IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 3, 2020, accepted June 10, 2020, date of publication June 15, 2020, date of current version June 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002191

A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods

G. SAMBASIVAM[1], J. AMUDHAVEL[2], (Member, IEEE), AND G. SATHYA[3]
[1]Faculty of Information and Communication Technology, ISBAT University, Kampala, Uganda
[2]School of Computer Science and Engineering, VIT Bhopal University, Bhopal 466114, India
[3]Indira Gandhi Government General Hospital and Post Graduate Institute, Puducherry 605001, India

Corresponding author: G. Sambasivam (gsambu@gmail.com)

ABSTRACT Vitamin D Deficiency (VDD) is one of the most significant global health problem and there is a strong demand for the prediction of its severity using non-invasive methods. The primary data containing serum Vitamin D levels were collected from a total of 3044 college students between 18-21 years of age. The independent parameters like age, sex, weight, height, body mass index (BMI), waist circumference, body fat, bone mass, exercise, sunlight exposure, and milk consumption were used for prediction of VDD. The study aims to compare and evaluate different machine learning models in the prediction of severity in VDD. The objectives of our approach are to apply various powerful machine learning algorithms in prediction and evaluate the results with different performance measures like Precision, Recall, F1-measure, Accuracy, and Area under the curve of receiver operating characteristic (ROC). The McNemar's test was conducted to validate the empirical results which is a statistical test. The final objective is to identify the best machine learning classifier in the prediction of the severity of VDD. The most popular and powerful machine learning classifiers like K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), AdaBoost (AB), Bagging Classifier (BC), ExtraTrees (ET), Stochastic Gradient Descent (SGD), Gradient Boosting (GB), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) were implemented to predict the severity of VDD. The final experimentation results showed that the Random Forest Classifier achieves better accuracy of 96 % and outperforms well on training and testing Vitamin D dataset. The McNemar's statistical test results support that the RF classifier outperforms than the other classifiers.

INDEX TERMS Machine learning algorithms, random forest classifier, severity of VDD, vitamin D.

I. INTRODUCTION

Vitamin D is an essential vitamin that has powerful influence on several parts of the human body. Nearly one billion people highly suffered from Vitamin D Deficiency across the globe [1]. Vitamin D Deficiency is associated with several auto immune disorders like cardiovascular disease, diabetes mellitus and breast cancer [2]–[4]. Though, the enormous amount of data is collected every day in the medical field, processing of the large datasets will be challenging with the traditional approaches and recent studies proved that applying machine learning models will yield better results [5]. Machine learning models will be useful in discovering new patterns of the etiology and thus preventive public health measures can be applied effectively [6]. VDD diagnosis by using machine learning models will prove to be economically efficient for better treatment.

The traditional severity prediction of VDD have used questionnaires [7] with statistical models such as Linear Regression (LR), Multivariable Adaptive regression Spline, Support Vector Regression (SVR) and support vector regression classifiers [8] to predict the severity of VDD. In previous studies, the results were compared between the statistical models and they have not used the machine learning algorithms for the severity prediction. The traditional statistical model like LR is used to predict the severity of VDD but its performance is deprived due to its predictive performance limit and many parameters. Currently, the analysis of Vitamin D status is highly expensive, and it is identified using the biochemical methods. The research gap identified urges to condense

The associate editor coordinating the review of this manuscript and approving it for publication was Haruna Chiroma.

TABLE 1. Analysis for Vitamin D deficiency prediction with various features.

S.no	Author's & Year	Classifier used	Parameters used	Sty	Spy	ACC	F1	AUC-ROC	CC	EM	CV
1	Souad Bechrouri, *et.al.*, 2019 [8]	Linear regression, Random Forest, Multivariable Adaptive Regression Spline and Support Vector Regression	age, calcium, chlorine, LDL-cholesterol, HDL-cholesterol, sex, uric acid, albumin, C-reactive protein CRP, Gamma-Glutamyl-Transferase, glucose, alkaline phosphatase, total protein, sodium, season and phosphorus, potassium	-	-	-	-	-	-	RMSE & MAE	-
2	*Gonoodi,K ,et.al.*, 2019 [9]	Decision Tree	age, waist circumference, waist to hip ratio, zinc, calcium, superoxide dismutase, full blood count, high-density lipoprotein cholesterol, red blood cells, mean corpuscular volume, mean corpuscular Hb concentration and hematocrit.	-	✓	✓	-	✓	-	-	✓
3	Akiko Kuwabara, *et.al.* 2018 [11]	Univariate and multivariate logistic regression	age, sex, season of blood drawn, exercise habit, sunscreen use, sun exposure and fish intake.	✓	✓	✓	-	✓	Chi-squared test	-	-
4	*Carlberg C, et.al., 2018* [15]	K-means & Random forests	vitamin D receptor, chromatin	-	-	-	-	-	-	-	-
5	Merlijn T, *et.al.*, 2018 [22]	Multivariable logistic regression	age, BMI, vitamin D supplement, multivitamin supplement, calcium supplementation, daily use of margarine, Fish intake, outdoors in summer, blood sampling, walking, and smoking.	✓	✓	-	-	✓	-	R^2	-
6	*Bjørn Jensen C, et.al., 2013* [13]	Univariate Regression Analysis	Dietary Intake, BMI, social occupation status, blood, parity, bed use, sunny destination travel, Ultra-violet radiation, Fish intake, intake from supplements, smoking and maternal birth.	-	-	-	-	-	PCC, Cohen's weighted kappa coefficient	-	✓
7	*Guo,Shuyu ,et.al.,* 2013 [7]	Support Vector Regression and Multiple Linear Regression	Physical activity, sun exposure, sun protection, smoking habits, diet, supplements, types of skin, height, weight, waist, and hip circumference.	✓	✓	✓	-	✓	PCC	MAD	✓
8	Tran, Bich, *et.al.*, 2013 [23]	Multivariable logistic regression	sex, ethnicity, skin color, BMI, state, residential location, ambient, physical activity, time in outdoors, Vitamin D intake and smoking.	✓	✓	✓	-	✓	Regression Coefficient	MAE	-
9	Vande Luijtgaarden, Koen, *et.al.*, 2012 [24]	Multivariable regression analyses	age, gender, ischaemic, heart failure, diabetes, hypertension, heart disease, cerebrovascular disease, renal function, smoking, and arterial disease types.	-	-	-	-	-	Chi-squared test	-	-

Abbreviations: Sty-Sensitivity; Spy-Specificity; ACC- Accuracy; F1-F1 Measure; EM-Error Measures; CV-Cross Validation; CC-Correlation Coefficient; MAD-Mean Absolute Difference; PCC-Pearson correlation coefficient.

the cumbersome analytical procedures in identifying VDD among the patients. So, for predicting the severity of VDD among the patients, we have used various machine learning models. The present study throws light on the way of identification and categorization of severity of VDD, Insufficiency and sufficiency.

Machine Learning is one of the fastest emerging recent technologies which is used in various fields due to its high performance and ease in applicability. In recent years, the applications and usage of machine learning in the medical field is very high. The main objective of machine learning is to learn from the input data which is usually called training data and make future predictions with the new data.

The previous studies used only the traditional statistical models to predict the severity of deficiency in Vitamin D datasets. The traditional works applied the statistical models on vitamin datasets with a smaller size [9]. Probably if the traditional methods applied to larger datasets then there is a chance of degradation of the performance. To our knowledge, this is the first study to predict the severity of VDD which compared it with various ML models. Furthermore, there will be variation in its etiology of different geographical locations due to diversified climatic conditions.

The objectives of the current study is to predict the severity of VDD datasets by using various types of machine learning models like Linear Regression, k-nearest neighbor, Gaussian Naïve Bayes, Decision Tree, Random Forest classifier, Multilayer Perceptron, AdaBoost Classifier, Stochastic Gradient Classifier, Boosting Classifier, Linear Discriminant Analysis, Support Vector Machine, and Gradient Boosting classifier. Secondly, to compare the results of machine learning models with various performance measures like Precision, Recall, F1-measure and ROC curves [10] in predicting Vitamin D deficiency severity as well as with different error measures, Cohen's Kappa and correlation coefficient to identify the best machine learning classifier in the prediction severity of VDD.

The primary objective of the study is to predict the severity of VDD using ML classifiers and validate the empirical results with statistical significance test as well using the error measures. The final objective is to identify the best ML classifier in the prediction of severity in VDD.

The presentation of the paper is organized as follows: related works of this study are presented in Section II. The Machine Learning Models used in this research work were described in Section III. The Experimentation and result analysis are presented in Section IV. Finally, Section V concludes this article.

II. RELATED WORKS

In this section, we are going to discuss mainly on the previous works on VDD. The traditional work mostly used multivariate regressions classifier and least used models are random forest tree, k-means, support vector regression, and decision tree. All the previous studies used different types of parameters to predict vitamin D deficiency. From the previous study, so far there is no research have been observed on VDD severity as multiclass classification. In Table 1, We have presented a survey related to vitamin D deficiency with parameters, performance & statistical measures. They have used the Pearson correlation coefficient to predict the

correlation between the actual and predicted serum values. Kuwabara *et al.* [11], used a list of questionnaires to predict vitamin d deficiency severity. The prediction model was developed by using logistic regression. They have used sensitivity, specificity and Area Under the curve (AUC) for evaluating the model [12]. Gonoodi *et al.* [9], used decision tree models to assess the risk factors correlated with Vitamin D deficiency. They have used accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curve. Bjorn Jensen used serum 25-hydroxy-vitamin D to predict Vitamin D Deficiency using Univariate regression algorithm [13]. They have used the Pearson correlation coefficient as well as Cohen's weighted kappa coefficient to evaluate the model.

