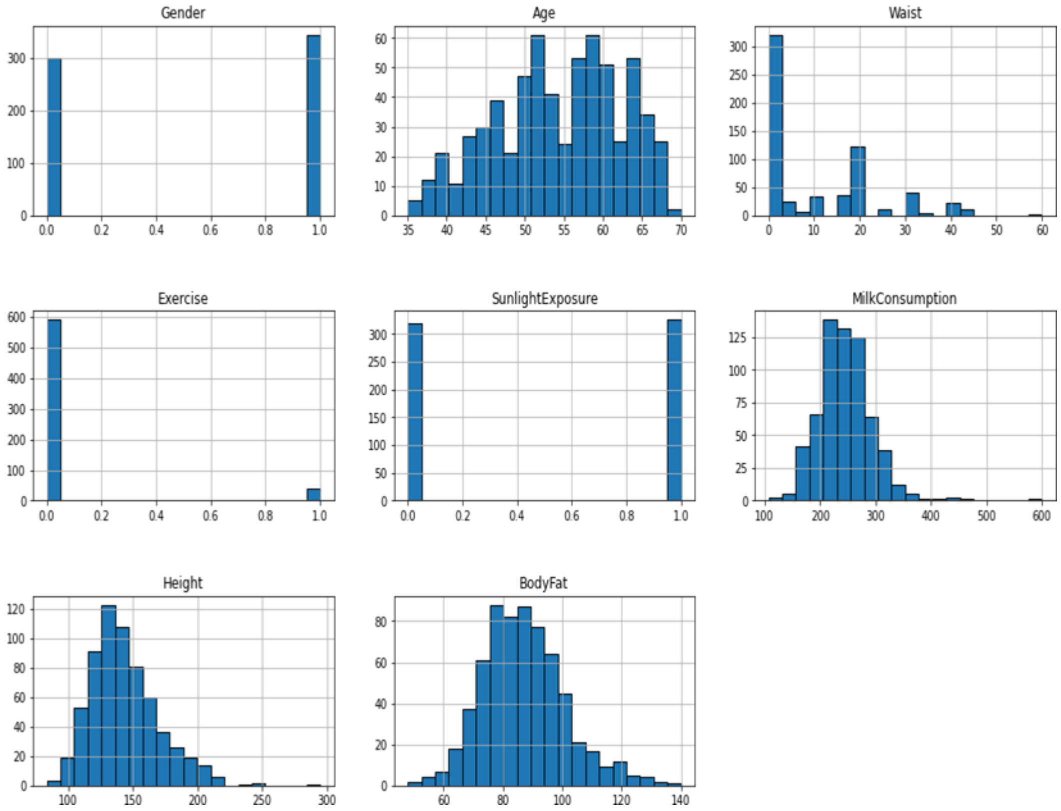


Figure 3. Graph Visualization of different parameters

Above figure 3 is consists of different graph visualization of different parameters. On the gender graph, the X-axis indicates female by 1.0 and male by 0.0. When compared to males, females outnumber males. The x axis of the age graph indicates the age of the people, while the y axis indicates the number of people. This graph depicts the number of people present at a given age. The x axis of the waist graph represents the size of the waist, and the y axis represents the number of people. In the exercise graph, the x axis represents the values that indicate whether the people's lives are close to the exercise routine or not by 0 and 1, and the y axis represents the number of people. This graph depicts the number of people who exercise and the number of people who do not exercise. The x-axis in the sunlight exposure graph indicates the values 0 and 1, where 1 indicates, they are not exposed to any sunlight and 0 indicates they are exposed to sunlight, and the y-axis indicates the number of people. This graph depicts the number of people who are exposed to sunlight and the number of people who are not. The x-axis of the Milk Consumption graph represents the quality of milk consumed, while the y-axis represents the people. According to the graph, many people consume 200-300ml of milk per day. The x-axis in the height graph represents the height values in centimetres, while the y axis represents the frequency of people. Many people are between the heights of 100 and 150 centimetres. The x-axis in the body fat graph represents the body fat range, and the y-axis represents the frequency of people.

Handling Imbalance data

Arrays of datasets are split into random subsets of trained sets and test sets. We used four variables for the output which are 'xtest', 'xtrain', 'ytest', 'ytrain'. Features of the training data are determined by xtrain. Features of testing data are determined by testing data. Dependent variables for training data are determined by 'ytrain' and independent variables of testing data are determined by 'ytest'. SMOTE issued for the imbalance classification of data. When working with imbalanced datasets the problem is that most machine learning techniques would disregard the class of the minority, resulting in poor results. So, to avoid this, the oversampling imbalanced dataset is called SMOTE (Synthetic Minority Oversampling Technique). Imblearn library is used for the implementation of SMOTE in python.

