A Predictive Analytic Multi-Model Approach for Estimating the Risk of Diabetic Forecast using Machine Learning Techniques

Mohamed Saleem Haja Nazmudeen[1], A. Sheik Abdullah[2,*], R. Priyadarshini[2]

[1]	Graduate Studies and Research Office, Universiti Teknologi Brunei, BE1410, Brunei
[2]	School of Computer Science Engineering, Vellore Institute of Technology, Rajan Nagar, Chennai, Tamil Nadu 600127, India

ABSTRACT

The realm of disease monitoring and analysis plays an important role in day-to-day human life. Analysing and estimating the risk for a specified disease becomes challenging from region to region. This research concentrates towards the development of a multi-model approach for estimating the diabetic risk factors using machine learning techniques. Different sorts of practices such as Decision trees, Support Classifiers, Random tree Forest and Neural Network (FF) has been used for model development and risk estimation. The result analysis shows that hyperparameter tuning with signified estimators and finite states in random forest provides highest accuracy level. Statistical analysis has also been made using spearman rank correlation analysis. The results proved that risk factors pertaining to PPG, FPG and MBG with a correlation value of about 0.79 is found to be significant and correlated. Higher level of R-square value is observed among the risk factors PPG and FPG respectively. The interpretation and evaluation have made in accordance with the medical experts along with the preparation of dirt chart advisor for significant varied analysis.

Keywords:
Diabetic prediction; predictive analytics; machine learning

1. Introduction

The science of translating data into information with the goal of enhancing output and profit is known as data analytics. Businesses employ several techniques to extract and separate data in order to recognize and assess behavioural data and patterns. Data analytics are used in a variety of industries to help businesses make better commercial decisions, as well as in science to confirm or refute existing models or ideas. Information dashboards with real-time data streams are common in today's data analytics [6]. There are various kinds of analytics. Predictive analytics is one of the most useful tools for predicting future events. Predictive analytics employs statistical knowledge, machine learning, and artificial intelligence. Trends in historical and transactional data can be utilized to predict future risks and opportunities and so on are examples of predictive analytics applications [7].

Mainly Healthcare workers utilize predictive analytics to evaluate patient data, anticipate the likelihood of disease. Risk classification will help prioritise clinical progress, reduce waste in the system, and ensure cost-effective population management. Significant analysis focusing on risk analysis with a keen focus on low, moderate and high levels will be considered to be crucial in certain situations [8]. By analysing the risk score and its estimating factors certain pre-emptive analysis can be made upon medical reconciliation [24].

Now-a-days Diabetes has become the most prevalent disease among adults and the elderly. In 1980, 108 million people had diabetes; by 2014, that figure had risen to 422 million. Adults with diabetes account for 8.5 percent of this group. Diabetes Type 1 and Type 2 are two distinct types of diabetes. A shortage of insulin synthesis causes Type I diabetes, but Type 2 diabetes affects most people globally. Most of the high glucose-related deaths targets towards diabetes and the intake of insulin at appropriate levels. In 2030, it has been estimated that diabetes will be considered as one of the seventh leading cause towards mortality among people [9].

The number of features used in machine learning and pattern recognition applications has increased dramatically in recent years, from tens to hundreds. When the number of characteristics is increased, the computation time increases as well. Many strategies have been developed to address the challenge of minimizing inappropriate and superfluous variables that cause problems on difficult assignments [10]. Feature selection aids in the comprehension of data, the reduction of computing time, the reduction of the effect of dimensionality irritation, and the improvement of predictor performance. Feature selection does not introduce additional functionality; somewhat, it decreases the number of input variables. It also affects classification analysis and necessitates the elimination of variables or features. The major goal is to use an efficient algorithm to draw a multi-model feature analysis pattern to evaluate diabetic condition [11].

2. Literature Review

Based on artificial neural networks, the authors [1] suggested the PE-CMR approach for predicting cardio metabolic risk (ANN). The sample size is 1281 people, with 692 males and 589 women ranging in age from 18 to 67. All of the tests were done in the morning (after a night of fasting) at the Clinical Centre of Vojvodina's Endocrinology, Diabetes, and Metabolic Disorders section in Novi Sad (Serbia). The following risk factors were investigated: Gender (GEN), Blood Pressure (BP), Waist in Height (WHtR), Body Mass (BMI), and Age (AGE) are the key determinants; secondary factors include Lipid status, Uric acid (UAC), Fibrinogen (FIBR), and Glycemia (GLY). In future, the investigation can be about optimality of the several ANN types and architectures.

The work by the authors [2] focuses towards the analysis of type II diabetes in Korean adults. The analysis concentrated in diseased vs normal cases using different forms of machine learning techniques. They have analysed phenotypes with cross-sectional study including 937 cases during the year 2006 to 2013 respectively. To create and analyse a generic phenotype across world populations, more research is needed. The authors' research revealed an issue in predicting the patient's survival [3]. The implementation has been made using ANIFS model development with the examination from 271 patients who are subjected to cancer radiation. The different stages for analysis have also been measured and analysed with accordance to the clinical values. Median years of age, gender, histology, AJCC stage, treatment, and Glasgow prognostic score are the variables used (GPS). Furthermore, whether or not to include a novel prognostic feature in clinical result prediction models is determined by the costs versus the potential for saving lives.

The authors [4] offered different sorts of NLP mechanisms, and keyword dictionaries in their paper. In the future, their dictionary lookup strategy will mostly rely on the training data set's narrow

keyword list. The i2b2 corpus data collection was used. The risks factors are studied in this research are CAD, Diabetes, Obese status, hypertension, Medication, smoking status, family history and hyperlipidaemia. The analysis has been with regard to 3 different forms of clinical entity with a signified threshold value. Calculating the value of such data in terms of improving system performance is fascinating. The findings from the annotated data suggest that medication kinds are actually linked to a specific form of clinical trait. Significant additional analysis has also been made with regard to the adherent factors that contribute to the disease.

The Tree-Lasso model was proposed by the authors [5] for medical data analysis. When compared to other techniques such as Gain ratio analysis, Relief, and test statistic, the execution of Lasso feature analysis is good when compared to other approaches. Also, the predictive power of Tree-Lasso is more adherable to most of the disease-specific analysis with the utilization of more machine learning models. A significant regional hospital in Australia provided the cancer statistics. This data record contains eleven distinct cancer kinds from patients who visited the hospital between 2010 and 2012. Electronic Medical Records (EMRs) are used to collect patient information (EMR). The dataset included 4293 patients, who were diagnosed with 439 different ICD codes. The same hospital in Australia also provided the Acute Myocardial Infarction (AMI) Dataset. There are 2941 people in the dataset, and their illnesses are characterized using 528 ICD codes. In the future, it would be grateful if the possibility of extending their framework to multi-level setting for sample analysis with the interpretation of ML models.

