Hybrid Pruning with Manhattan Metric Measures (HPMMM) for Diabetes Detection and Prediction

K. Kanmani,
Research scholar, Dr. Ambedkar Govt Arts College, Vyasarpadi , Chennai- 600039, India,
Assistant professor, Department of Computer Applications, College of Science and Humanities,
SRM Institute of Science and Technology, Kattankulathur- 603203. Chennai, TN, India

Dr. A. Murugan
Associate Professor and Head,
PG and Research Department of Computer Science, Dr. Ambedkar Govt. Arts College (Autonomous), Vyasarpadi,
Chennai – 600039, TN, India

Abstract
Diabetes mellitus has been identified as a keen disease which affects the human at any age. Predicting the disease at earlier would help the person in preventing and maintaining it. Towards predicting diabetes there are number of approaches available like Support Vector Machine (SVM), KNN, Genetic Algorithm (GA), Pattern mining and Decision Trees. The methods use clinical, lifestyle, professional, and other features in predicting the possibility of disease, but suffer to achieve higher prediction accuracy. This work proposes a method of Hybrid pruning tree with Manhattan Metric measures which is used to find for early diagnosis of diabetes. To start with, the method reads the PIMA data set and applies feature level normalization technique to remove noisy records and normalize the data set. Further, the features of the data set have been extracted and generate the tree according to the features extracted. Once the tree has been generated, then the method applies tree pruning and trims the tree to a simplified form. With the test sample, the features of the test sample are extracted and measures Multi Feature Manhattan Similarity (MFMS) against the tree available. According to the value of Manhattan distance, the method predicts the chance of person to get affect by diabetes and classifies the sample. The HPMMM model introduces higher prediction accuracy with less time complexity.

Keywords: Diabetes, Pruning, Manhattan Distance, MFMS, Diabetes Prediction, Decision Tree.

1. Introduction

The human society faces number of challenges and diseases in their lifetime. Some of the diseases produce permanent illness and some of them are curable. Also, some of the diseases are chronic and the person can only manage the disease for their lifetime. Diabetes is the one which is having higher impact in the human society. According to the statistics, the major population have been affected by diabetes and the rate of infection is getting increased every year. Such growing rate should be monitored and reduced for the improvement of public health. Towards this, decisive support systems are developed to support the prediction of diabetes with set of symptoms.

The decisive support systems are designed to assist the medical practitioner in making decision. The medical practitioner would have certain experience and there will be human error and they cannot mind everything to consider variety of factors in making decision about predicting the disease. The human population is getting increased and the decisive support system must use huge number of data sets with number of features [1]. The performance of decisive support systems is greatly depending on the set of features considered. Also, the performance of prediction can be improved by adapting different machine learning and artificial intelligence methods [2]. By incorporating such intelligence techniques, the quality of patient's life can be improved.

The growth of information technology has been adapted to various medical problems. For example, the medical practitioners of specific medical organization, who are geographically distributed would communicate with each other and share variety of information about the patient's history and kind of disease and symptoms they are identifying. Also, they share the kind of treatment given to different patients between them and their success rate. Further, the medical practitioners provide treatment through online like phone calls and E-mails [3]. Also, voice over call and video calls are used in providing telemedicine to provide better service [4].

Apart from predicting the disease, finding the type of diabetes is most important. There exist different types of diabetes namely, type-I, type-2, gestational, and other forms [8]. The young adults in the age of 30-40 suffer with Type-I diabetes, where the patient must inject insulin for their life time. But, in case of Type-2 diabetes, the patients at the age above 40 get into this type of disease. Also, it depends on the weight of the person [9]. The machine learning algorithms have great impact in variety of medical problems. In this way, there are number of methods available towards predicting diabetes with Support Vector Machine (SVM), Decision Tree, Genetic Algorithm, Artificial neural network, Bayesian Classification, KNN, and so on. Each method has its own merits and differ with each other in measuring similarity and set of features considered. The method of measuring similarity plays vital role in identifying the class of person. The SVM algorithm measures support value against various class to identify the class of person. Similarly, KNN algorithm computes Euclidean distance between various class of samples towards predicting the disease. The Decision tree algorithm computes the similarity according to the node similarity where genetic Bayesian classification algorithm computes similarity according to the rules available. Similarly, there exist numerous techniques to support the prediction of diabetes.

Number of Times pregnant
Plasma glucose concentration in an oral
Glucose Tolerance Test
Diastolic blood pressure
Triceps Skin fold thickness
2 – hour serum insulin
Body mass Index
Diabetes Pedigree function'
Age

Figure 1: Attributes involved in Diagnosis of Diabetes [1]

The problem of diabetes prediction has been approached with various approaches like ensemble learning, which maintains number of ensembles of diabetes persons and by measuring the similarity among ensembles, the disease prediction is performed. Similarly, genetic algorithm and support vector machines are used towards diabetes prediction according to fitness and support values. The Bayesian classifiers are used to support diabetes prediction which works according to the rule available. Similarly, there are number of approaches available towards the problem and they suffer to achieve expected performance. By considering all these, this article proposes a hybrid pruning tree based diabetes prediction model with Manhattan distance metric. The model performs disease prediction according to different features of PIMA data set and the detailed approach is presented in this section.

