**Research Gap Justification (Approaches and Techniques)**

**Type 2 diabetes mellitus prediction model based on data mining – Han Wu**

They used 9 factors as their parameters. These data contain personal health statistics as well as outcomes from health checkups. Also, they collected data from 768 patients. Those parameters are,

1. Number of times pregnant (preg)
2. Plasma glucose concentration at 2 h in an [oral glucose tolerance test](https://www.sciencedirect.com/topics/medicine-and-dentistry/oral-glucose-tolerance-test) (plas)
3. Diastolic blood pressure (pres)
4. Triceps skin fold thickness (skin)
5. 2-h serum insulin (insu)
6. Body mass index (bmi)
7. Diabetes pedigree function (pedi)
8. Age (age)
9. Class variable (class)

When focusing the techniques they’ve used, mainly they used WEKA toolkit, K-means algorithm and [logistic regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression) algorithm to application. After using all these techniques their experimental results get around 95% accuracy.

In our research, we considered about 9 factors which are directly connected with T2DM including some risk factors published by WHO. We take those 9 factors as our parameters to increase accurate of our application. Our parameters are,

1. Blood Pressure
2. Age
3. Diabetes Pedigree
4. Skin Thickness
5. Insulin
6. BMI
7. Gestational Diabetes
8. Glucose Level
9. Prediabetes

When focusing the techniques of our research we will hope to use Gradient Boosting Classifier, Random forest classifier, Decision Tree Classifier, Naïve Bayes Classifier, Gaussian Classification and XG-Boost. We hope to use these techniques instead of WEKA toolkit, K-means algorithm and [logistic regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression) algorithm. Because Gradient Boosting Classifier have around 79% accurate level, Random forest classifier have around 78% accurate level, Decision Tree Classifier have around 71% accurate level. By combining all of these we can achieve a good accurate level in our application.

**Prognosis of Diabetes Using Data Mining Approach-Fuzzy C Means Clustering and Support Vector Machine - Ravi Sanakal , Smt. T Jayakumari**

These days, data mining technique is applied in medical research to analyze large volume of medical data. In their research they used data mining technique to analyze the databank of Diabetes disease and diagnose the Diabetes disease. In this study they used FCM and SVM and testing it on a bunch of medical data related to diabetes diagnosis problem.

They got around 94% average accuracy and positive predictive value which is around 88% also Negative predictive value is around 97% when using Fuzzy C-means clustering (FCM). Also they got 59.5% accuracy when they using SMV.

In our research we do not hope to use Fuzzy C-means clustering (FCM) and 2 Support Vector Machine (SVM) because SVM have quite low accuracy level and we chose several suitable classifiers to our research. Gradient Boosting Classifier have around 79% accuracy, Random forest classifier have around 78% accuracy, Decision Tree Classifier have around 71% accuracy. By combining all of these we hope we are able to archive higher accuracy than others.

**A Tool for Diabetes Prediction and Monitoring Using Data Mining Technique - S. R. Priyanka Shetty, Sujata Joshi**

In this research they used huge number of factors which are directly and indirectly affect to the T2DM as their parameters. Those parameters are,

1. Age
2. Weight
3. Physical activity
4. Urination
5. Water consumption
6. Diet
7. Systolic blood pressure
8. Hypertension
9. Tiredness
10. Blurred vision
11. Wound healing
12. Sleepy/drowsy
13. Sudden weight loss
14. Heredity
15. Class

According to these parameters they cannot collect data easily because these parameters are not based on simple medical information. Because of that they are unable to collect huge amount of dataset. Low amount of data can be affected to the accuracy of their implementation.

In this research they did a diabetes prediction and monitoring system. When focusing the techniques, they used ID3 Classification algorithm. When using this ID3 algorithm, there are 94% correctly classified instance and 6% incorrectly classified instance.

In our research, we considered about 9 factors which are directly connected with T2DM including some risk factors published by WHO. Our parameters are,

1. Blood Pressure
2. Age
3. Diabetes Pedigree
4. Skin Thickness
5. Insulin
6. BMI
7. Gestational Diabetes
8. Glucose Level
9. Prediabetes

When we collect data form patients, we can easily collect required data because our parameters are based on medical information. So, we can collect huge amount of data for accurate our application.

When focusing the techniques of our research we will hope to use several techniques such as Gradient Boosting Classifier, Random forest classifier, Decision Tree Classifier etc. We hope to use these techniques instead of ID3 Classification algorithm. Because Gradient Boosting Classifier have around 79% accurate level, Random forest classifier have around 78% accurate level, Decision Tree Classifier have around 71% accurate level. By combining all of these we can achieve a good accurate level in our application.

**Reference**

1. Wu, H., Yang, S., Huang, Z., He, J. and Wang, X., 2017. Type 2 Diabetes Mellitus Prediction Model Based on Data Mining. [online] ScienceDirect.

Available at: <https://www.sciencedirect.com/science/article/pii/S2352914817301405>

[Accessed 15 November 2020].

1. Sanakal, R. and Jayakumari, S., 2020. Prognosis of Diabetes Using Data Mining Approach-Fuzzy C Means Clustering and Support Vector Machine. [online] Citeseerx.ist.psu.edu.   
   Available at: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1072.5839&rep=rep1&type=pdf>   
   [Accessed 15 November 2020].
2. Shetty, S. and Joshi, S., 2016. A Tool for Diabetes Prediction And Monitoring Using Data Mining Technique. [online] J.mecs-press.net.   
   Available at: <http://j.mecs-press.net/ijitcs/ijitcs-v8-n11/IJITCS-V8-N11-4.pdf>  
   [Accessed 15 November 2020].