This study focuses entirely on creating a multi-model framework for assessing and evaluating diabetic risk factors in a real-time environment. The analysis will be made in such a way that the predominant and adherent risk factors are suitably estimated with the machine learning approaches.

3. Proposed Methodology

The multi-model approach focuses on the evaluation of risk related to diabetes. In this research work we have collected real-time data pertaining to diabetes. It includes the complete analysis of different variants that relates to diabetes. Also, the model evaluates the time and space complexity measure of the algorithmic model used for evaluation. This is considered as important because the performance of the algorithm has to be estimated all the times when handling with real-time data. The workflow is depicted in Figure 1 below.
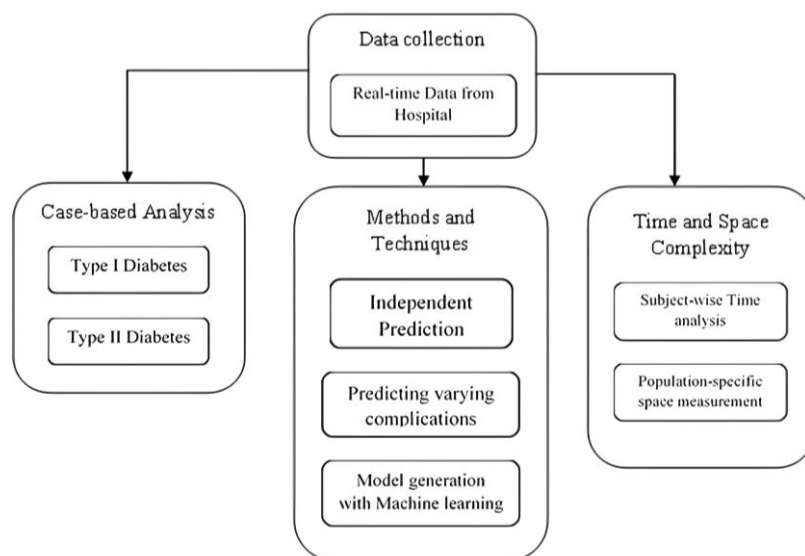

Fig. 1. Proposed workflow

3.1 Rationale Behind the Proposed Work

Data models with automated working environment are enabled in recent days for most of the engineering applications. The major focus lies at the data that we use and the format of the output that focus on the data. Many literatures have been derived with different sorts of machine learning approaches with the intention behind the deployment of the data model. Some may extensively find fruitful results in applications such as sensor, forecasting, biomedical appliances, and human computer interaction and so on. The data that is used for experimentation has to be focused clearly with the maximum and minimum level of significance of each attribute. Once the level of significance is understood with its rapidly changing values in most of the real-time applications, then the occurrence of those notable variables can be estimated more easily. For the applications that focus on the development of data models in biomedical engineering the authors do not focus on the level of significance rather they focus on the existence of the attribute alone. In real-time data processing applications measuring the level of significance of the attributes makes the user/people to have awareness with regard to the cause and effect among diseased cases. This made as a motto to observe the notable significance level among the diseased cases for real-time data analysis [12].

3.2 Methodological Workflow

The proposed work focusses to forecast the diabetic risk and its complications using machine learning models. In this research, we have incorporated the models to best estimate the risk that pertains to diabetes. The entire methodology is calibrated into external complications and internal complications. Based on the categorization of risk the model has been derived [13]. The patterns that focusing to eye disease, skin problems, nutritional disorders and significant infections are categorized as external complications.

If the notable conditions that falls under internal complications, they are derived into that category of risk analysis [14]. To the best of execution process a set of visual analytics is made in order to exactly measure the incorporation of the risk factors that focus towards the disease. Once the analysis is complete then the investigation by medical experts will be made to evaluate the conditions and complications that arise with the disease conditions and actions. The following Figure 2 provides the execution phases in detail.
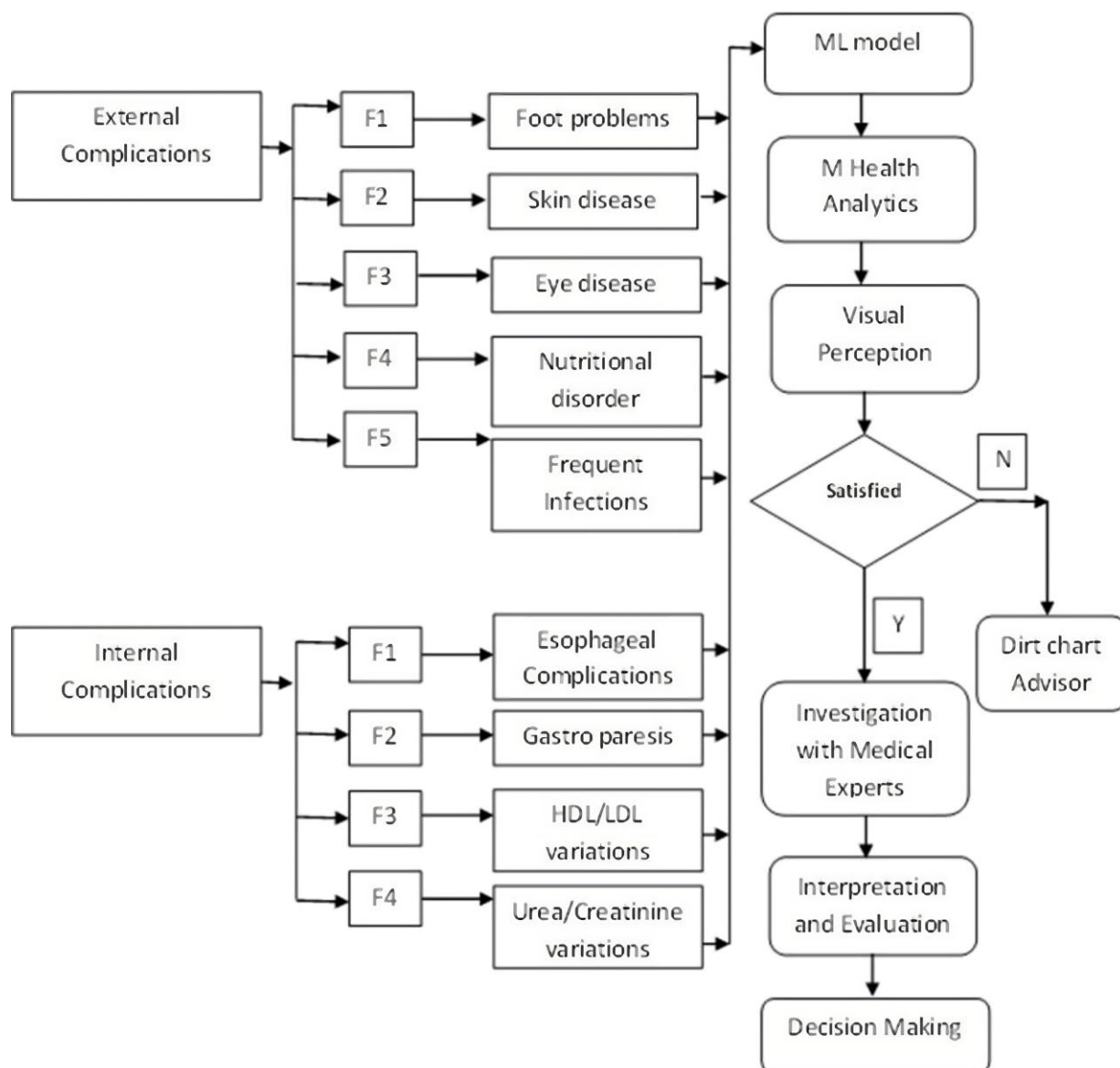
Fig. 2. Multi-model execution phases