2. Review of Literature

There exist numbers of approaches towards predicting diabetes in literature. A set methods are discussed around the problem in this section.

Towards finding the risk associated with diabetes and predicting risk factors of diabetes, an risk analysis model is presented in [4,5] which finds the risks according to the factors and symptoms. On the other side, a early prediction model with economic burden is presented in [6], to support the public health and clinical practice [7].

A high complex screening model is presented in [8], which finds a sub set of population have higher risk and support preventing diabetes. A prevention strategy is presented in [9,10] which consider the quality of life and present a prevention strategy. A lifestyle-based diabetes prediction and recommends a modification scheme in preventing the disease [11].

Towards predicting gestational diabetes, a risk analysis model is presented in [6, 7, 12], which consider lifestyle, history, food habits, aging, ethnic group, and so on. An efficient approach is presented in [13,14], which consider BMI, age, gender, metabolic status in predicting the disease.

An attribute score-based prediction model and risk analysis model is presented in [14], [15]. The method considers various risk factors of individuals and enforces effective feature extraction method to support efficient diabetes prediction. A symptom orient prediction and risk analysis model is presented in [15], which consider various symptoms in predicting the risk. An efficient diabetes prediction model is presented in [18], which consider different categorical features to improve the performance. The complexity of the model in predicting diabetes is reduced by reducing input factors in [19].

The performance of disease prediction is greatly depending on the kind of similarity measures used. The diabetes prediction is performed by using Hamming, Euclidean, Manhattan distance, and Minkowski distance [20], [21]. The application of jacquard coefficient and Euclidean distance in classification and noise prediction is presented in [22]. Towards performing image classification, a weighted Chebyshev distance-based model is presented in [23]. The objective distance is used in different personalized learning in [24], which has been used to measure the distance between the current competency of a student and the expected level to attain learning expectation. To handle the outlier problem, an optimized approach to support healthcare is presented in [25], which handles the outlier problem and improves classification performance.

3. Methodology

The proposed hybrid pruning tree with Manhattan similarity-based diabetes prediction model (HPMMM) predict the disease according to the data set given and the set of symptoms available. Initially, the method applies pre-processing using feature level normalizing technique. The pre-processing algorithm eliminates the noisy data points and normalizes the feature values. The pre-processed data set has been used towards feature extraction and generates the diabetes tree (DT) with number of branches and leafs. A single branch in the tree denotes the class of the diabetes which includes type 1, type 2, gestational, normal and so on. Each branch would have number of siblings which denotes each data point. Similarly, each branch has number of siblings and nodes where the nodes contain the features and values. Generated tree has been pruned at sibling level to trim the tree to support effective disease prediction. Finally, the method applies Leaf level Diabetes prediction algorithm which uses Manhattan distance as a metric for measuring similarity among the data points. The prediction algorithm computes multi feature Manhattan Similarity (MFMS) towards various class of diabetes and disease to perform disease prediction. The detailed approach is presented in this section.