Minority class input vector is chosen, K nearest neighbours is found out by specifying k-neighbours as an argument in the SMOTE function Select one of these neighbours and place synthetic point anywhere on the line connecting the point under consideration and its selected neighbour. This process is repeated until the data gets balanced.

Classification Algorithms

Multilayer Perceptron (ANN): An MLP in a directed graph is composed of multiple layers of nodes, each completely connected to the next. A feed-forward artificial neural network model called a Multilayer Perceptron (MLP) maps input parameter to acceptable output data sets. Back propagation is used by MLP to train the network. Multiple layers of computational units are interconnected in a feed-forward fashion in this type of network. Some conditions in this classifier terminate because the error must be less than a certain threshold or condition, the error on a separate validation set is less than a predefined threshold, and iterations are limited to a certain number of times.

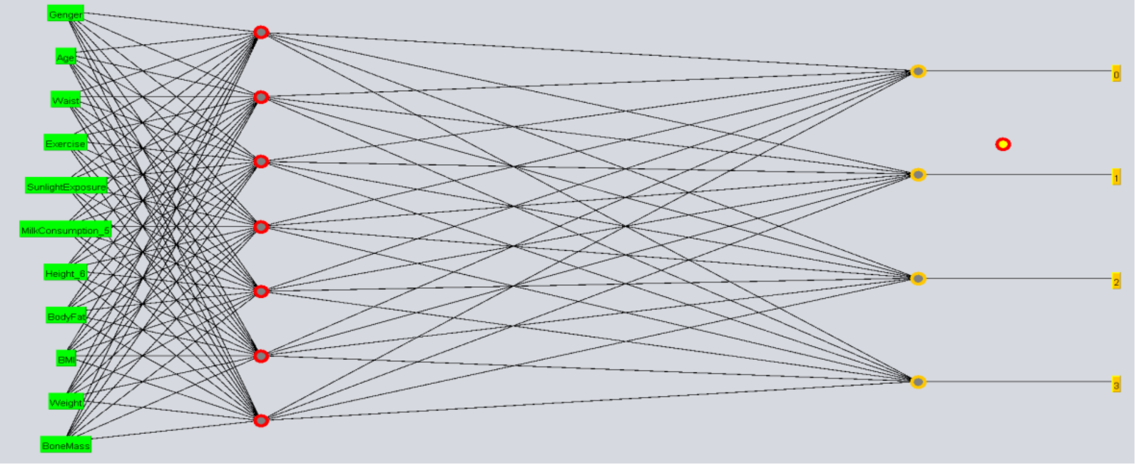


Figure 4. MLP-Neural Network representation w.r.t. 11 attributes

The above figure 4 represents a neural network depicted with respect to the eleven attributes of the VDD dataset. It includes the hidden layers as well as the outcome.

Extra Tree Classifier: The ET algorithm uses standard top-down technique to build a set of unpruned decision or regression trees. Divide nodes by selecting random cut pointed spots and grow trees using the whole learning sample, it differs in two ways from other tree-based ensemble methods: (rather than a bootstrap replica).

Gradient Boosting Classifier: Each predictor in GB corrects the error of its predecessor. Unlike Adaboost, the training instance weights are not adjusted. Each predictor is instead trained with the residual errors of the predecessor as markers. CART (Classification and Regression Trees) is the foundation learner in a methodology called Gradient Boosted Trees. The function matrix X and the labels y are used to train Tree1. The training set residual errors r1 are calculated using the predictions labelled y1(hat). The X function matrix and Tree1's residual error r1 will then be used as Tree2 labels. The residual r2 is calculated using the expected resultsr1(hat). The process continues until all the N trees of the ensemble are trained.

Shrinkage is a critical parameter to consider when using this technique. After the learning rate (eta), which varies from 0 to 1, is increased to multiply every tree in the ensemble, the prediction of each tree is shrunk. There is a trade-off between eta and the number of estimators, resulting in a reduction in learning.

To achieve certain model efficiency, rate must be compensated with increasing estimators. Predictions can now be made because all trees have been conditioned. Each tree predicts a label, with the formula providing the final prediction. In the ensemble is shrunk. There is a trade-off between eta and the number of estimators, resulting in a reduction in learning. To achieve a certain model efficiency, rate must be compensated with increasing estimators.

Random Forest Classifier: Leo Breiman's Random Forest is a random sample selection of samples from unpruned classifying or regression trees. The following methods are used to obtain each tree. If N is the total number of cases in the training set, and the training set replaces the original data. By randomly sampling these N cases and using them as the tree growth workout set.