4. Multi-Model Development using Data Classification Algorithms
4.1 Dataset Description

The implementation proceeds by evaluating the model using machine learning algorithms. Since most of the algorithms performs differently on different datasets the evaluation has been made using significant notable algorithms such as SVM, decision trees, random tree forests and neural network classifiers [15]. The dataset used has been collected from a hospital where there occur prevalent cases of diabetes treatment. The observed dataset consists of about 13 attributes and a label for classification process.

The observed raw dataset contains missing values and duplicates with regard to the nature of attributes and its associated records. In order to overcome this problem, we have deployed different forms of feature enhancement techniques with different parametric evaluations [16]. The Table 1 provides the data description.

Table 1
Dataset description

Attributes	Description
Patient Name	Name of the patient
Patient ID	Unique Id is given to each patient
Gender	1- Male
	2- Female
Age	Age of the patient is given
FPG	Sugar level in blood before eating food
	range - 70- 110 mg
PPG	Sugar level in blood after eating food
	Normal range- 120- 150 mg
Cholesterol	It depicts the total value of cholesterol and it is of two types HDL and LDL
levels	range- 140- 200 Mg/dl
TGL	The amount of excess fat is burnt into calories and stored as TGL in human body.
	range- 150 Mg/dl
HDL	It is defined as the good level cholesterol
	range- 45- 50 Mg/dl
LDL	It is defined as the bad cholesterol level
	range- < 100 Mg/dl
VLDL	It is the high representation of tri-glycerides
	range- 5 – 20 Mg/dl
Urea	The excretion and functioning of kidney is completely measured using the urea level. Normal
	range- 14- 40 Mg/dl
HB	It is a protein in red blood cells
	range- 12- 16 gms

The dataset collected has been carefully analysed and the same has been processed accordingly. The dataset includes class labels of 3 varying types signifying the level 1, level 2 and level 3 of the disease as specified. The level 3 denotes the higher occurrence and the level 1 denotes the lower level of occurrence respectively [17]. Here in order to unify the data records the mechanism of min-max normalization is used, since each of the attribute obviously will have a minimum and a maximum value of representation based on the observed medical factors. The expression evaluation for min-max is given in the following Eq. (1) as:

$$V^{'} = \frac{v - \min_A}{\max_A - \min_A}\left(new_{\max_A} - new_{\min_a}\right) + new_{\min_A}$$

(1)

where,
$\min_A$– minimum value of occurrence
$\max_A$ – maximum value of occurrence
$new_{\max A}$– new minimum value set by the user
$new_{\min A}$– new maximum value set by the user

This process may go in an error state if the future value falls outside the range as mentioned. Therefore, all the observed records and variations must be considered based on the minimum and the maximum limits as mentioned.

4.2 Algorithms Involved in Model Development
4.2.1 Decision trees

In data classification process, there exists the mechanism of supervised learning scheme which evaluates the dataset based on the presence of class label. Decision tree is one among them which suitably evaluates the dataset and its corresponding attributes by creating a tree like structure containing the representation of nodes for each of the attribute. The tree structure depicts the representation as like root and subsequent leaf accordingly based on the attribute list. The decision at each level will be made in accordance with the splitting measure used [18]. In this multi-model assessment, we have used gain ratio to determine the best split among the attributes that has been used for evaluation.

The following equation is used to calculate the gain value as depicted by,

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

(2)

The gain ratio is given as,

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

(3)

The evaluation has been made in such a way that the generation of tree with accordance to that of the splitting criterion has been made with the attributes used for evaluation. The decision is based on the features associated with the dataset.

4.2.2 Random forest

The mechanism of random forest algorithm works on the basis of ensemble learning scheme. Here the output reached by multiple classifiers are taken into consideration for determining the evidence result from the learning mechanism. Thereby problems falling under complex categories are improvised from the selection of best results as produced from the product ensemble. Instead focusing on a single decision, we can analyse the performance of the model from the patterns which has been derived as from majority votes and thereby predicting the final output. Hence the accuracy of the model gets improved from time to time upon generation of the model in order to avoid the mechanism of overfitting [19].

4.2.3 Support vector machine classifiers

The intention behind the SVM algorithm is to best determine the separation which segregates the given dimension in a n-dimensional hyperplane. The decision boundary determines the split incurred among the attributes for the ranges of class label. The points chosen are considered to be the vectors for the determination of best split in the given hyperplane. In a given hyperplane there may be multiple decision boundaries to classify the data points [20]. The best identifiable line of separable is found to be the hyperplane among the given class labels. The hyperplane is set to the

level of maximum margin determines the maximum distance that is set among the data points. The Table 2 depicts the parameter setting specified for model development.

Table 2
Parameter setting for the classification techniques

Classification Algorithm	Incorporations/Execution parameter setting
Decision tree algorithm	In this algorithm gain ratio is used as the splitting measure for evaluating the data records with regard to testing and training of data tuples.
Support Vector Machine	Upon considering the estimator and the function value score we have used linear kernel function for evaluating the data records. The variation has been made with accordance to the parametric values focusing linear kernel observations.
Random Forest	In this algorithm the parameter setting has been made in accordance with the majority voting mechanism. The generated set of trees has been set to a maximum of 200 with an initiation towards hyperparameter tuning mechanism.
Neural Network	Here we have used up FF Neural Network for evaluation. The modification has been made in accordance with the approximation function for assigning the input x to that of the output y. finally, the best value for the approximation function is considered as the value for evaluating the tuple of records.