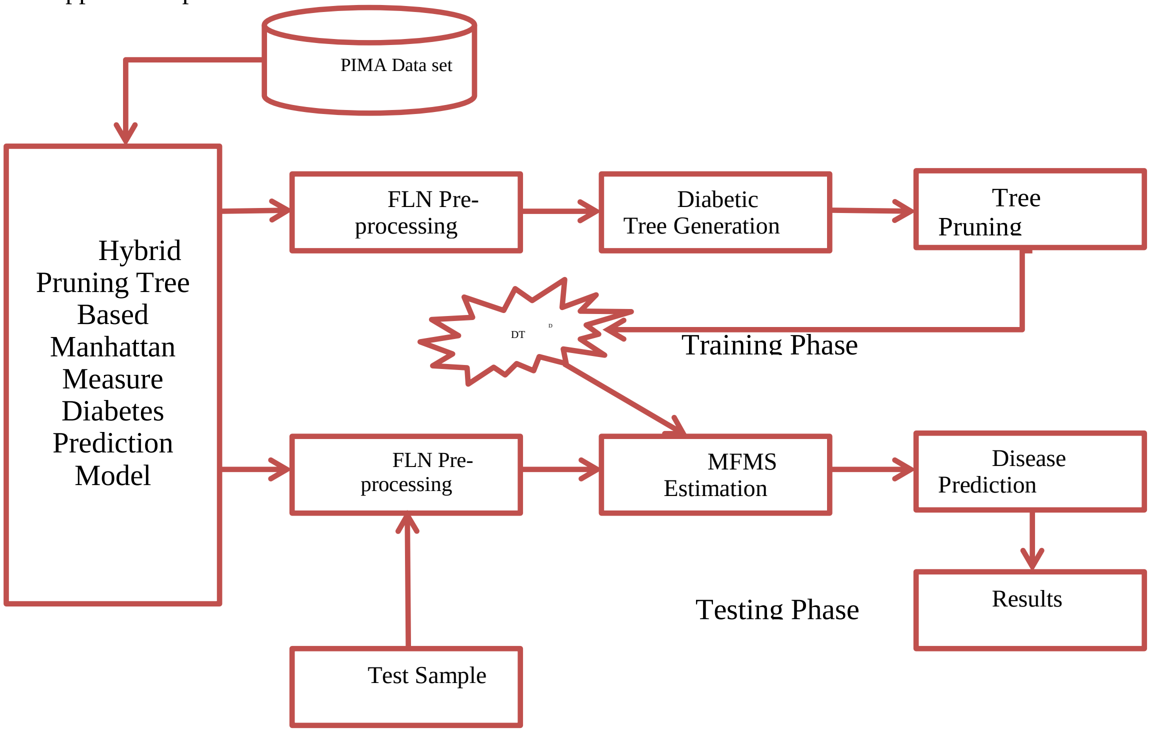


Figure 1: Architecture of proposed diabetes prediction model (HPMMM)

The functional architecture of proposed HPMMM model is presented in Figure 1, where the functional components are discussed in detail in this section.

Feature Level Normalization (FLN) Technique:

The data set considered for pre-processing would contain number of features and large number of tuples. It would contain missing features and values which are considered as noisy records. In order to stimulate the performance of disease prediction, the feature level normalization technique reads all the tuples and identifies the set of features available in the data set. Further, for each feature in the data set, according to the category like binary, numeric, the method computes the value of Adaptive Mean value (AMV) based on the value of the feature in different tuples. Using the value of AMV, the method traverse through each tuple and the presence of missing value has been adjusted with the concern AMV value and for the binary valued features; the method generates random value and replace it. Similarly, the tuples of the data set have been traversed and normalized to support effective disease prediction.

```
Algorithm:
Given: Data Set Pima
Obtain: Pre-processed set Prs
Start
    Read Pima.
    Initialize feature set Fs.
    For each trace Ti
        Feature list Fl = $\sum Features \in Ti$
        For each feature f
            If f∋ $Fs$ then
                Fs = fs $\cup f$
            End
        End
    End
    For each feature f
        If f.type!=binary then
            Compute Adaptive mean value AMV = $\frac{\sum_{i=1}^{size(Pima)} Pima(i).f.value\ ?Pima(i).f.value!=null}{size(Pima)}$
        End
    End
    For each trace Ti
        For each feature f
            If f.type!=binary && f.value==null then
                Ti(f)=f.Amv
            Elseif f.type==binary && f.value==null then
                Ti(f)=Random (0,1)
            End
        End
        Add to pre-processed set Prs = $\sum(Tk \in Prs) \cup Ti$
    End
    Stop
```

The feature level normalization algorithm eliminates the noisy data points and normalizes the traces and tuples to support disease prediction.

Diabetes Tree Generation:

The proposed model transforms the data set in form of diabetes tree which is hierarchical in nature. The method generates three different branches where each belongs and denotes a specific type of diabetes. Also, the method generates number of siblings according to deciding factors like age, BMI, insulin, blood pressure, serum insulin, glucose tolerance test and so on. According to various factors, the method generates number of siblings and at each sibling; the method generates a node to represent the tuple from the data set. The nodes of a sibling have similar feature values and the siblings of a branch represent similar class of diabetes. Even though the tuples of a diabetes class have same label, their feature values and symptoms would be different. This kind of tree generation helps the classification and disease prediction to be performed effectively.

Algorithm:

```
Given: Pre-processed data set Pds.
Obtain: Diabetic Tree DT.
Start
    Read Pds.
    Initialize diabetic tree DT.
    For each class of diabetes Dc
        Generate a branch B.
        Add branch B to DT.
        Class trace set CTS = $\sum_{i=1}^{Size(Pds)} Pds(i).class == Dc$
        Generate random sibling and add to branch B.
        Random tuples set RTS = $Random(CTS,k)$
        For each tuple T
            If size of sibling ==0 then
                Add t to sibling s.
                S = $\sum(leafs \in s) \cup T$
            Else
                Compute Sibling Match Score SMS.
                SMS = $\frac{{}_{i=1}^{size(s)} \sum_{j=1}^{size(s(i))} s(i)(j)==T(j) \Big/ size(T)}{size(s)}$
                If SMS> Th then
                    Add t to s.
                    S = $\sum(leafs \in s) \cup T$
                Else
                    Create new sibling s1.
                    S1 = $\sum(leafs \in s1) \cup T$
                    Add to branch B.
                End
            End
        End
    End
Stop
```

The diabetes tree generation algorithm, generates the diabetes tree from the data set given where the tuples are indexed in the tree according to the SMS score measured against various siblings and branches. Such generated tree has been used to predict the diabetes disease.