Tamune *et al.* [14], used random forest classifier to predict the vitamin B deficiency in patients who had psychiatric symptoms and they have used sensitivity, specificity, and Area Under the curve (AUC) for evaluating the model. Carlberg *et al.* [15], claimed the usage of supervised and unsupervised machine learning algorithms for inferring VDD. Mainly random forests tree had been used for the classification of constant data. Random forest classifier includes the knowledge of the examined biological system [16].

The multivariate and time series analysis was used for the detection of vitamin D levels in patients [17]. The biochemical parameters such as age, calcium, chlorine, LDL-cholesterol, HDL-cholesterol [8], BMI, and WC [18] in the prediction of VDD using statistical models like Random forests, Support vector regression, linear regression, and Multivariable Adaptive Regression Spline. They have used error measures like Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) for comparison and there is a weak correlation between the vitamin D and biochemical parameters [8]. Also, the authors have compared the correlation of biochemical parameters using a heat map. The prediction model was developed by using multivariable logistic regression. They have used sensitivity, specificity, and Area Under the ROC curve for evaluating the model.

In earlier studies there has been many literatures found on class imbalance [19]. Class imbalance is denoted as one class is significantly higher than other class. We have used the imbalanced dataset which the class distribution was shown in Table 2. If there is a class imbalance, the accuracy of the classifier decreases. To overcome this problem, we have used oversampling method called Synthetic Minority Oversampling technique (SMOTE) to balance our dataset [20]. Too much oversampling results in overfitting problem, so that we have not applied SMOTE to the test set. The prediction model was developed by using multivariate logistic regression [8]. To estimate the best prediction model, statistical test like McNemar's test were conducted [21]. This will help us to validate the empirical results. Cross-validation is one of the important resampling technique to evaluate the skill of ML models. It has the parameter k which is used to split the datasets into groups. In our datasets, we have used k=10-fold cross-validations which is most reliable to evaluate the models and most of the previous study reported to use

TABLE 2. Vitamin D Deficiency Severity Level in dataset.

Deficiency Severity	Description	Frequency	Percentage (%)
Level 1	Sufficiency	1329	43.65
Level 2	Insufficiency	957	31.43
Level 3	Deficiency	575	18.88
Level 4	Severe Deficiency	183	6.01
Overall	-	3044	100

10-fold cross-validation [38]. The procedure of the k-Fold cross validation is to split the datasets into training and testing datasets into k groups. The ML models are fitted into the training sets and evaluate it with the testing set to find out its accuracy of the specific model.

III. METHODOLOGY

A. STUDY POPULATION

For our research work, we have obtained the datasets and the study protocol was approved by the Institute Ethical Committee of Bharathidasan Govt. College for Women, Puducherry, India. A Vitamin D dataset which was collected from a total of 3044 college students aged between 18-21 years was used as primary data. In this study, we used 11 input parameters like Age (18-21), Sex (male/female), Weight (61-91 kgs), Height (1.48-1.73m), BMI (25.94-34.81 kg/m^2), Waist Circumference (58-92 cm), Body Fat (21.60-41.20 %), Bone Mass (2.00-3.60), Exercise (yes/no), Sunlight Exposure/Day (5.0-30 hrs) and Milk Consumption (0-500).

The p-value have computed for the used variables using one sample T-test and the values are Age (=1.00), Sex (0.00041), Weight (0.0207), Height (0.0207), BMI (0.0041), Waist Circumference (0.003), Body Fat (0.0978), Bone Mass (0.00013), Exercise (=1.00), Sunlight Exposure/Day (0.0033) and Milk Consumption (0.0418). The p-value less than 0.05 is statistically significant. The Vitamin D deficiency severity data consists of four levels which is shown in Table 2. Level 1 is Sufficiency class which accounts nearly 43.65 % of the data. Level 2 is Insufficiency class which accounts nears to 31.43 % of the data. Level 3 is the Deficiency class which accounts nears to 18.88 % of the data. Level 4 is Severe Deficiency class which accounts nears to 6.01 % of the data.

B. DATA PREPROCESSING

For any research work, data is the main component for analysis and prediction to meet the demands. Data preprocessing is the technique that ensures the datasets have all the required data in an appropriate format without any missing values [25]. This data preprocessing is achieved through a sequence of steps to ensure that the datasets are validly shown in Fig. 1.

The steps involved in data preprocessing are the required libraries should be imported, reading the data, checking for the null values, categorizing the data, standardize the data and finally splitting the data. The data will be split into two

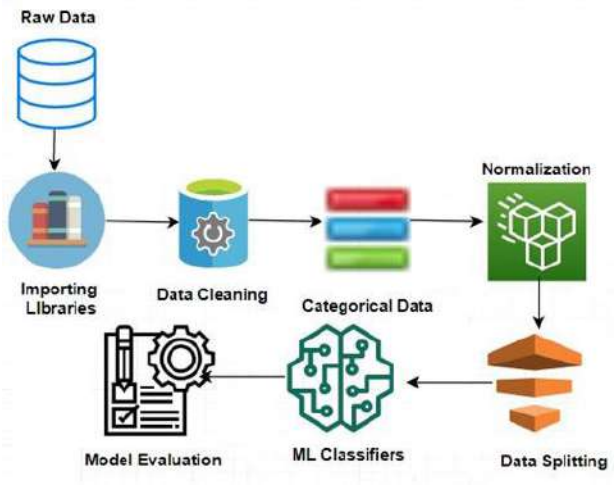

FIGURE 1. Data processing framework for VDD.

phases namely training data and test data. The ML models are applied to both training and testing data to check the accuracy. In the process of data cleaning, we concentrated on two issues one is null values and outliers [25].

The null values are identified with the use of python preprocessing libraries with a class called imputer. The imputer parameter handles the null values in the vitamin D datasets. The typical way to identify the null values is to get the mean of the appropriate column in the dataset and to replace the missing data. The next step is to deal with the categorical data which has text values and it affects the performance of the classifiers. So, the next step is to encode the categorical values by using the Label Encoder class in the scikit library. The feature scaling is the process of applying normalization to the independent variables within the range of 0 to 1 [26]. The normalization formula is in Equation (1) is shown, where x is the original value and N is the normalized value [27].

$$N = \frac{x - \min(x)}{max(x) - \min(x)} \tag{1}$$

Features plays an important role in machine learning classification and it will remove the irrelevant features. The use of feature selection helps in reducing the training time, decreases overfitting and increases the performance of the classifier.

We have used the Recursive Feature Elimination (RFE) for feature selection by using the class sklearn.feature_selection. RFE [28] and 11 features were extracted. The RFE method eliminate the least used features in a recursive manner and the results were presented in the Table 3.

C. MACHINE LEARNING MODEL BUILDING

The main purpose of this study is to detect the deficiency severity in VDD from different parameters accurately with the use of best machine learning models.

a) Logistic Regression
Logistic Regression (LR) is a statistical model that comes under a supervised ML technique which uses the logistic function to solve multiclass classification problems. We implemented LR using the multiclass parameters [29], [30] by using sklearn.linear_model.LogisticRegression. Let us consider an example $x \in R^m$, then the score represented using y= M^Rx+f where matrix $V \in S^{C*M}$ and vector $f \in S^C$ are variables learned from the datasets. The Vitamin D deficiency severity of each class is given by the sigmoid of each individual class score $z(y_c = 1) = \sigma(y_c) = \sigma\left(V_C^T x + f\right)$. We have used One-vs-rest classifier for training sets and testing datasets which trains single classifier for each deficiency severity class.

b) K-Nearest Neighbor
The K-Nearest Neighbor (KNN) algorithm is a supervised machine learning algorithm [31], [32] which is used for classification and regression problems. The KNN algorithm looks for close proximity to the datasets. Initially, it calculate the mathematical values of the nearest data points and the nearest neighbor contributes more than the distant ones. The test data will be compared with the data points of the training sets then it finds the probability of similar data points. The algorithm classifies the output based on the points which have the highest probability. The irrelevant features may affect the performance of the KNN, so the relevant features are considered from the datasets. We have implemented KNN using import sklearn.neighbors. KNeighbors Classifier for the experiment.

c) Gaussian Naive Bayes
Gaussian Naïve Bayes is a special type supervised machine learning classifier used for classification problem which follows the probabilistic method. We have implemented GNB using sklearn.naive_bayes import GaussianNB for the experiment. It follows the preceding and future probability of the different severity classes in the vitamin D training and testing dataset respectively.

$$p(x = v|S_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{2}$$

where, S-Different Severity measures; μ_{k-} Mean of the different variables; v- probability distribution; σ_k^2 – Variance of the values.

TABLE 3. Feature selection by RFE.