To specify m«M for each node, the variable m is chosen for the M number of input variables, m variables are chosen at random from the M, and the best divisions in those m are used for the node division. The value of m remains constant during the growth of the forest. The tree is grown for as long as possible. There will be no pruning.

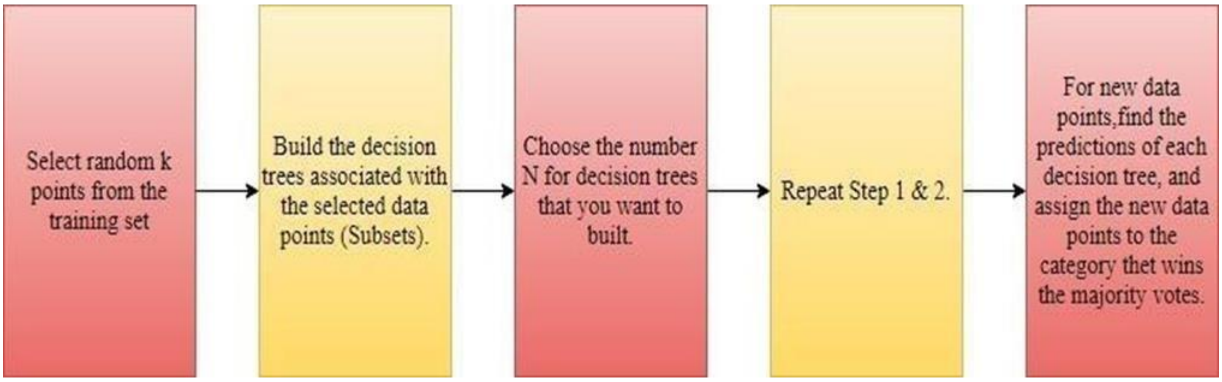
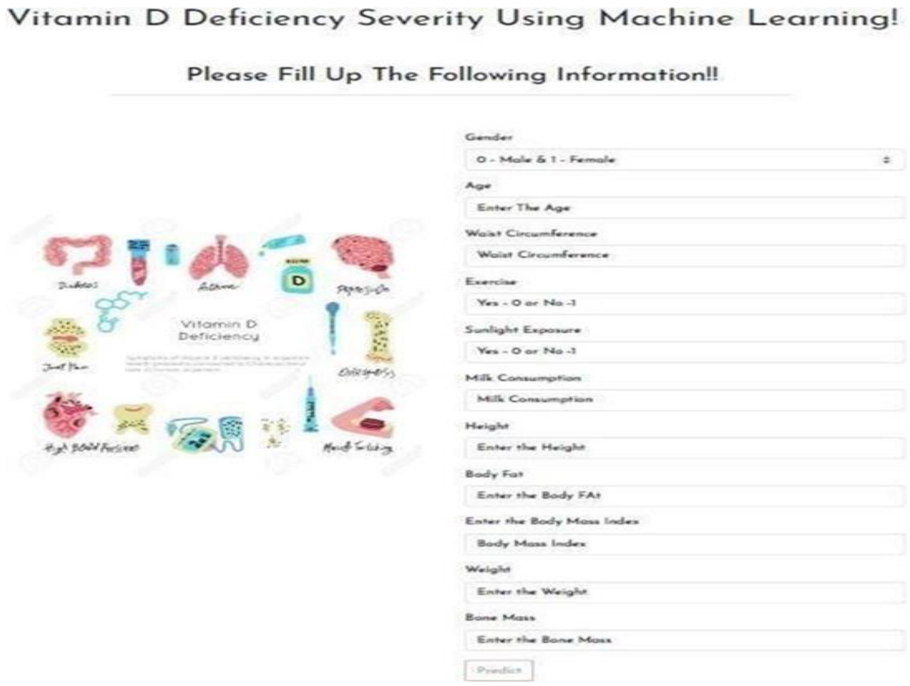

Select random k points from the training set
Build the decision trees associated with the selected data points (Subsets).
Choose the number N for decision trees that you want to built.
Repeat Step 1 & 2.
For new data points,find the predictions of each decision tree, and assign the new data points to the category thet wins the majority votes.

Figure 5. Sequence depicting the process of Random Forest classifier algorithm

Sequence showing the process of random forest classifier algorithm is shown in above figure 5. k data points are chosen at random from the training set. Decision trees are constructed using this subset of selected data points. Then we choose N number of decision trees to construct. Two previous steps are repeated. For new data points, the prediction of each decision tree is determined, and a new data point is assigned to the category that receives most votes.

4. Result Analysis and Implementation



Vitamin D Deficiency Severity Using Machine Learning!

Please Fill Up The Following Information!!