4.2.4 Neural networks

The utilization of neural network algorithm came into existence since 1969 with its applications towards diverse field of actions. The neural networks are designed and modelled in such a way that the connections have 3 different layers for analysis. In each of the linking to its input a weight is assigned along its path till the output is assigned and generated. Each of the value along with its associated node is gets manipulated with the weight value in order to generate a resulting value to its signified output. If the generated value is below than that of the assigned threshold value then the data with that layer is stopped for proceeding to the next layer for analysis. If the generated value is higher than that of the threshold then the data gets passed for further processing along with the path that is assigned with the network layer. Initially the weights and the nature of threshold values are set to random generation of values within the limits as specified. In the training phase the value of the weights and threshold are modified with accordance to the performance of the algorithm in order to improve consistency and efficiency.

5. Experimental Analysis and Discussion Results

The experimentation is made accordance with different sorts of dataset along with the real-time data that has been collected from the hospital.

5.1 Metrics for Evaluation

The metrics used to assess the suggested model's performance are listed below:

i. <u>Accuracy:</u> It is defined as the number of tuples that are correctly classified by the model. The evaluation is defined in Eq. (4) as,

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN} \right)$$

(4)

ii. <u>Error rate:</u> The total number of erroneously classified data during evaluation is the error rate of a predictive model. It is calculated using Eq. (5) as,

$$E_i = \left(\frac{n}{N} \right)$$
(5)

iii. <u>Kappa measure:</u> It is the grade to which nonrandom agreement can be measured. It is calculated using Eq. (6) as,

$$K = \left(\frac{p(A) - p(E)}{1 - p(E)} \right)$$
(6)

iv. <u>Recall:</u> It is the percentage of relevant occurrences that are returned. It is expressed in Eq. (7) as,

$$recall = \left(\frac{TP}{TP + FN} \right)$$
(7)

v. <u>Precision:</u> It is defined as the proportion of relevant retrieved instances. It is expressed in Eq. (8) as,

$$precision = \left(\frac{TP}{TP + FP} \right)$$
(8)

vi. <u>Spearman Rho:</u> It is defined as the non-parametric measure of rank correlation which determines the statistical dependence and ranking among the two defined variables. If the observed correlation between the two variables is said to be high then they have similar means of correlation and the rank is said to be 1. Otherwise, the correlation is said to be -1.

vii. <u>Root Mean Squared error:</u> It is also known as the quadratic mean, and it is the measure of altering the size of a changing statistical measure as in Eq. (9) as,

$$t_{RMS} = \sqrt{\frac{\sum_{i=1}^{n} t_i^2}{n}}$$
(9)

viii. <u>Confusion Matrix:</u> The confusion matrix shows how many correct and wrong predictions the model made in comparison to the test data classifications. It is represented in the following Table 3.

Current class used	Class Predicted		
	X	True class value	False class value
	True class value	TPV	FPV
	False class value	TNV	FNV

The variation has been observed with accordance to the parametric values, the number of iterations and the accuracy. The performance of the multi-model evaluations has been summarized in Table 3.

Table 3
Evaluation results of classification algorithms

Algorithms Metrics	Decision tree	Support Vector Machine	Random forest	Neural Networks
Accuracy	97.57	75.80	98.38	97.58
Error	2.43	25.20	1.62	2.42
Kappa Statistics	0.933	0.00	0.956	0.936
Weighted mean Recall	64.65	33.33	65.38	65.02
Weighted mean Precision	65.04	25.27	65.67	64.43
Spearman rho	0.941	0.00	0.962	0.919
Root means square error	0.129	0.492	0.138	0.136

From the Table 3, it has to be observed that the performance of the random forest algorithm has found to be good when compared to the other schemes. Even for SVM the total number of vectors is set to 248 and the offset bias value has been assigned to 0.626 with vectors assigned with 59 for class 1, 188 for class 2 and 1 for class 3 respectively.

Upon considering the results observed for decision tree algorithm, the tree assigned PPG as its root node by evaluating through gain ratio as the splitting criterion. The generation of decision tree is depicted in Figure 3. The tree has its consecutive branches focusing on attributes such as FPG, which then denotes that PPG and FPG are considered to be the predominant risk factors with diabetic prediction. In the execution of the random forest algorithm the main consequence is the tuning of the hyper parameters with different level of variations. The following are the important parametric constraints that can maximize the performance of the random forest algorithm [21].