ML Classifier	Index for the selected features
ET	[9 8 4 2 6 1 0 3 5 7 2]
LR	[9 8 4 2 6 1 0 3 5 7 2]
RF	[7 9 1 2 5 1 0 3 8 4 6]
GB	[9 8 7 2 2 1 0 4 6 5 3]
SGD	[2 7 6 8 1 5 4 1 1 3 9]
AB	[9 8 7 2 1 3 0 6 2 5 4]
DT	[9 8 4 2 6 1 0 5 2 3 7]

d) AdaBoost Classifier
AdaBoosting Classifier is an ensemble machine learning approach which makes itself a strong learner by combining the weak learner model. A machine learning algorithm will act as a base learner when it accepts the weights from the training data. We have implemented AdaBoost Classifier by using sklearn.ensemble import AdaBoostClassifier. Initially, the training set will be selected randomly, and the model trains it iteratively. The misclassified observations are assigned with higher weight and it will get a higher probability in the next process of iteration. This procedure will continue until the training set data fits in the model without any fault.

e) Bagging Classifier
A Bagging Classifier (BC) is an ensemble machine learning classifier which the original dataset is divided into training and testing data. The existing base classifiers shall be fitted on the random subsets of the training data and finally the individual prediction shall be aggregated to make a final prediction. The Bagging classifier is implemented by using sklearn.ensemble import BaggingClassifier.

f) ExtraTrees Classifier
The ExtraTrees Classifier (ET) is an ensemble machine learning method that uses meta-estimator which fits into a randomized on a various subset of the training dataset and the prediction accuracy is enhanced and the overfitting is controlled [33]. The ExtraTrees classifier is implemented by using sklearn.ensemble import ExtraTreesClassifier. By default, the bootstrap will be set to false to build several trees. The training datasets will be randomly split into subsets and the randomness is obtained from the randomly split subsets and not from the bootstrapping of the data.

g) Stochastic Gradient Descent
Stochastic Gradient Descent is machine learning algorithm which is used for discriminative study of direct classifier which uses convex loss function.The advantage of using SGD is to easy implementation and efficiency is improved. The stochastic gradient descent is implemented using SGD classifier which helps in finding various loss function and penalties. The SGD classifier is implemented using from sklearn.linear_model import SGDClassifier.

h) Decision Tree
Decision Tree Classifier is an eminent supervised ML tool that is used for solving classification problems and it has a tree-like model or graphs. The DT can capture the decision-making knowledge from the given data. We have used DT using sklearn.tree import DecisionTreeClassifier for the experiment. In DT that every branch indicates the output of the test set and every leaf node represents the particular label. The classification rules are represented by the path from the root node to the leaf node. For our vitamin D deficiency severity modeling, each node in the tree predicts the deficiency severity and each branch indicates the states of the variable. The Vitamin D dataset has the four deficiency severity types as the outcome and has many independent variables, (c,T) $= (c_1, c_2, c_3, c_4 \ldots, \mathrm{T})$, where, T is the deficiency severity variable and the vector c is comprised of several independent variables like $c_1, c_2, c_3, c_4 \ldots c_n$ used for classification.

i) Random Forest Classifier
Random forest classifier (RF) as proposed by Breiman [34], is an ensemble machine learning method used for solving classification problems. RF constitutes many decision trees randomly from the training set and then it aggregates the values from different decision trees and predicts final severity deficiency as the outcome. The parameters considered for RF are n_estimators (n=10), criterion, minimum samples split (split=2) and minimum sample leaf (leaf=1) in the dataset. The previous studies showed that the RF classifier outperforms well with the other classifiers [41]. We have used RF using sklearn.ensemble import RandomForestClassifier for the experiment.

j) Multi-Layer Perceptron Classifier
Multi-Layer perceptron (MLP) is a feed-forward artificial neural network and it is inspired by the biological brain, that tries to mathematically express the real brain that maps the set of inputs to the corresponding output. The MLP has three layers namely input, hidden and output layer. It has one input and output layer, but it has one or more hidden layers. The perceptron has inputs$\{x_1, x_2, x_3, \ldots x_n)$ and each input has corresponding weights $\{w_1, w_2, w_3, \ldots w_n)$. It has a summation processor with function $(g\,(x) = \sum_{i=0}^{n} w_i.x_i)$ and it has an activation function. It has parameters like iterations, learning rate, input/output layers, different deficiency severity classes, and activation function. The training data and labels will be provided to the classifier for training. We have used MLP using sklearn.neural_network import MLPClassifier for the experiment. The MLP models use the log loss function to predict the accuracy and try to minimize the values where the good model has a log loss of 0. If the log loss increases, then there is a divergence in the prediction with the actual label.

k) Linear Discriminant Analysis
Linear Discriminant Analysis (LDA) is an unsupervised machine learning algorithm used for finding the linear combinations of the parameter for the multiclass classification problem [35]. It calculates the differences in the mean of different classes (Sufficiency, Insufficiency, Deficiency and Severe Deficiency) from Vitamin D dataset. Secondly, it calculates the distance between the mean and sample of a particular class. The dimensionality is the third step which increases and decreases between the class variance. We have

TABLE 4. Comparison of dataset splits of. Mean, Min, Max, Std Dev for original vs training set vs testing set.

Dataset Split up	Statistical Analysis	AGE	SEX	WEIGHT	HEIGHT	BMI	WC	BF	BM	EXERCISE	SUNLIGHT	MC
Original Dataset (100 %)	Mean	18.28	0.41	78.10	1.62	29.68	79.55	29.17	2.72	0.24	17.07	211.36
	Minimum	18.00	0.00	61.00	1.48	25.94	58.00	21.60	2.00	0.00	5.00	0.00
	Maximum	21.00	1.00	92.00	1.73	34.81	92.00	41.20	3.60	1.00	30.00	500.00
	Standard deviation	0.45	0.49	7.24	0.06	2.47	7.60	3.66	0.46	0.43	5.54	83.51
Training Set (80 %)	Mean	18.27	0.37	77.70	1.62	29.46	79.95	29.51	2.76	0.31	17.58	212.80
	Minimum	18.00	0.00	61.00	1.48	25.94	58.00	21.60	2.00	0.00	5.00	0.00
	Maximum	19.00	1.00	92.00	1.73	34.81	92.00	41.20	3.60	1.00	30.00	500.00
	Standard deviation	0.45	0.48	7.49	0.06	2.44	6.68	3.38	0.45	0.46	5.35	80.08
Testing Set (20 %)	Mean	18.33	0.55	79.69	1.62	30.55	77.99	27.82	2.55	0.00	15.06	205.69
	Minimum	18.00	0.00	69.00	1.48	26.01	58.00	23.00	2.00	0.00	10.00	0.00
	Maximum	19.00	1.00	87.00	1.70	33.98	91.00	38.80	3.20	0.00	25.00	400.00
	Standard deviation	0.47	0.50	5.93	0.06	2.37	10.31	4.33	0.49	0.00	5.80	95.72

used LDA using *sklearn.discriminant_analysis*for our experiment.

l) Support Vector Machine
Support Vector Machine (SVM) is best known supervised machine learning classifier as proposed by Cortes &Vapnik [36] used to solve classification and regression problem The purpose of SVM is to find the hyperplane with the number of given independent variables and it distinctly categorizes the data points. The implementation of SVM is done by using *sklearn.svm import SVC* for our experiment. The distance between the classes should be maximized and we have use linear SVM. Hyperplane are used to classify the data points which is considered as decision boundaries. The points which falling both sides of the hyperplane are considered as different classes. The hyperplane dimension will depend on the number of features. The data points which nearer to the hyperplane are called support vectors. By using a support vector, the margin of the classifier is maximized and as well as it changes the hyperplane position.

m) Gradient Boosting
Gradient Boosting (GB) [37] is an ensemble machine learning technique that is used to solve classification and regression problems. It collects the weak points and ensemble them into strong points. We implemented GB using sklearn.ensemble import GradientBoosting-Classifier. GB uses three features like loss function, weak learner, and additive model. The loss function is used to measure how the coefficient are fitting with the training and testing data. The loss function depends on the problem where classification uses logarithmic loss and regression uses squared error.GB identifies the shortcomings by using the high weighted data points along the loss functions.

IV. EXPERIMENTATION AND RESULTS ANALYSIS

A. EXPERIMENTS

The VDD severity datasets were divided randomly into training (80 %) and testing (20 %) dataset with the ratio of 4:1. The different machine learning and statistical prediction algorithms were applied to the training dataset. After training the models with the training dataset, the trained models were applied to predict the deficiency severity with the testing dataset. We have done 10 fold cross-validation experiments with different samples in training and testing datasets to minimize the prejudice related to the randomly separated dataset. The most significant task is to predict the performance of the important indicator in the severity levels (1-4). For comparison analysis of our experiment the standard deviation and mean classification of fifteen estimates were considered for better prediction [39].

The experiments were set up for the evaluation of the different ML (KNN, DT, SVM, MLP, GNB, GB, AB, BC, ET, SGD & RF) classifier over the vitamin D datasets. The experiments were carried out using python computer programming language. We have used scikit-learn library in python for implementing the machine learning algorithms which we have used for our research. In Table 4 and Fig. 2, we have shown the different statistical information like min, max, mean, and standard deviation of the training and testing dataset. There is a variation in the standard deviation of all the features in the training and testing datasets and this variation in both datasets helps us to identify the best ML models.

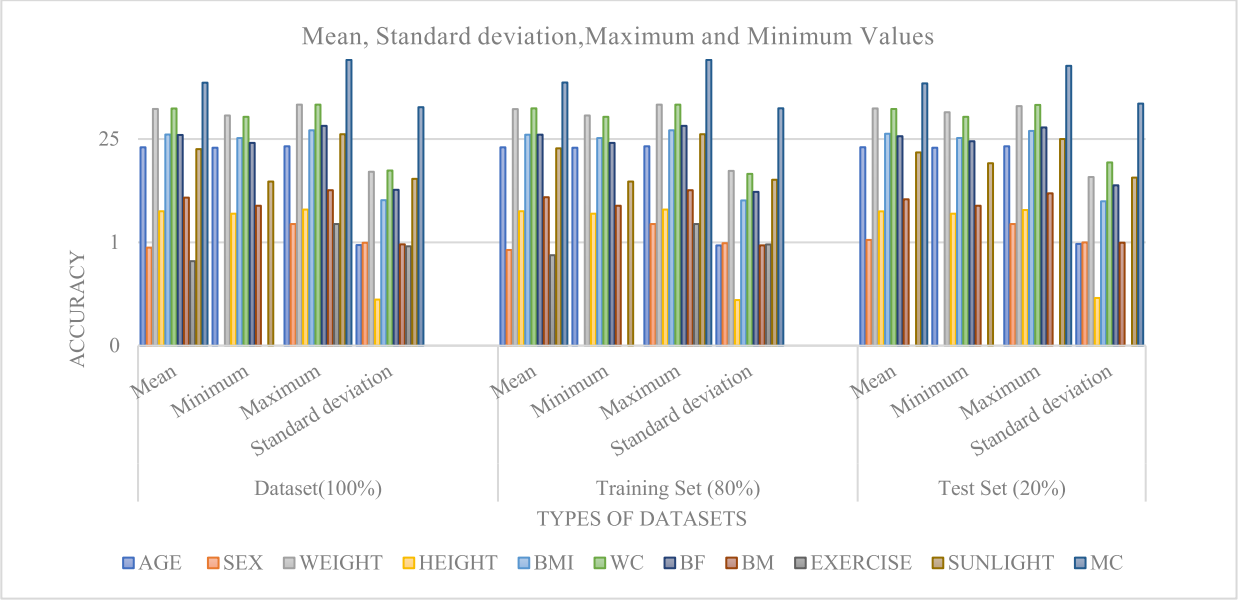

FIGURE 2. Comparison of Dataset Splits of. Mean, Min, Max, Stdev for Original vs Training set vs Testing set.