Gender
0 - Male & 1 - Female

Age
Enter The Age

Waist Circumference
Waist Circumference

Exercise
Yes - 0 or No -1

Sunlight Exposure
Yes - 0 or No -1

Milk Consumption
Milk Consumption

Height
Enter the Height

Body Fat
Enter the Body FAt

Enter the Body Mass Index
Body Mass Index

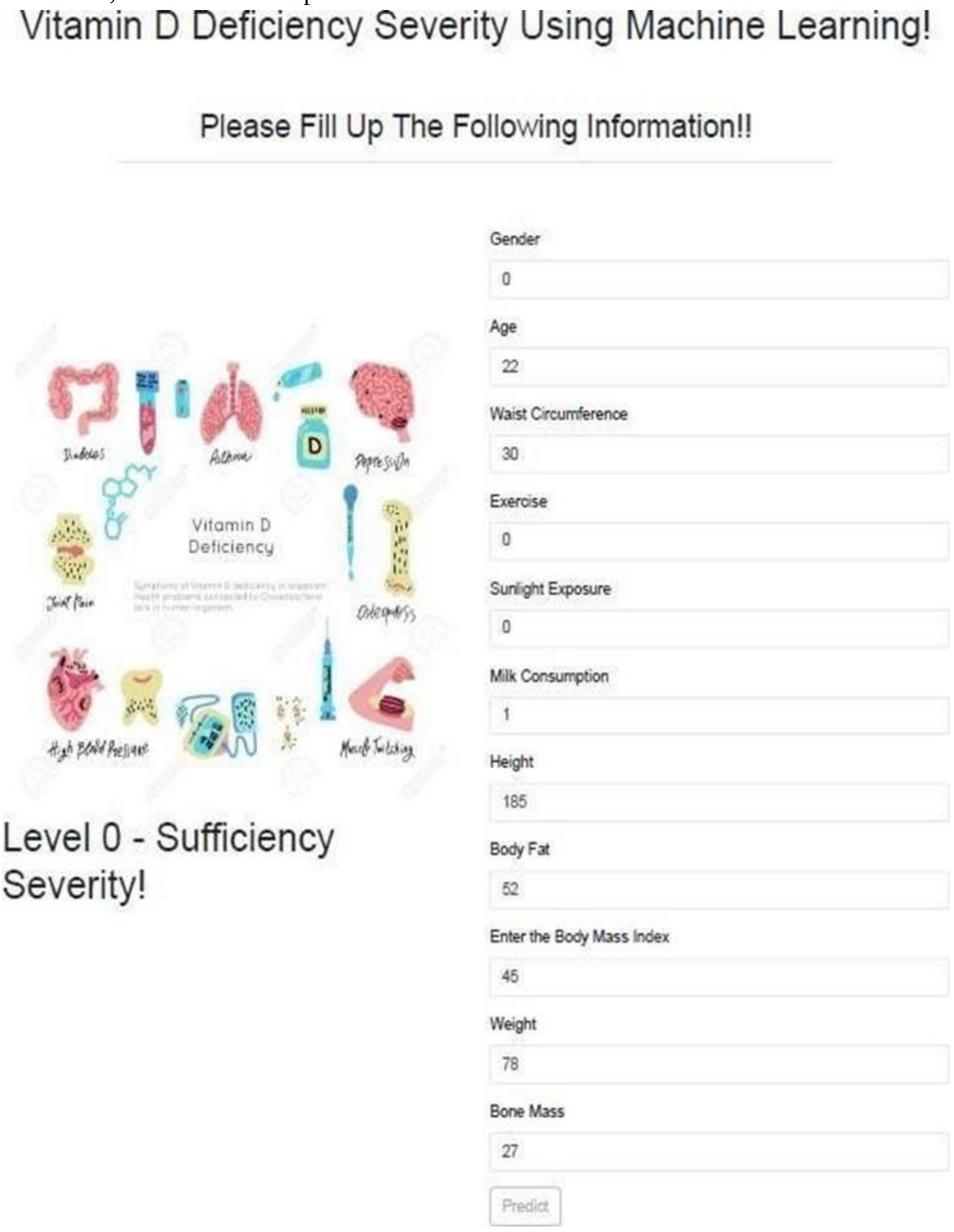
Weight
Enter the Weight

Bone Mass
Enter the Bone Mass

Predict

Figure 6. Interface diagram developed to predict VDD in a new patient

Figure 6 above is the interface diagram developed to predict VDD in a new patient. Gender input values can be 0.0 or 1. 0 denotes male and1 denotes female. The patient's age is entered into the Age input box. Waist Circumference measures a patient's waist circumference. In the exercise input box, 0 indicates that exercise is part of the patient's routine, and 1 indicates that exercise is not part of the patient's routine. In the sunlight exposure input box, 0 indicates that sunlight exposure is part of the patient's routine, and 1 indicates that sunlight exposure is not part of the patient's routine. Milk consumption is specified as a quantity of milk. Height uses the patient's height in centimetres as input. Body fat, bone mass, weight, and body mass are calculated using the patient's input values. A client can run the interface and fill in the required details like Gender, Age, Waist circumference, Exercise, Sunlight exposure, Milk consumption, Height, Body Fat, Body mass, index, Weight and Bone mass as shown below. The test can be repeated ordinarily as to date and time. During the test once all the details are filled in, one can click on predict.



Vitamin D Deficiency Severity Using Machine Learning!

Please Fill Up The Following Information!!

Level 0 - Sufficiency Severity!

Gender
0

Age
22

Waist Circumference
30

Exercise
0

Sunlight Exposure
0

Milk Consumption
1

Height
185

Body Fat
52

Enter the Body Mass Index
45

Weight
78

Bone Mass
27

Figure 7. Output Interface

Contingent upon the alternatives finished in the test, the outcome leaves the calculation of the classifier algorithms applied. The assessment outcomes are Level 0 - Sufficiency Severity, Level 1 - Insufficiency Severity, Level2 - Deficiency Severity and Level 3 - Severe Deficiency observed in the client as shown in Figure 7.

The performance measures of different machine learning models are discussed here below with table 2.

Performance Measures

Precision: The fraction of correctly categorized instances or samples among those classified as positives is calculated by precision. As a result, the precision formula is as follows:

$$\text{Precision} =$$

$$\text{TP TP} + \text{FP}$$