    i.   Max_features
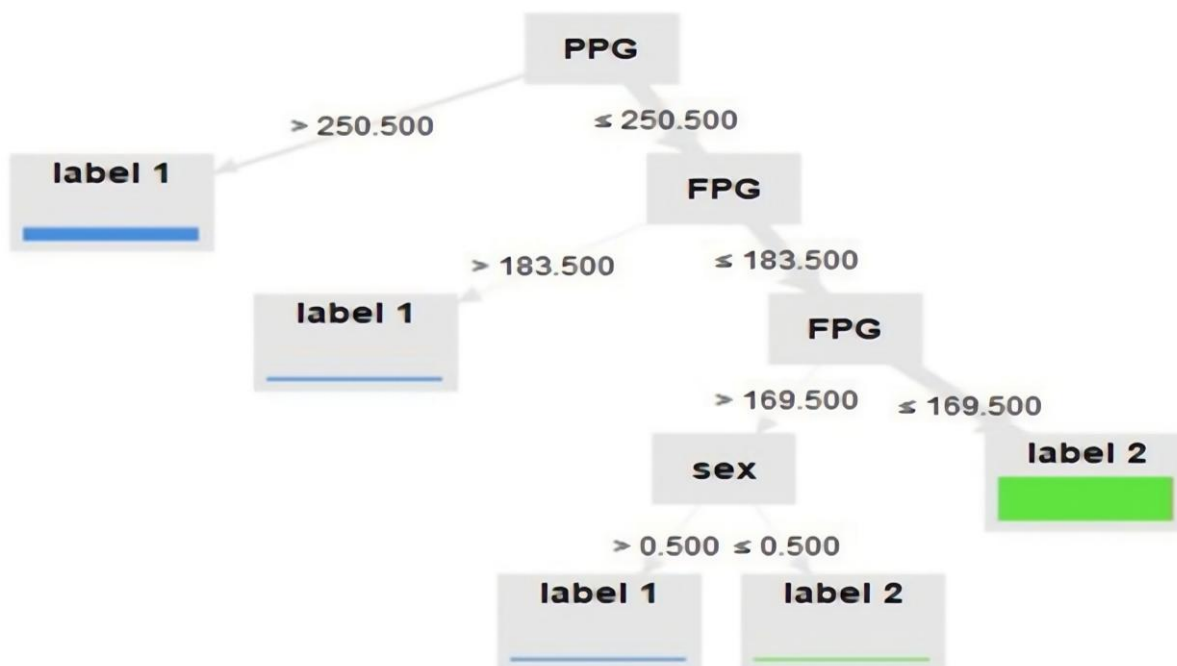   ii.   N_estimators
  iii.   Random_states
  iv.   N_jobs
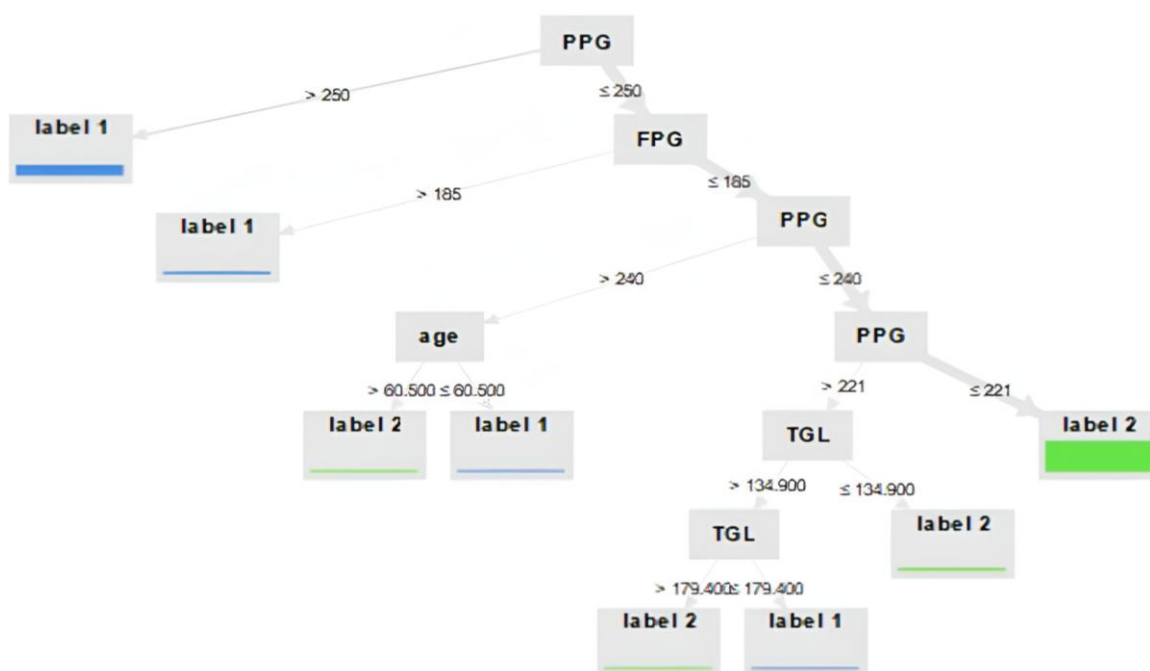
Fig. 3. Generated decision tree

Fig. 4. Generated forest (highest votes – majority voting procedure)

Based on the evaluator states of the above parameters the performance of the algorithms gets varied and the execution of number of trees will be fixed. The following Table 4 depicts the metrics used for evaluating the improvisations in the random forest algorithm.

Table 4
Parametric values for random forest algorithm

Parameter	Value
Number of trees generated	200
Splitting measure used	Gain ratio
The maximum dept as mentioned	10

Set confidence level	0.1
Gain which is of minimal	0.01
Leaf size which is of minimal	02
Voting principle used	Majority voting procedure

The mechanism of majority voting is used to combine the predictions from multiple learned models. Thereby the predictions can be improved significantly. Here, the mode is distributed with a learning time and rate of zero level convergence for the models trained during the evaluation. The observed confusion matrix as per the label is depicted in Table 5.

Table 5
Observed confusion matrix for random forest algorithm

Category	label 1	label 2	label 3	class precision
pred. label 1	57	1	0	98.28%
pred. label 2	2	187	1	98.42%
pred. label 3	0	0	0	0.00%
class recall	96.61%	99.47%	0.00%	98.38%

The performance of the proposed approach has also been tested using other benchmark datasets. The algorithmic model with the same sort of parametric evaluations has been used up for execution [22]. As we know that different algorithms perform differently on different dataset the efficacy has been determined in accordance with the dataset used. The following Table 6 summarizes the performance estimation with another benchmark dataset.

Table 6
Performance evaluation using Benchmark dataset (diabetic data)

Metrics Algorithm	Accuracy	Error	Kappa Statistics	Weighted mean Precision	Weighted mean Recall	F Measure
Decision Tree	77.47	22.56	0.471	0.771	0.775	0.765
Support Vector Machines	57.03	42.96	0.292	0.549	0.570	0.559
Random forest	79	21	0.625	0.793	0.792	0.79
Neural Network	57.77	42.22	0.299	0.560	0.578	0.569

With regard to the benchmark data, it has been found that the random forest algorithm has given a good accuracy level of about 79% when compared to SVM, Decision trees and Neural Network. Hence this algorithm has a significant improvisation with regard to medical data and its contributing risk factors [23]. The risk factors pertaining to any sort of disease surely focus on the adherent and predominant risk factors with regard to the disease concerned. From this research with the utilization of different classification schemes it has been found that PPG, FPG and MBG are found to be the most contributing factors related to the disease.