B. RESULT ANALYSIS

In this section, we discuss about the performance measures, statistical test and error measures with the comparison of accuracy of training versus testing datasets and different error measures for different machine learning models that we applied in our of Vitamin D dataset [40].

1) PERFORMANCE MEASURES

a: PRECISION

The precision is defined as the actually predicted true positive values to the overall sum for true predicted positive values and false positive values.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

w*here, TP-True Positive; FN-False Positive*.

b: RECALL

The recall is defined as the actually predicted true positive values to the overall sum for true predicted positive values and false negative values.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

where, TP- True Positive; FN -False Negative.

c: F1-MEASURE

The F1 measure is defined as the harmonic mean of precision and recall

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

d: ACCURACY

$$Accuracy = \frac{1}{n_s} \sum_{i=0}^{n_s} (yi = yx) \tag{6}$$

where yi -predicted value of i_{th} sample;yx – Corresponding true value n_s – Correct predictions of n samples.

e: RECEIVER OPERATING CURVE (ROC)

ROC curves are used to evaluate the performance of multi-class classifier problems. The ROC curve has false positive rate on X-axis and true positive rate on Y-axis. In the ROC curve topmost left edge of the plot considered to be the ideal point and the steepness of the curve also very important since the TPR value should maximize and the FPR should be minimized. The ROC curve for multiclass (Sufficiency-0, Insufficiency-1, Deficiency-2 and Severe Deficiency-3) for all the ML algorithms (KNN, DT, SVM, MLP, GNB, GB, AB, BC, ET, SGD & RF) as shown in the Fig. 3-15. The RF and DT classifier achieves the highest performance and GNB gets the lowest performance. The Random forest ROC curve achieved the best performance when compared to other models. The ROC curve of Random forest classifier (98 %) accurately predicts the vitamin D deficiency severity shown in Fig. 11. The next to RF is KNN and GB classifier. The ROC curve of the GNB classifier is very low.

Table 5 and Fig. 16 show precision, recall, and F1-Measure of multiclass for different ML models. In our dataset, we have multiclass (Deficiency, Severe, Sufficiency and Insufficiency). We have used the performance metrics such as precision, recall, F1-Measure to evaluate the performance of different classes in the Vitamin D datasets. The RF model has the highest precision accuracy of 99 % for all classes. The LDA model has the lowest accuracy of 83 %. The

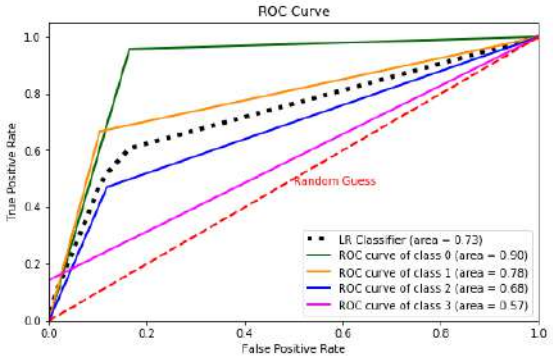

FIGURE 3. ROC curve of LR.

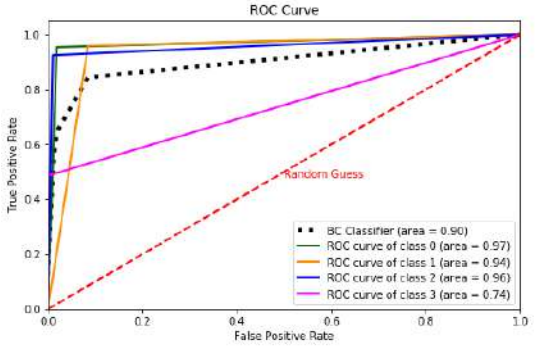

FIGURE 4. ROC curve of BC.

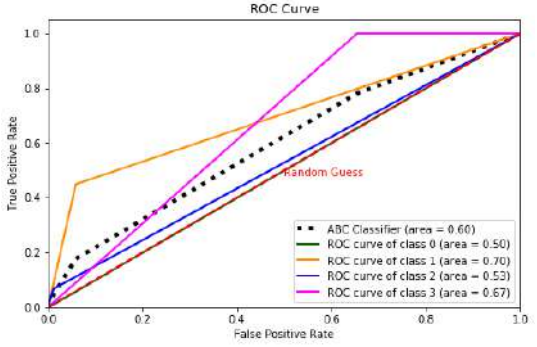

FIGURE 5. ROC curve of AB.

GNB model has the highest precision accuracy of 81% for the class Deficiency and the LDA model has the lowest accuracy of 61 %. The KNN, DT, RF, LDA model has the highest precision accuracy of 96 % for the class Severe Deficiency.

The RF model has the highest recall accuracy of 99 % for the class Insufficiency and the GNB model has the lowest accuracy of 65 %. The DT model has the highest recall accuracy of 96 % for the class sufficiency and the GNB model has the lowest accuracy of 49 %. The DT, GB model has the highest recall accuracy of 96 % for the class Deficiency and LR.LDA model has the lowest accuracy of 62 %. The KNN, GNB, DT, RF, GB model has the highest recall accuracy

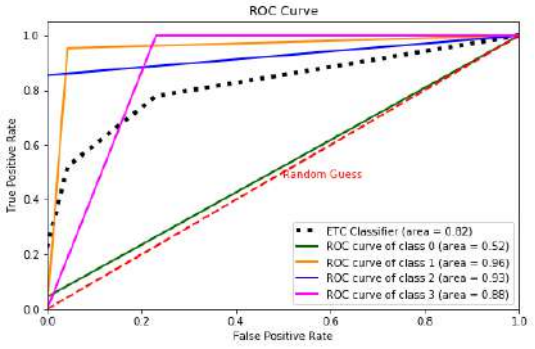

FIGURE 6. ROC curve of ET.

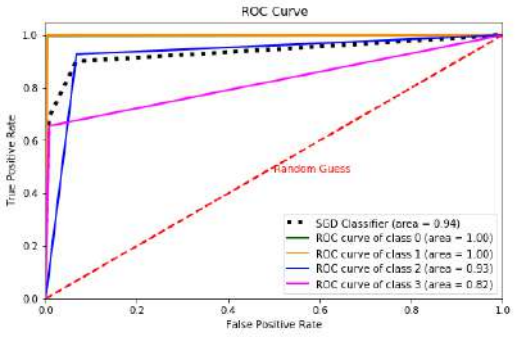

FIGURE 7. ROC curve of SGD classifier.

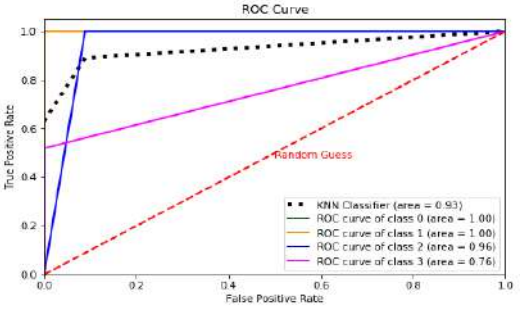

FIGURE 8. ROC curve of KNN classifier.

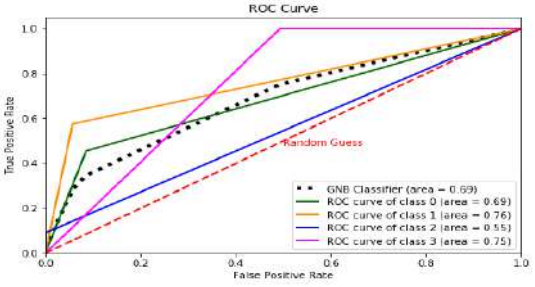

FIGURE 9. ROC curve of GNB.

of 98 % for the class Severe Deficiency and LR, LDA model has the lowest accuracy.

The RF model has the highest F1-measure accuracy of 94% for the class Insufficiency and the GNB model has the

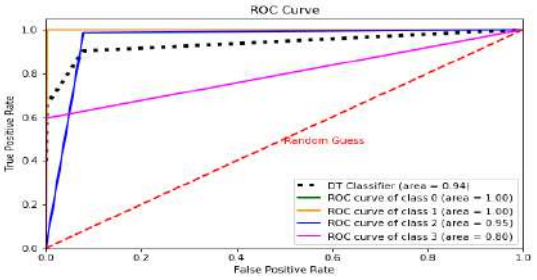

FIGURE 10. ROC curve of DT.

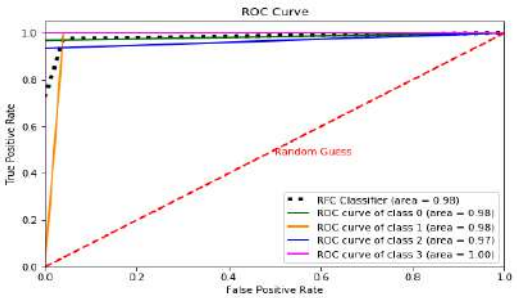

FIGURE 11. ROC curve of RF.