Recall: Recall is calculated by splitting the number of true positives by the total of true positives and false negatives in a two-class imbalanced classification problem. The result is a value between 0.0 for no recall and 1.0 for full or perfect recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 measure: F-Measure is a method for combining precision and recall into a single measure that encompasses both. This is how the traditional F calculation is calculated.

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 2. Performance Measures of Different Machine Learning Models

Machine Learning Classifier Models	Multiclass (n=4)	Precision	Recall	F1-Score	Support
Support Vector Machine	Sufficiency Severity	0.81	0.07	0.83	973
	Insufficiency Severity	0.33	0.24	0.27	187
	Deficiency Severity	0.06	0.04	0.05	52
	Severe Deficiency	0.00	0.00	0.00	54
Random Forest Classifier	Sufficiency Severity	0.77	0.66	0.71	973
	Insufficiency Severity	0.33	0.24	0.27	187
	Deficiency Severity	0.06	0.04	0.05	52
	Severe Deficiency	0.00	0.00	0.00	54
Decision Tree Classifier	Sufficiency Severity	0.77	0.66	0.71	973
	Insufficiency Severity	0.21	0.25	0.23	187
	Deficiency Severity	0.04	0.08	0.05	52
	Severe Deficiency	0.02	0.04	0.03	54
K Neighbors Classifier	Sufficiency Severity	0.82	0.44	0.57	973
	Insufficiency Severity	0.24	0.44	0.31	187
	Deficiency Severity	0.07	0.25	0.11	52
	Severe Deficiency	0.04	0.17	0.07	54
Logistic Regression	Sufficiency Severity	0.81	0.2	0.32	973
	Insufficiency Severity	0.22	0.51	0.31	187
	Deficiency Severity	0.06	0.33	0.1	52
	Severe Deficiency	0.05	0.331	0.09	54

MLP-ANN Classifier	Sufficiency Severity	0.78	0.37	0.51	973
	Insufficiency Severity	0.37	0.27	0.31	187
	Deficiency Severity	0.05	0.35	0.08	52
	Severe Deficiency	0.03	0.19	0.06	54
SGD Classifier	Sufficiency Severity	0.77	0.98	0.86	973
	Insufficiency Severity	0.33	0.01	0.01	187
	Deficiency Severity	0.00	0.00	0.00	52
	Severe Deficiency	0.04	0.02	0.03	54
Gaussian NB	Sufficiency Severity	0.73	0.09	0.15	973
	Insufficiency Severity	0.31	0.38	0.34	187
	Deficiency Severity	0.03	0.13	0.05	52
	Severe Deficiency	0.04	0.48	0.07	54
Extra Trees Classifier	Sufficiency Severity	0.79	0.91	0.85	973
	Insufficiency Severity	0.37	0.21	0.27	187
	Deficiency Severity	0.00	0.00	0.00	52
	Severe Deficiency	0.00	0.00	0.00	54
Gradient Boosting Classifier	Sufficiency Severity	0.79	0.84	0.81	973
	Insufficiency Severity	0.3	0.26	0.28	187
	Deficiency Severity	0.09	0.08	0.08	52
	Severe Deficiency	0.04	0.02	0.02	54
Ada Boost Classifier	Sufficiency Severity	0.8	0.62	0.7	973
	Insufficiency Severity	0.27	0.45	0.34	187
	Deficiency Severity	0.06	0.13	0.08	52
	Severe Deficiency	0.06	0.09	0.08	54