5.2 Statistical Analysis using Spearman Correlation Coefficient
5.2.1 Visual perception

The risk factors pertaining to PPG, MBG, FPG and cholesterol levels has been considered for a significant visual analysis among the values. When comparing PPG and FPG the R square value has been found to 0.74 with good insights in $P < 0.001$ respectively. The illustration for the same has been depicted in Figure 5. Similarly, significant analysis has been made for the factors that contribute towards the most predominant risk for diabetic prediction.
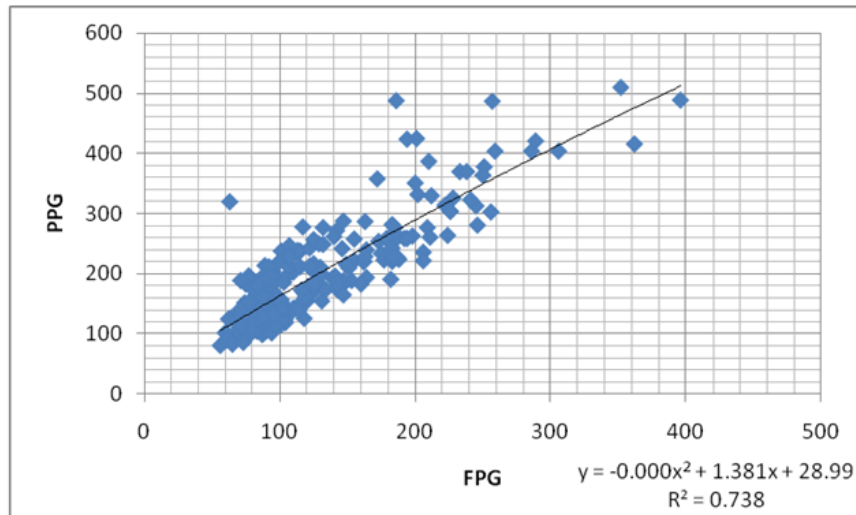
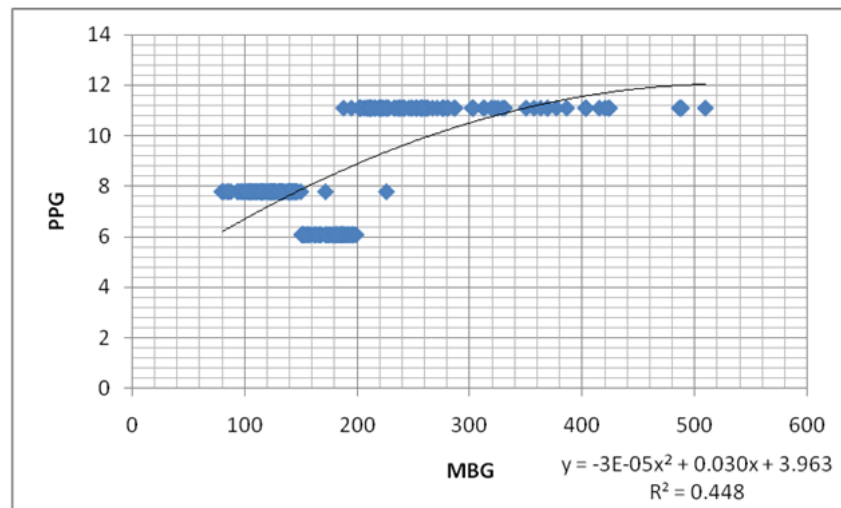**Fig. 5.** Feature analysis among FPG and PPG

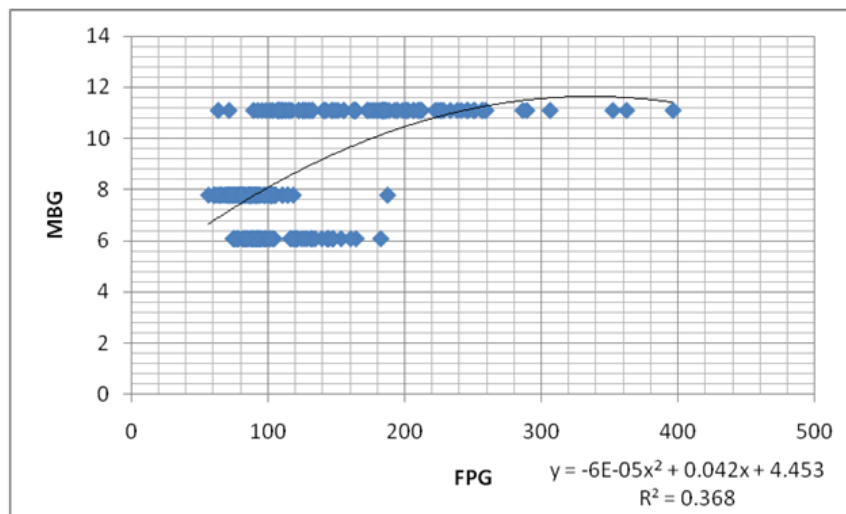**Fig. 6.** Feature analysis among MBG and PPG

Fig. 7. Feature analysis among FPG and MBG

According to the findings of the investigation, the feature focusing on FPG and PPG have found to be strongly correlated with a good impedance and P value respectively.

5.2.2 Correlation analysis

The following is the formula which is used to calculate the impedance among the unpaired variables:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \tag{10}$$

Where n signifies the count of the data points between the two variables and d_i represents the rank difference in the ith element. The following formulae is used to calculate the relation among the paired variables:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{11}$$

where i represents the score of paired variables.

Hence for the risk factors FPG and PPG that is of paired we are calculating the correlation coefficient as follows:

$$R = \frac{Cov}{(XR * YR)} \tag{12}$$

XR – Ranks corresponding to X values
YR – Ranks corresponding to Y values
Mx – Minimum of ranks corresponding to x
My – Minimum of ranks corresponding to y
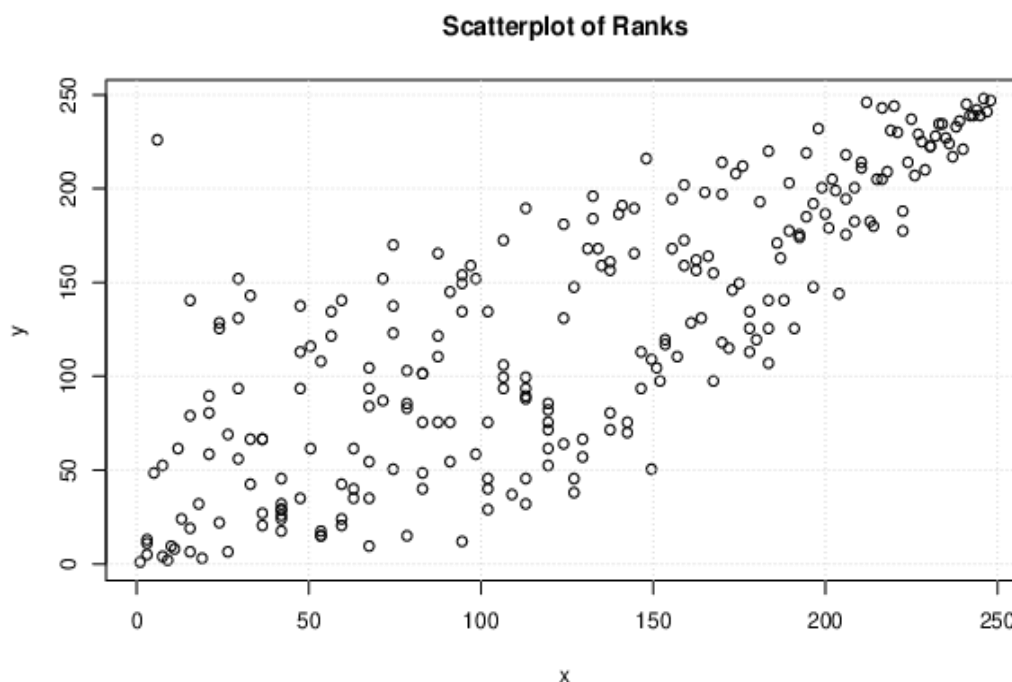Sum_Diffrs = (XR - Mx) * (YR - My)