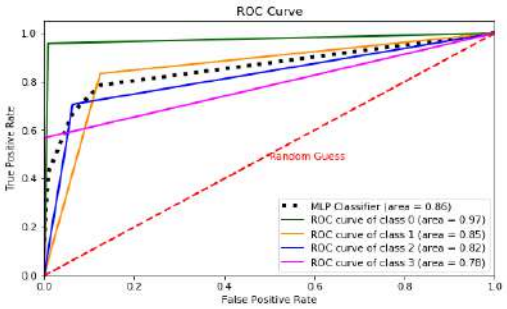

FIGURE 12. ROC curve of MLP.

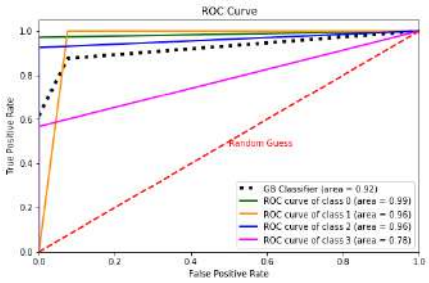

FIGURE 13. ROC curve of GB.

lowest accuracy of 72 %. The RF, GB model has the highest F1-measure accuracy of 97 % for the class sufficiency and the GNB model has the lowest accuracy of 63 %. The KNN, RF model has the highest F1-measure accuracy of 97 % for the class Deficiency and the GNB model has the lowest accuracy of 17 %. The KNN, DT, RF model has the highest F1-measure accuracy of 98 % for the class Severe Deficiency and LR, GNB model has the lowest accuracy of 13 %.

In Table 6 and Fig. 17, we have shown the classification accuracy of different models with respective different performance measures. The RF model has the highest precision

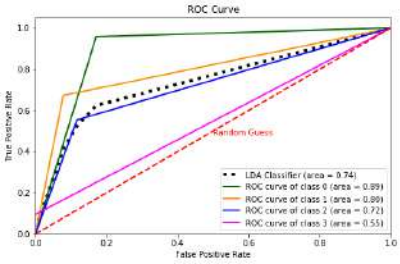

FIGURE 14. ROC curve of LDA.

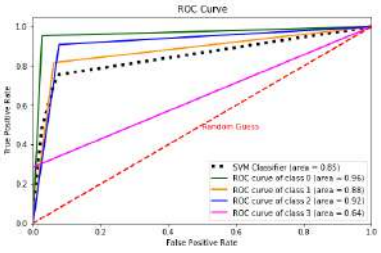

FIGURE 15. ROC curve of SVM.

with overall accuracy of 96 % for the Vitamin D datasets and LR model has the lowest accuracy of 75 %. The RF model has the highest recall accuracy of 96 % for the class sufficiency and the GNB model has the lowest accuracy of 45 %. The RF model has the highest F1-Measure accuracy of 96 % and the GNB model has the lowest accuracy of 53 %. The KNN, RF, GB model has the highest ROC accuracy of 97 % and GNB model has the lowest accuracy of 70 %. When compared to other ML models RF model achieves the highest accuracy of 94 % in the vitamin D datasets. The RF does not make any conventions about the functional attributes of the data and it need less variables to achieve highest accuracy in the prediction of severity in VDD.

The comparison of training dataset and test dataset with different ML models were shown in Table 7 and Fig. 18. The overall prediction accuracy of training models was ranging from 45.82 % to 96.40 %. The overall prediction accuracy of training models was ranging from 43.55 % to 94.89 %. RF, GB, and KNN perform well on the training and testing dataset but when comparing the range between the training and testing RF algorithms outperforms first [41] and the second, third place goes to GB and KNN.

The Gaussian Naive Bayes algorithms performed very poorly in both training and test vitamin D dataset. When comparing the difference between the training and testing dataset with GNB is good and it fails with the overall accuracy. All the other algorithms have the same difference in both training and testing dataset and some of the algorithms have high accuracy. We have classified the algorithms into four categories (Excellent, Good, better, and average) based upon the performance accuracy with the training and test dataset. The RF,GB,KNN,ET & SGD (96.4 % to 95.21 %) are the excellent algorithms and DT,BC,MLP,SVM (92.07 % to 83.90 %) are good and LR,LDA (78.48 % to 78.16 %) are better than GNB and AB (45.82 % to 51.04 %) which

TABLE 5. PRECISION, RECALL AND F1-Measure l predicted by different ML models under a different classification.

Machine Learning Models	Multiclass (n=4)	Precision	Recall	F1-Measure
LR	Sufficiency	0.83	0.95	0.88
	Insufficiency	0.77	0.73	0.75
	Deficiency	0.68	0.62	0.65
	Severe	0.00	0.00	0.00
KNN	Sufficiency	0.97	0.94	0.96
	Insufficiency	0.87	0.94	0.90
	Deficiency	0.99	0.95	0.97
	Severe	0.98	0.97	0.97
BC	Sufficiency	0.98	0.95	0.96
	Insufficiency	0.97	0.94	0.92
	Deficiency	0.89	0.93	0.95
	Severe	0.85	0.92	0.94
AB	Sufficiency	0.88	0.49	0.63
	Insufficiency	0.82	0.65	0.72
	Deficiency	0.98	0.09	0.17
	Severe	0.63	0.65	0.13
SGD	Sufficiency	0.87	0.48	0.63
	Insufficiency	0.81	0.62	0.71
	Deficiency	0.99	0.08	0.16
	Severe	0.05	0.75	0.14
ET	Sufficiency	0.97	0.95	0.95
	Insufficiency	0.89	0.94	0.91
	Deficiency	0.90	0.95	0.93
	Severe	0.91	0.48	0.62
SGD	Sufficiency	0.97	0.95	0.92
	Insufficiency	0.89	0.94	0.91
	Deficiency	0.9	0.95	0.93
	Severe	0.91	0.48	0.62
GNB	Sufficiency	0.89	0.49	0.63
	Insufficiency	0.82	0.65	0.72
	Deficiency	0.83	0.09	0.17
	Severe	0.07	0.68	0.13
DT	Sufficiency	0.95	0.96	0.96
	Insufficiency	0.90	0.82	0.86
	Deficiency	0.88	0.96	0.92
	Severe	0.98	0.97	0.97
RF	Sufficiency	0.99	0.95	0.97
	Insufficiency	0.89	0.99	0.94
	Deficiency	0.99	0.95	0.97
	Severe	0.99	0.99	0.99
MLP	Sufficiency	0.97	0.95	0.96
	Insufficiency	0.89	0.94	0.91
	Deficiency	0.9	0.95	0.93
	Severe	0.91	0.48	0.62
LDA	Sufficiency	0.83	0.95	0.89
	Insufficiency	0.78	0.73	0.76
	Deficiency	0.67	0.62	0.64
	Severe	0.75	0.10	0.17
SVM	Sufficiency	0.89	0.94	0.91
	Insufficiency	0.85	0.76	0.8
	Deficiency	0.74	0.84	0.79
	Severe	0.00	0.84	0.82
GB	Sufficiency	0.99	0.94	0.97
	Insufficiency	0.88	0.94	0.91
	Deficiency	0.94	0.96	0.95
	Severe	0.95	0.95	0.94

is considered to be the average. The performance accuracy of training and testing dataset with different ML models as shown in the Fig. 14.

2) STATISTICAL SIGNIFICANCE TEST

McNemar's is a paired hypothetical test that is applied to both training and testing set with respect to different machine learning models [42]. The RF classifier empirical result (accuracy=96.40) and p-value ($p<0.035$) results were validated with statistical a hypothesis test. Thus, the statistical test was used to validate the empirical results. The empirical results and statistical hypothesis test confirm the RF is the best classifier in the prediction of severity in VDD. The statistical hypothesis test helps in comparing and choosing the

TABLE 6. Various Performance measures of different machine learning models.

Machine Learning Models	Precision	Recall	F1-Measure	ROC	Accuracy
LR	0.75	0.78	0.76	0.73	0.74
KNN	0.95	0.95	0.95	0.93	0.93
BC	0.89	0.95	0.92	0.90	0.93
AB	0.77	0.62	0.62	0.62	0.51
ET	0.88	0.95	0.92	0.98	0.94
SGD	0.87	0.94	0.91	0.94	0.93
GNB	0.87	0.45	0.53	0.69	0.43
DT	0.92	0.92	0.92	0.93	0.90
RF	0.96	0.96	0.96	0.98	0.94
MLP	0.93	0.83	0.86	0.84	0.91
GB	0.95	0.95	0.95	0.91	0.93
LDA	0.79	0.78	0.77	0.74	0.76
SVM	0.81	0.84	0.82	0.85	0.81