Accuracy: One metric for assessing classification models is accuracy. Informally, accuracy refers to the percentage of correct predictions made by our model. The following is a formal description of accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

TP – True Positives
TN – True Negatives
FP – False Positives
FN – False Negatives

An accuracy range of different machine learning classification algorithms is obtained as shown in the below table. Support vector machine algorithm has an accuracy of about 17.9 percent. The Random Forest algorithm has a 72.1 percent accuracy. The decision tree algorithm has a 41.7 percent accuracy. The logistic Regression algorithm has a 24.1 percent accuracy. The accuracy of the Multilayer Perceptron is 45.3 percent. The Gaussian Naive Bayes algorithm has a 14.8 percent accuracy.

Table 3. Accuracy

Machine Learning ClassifierModels	Accuracy
Support Vector Machine	17.9
Random Forest Classifier	72.1
Decision Tree Classifier	56.1
K Neighbors Classifier	41.7
Logistic Regression	24.1
MLP-ANN Classifier	45.3
SGD Classifier	21.3
Gaussian NB	14.8
Extra Trees Classifier	73.3
Gradient Boosting Classifier	68.4
Ada Boost Classifier	55.5

The stochastic gradient descent algorithm has 21.3 percent accuracy. The Extra Tree Classifier algorithm is 73.3% accurate. The Gradient Boosting algorithm is approximately 68.4% accurate, while the AdaBoost Classifier algorithm is approximately 55.5% accurate. RF Algorithm and Extra Tree Classifier Algorithm outperformed with all the other algorithms and with a high accuracy of 72.1% and 73.3% as shown in table 3.

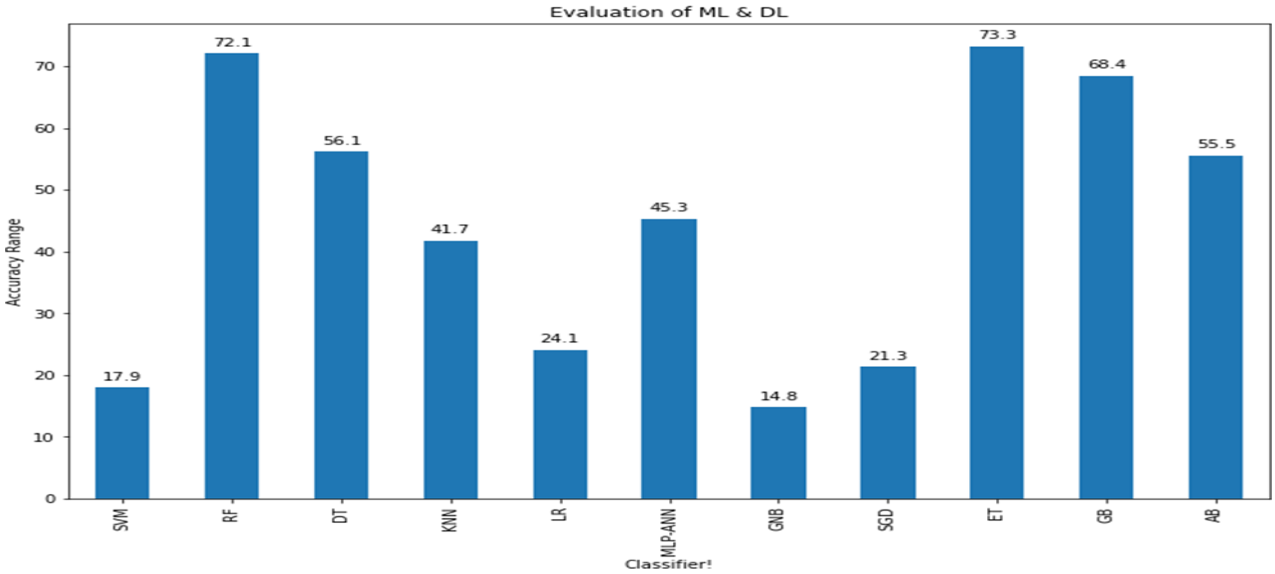


Figure 8. Evaluation of Accuracy range of different ML Classifier Algorithms

The graph above in the figure 8 depicts the comparison between the classification's algorithms used based on the accuracy of each model, respectively. After a significant comparison we have found that Extra Tree classifier has the maximum accuracy of 73.3%, hence we select the ET classifier algorithm to further analyse and show the result.