Scatterplot of Ranks

Fig. 8. FeScatterplot of Ranks

From the formulation the correlation coefficient value was discovered to be 0.799 for factors FPG and PPG with two-sided significant P-value of about 2.015. Thereby the risk pertaining to varying complications with regard to time as an important factor relies mostly on the nature of risk that occurs along with its co-morbidities.

5.2.3 Statistical analysis using Fisher's Z – distribution

It is observed that from the observed population sample of estimate a correlation coefficient if 0.72 is obtained from a sample of 239 observations. Upon considering Snedecor's F-distribution with $(v1, v2)$ d.f., we put.

$$Z = \frac{1}{2} \log_e F \tag{13}$$

The distribution of Z becomes,

$$2 \frac{(v_1 / v_2)}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \frac{e^{vz}}{\left[1 + \frac{v_1}{v_2} e^{2z}\right]^{(v_1 + v_2)/2}} \tag{14}$$

The objective is to determine that the sample has been drawn for correlation value 0.8 with 95%

Hypothesis:

H_0 - there exists no difference between the observed r value = 0.72 and ρ value 0.80 therefore the sample can be considered from a bivariate sample.

Here,

$$Z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

(15)

$$=1.15\log_{10}6.1$$

$$=0.907$$

$$\xi = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$$

(16)

$$=1.16 \; X \; 0.9$$

$$=1.044$$

Under the relevance of H_0 test analysis

$$U = \frac{Z-\xi}{1\sqrt{n-3}} \sim N(0,1)$$

(17)

$$U = \frac{(0.907-1.100)}{0.196}$$

$$= -0.985$$

From the results it is observed that $|U < 1.96|$ is having its 5% level and H_0 is accepted. Therefore, the sample corresponds to bivariate i.e normal population statistic. The value of 95% of limits for ρ is evaluated by the input information to the sample focusing on the following derivation as follows:

$$|U|<1.96 \;\; |z-\xi|<1.96X\frac{1}{\sqrt{n-3}}=1.96 \; X \; 0.196 \;\; 0.523\leq\xi\leq1.291 \;\; 0.4543\leq\log_{10}\left(\frac{1+\rho}{1-\rho}\right)\leq1.1213$$

Now we get, $\log_{10}\left(\frac{1+\rho}{1-\rho}\right)=0.4543 \quad and \quad og_{10}\left(\frac{1+\rho}{1-\rho}\right)=1.1213$. Therefore, the value of ρ

becomes, $\rho=\frac{2.846-1}{2.846+1}=\frac{1.846}{3.846}=0.4799 \quad and \quad \rho=\frac{13.22-1}{13.22+1}=\frac{12.22}{14.22}=0.86$

Upon substitution we get, $0.48\leq\rho\leq0.86$. Hence the observe sample of data is found to be significant for the values as observed. Thereby the model evaluated has its equivalence towards the determination of risk factors that contribute towards the disease.

6. Conclusion and Future Work

In most technical and medical applications, data analysis and classification are critical. The process behind this is to utilize the right algorithm or model for the correct data to infer the pattern that exist between them. In recent days, the mechanism of real-time data analysis is found to be challenging with improved accuracy along with explicitly distinguishable patterns. This research focus on the development of a multi-model approach for predicting the risk that pertains to diabetes using machine learning algorithms. It considered different impacts that has been observed among the

cases that relates to diabetes and the process has been framed with significant rules in accordance to the adherent and prevalent risk factors.

The model development incorporates the utilization of different classification approaches such as Decision trees, Support Vector Machine classifiers, Random Forests, and Neural Networks. At different levels the risk estimation has been done along with the visual perception of the data tuples. The model development with random forest using hyper parameter tuning with its estimators and varied states gives a highest accuracy level of about 98.38% respectively. The same illustrative mechanism has also been evaluated with benchmark datasets using the different algorithms in machine learning. It also signifies the utilization of random forest algorithm for diabetic risk level prediction provides a highest accuracy and the best segregation in predominant and adherent risk factors. Significant statistical evaluation is made by spearman rank correlation and the values are found to be 0.79 which in turn also found to be significant. Among the factors pertaining disease the most predominant and adherent factors are found to be PPG, FPG and MBG respectively.

The future work can be made in progression with other such data sets with a real-time impact and its adherent risk estimation. We have initiated the future analysis for the multi-model development for kidney disease prediction and estimation among the rural people.

Acknowledgement

References

[1] Kupusinac, Aleksandar, Rade Doroslovački, Dušan Malbaški, Biljana Srdić, and Edith Stokić. "A primary estimation of the cardiometabolic risk by using artificial neural networks." *Computers in biology and medicine* 43, no. 6 (2013): 751-757. https://doi.org/10.1016/j.compbiomed.2013.04.001

[2] Lee, Bum Ju, and Jong Yeol Kim. "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning." *IEEE journal of biomedical and health informatics* 20, no. 1 (2015): 39-46. https://doi.org/10.1109/JBHI.2015.2396520

[3] Wang, Chang-Yu, Jinn-Tsong Tsai, Chun-Hsiung Fang, Tsair-Fwu Lee, and Jyh-Horng Chou. "Predicting survival of individual patients with esophageal cancer by adaptive neuro-fuzzy inference system approach." *Applied Soft Computing* 35 (2015): 583-590. https://doi.org/10.1016/j.asoc.2015.05.045

[4] Yang, Hui, and Jonathan M. Garibaldi. "A hybrid model for automatic identification of risk factors for heart disease." *Journal of biomedical informatics* 58 (2015): S171-S182. https://doi.org/10.1016/j.jbi.2015.09.006

[5] Kamkar, Iman, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. "Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso." *Journal of biomedical informatics* 53 (2015): 277-290. https://doi.org/10.1016/j.jbi.2014.11.013