Tree Pruning:

Tree pruning is the process of trimming the tree for its shape and dimension. In this stage, the method traverses through each branch, sibling and leaves. By analysing the features of the samples indexed in the tree, the method computes sibling fitness score (SFS). The value of SFS is measured according to the number features present in the leaves of the sibling and number of them have importance in the disease class according to their feature and value. According to SFS value, a subset of feature is identified and finds the leaves to prune the tree. Such pruned tree has been used to perform disease prediction.

```
Algorithm:
Given: diabetes Tree DT.
Obtain: Diabetes Tree DT.
Start
    Read DT.
    For each branch B
        For each sibling s
            Identify set of features Fs = $\sum_{i=1}^{size(s)} Features \in s(i)$
            For each feature F
                Compute sibling fitness score (SFS).
                SFS = $\frac{\sum_{i=1}^{size(s)} S(i).f.value>Th\ OR\ S(i).f.value==1}{size(S)}$
                If SFS>Th then
                    Trim the sibling for the feature f
                End
            End
        End
    End
Stop
```

The tree pruning algorithm estimates the sibling fitness score for any tree. According to the value of SFS, the method trims the tree at sibling level to support disease prediction in more efficient manner.

Disease Prediction:

The proposed hybrid pruning with Manhattan metric measure-based approach performs disease prediction based on the tree generated. Towards this, the input test sample are read and based on the feature present in the test sample, the multi feature Manhattan similarity (MFMS) is measured towards various class of diseases. According to MFMS, the class of tuple is identified as result.

```
Algorithm:

Given: Test sample Ts, Diabetes Tree DT.
Obtain: Class C.
Start
    Read TS and DT.
    For each branch b
        For each sibling s
            Compute Multi feature manhattan similarity MFMS.
            MFMS = $\frac{\overset{size(s)}{\underset{i=1}{\overset{size(features)}{\underset{j=1}{\sum_{k=1}^{size(s)} Dist(features(k(j)),Ts(j)) \Big/ _{Size(s)}}}}}}{Size(S)}$
        End
        Compute cumulative MFMS value.
            CMFMS = $\frac{\sum MFMS}{size(Sibling)}$
    End
    Class C = choose the class with maximum CMFMS.
Stop
```

Disease prediction algorithm performs disease prediction according to the diabetic tree generated. The proposed method computes the value of multi feature Manhattan similarity for various classes of siblings and leafs. Based on the value of MFMS, the method computes the value of CMFMS and identifies a class with maximum CMFMS value. Identified disease class has been returned as result.

Results and Discussion:

The Hybrid Prune based Manhattan Metric measure (HPMMM) based diabetic prediction has been implemented using R language with the PIMA data set. The method is evaluated for its performance on different factors and compared with other approaches. Obtained results have been presented in this section in detail.

Parameter	Value
Data Set	PIMA
Number of Features	11
Number of tuples	605
Tool used	R-language

Table 2: Evaluation Details

The evaluation details used towards performance evaluation has been presented in Table 1, which has been measured under various parameters and analysed with others.

Disease Prediction Accuracy %			
	1000 Tuples	3000 Tuples	5000 Tuples
CART	78	81	85
Ensemble Learning	81	84	89
SVM	83	87	92
HPMMM	87	91	96

Table 3: Performance on Disease Prediction Accuracy %

The accuracy in predicting the disease is measured for proposed and other methods and displayed in Table 3, where the HPMMM approach produces higher accuracy than other approaches.

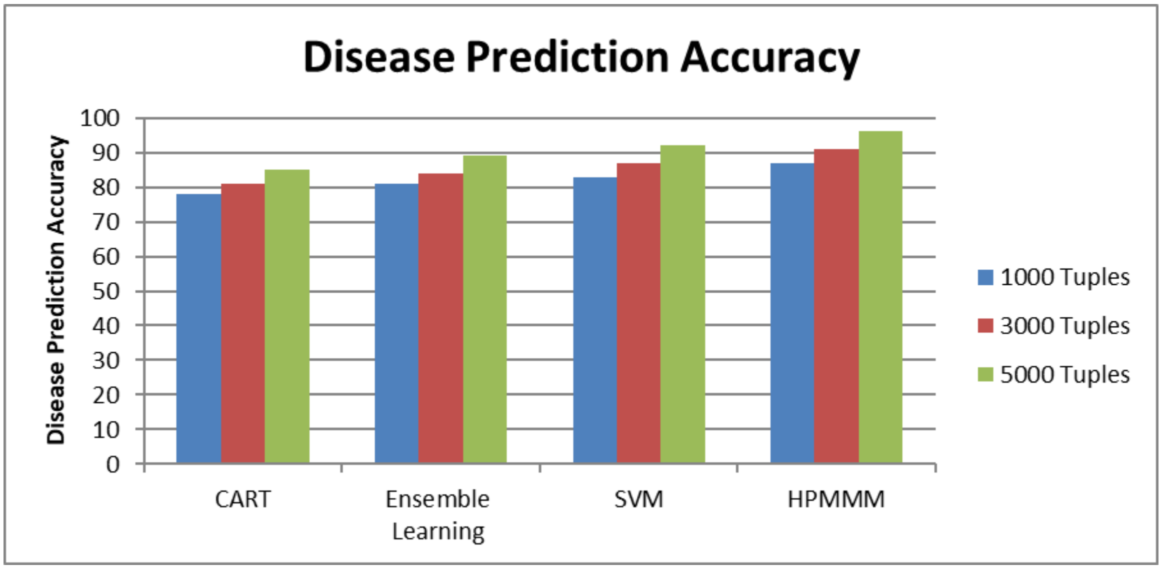


Figure 2: Analysis on Disease Prediction Accuracy

The performance in disease prediction produced by various approaches were measured and presented in Figure 2. The proposed HPMMM algorithm has produces higher accuracy than other methods.