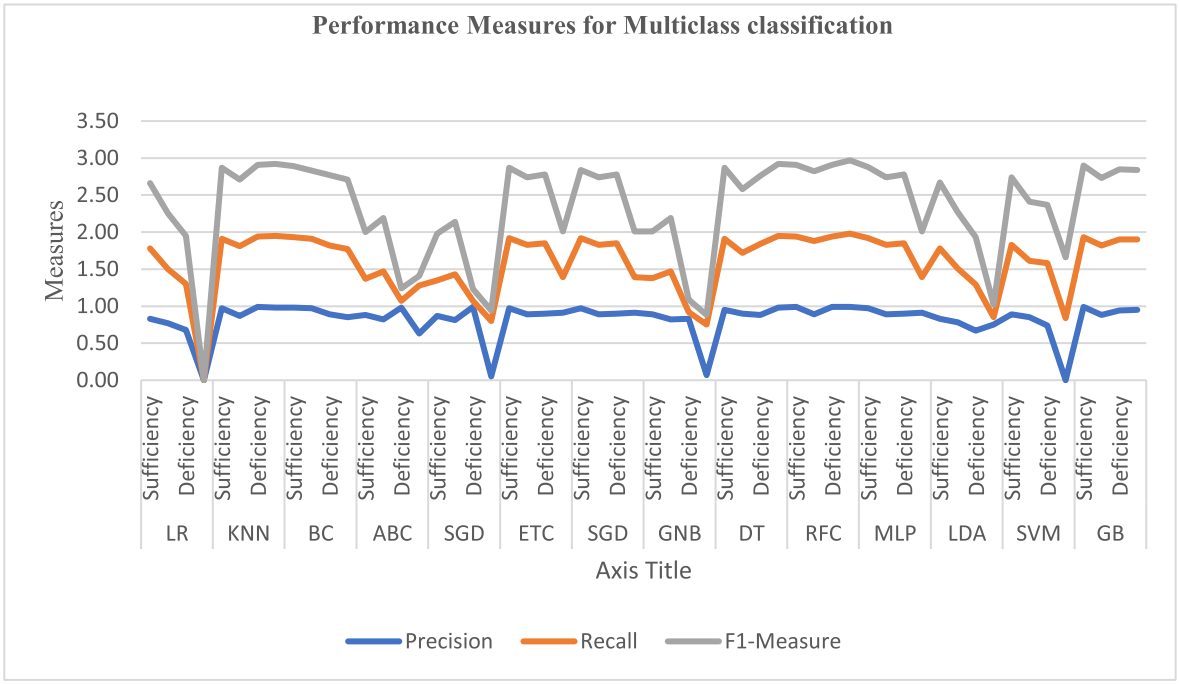

FIGURE 16. Precision, Recall, and F1-Measure predicted by different ML models under different classification.

best classifiers. It is used to check the statistical validity of the results obtained from different machine learning models. The obtained p-value using the hypothetical test can be interpreted using p>alpha which fails to reject H0 and has no difference. On the other hand, p<=alpha which rejects H0, that there is a significant difference. The VDD severity prediction is found to be statistically different when the p-value is lesser than 0.05 from the ground truth by using McNemar's test. The McNemar's model finds the errors made by the classifiers using the continency table which has cell values of No/Yes and Yes/No. This tests probably check the significant difference in the counts in these cells. As a conclusion, if the counts are similar then both classifiers have the same error with same proportion and in this case the null hypothesis will not be rejected. Hence, we can classify the output from this statistical test as classifiers have a similar proportion of error and different proportion of errors on the test set. We can classify the output from this statistical test as the cell count is not in a similar proportion of error then the null hypothesis will be rejected. We have implemented McNemar's test using python by mcnemar() Statsmodels function and it takes contingency table as a parameter which will return the test statistic and p-value. The McNemar's test statistic and p-value were shown in Table 8 and the p-value of RF,ET,AB, and GB

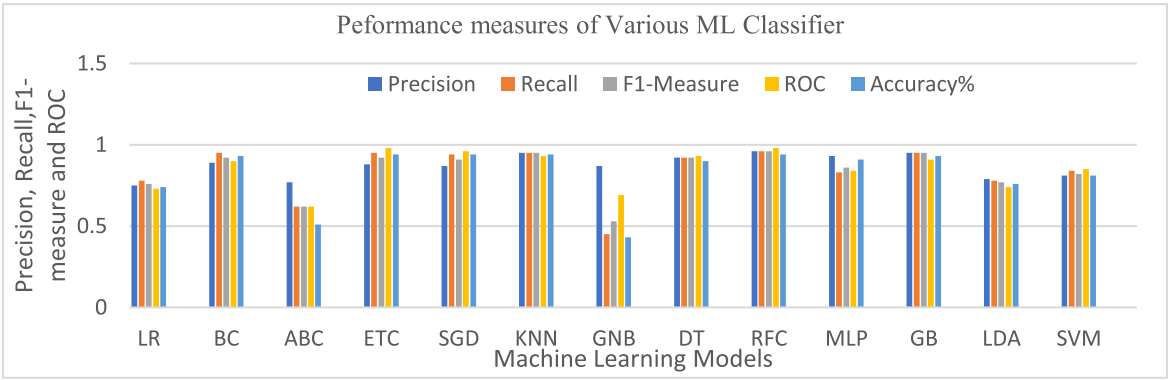

FIGURE 17. Performance measures of different machine learning models.

TABLE 7. Training accuracy and Testing accuracy of different types of models.

Machine Learning Model's	Accuracy on Training Set (%)	Accuracy on Test Set (%)
LR	0.78	0.74
KNN	0.92	0.90
BC	0.91	0.89
AB	0.51	0.48
ET	0.92	0.86
SGD	0.92	0.87
GNB	0.45	0.40
DT	0.90	0.87
RF	**0.96**	**0.94**
MLP	0.91	0.87
GB	0.92	0.89
LDA	0.78	0.73
SVM	0.83	0.79

classifiers has different proportions of errors (reject H0) and other classifiers yield same proportions of errors (fail to reject H0). In table 8, p-value and T-statistic values of McNemar's test were presented to examine the performance of Random Forest classifier is statistically different from other classifiers.

3) CORRELATION AND ERROR MEASURES

In this section correlation and error measures are discussed with various machine learning classifier and it has been shown in Table 9 and Fig. 19.

a: MATHEWS' CORRELATION COEFFICIENT (MCC)

The Mathew correlation coefficient will be used as a measure of predicting the quality of a multiclass classification [8], [43].

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

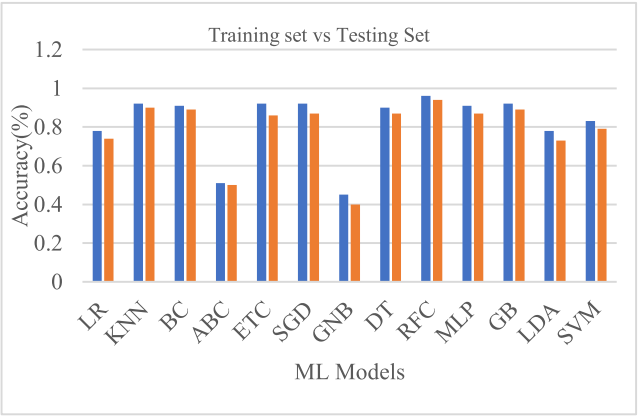

FIGURE 18. Training accuracy and Test accuracy of different types of models.

From Eqn.4 we have calculated the MCC which is an important aspect to measure the quality in prediction for multi-class problems. It takes the true positive and true negative and positive values into the account and the measure is suitable for multiclass problems even the datasets are in different sizes. The correlation coefficient of target and prediction values lies between the -1 to $+1$ where $+1$ denotes a good prediction and -1 denotes inverse prediction and 0 denotes the random average prediction. In Table 9, MCC shown for the different models, the overall MCC values lie between 0.40 to 0.94 for all our used machine learning models. The MCC values of RF models have obtained 0.94 which is near to $+1$ it has the highest correlation between the predicted and target and GNB models gain the least MCC as 0.40.

b: HAMMING LOSS

We have calculated the HL which is an important aspect to calculate the loss generated during the prediction of correct labels in a multi-class problem.

$$L_{Hamming(y,x)} = \frac{1}{n_{labels}} \sum_{j=0}^{n_{lables}-1} 1 * (y_j \neq x_j) \tag{8}$$

where, x_j -predicted value of j^{th} label; y_j-corresponding true value; n_{labels}-no. of classes.

The HL values lies between the 0 to 1 where lesser the value denotes a good classifier and the value greater denotes

TABLE 8. Statistical significance test using McNemar's test.

Machine Learning Models	McNemar's test statistic T-Statistic	McNemar's test p-value
LR	14	1.000
LDA	13	1.000
KNN	07	0.930
BC	04	0.180
AB	31	0.040
ET	03	0.013
SGD	07	0.093
DT	10	0.830
RF	05	0.035
MLP	04	0.069
GB	02	0.048
GNB	07	0.093
SVM	10	0.424

the worst classifier. In Table 9, HL shown for the different models, the overall HL values lies between 0.03 to 0.54 for all our used machine learning models. The MCC values of RF models has obtained 0.03 which has the highest HL between the predicted and target and GNB models has 0.54 the worst HL values.

c: MAE

It signifies the distinction between the trained dataset and the predicted values of the trained dataset that is separated from the mean difference over the dataset [8]. It can also be called mean of absolute error.

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_i - x_i| \tag{9}$$

where, y_i = *Predicted value from testing dataset;* $x_{i=Original}$ *value from the training dataset.*

In Table 9, MAE shown for the different models, the overall MAE values lies between 0.04 to 1.07 for all our used machine learning models. The MAE values of RF models have obtained 0.04 which is lesser it has the best MAE between the predicted and target and GNB models gain the highest MAE as 1.07.

d: MSE

It signifies the distinction between the trained dataset and the predicted values of the trained dataset that is separated from the squared the mean difference over the dataset. The smaller value of MSE will considered to be the best and raises the sensitivity of the model. So, in our estimation, we have smaller values for all the models except the GNB (1.59 %) and LR (0.79 %) and we conclude that all the other models are high in sensitivity.

$$MSE = \frac{1}{n}\sum_{k=1}^{n}|y_i - x_i|^2 \tag{10}$$

where, $|y_i - x_i|^2$ – Squares of the error; y_i = Predicted value from testing dataset; x_i = Original value from the training dataset.