[6] Cox, Andrew Paul, Mireia Raluy-Callado, Meng Wang, Abdel Magid Bakheit, Austen Peter Moore, and Jerome Dinet. "Predictive analysis for identifying potentially undiagnosed post-stroke spasticity patients in United Kingdom." *Journal of biomedical informatics* 60 (2016): 328-333. https://doi.org/10.1016/j.jbi.2016.02.012

[7] Ibrahim, Heba, Amr Saad, Amany Abdo, and A. Sharaf Eldin. "Mining association patterns of drug-interactions using post marketing FDA's spontaneous reporting data." *Journal of biomedical informatics* 60 (2016): 294-308. https://doi.org/10.1016/j.jbi.2016.02.009

[8] Ng, Kenney, Amol Ghoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, and Jimeng Sun. "PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records." *Journal of biomedical informatics* 48 (2014): 160-170. https://doi.org/10.1016/j.jbi.2013.12.012

[9] Kourou, Konstantina, George Rigas, Konstantinos P. Exarchos, Yorgos Goletsis, Themis P. Exarchos, Steven Jacobs, Bart Meyns, Maria-Giovanna Trivella, and Dimitrios I. Fotiadis. "Prediction of time dependent survival in HF patients after VAD implantation using pre-and post-operative data." *Computers in biology and medicine* 70 (2016): 99-105. https://doi.org/10.1016/j.compbiomed.2016.01.005

[10] Nguyen, Thanh, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi. "Classification of healthcare data using genetic fuzzy logic system and wavelets." *Expert Systems with Applications* 42, no. 4 (2015): 2184-2197. https://doi.org/10.1016/j.eswa.2014.10.027

[11] Abdullah, A. Sheik. "Assessment of the risk factors of type II diabetes using ACO with self-regulative update function and decision trees by evaluation from Fisher's Z-transformation." *Medical & Biological Engineering & Computing* 60, no. 5 (2022): 1391-1415. https://doi.org/10.1007/s11517-022-02530-2

[12] Sheik Abdullah, A., S. Selvakumar, and M. Venkatesh. "Assessment and evaluation of CHD risk factors using weighted ranked correlation and regression with data classification." *Soft Computing* 25, no. 6 (2021): 4979-5001. https://doi.org/10.1007/s00500-021-05663-y

[13] Sheik Abdullah, A., and S. Selvakumar. "Assessment of the risk factors for type II diabetes using an improved combination of particle swarm optimization and decision trees by evaluation with Fisher's linear discriminant analysis." *Soft Computing* 23, no. 20 (2019): 9995-10017. https://doi.org/10.1007/s00500-018-3555-5

[14] Olago, Victor, Mazvita Muchengeti, Elvira Singh, and Wenlong C. Chen. "Identification of malignancies from free-text histopathology reports using a multi-model supervised machine learning approach." *Information* 11, no. 9 (2020): 455. https://doi.org/10.3390/info11090455

[15] Hamann, Hendrik F. *A multi-scale, multi-model, machine-learning solar forecasting technology*. No. DE-EE-0006017. IBM, Yorktown Heights, NY (United States). Thomas J. Watson Research Center, 2017. https://doi.org/10.2172/1395344

[16] Selvakumar, S., A. Sheik Abdullah, and R. Suganya. "Decision support system for type II diabetes and its risk factor prediction using bee-based harmony search and decision tree algorithm." *International Journal of Biomedical Engineering and Technology* 29, no. 1 (2019): 46-67. https://doi.org/10.1504/IJBET.2019.10017862

[17] Suganya, R., S. Rajaram, A. Sheik Abdullah, and V. Rajendran. "Prediction of heart diseases using hybrid feature selection and modified Laplacian pyramid non-linear diffusion with soft computing methods." *International Journal of Biomedical Engineering and Technology* 25, no. 1 (2017): 30-43. https://doi.org/10.1504/IJBET.2017.086550

[18] Abdullah, A. Sheik, A. Manoj, GT Tarun Kishore, and S. Selvakumar. "A New Approach to Remote Health Monitoring using Augmented Reality with WebRTC and WebXR." In *2021 22nd International Arab Conference on Information Technology (ACIT)*, pp. 1-5. IEEE, 2021.

[19] Abdullah, A. Sheik, A. Manoj, and S. Selvakumar. "Assessment and Evaluation of cancer CT images using deep learning Techniques." In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, pp. 399-403. IEEE, 2021. https://doi.org/10.1109/ICSCCC51823.2021.9478176

[20] Abdullah, A. Sheik, R. Parkavi, P. Karthikeyan, and S. Selvakumar. "A Text Analytics-based E-Healthcare Decision Support Model Using Machine Learning Techniques." In *Smart Computational Intelligence in Biomedical and Health Informatics*, pp. 169-182. CRC Press, 2021. https://doi.org/10.1201/9781003109327-12

[21] Abdullah, A. Sheik, A. Manoj, and S. Selvakumar. "A Hybrid Data Analytic Approach to Evaluate the Performance of Stirling Engine using Machine Learning Techniques." In *2021 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1-4. IEEE, 2021. https://doi.org/10.1109/IBSSC53889.2021.9673219

[22] Harsheni, S. K., S. Souganthika, K. Gokul Karthik, A. Sheik Abdullah, and S. Selvakumar. "Analysis of the Risk factors of Heart disease using Step-wise Regression with Statistical evaluation." In *Emerging Trends in Computing and Expert Technology*, pp. 712-718. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-32150-5_70

[23] Mahmood, Sozan Abdullah, and Hunar Abubakir Ahmed. "An improved CNN-based architecture for automatic lung nodule classification." *Medical & Biological Engineering & Computing* 60, no. 7 (2022): 1977-1986. https://doi.org/10.1007/s11517-022-02578-0

[24] Van Nieuwenhuyse, Enid, Sander Hendrickx, Robin Van den Abeele, Bharathwaj Rajan, Lars Lowie, Sebastien Knecht, Mattias Duytschaever, and Nele Vandersickel. "DG-Mapping: a novel software package for the analysis of any type of reentry and focal activation of simulated, experimental or clinical data of cardiac arrhythmia." *Medical & Biological Engineering & Computing* 60, no. 7 (2022): 1929-1945. https://doi.org/10.1007/s11517-022-02550-y

[25] Cao, Peng, Guangqi Wen, Xiaoli Liu, Jinzhu Yang, and Osmar R. Zaiane. "Modeling the dynamic brain network representation for autism spectrum disorder diagnosis." *Medical & Biological Engineering & Computing* 60, no. 7 (2022): 1897-1913. https://doi.org/10.1007/s11517-022-02558-4