False Prediction Ratio %			
	1000 Tuples	3000 Tuples	5000 Tuples
CART	22	19	15
Ensemble Learning	19	16	11
SVM	17	13	8
HPMMM	13	9	4

Table 4: False Prediction Ratio

The false prediction ratio produced by various approaches are measured and presented in Table 4, where the proposed HPMMM model has produces less false ratio compare to others.

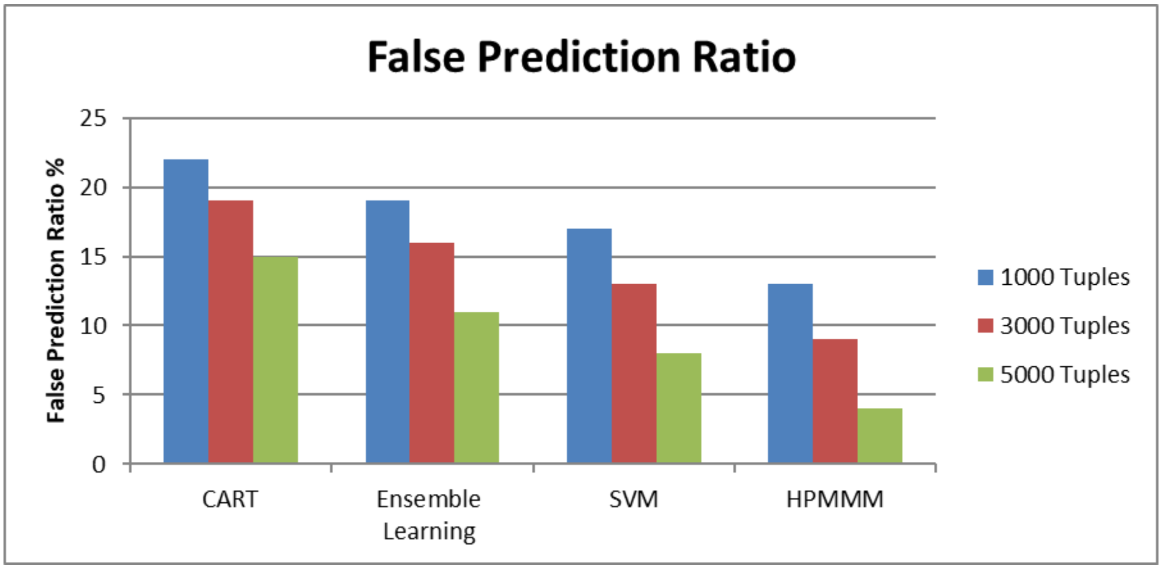


Figure 3: Analysis on False Classification Ratio

The false prediction ratio introduced by different approaches are measured and compared in Figure 3, where the proposed HPMMM algorithm has produced less false ratio compare to other approaches.

Time Complexity in Disease Prediction			
	1000 Tuples	3000 Tuples	5000 Tuples
CART	38	61	85
Ensemble Learning	35	58	79
SVM	31	52	72
HPMMM	17	31	46

Table 5: Analysis on Time Complexity in seconds

The time complexity introduced by various approaches in predicting the disease has been measured and presented in Table 5, where the proposed HPMMM model has produced less time complexity than other methods.