In Table 9, MSE shown for the different models, the overall MSE values lies between 0.04 to 2.54 for all our used machine learning models. The MSE values of RF models have obtained 0.04 which is lesser which indicates the best MSE between the predicted and target and GNB models gains the highest MSE as 2.54.

e: RMSE

It signifies the actual error rate by the square root of RMSE [9]. Here smaller value of MSE is considered to be the best and raises the sensitivity of the model. So, in our estimation we have smaller values for all the models except the GNB (1.59 %) and LR (0.79 %) and we conclude that all the other models are high in sensitivity.

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_i - x_i)^2} \tag{11}$$

where n= number of predictions.

y_i = Predicted value from testing dataset; I = Original value from the training dataset.

In Table 9, RMSE shown for the different models, the overall RMSE values lie between 0.21 to 1.59 for all our used machine learning models. The RMSE values of RF models have obtained 0.21 which is lesser it has the best RMSE between the predicted and target and GNB models gain the highest RMSE as 1.59.

f: R^2

It signifies (Coefficient of determination) that how good the coefficient is actually fit with the training dataset values. It ranges from 0 to 1 and if the values are higher, then it is the best model. If the R^2

$$R^2 = 1 - \frac{\sum\left(y_i - \hat{y}\right)^2}{\sum\left(y_i - \hat{yr}\right)^2} \tag{12}$$

where $\hat{y}$ – predicted value; $\bar{y}$ – Mean Values.

In Table 9, R^2 shown for the different models, the overall R^2 values lie between -2.11 to 0.94 for all our used machine learning models. The R^2 values of RF models have obtained 0.94 which is lesser it has the best R^2 between the predicted and target and GNB models gain the highest negative value R^2 as −2.11.

g: MSLR

It signifies the ratio log between the actual and predicted values.

$$MSLR = log((y_i + 1)/(x^i + 1)) \tag{13}$$

y_i = Predicted value from testing dataset; x_i = Original value from the training dataset.

In Table 9, MSLR shown for the different models, the overall MSLR values lie between 0 to 0.22 for all our used

TABLE 9. Correlation and error measures of different ml algorithms.

Machine Learning Models	Matthews correlation coefficient (MCC)	Hamming Loss (HL)	Mean Absolute Error (MAE)	Coefficient of determination (R^2)	Root Mean Square Error (RMSE)	Mean Squared Error (MSE)	Mean Squared Log Error (MSLR)	Cohen's kappa (Ck)
LR	*0.65*	*0.21*	*0.3*	*0.39*	*0.70*	*0.49*	*0.05*	*0.64*
LDA	*0.73*	*0.17*	*0.25*	*0.43*	*0.66*	*0.04*	*0.04*	*0.72*
KNN	*0.92*	*0.05*	*0.05*	*0.93*	*0.22*	*0.05*	*0.00*	*0.91*
BC	*0.92*	*0.04*	*0.05*	*0.92*	*0.24*	*0.06*	*0.00*	*0.91*
AB	*0.31*	*0.47*	*0.49*	*0.34*	*0.73*	*0.06*	*0.06*	*0.30*
ET	*0.93*	*0.04*	*0.04*	*0.94*	*0.20*	*0.04*	*0.00*	*0.92*
SGD	*0.93*	*0.04*	*0.04*	*0.94*	*0.21*	*0.00*	*0.00*	*0.92*
DT	*0.88*	*0.77*	*0.08*	*0.88*	*0.30*	*0.09*	*0.01*	*0.86*
RF	*0.94*	*0.03*	*0.04*	*0.94*	*0.21*	*0.04*	*0.00*	*0.94*
MLP	*0.89*	*0.06*	*0.07*	*0.89*	*0.29*	*0.08*	*0.00*	*0.88*
GB	*0.92*	*0.05*	*0.05*	*0.91*	*0.25*	*0.06*	*0.00*	*0.91*
GNB	*0.4*	*0.54*	*1.07*	*-2.11*	*1.59*	*2.54*	*0.22*	*0.31*
SVM	*0.82*	*0.11*	*0.17*	*0.61*	*0.54*	*0.30*	*0.03*	*0.74*

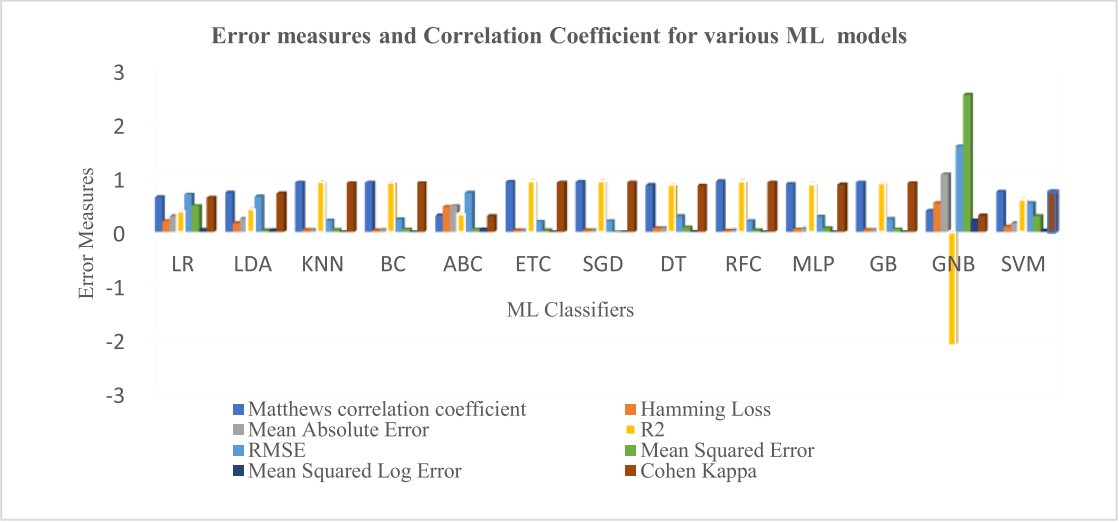

FIGURE 19. Correlation and Error Measures of different ML algorithms.

machine learning models. The MSLR values of KNN, RF, MLP, and GB models have obtained 0 which is lesser MSLR to the predicted and target and GNB models gain the highest MSLR as 0.22.

If RMSE ≥ MAE, the values of RMSE will be greater than MAE as shown in the table and if the errors are on the same scale then RMSE=MAE. (i.e. RMSE will be equal to MAE). The only disadvantage of MAE over RMSE is that it takes only the absolute value.

h: COHEN'S KAPPA (CK)

Cohen's kappa is a statistical measure used to calculate the inter-annotator agreement for the multi-classification problem [13]. It is calculated by

$$Ck = \left(\frac{p_0 - p_e}{1 - p_e}\right) \tag{14}$$

where p_0 is called as the empirical probability ratio to the sample and p_e denotes the expected value which is calculated using for each annotator empirical preceding over the output labels. The Cohen's kappa results were given in the Table 9 for all the classifiers. The values lies between 0 to 1. If the values lies 0 or less than 0 then the classifier performance is poor. The classifier RF obtained the value 0.94 closer to 1, so it practically perfect than the other classifier.

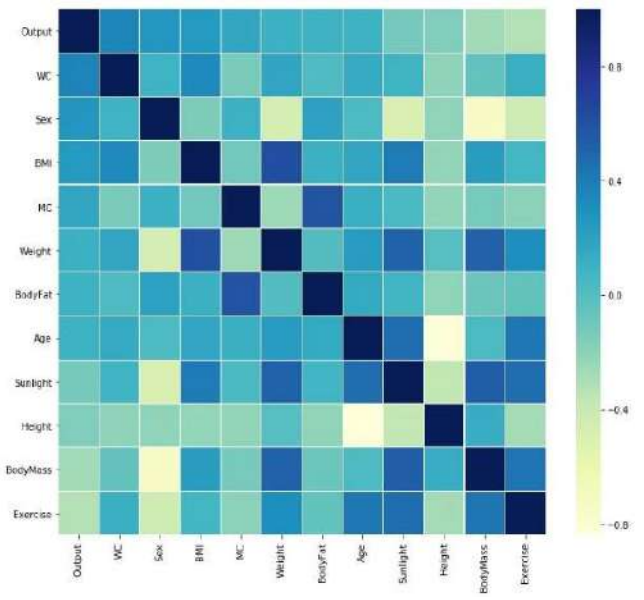

FIGURE 20. Correlation between different variables in the datasets.

In Fig. 20, the correlation between the variables using the heat map is presented. Correlation matrix which gives the correlation between the two parameters of the given dataset. In exploratory data analysis, the heat map is an important part of checking correlations. The purpose of the heat map is to decide which parameters influence the output variable and it also used to visualize the correlation matrices. From the heat map the Parameters Like BMI, Waist Circumference, Body Fat, Bone Mass, Exercise and Sunlight Exposure. From the Fig. 20, it is noted that there is a strong correlation between the BMI to body fat and weak correlation between age, height. The heat map values lie between $+1$ to -1, where the positive measure indicates positive correlation and negative measure indicates a negative correlation. The data points closer to $+1$ will have stronger linear association whereas values closer to 0 will have a weaker linear association.

V. CONCLUSION

The main objective of this study is to identify the best machine learning model in the prediction of VDD severity. The prediction accuracy was calculated and compared with the training and testing set. For this study, we have used 11 machine learning models and performance measures like precision, recall, F1-measure, and accuracy. We have used 11 parameters in the severity prediction and RFE [28] technique is used for feature selection. McNemar's statistical significance test is used to validate the empirical results. From McNemar's test, it is undoubtedly RF scores high in prediction when compared to different models and the Pearson's correlation coefficient and error measures result concluded the same.