Support: The number of actual occurrences of the class in the listed dataset is known as support. Imbalanced support in the training data may imply systemic deficiencies in the classifier's recorded ratings, necessitating stratified sampling or re-balancing. Support does not adjust depending on the model; instead, it diagnoses the assessment process.

$$\text{Support} = \frac{\sigma(\mathrm{X} + \mathrm{Y})}{\text{Total}}$$

5. Conclusions

Our project's main goal is to find the most accurate machine learning classification model and incorporate it into a predicting system. The different performance measures used in determining accuracy are support, F1-score, recall, and precision. The precision of eleven machine learning classification models has been determined. Algorithms such as Gaussian Naive Bayes algorithm, Stochastic Gradient Descent classifier algorithm, Multilayer Perceptron ANN classifier, Logistic Regression, K Neighbours Classifier, and the Support Vector Machine performed poorly. The accuracy of the decision tree classifier and the AdaBoost classifier algorithm was 56% and 55%, respectively. The accuracy of the Gradient Boosting Classifier, Extra Trees Classifier, and Random Forest algorithm was 68%, 73.3%, and 72.1%, respectively. This Extra Trees Classifier with the highest accuracy is used in the dashboard to predict the level of vitamin D deficiency in a new patient. This dataset does not consider all age groups. As a result, the model shall be validated for all age group datasets in the future study. Such a framework should be accessible online for self-prediction and prevention from getting severe deficiency of Vitamin D.

References

1. S. Guo, R. Lucas, and A. Ponsonby, *"A novel approach for prediction of vitamin D status using support vector regression"*, PLoS ONE, vol. 8, no. 11, Nov. 2013, Art. no. e79970. DOI.org/10.1371/journal.pone.0079970

2. Kayhan Gonoodi, Maryam Tayefi, Maryam Saberi-Karimian, Alireza Amirabadi zadeh, Susan Darroudi, Seyed Kazem Farahmand, Zahra Abasalti, Alireza Moslem, Mohsen Nematy, Gordon A. Ferns, Saeid Eslami, Majid Ghayour Mobarhan, *"An assessment of the risk fac- tors for vitamin D deficiency using a decision tree model"*, Diabetes Metabolic Syndrome: Clinical Research Reviews, Vol- ume 13, Issue 3,2019, Pages 1773-1777, ISSN 1871- 4021, DOI.org/10.1016/j.dsx.2019.03.020.

3. Al-Asoom, Lubna Al-Hariri, Mohammed. (2017), *"The association of adiposity, physical fitness, vitamin D levels and haemodynamic parameters in young Saudi females"*, Journal of Taibah University Medical Sciences. 13. 10.1016/j.jtumed.2017.05.004.

4. Merlijn T, Swart KMA, Lips P, et al. "*Prediction of insufficient serum vitamin D status in older women: a validated model*". Osteoporosis Int.2018;29(7):1539-1547. DOI:10.1007/s00198-018-4410-3.

5. S. Alghunaim and H. H. Al-Baity, *"On the scalability of machine- learning algorithms for breast cancer prediction in big data context"*, IEEE Access, vol. 7, pp. 91535–91546, 2019.

6. Juan-Jose Beunza, Enrique Puertas, Ester García-Ovejero, Gema Vil- lalba, Emilia Condes, Gergana Koleva, Cristian Hurtado, Manuel F. Landecho, *"Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)"*, Journal of Biomedical Informatics, Volume 97, 2019, 103257,ISSN 1532-0464, DOI.org/10.1016/j.jbi.2019.103257.

7. Jorge, V. M. Castro, A. Barnado, V. Gainer, C. Hong, T. Cai, T. Cai, R. Carroll, J. C. Denny, L. Crofford, K. H. Costenbader, K. P. Liao, E. W. Karlson, and C. H. Feldman, *"Identifying lupus patients in electronic*

*health records: Development and validation of machine learning algorithms and application of rule-based algorithms",*Seminars Arthritis Rheumatism, vol. 49, no. 1, pp. 84–90, Aug. 2019.

8. Tamune, Hidetaka Ukita, Jumpei Hamamoto, Yu Tanaka, Hiroko Narushima, Kenji Yamamoto, Naoki, *"Efficient Prediction of Vita- min B Deficiencies via Machine-Learning Using Routine Blood Test Results in Patients with Intense Psychiatric Episode",* 2020 Frontiers in Psychiatry. DOI:10.3389/fpsyt.2019.01029.