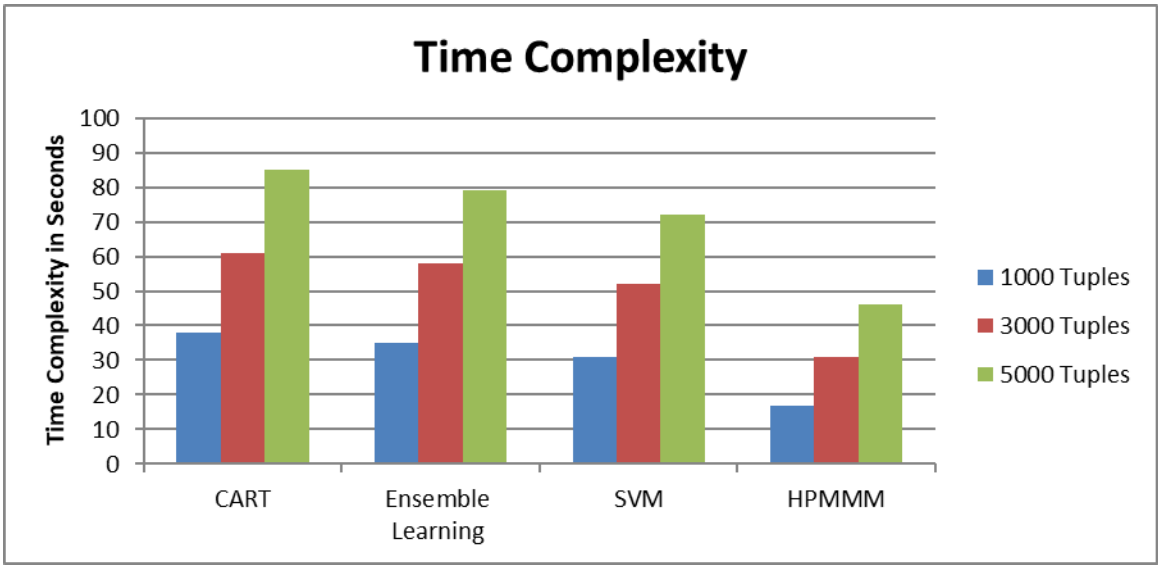
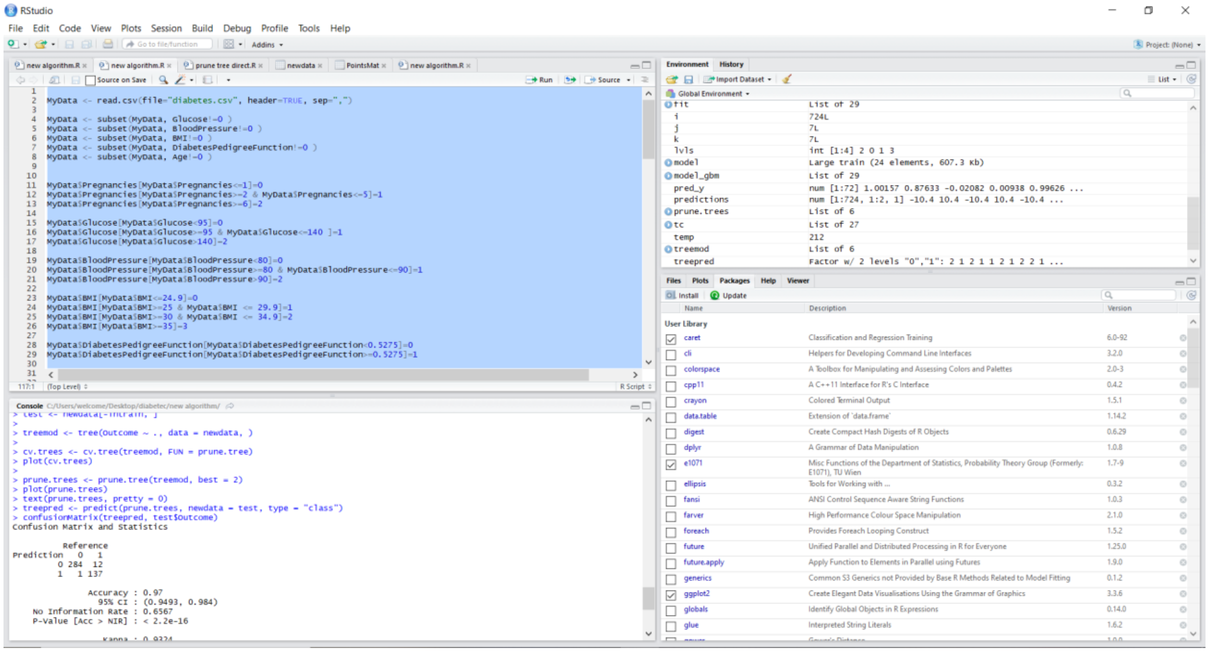


Figure 4: Analysis on Time Complexity

The value of time complexity towards disease prediction has been measured and compared with the result of other methods in Figure 4. The proposed HPMMM approach has produced less time complexity compare to other methods.



```

MyData <- read.csv(file="diabetes.csv", header=TRUE, sep=",")

MyData <- subset(MyData, Glucose!=0 )
MyData <- subset(MyData, BloodPressure!=0 )
MyData <- subset(MyData, BMI!=0 )
MyData <- subset(MyData, DiabetesPedigreeFunction!=0 )
MyData <- subset(MyData, Age!=0 )


MyData$Pregnancies[MyData$Pregnancies<=1]=0
MyData$Pregnancies[MyData$Pregnancies>=2 & MyData$Pregnancies<=5]=1
MyData$Pregnancies[MyData$Pregnancies>=6]=2

MyData$Glucose[MyData$Glucose<95]=0
MyData$Glucose[MyData$Glucose>=95 & MyData$Glucose<=140 ]=1
MyData$Glucose[MyData$Glucose>140]=2

MyData$BloodPressure[MyData$BloodPressure<80]=0
MyData$BloodPressure[MyData$BloodPressure>=80 & MyData$BloodPressure<=90]=1
MyData$BloodPressure[MyData$BloodPressure>90]=2

MyData$BMI[MyData$BMI<=24.9]=0
MyData$BMI[MyData$BMI>=25 & MyData$BMI <= 29.9]=1
MyData$BMI[MyData$BMI>=30 & MyData$BMI <= 34.9]=2
MyData$BMI[MyData$BMI>=35]=3

MyData$DiabetesPedigreeFunction[MyData$DiabetesPedigreeFunction<0.5275]=0
MyData$DiabetesPedigreeFunction[MyData$DiabetesPedigreeFunction>=0.5275]=1