The machine learning methods could be used as a substitute for the efficient prediction of severity of VDD with high accuracy. The results of this research work proved that the machine learning models especially the random forest classifier accurately predict the severity of Vitamin D deficiency. In particular, the Random forest classifier achieved the highest accuracy (96%) and outperforms well than other classifiers. This machine learning classifier will have a greater opportunity in the real-world medical domain which would assist experts to efficiently identify the severity of VDD. The major advantage of this study is that it has explored a new approach for the prediction of VDD severity using the Random Forest model and it has evaluated the results of the machine learning models using various performance measures accurately among the adolescents. So, the study claims that the Random forest model can be used to predict the severity of VDD with high accuracy than the other models. The future direction of our research is to validate the model with a different type of Vitamin D datasets of all age groups.

REFERENCES

[1] M. Holick, "Vitamin D deficiency," *New England J. Med.*, vol. 357, no. 3, pp. 266–281, 2007.
[2] I. R. Reid and M. J. Bolland, "Role of vitamin D deficiency in cardiovascular disease," *Heart*, vol. 98, no. 8, pp. 609–614, Apr. 2012.
[3] B. Schöttker, C. Herder, D. Rothenbacher, L. Perna, H. Müller, and H. Brenner, "Serum 25-hydroxyvitamin D levels and incident diabetes mellitus type 2: A competing risk analysis in a large population-based cohort of older adults," *Eur. J. Epidemiol.*, vol. 28, no. 3, pp. 267–275, Mar. 2013.
[4] S. B. Mohr, E. D. Gorham, J. E. Alcaraz, C. J. Kane, C. A. Macera, J. K. Parsons, D. L. Wingard, and C. F. Garland, "Serum 25-hydroxyvitamin D and prevention of breast cancer: Pooled analysis," *Anticancer Res.*, vol. 31, no. 9, pp. 2939–2948, 2011.
[5] Y. Lee, R.-M. Ragguett, R. B. Mansur, J. J. Boutilier, Z. Pan, D. Fus, J. D. Rosenblat, A. Trevizol, E. Brietzke, K. Lin, M. Subramaniapillai, T. C. Y. Chan, C. Park, N. Musial, H. Zuckerman, V. C.-H. Chen, R. Ho, C. Rong, and R. S. McIntyre, "Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review," *J. Affect. Disorders*, vol. 241, pp. 519–532, Dec. 2018.
[6] 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.
[7] 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.
[8] S. Bechrouri, A. Monir, H. Mraoui, E. H. Sebbar, E. Saalaoui, and M. Choukri, "Performance of statistical models to predict vitamin D levels," in *Proc. New Challenges Data Sci., Acts 2nd Conf. Moroccan Classification Soc. ZZZ (SMC)*, New York, NY, USA, 2019, pp. 1–4.
[9] K. Gonoodi, M. Tayefi, M. Saberi-Karimian, A. A. Zadeh, S. Darroudi, S. K. Farahmand, Z. Abasalti, A. Moslem, M. Nematy, G. A. Ferns, S. Eslami, and M. G. Mobarhan, "An assessment of the risk factors for vitamin D deficiency using a decision tree model," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 13, no. 3, pp. 1773–1777, May 2019.
[10] J.-J. Beunza, E. Puertas, E. García-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, and M. F. Landecho, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)," *J. Biomed. Informat.*, vol. 97, Sep. 2019, Art. no. 103257.
[11] A. Kuwabara, N. Tsugawa, K. Mizuno, H. Ogasawara, Y. Watanabe, and K. Tanaka, "A simple questionnaire for the prediction of vitamin D deficiency in Japanese adults (Vitamin D deficiency questionnaire for Japanese: VDDQ-J)," *J. Bone Mineral Metabolism*, vol. 37, no. 5, pp. 854–863, Sep. 2019.
[12] A. 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.

[13] C. B. Jensen, A. L. Thorne-Lyman, L. V. Hansen, M. Strøm, N. O. Nielsen, A. Cohen, and S. F. Olsen, "Development and validation of a vitamin D status prediction model in Danish pregnant women: A study of the Danish national birth cohort," *PLoS ONE*, vol. 8, no. 1, Jan. 2013, Art. no. e53059.
[14] H. Tamune, J. Ukita, Y. Hamamoto, H. Tanaka, K. Narushima, and N. Yamamoto, "Efficient prediction of vitamin B deficiencies via machine-learning using routine blood test results in patients with intense psychiatric episode," *Frontiers Psychiatry*, vol. 10, p. 1029, Feb. 2020.
[15] C. Carlberg and A. Neme, "Machine learning approaches infer vitamin D signaling: Critical impact of vitamin D receptor binding within topologically associated domains," *J. Steroid Biochemistry Mol. Biol.*, vol. 185, pp. 103–109, Jan. 2019.
[16] N. Altman and M. Krzywinski, "Ensemble methods: Bagging and random forests," *Nature Methods*, vol. 14, no. 10, pp. 933–934, Oct. 2017.
[17] J. Š. Benth, K.-M. Myhr, K. I. Løken-Amsrud, A. G. Beiske, K. S. Bjerve, H. Hovdal, R. Midgard, and T. Holmøy, "Modelling and prediction of 25-hydroxyvitamin D levels in Norwegian relapsing-remitting multiple sclerosis patients," *Neuroepidemiology*, vol. 39, no. 2, pp. 84–93, 2012.
[18] L. I. Al Asoom and M. T. Al Hariri, "The association of adiposity, physical fitness, vitamin D levels and haemodynamic parameters in young Saudi females," *J. Taibah Univ. Med. Sci.*, vol. 13, no. 1, pp. 51–57, Feb. 2018.
[19] R. Longadge, S. S. Dongre, and L. Malik, "Class imbalance problem in data mining: Review," *Int. J. Comput. Sci. Netw.*, vol. 2, no. 1, pp. 1552–1563, Feb. 2013.
[20] A. I. Saad, Y. M. K. Omar, and F. A. Maghraby, "Predicting drug interaction with adenosine receptors using machine learning and SMOTE techniques," *IEEE Access*, vol. 7, pp. 146953–146963, 2019.
[21] D. Tay, C. L. Poh, E. Van Reeth, and R. I. Kitney, "The effect of sample age and prediction resolution on myocardial infarction risk prediction," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 1178–1185, May 2015.
[22] T. Merlijn, K. M. A. Swart, P. Lips, M. W. Heymans, E. Sohl, N. M. Van Schoor, C. J. Netelenbos, and P. J. M. Elders, "Prediction of insufficient serum vitamin D status in older women: A validated model," *Osteoporosis Int.*, vol. 29, no. 7, pp. 1539–1547, Jul. 2018.
[23] B. Tran, K. B. Armstrong, K. McGeechan, R. P. Ebeling, R. D. English, G. M. Kimlin, R. Lucas, C. J. van der Pols, A. Venn, C. V. G. D. Whiteman, M. P. Webb, and E. R. Neale, "Predicting vitamin D deficiency in older Australian adults," *Clin. Endocrinol.*, vol. 79, no. 5, pp. 631–640, 2013.
[24] 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.
[25] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, Mar. 2010.
[26] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 1577–1580.
[27] Y.-P. Huang and and M.-F. Yen, "A new perspective of performance comparison among machine learning algorithms for financial distress prediction," *Appl. Soft Comput.*, vol. 83, pp. 1056–1063, Oct. 2019.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[29] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[30] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines," *J. Amer. Stat. Assoc.*, vol. 99, no. 465, pp. 67–81, 2004.
[31] C. Elkan. (Jan. 2011). *Nearest Neighbor Classification*. [Online]. Available: https://scholar.google.com/scholar?cluster=1396704159804385485&hl=en&as_sdt=0,5&sciodt=0,5
Article Location: http://cseweb.ucsd.edu/ elkan/151/nearestn.pdf
[32] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. TIT-13, no. 1, pp. 21–27, Jan. 1967.
[33] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[35] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, Sep. 1968.
[36] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, pp. 1303–1308, Mar. 1995.
[37] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
[38] A. Koivu, T. Korpimäki, P. Kivelä, T. Pahikkala, and M. Sairanen, "Evaluation of machine learning algorithms for improved risk assessment for down's syndrome," *Comput. Biol. Med.*, vol. 98, pp. 1–7, Jul. 2018.
[39] 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.
[40] T. Turki and Y.-H. Taguchi, "Machine learning algorithms for predicting drugs–tissues relationships," *Expert Syst. Appl.*, vol. 127, pp. 167–186, Aug. 2019.
[41] C. C. Wu, W. C. Yeh, W. D. Hsu, M. M. Islam, P. A. Nguyen, T. N. Poly, Y. C. Wang, H. C. Yang, Y. C. Li, "Prediction of fatty liver disease using machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 170, pp. 23–29, Mar. 2019.
[42] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.
[43] F. Firouzi, M. Rashidi, S. Hashemi, M. Kangavari, A. Bahari, N. E. Daryani, M. M. Emam, N. Naderi, H. M. Shalmani, A. Farnood, and M. Zali, "A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software," *Eur. J. Gastroenterol. Hepatol.*, vol. 19, no. 12, pp. 1075–1081, Dec. 2007.

G. SAMBASIVAM received the Ph.D. degree in computer science and engineering from Pondicherry University, Puducherry, India. He is currently working as a Senior Lecturer with the Faculty of Information and Communication Technology, ISBAT University Kampala, Uganda. His research interests include artificial intelligence, machine learning, deep learning, and Web service computing.

J. AMUDHAVEL (Member, IEEE) received the Ph.D. degree in computer science and engineering from Pondicherry University, Puducherry, India. He is currently working as a Senior Assistant Professor with the School of Computer Science and Engineering, VIT Bhopal University, India. His research interests include artificial intelligence, machine learning, bio-inspired algorithm, evolutionary computing, and distributed systems.

G. SATHYA received the Ph.D. degree in home science from Pondicherry University, Puducherry, India. She is currently working as a Dietician with the Indira Gandhi Government General Hospital and Post Graduate Institute, Puducherry. Her research interests include vitamin D deficiency and clinical nutrition, and dietetics.