9. Souad Bechrouri, Abdelilah Monir, Hamid Mraoui, El Houcine Sebbar, Ennouamane Saalaoui, Mohamed Choukri, *"Performance of Statistical Models to Predict Vitamin D Levels"*, 2019, Acts of the Second Conference of the Moroccan Classification Society, DOI.org/10.1145/3314074.3314076.

10. Ishir Bhan, Sherri-Ann M. Burnett-Bowie, Jun Ye, Marcello Tonelli, Ravi Thadhani, *"Clinical Measures Identify Vitamin D Deficiency in Dialysis"*, CJASN February 2010, DOI:10.2215/CJN.06440909.

11. N. Altman and M. Krzywinski, *"Ensemble methods: Bagging and random forests"*, Nature Methods, vol. 14, no.10, pp. 933–934, Oct. 2017.

12. M. Holick, *"Vitamin D deficiency"*, New England J. Med., vol. 357, no. 3, pp. 266–281, 2007.

13. K. M. van de Luijtgaarden, M. T. Voûte, S. E. Hoeks, E. J. Bakker, M. Chonchol, R. J. Stolker, E. V. Rouwet, and H. J. M. Verhagen, *"Vitamin D deficiency may be an independent risk factor for arterial disease"*, Eur. J. Vascular Endovascular Surg., vol. 44, no. 3, pp. 301– 306, Sep. 2012.

14. G. Sambasivam, J. Amudhavel and G. Sathya, *"A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods"*, in IEEE Access, vol. 8, pp. 109492- 109507, 2020, DOI:10.1109/ACCESS.2020.3002191.

15. Y. Li, D. Wang, Y. Yu, and L. Jiao, 2016, *"An improved artificial immune network algorithm for data clusteringbased on secondary competition selection"*, 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, pp. 2744-2751. DOI: 10.1109/CEC.2016.7744135.

16. Y.-P. Huang and M.-F. Yen, *"A new perspective of performance comparison among machine learning algorithms for financial distress prediction"*, Appl. Soft Compute., vol. 83, pp. 1056–1063, Oct. 2019.

17. Carlberg and A. Neme, *"Machine learning approaches infer vitamin D signalling: Critical impact of Vitamin D receptor binding within topologically associated domains"*, J. Steroid Biochemistry Mol. Biol., vol. 185, pp. 103–109, Jan. 2019.

18. C. Wu, W. C. Yeh, W. D. Hsu, M. M. Islam, P. A. Nguyen, T. N. Poly, Y. C. Wang, H. C. Yang, Y. Li, *"Prediction of fatty liver disease using machine learning algorithms"*, Compute. Methods Programs Biomed., vol. 170, pp. 23–29, Mar. 2019.

19. K. M. van de Luijtgaarden, M. T. Voûte, S. E. Hoeks, E. J. Bakker, M. Chonchol, R. J. Stolker, E. V. Rouwet, and H. J. M. Verhagen, *"Vitamin D deficiency may be an independent risk factor for arterial disease"*, Eur. J. Vascular Endovascular Surg., vol. 44, no. 3, pp. 301– 306, Sep. 2012.

20. Merlijn T, Swart KMA, Lips P, et al. *"Prediction of insufficient serum vitamin D status in older women: a validated model"*, Osteoporosis Int.2018;29(7):1539-1547. DOI:10.1007/s00198-018-4410-3.

21. J. Zhang, Z. Li, Z. Pu, and C. Xu, *"Comparing prediction performance for crash injury severity among various machine learning and statistical methods"*, IEEE Access, vol. 6, pp. 60079–60087, 2018.

22. B Padmaja, V V Rama Prasad, K V N Sunitha (2016), "*TreeNet analysis of human stress behavior using socio-mobile data*," in Journal of Big Data, 3(1), pp. 1-15, 2016, Springer.

23. B Padmaja, Myneni Madhu Bala, E Krishna Rao Patro (2020), "*A Comparison on visual prediction models*

for MAMO(multi activity multi-object) recognition using Deep Learning,” in Journal of Big Data, 7(24), pp. 1-15, Springer.

24. B Padmaja, V V Rama Prasad, K V N Sunitha (2020), “ *A Novel Random Split Point Procedure using Extremely Randomized Trees Ensemble Method for Human Activity Recognition*,” in EAI Endorsed transactions on Pervasive Health and Technology, 6(22), PP. 1-10.

25. B Padmaja, V V Rama Prasad, K V N Sunitha (2018), “*Machine Learning Approach for Stress Detection using Wireless Physical Activity Tracker*,” in International Journal of Machine Learning and Computing, 8(1), pp. 33-38.