```

Console C:/Users/welcome/Desktop/diabetec/new algorithm/

```
> test <- newdata[-intrain, ]
>
> treemod <- tree(Outcome ~ ., data = newdata, )
>
> cv.trees <- cv.tree(treemod, FUN = prune.tree)
> plot(cv.trees)
>
> prune.trees <- prune.tree(treemod, best = 2)
> plot(prune.trees)
> text(prune.trees, pretty = 0)
> treepred <- predict(prune.trees, newdata = test, type = "class")
> confusionMatrix(treepred, test$Outcome)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 284  12
         1   1 137

               Accuracy : 0.97
                 95% CI : (0.9493, 0.984)
    No Information Rate : 0.6567
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9324
```

fit	List of 29
i	724L
j	7L
k	7L
lvls	int [1:4] 2 0 1 3
model	Large train (24 elements, 607.3 Kb)
model_gbm	List of 29
pred_y	num [1:72] 1.00157 0.87633 -0.02082 0.00938 0.99626 ...
predictions	num [1:724, 1:2, 1] -10.4 10.4 -10.4 10.4 -10.4 ...
prune.trees	List of 6
tc	List of 27
temp	212
treemod	List of 6
treepred	Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 2 2 1 ...

	Name	Description	Version
	User Library		
☑	caret	Classification and Regression Training	6.0-92
□	cli	Helpers for Developing Command Line Interfaces	3.2.0
□	colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	2.0-3
□	cpp11	A C++11 Interface for R's C Interface	0.4.2
□	crayon	Colored Terminal Output	1.5.1
□	data.table	Extension of `data.frame`	1.14.2
□	digest	Create Compact Hash Digests of R Objects	0.6.29
□	dplyr	A Grammar of Data Manipulation	1.0.8
☑	e1071	Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien	1.7-9
□	ellipsis	Tools for Working with ...	0.3.2
□	fansi	ANSI Control Sequence Aware String Functions	1.0.3
□	farver	High Performance Colour Space Manipulation	2.1.0
□	foreach	Provides Foreach Looping Construct	1.5.2
□	future	Unified Parallel and Distributed Processing in R for Everyone	1.25.0
□	future.apply	Apply Function to Elements in Parallel using Futures	1.9.0
□	generics	Common S3 Generics not Provided by Base R Methods Related to Model Fitting	0.1.2
☑	ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	3.3.6
□	globals	Identify Global Objects in R Expressions	0.14.0
□	glue	Interpreted String Literals	1.6.2

Figure 5. Implementation HPMMM in R - Programming

The snapshot of disease prediction with proposed HPMMM algorithm has been presented in Figure 5.

5. Conclusion

This study proposes a hybrid prediction method, called Hybrid Pruning with Manhattan Metric Measures (HPMMM), for type 2 diabetes prediction. The proposed Hybrid Pruning tree with Manhattan similarity-based diabetes prediction model (HPMMM) predict the disease according to the data set given and the set of symptoms available. Initially, the method applies pre-processing using feature level normalizing technique. The pre-processing algorithm eliminates the noisy data points and normalizes the feature values. The pre-processed data set has been used towards feature extraction and generates the Diabetes Tree (DT) with number of branches and leaves. A single branch in the tree denotes the class of the diabetes which includes type 1, type 2, gestational, normal and so on. Each branch would have number of siblings which denotes each data point. Similarly, each branch has number of siblings and nodes where the nodes contain the features and values. Generated tree has been pruned at sibling level to trim the tree to support effective disease prediction. Finally, the method applies Leaf level Diabetes prediction algorithm which uses Manhattan distance as a metric for measuring similarity among the data points. The prediction algorithm computes Multi Feature Manhattan Similarity (MFMS) towards various class of diabetes and disease to perform disease prediction. The open data set named Pima Indians Diabetes dataset were studied for the experiment, which contains a greater number of records for each set. Implementation of this HPMMM using R programming which provides 97% accuracy. Also, the proposed model reduces the false prediction ratio and time complexity as well.

REFERENCES

[1] Types of diabetes. https:// www. idf. org/ about diabetes/ what- is- diabetes. html.

[2] International Diabetes Federation—facts and figures. https:// www. idf. org/ about diabetes/ what- is- diabetes/ facts- figures. Html

[3] Alghamdi M., Al-Mallah M., Keteyian S., Brawner C., Ehrman J., Sakr S. “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach”, The Henry Ford Exercise Testing (FIT) project. PLoS ONE. 2017;12:e0179805. doi: 10.1371/journal.pone.0179805

[5] Nguyen B.P., Pham H.N., Tran H., Nghiem N., Nguyen Q.H., Do T.T., Tran C.T., Simpson C.R. “Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records”. Comput. Methods Programs Biomed. 2019;182:105055. doi: 10.1016/j.cmpb.2019.105055.

[6] Habibi S., Ahmadi M., Alizadeh S.” Type 2 diabetes mellitus screening and risk factors using decision tree: Results of data mining”. Glob. J. Health Sci. 2015; 7:304. doi: 10.5539/gjhs.v7n5p304.

[7] Ryden L., Standl E., Bartnik M., Van den Berghe G., Betteridge J., De Boer M.J., Cosentino F., Jönsson B., Laakso M., Malmberg K., et al. “Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: Executive summary” The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD) Eur. Heart J. 2007;28:88–136.

[8] Tuso P.” Prediabetes and lifestyle modification: Time to prevent a preventable disease”. Perm. J. 2014;18:88. doi: 10.7812/TPP/14-002

[9] IDF Clinical Guidelines Task Force Global Guideline for Type 2 Diabetes: Recommendations for standard, comprehensive, and minimal care. Diabet. Med. 2006;23:579–593. doi: 10.1111/j.1464-5491.2006.01918.x.

[10] Gregg E.W., Geiss L.S., Saaddine J., Fagot-Campagna A., Beckles G., Parker C., Visscher W., Hartwell T., Liburd L., Narayan K.V., et al. Use of diabetes preventive care and complications risk in two African-American communities”. Am. J. Prev. Med. 2001;21:197–202. doi: 10.1016/S0749-3797(01)00351-8.

[11] Knowler W.C., Barrett-Connor E., Fowler S.E., Hamman R.F., Lachin J.M., Walker E.A., Nathan D.M., “Diabetes Prevention Program Research Group Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin”. N. Engl. J. Med. 2002;346:393–403.

[12]Wild S., Roglic G., Green A., Sicree R., King H. “Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. Diabetes Care.” 2004;27:1047–1053. doi: 10.2337/diacare.27.5.1047.

[13] Engelgau M.M., Narayan K., Herman W.H.” Screening for type 2 diabetes. Diabetes Care.” 2000;23:1563–1580. doi: 10.2337/diacare.23.10.1563.

[14] Rolka D.B., Narayan K.V., Thompson T.J., Goldman D., Lindenmayer J., Alich K., Bacall D., Benjamin E.M., Lamb B., Stuart D.O.,. “Performance of recommended screening tests for undiagnosed diabetes and dysglycemia.” Diabetes Care. 2001;24:1899–1903. doi: 10.2337/diacare.24.11.1899]

[15] Ismail, L., Materwala, H., Tayefi, M... Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. *Arch Computat Methods Eng* **29,** 313–333 (2022). https://doi.org/10.1007/s11831-021-09582-x

[16] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari and M. A. Abdul-Ghani, "Advanced techniques for predicting the future progression of type 2 diabetes", *IEEE Access*, vol. 8, pp. 120537-120547, 2020.

[17] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension", *IEEE Access*, vol. 7, pp. 144777-144789, 2019.

[18] F. Aofa, P. S. Sasongko, Sutikno, Suhartono and W. A. Adzani, "Early detection system of diabetes mellitus disease using artificial neural network backpropagation with adaptive learning rate and particle swarm optimization", *Proc. 2nd Int. Conf. Informat. Comput. Sci. (ICICoS)*, pp. 1-5, Oct. 2018.

[19] M. T. Mira Kania Sabariah, S. T. Aini Hanifa and M. T. Siti Sa’adah, "Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART)", *Proc. Int. Conf. Adv. Inform.: Concept Theory Appl. (ICAICTA)*, pp. 238-242, Aug. 2014.

[20] T. M. Le, T. M. Vo, T. N. Pham and S. V. T. Dao, "A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic", *IEEE Access*, vol. 9, pp. 7869-7884, 2020.

[21] A. Dhar, N. Dash and K. Roy, "Classification of text documents through distance measurement: An experiment with multi-domain Bangla text documents", *Proc. 3rd Int. Conf. Adv. Comput. Commun. Autom. (ICACCA) (Fall)*, pp. 1-6, Sep. 2017

[22] G. Latifa, M. Jazouli, N. Es-Sbai, A. Majda and A. Zarghili, "Comparison between Euclidean and Manhattan distance measure for facial expressions classification", *Proc. Int. Conf. Wireless Technol. Embedded Intell. Syst. (WITS)*, pp. 1-4, Apr. 2017.

[23]N. Kwon, J. Lee, M. Park, I. Yoon and Y. Ahn, "Performance evaluation of distance measurement methods for construction noise prediction using case-based reasoning", *Sustainability*, vol. 11, no. 3, pp. 871, Feb. 2019

[24] S. Chaising and P. Temdee, "Determining recommendations for preventing elderly people from cardiovascular disease complication using objective distance", *Proc. Global Wireless Summit (GWS)*, pp. 151-155, Nov. 2018.

[25] K. Kanmani, Dr. A. Murugan, “Diabetes Prediction Medicament using Optimized SVM algorithm with Outlier detection and removal” IJCSNS, Vol. 22 No. 5 pp. 539-544. , 2022