Existential Methods on Diabetes Detection using Machine Learning

Vaishali Yogesh Baviskar

Abstract: Nowadays, a lot of research is going on in healthcare. One of the significant diseases increased all over the world is Diabetes Mellitus (DM). In this paper, the literature review is done on diabetes prediction using Machine Learning and Deep Learning techniques. Various ML algorithms are used using PIDD (Pima Indian diabetes dataset), and improved k-means using logistic regression among all algorithms achieved the highest accuracy. DL algorithms like CNN and LMST used in diabetic retinopathy images.

Keywords: SVM, NN, Naïve Bayes, KNN, Diabetes

I. INTRODUCTION

Recently diabetes is the major cause of death for all humans. In 2000, 171 million people were predicted, which can increase by 2040 up to 642 million all over the world. This increase in figure needs to pay attention to this disease. Many healthcare institutions across the globe spend billions of dollars on diabetes healthcare. Diabetes patients are categorized into four types as Type1 diabetic, pre-diabetic, Type 2 diabetic, and Gestational. Type 1 occurred due to a lack of insulin in youngsters and grownups. Pre-diabetic is the phase before Type2 and Gestational diabetes occurs in ladies during pregnancy. The diagnosis levels of all these patients can be done on various blood glucose sugar level tests. A1C means higher glucose levels test is done to detect Type1 and pre-diabetes diagnosis. Fasting glucose test is done to detect Type1, Pre diabetes and Type2 diagnosis. OTG- Oral glucose test is done to diagnose pre diabetes, Type2 and gestational disease. High level of glucose can affect on human health and leads to severe conditions like loss of vision, Kidney Neuropathy, Liver problems, Heart problems, and foot issues. Due to high sugar levels, diabetes retinopathy is required to diagnose, which can further cause for vision loss and night blindness.

II. RELATED WORK

A. Machine Learning Algorithms

Authors have shown an analysis by using a Decision tree, K-nearest neighbor, random forest, and support vector machine classifiers. Before preprocessing J48 showed the highest efficiency while after preprocessing, KNN and random forest showed the highest accuracy. An analysis is done on PIDD before and after preprocessing [1].

In [2], the authors proposed a framework to predict disease using machine learning and deep learning techniques on the PIDD dataset. Artificial Neural network (ANN) has got the highest accuracy as deep learning technique, and the Random Forest technique has got the highest precision in machine learning techniques.

In [3], the authors compared multiple prediction models using health checkups, and insurance claims data. Yearly health checks up and health insurance dataset from japan is used. XGBoost algorithm is used to predict Type2 diabetes and has got the highest accuracy.

In [4], the authors discussed diabetic research on 1) prediction and diagnosis, 2) Diabetic complications, 3) Genetic background, and environment and 4) Health care and management. Various methodologies are used as feature extraction and reduction using LDA (Linear Discriminant analysis) - MWSVM (Morlet Wavelet Support vector machine) for diabetes diagnosis, Ant colony classification used set of fuzzy rules to extract features, multivariate regression using support vector regression, fuzzy ontology-based case reasoning, multilayer classification and rotation forest on various datasets like clinical and biological datasets, gut microbiota, Electronic measurements of saliva, demographic, Anthropometric, diagnostic and clinical laboratory measurements and it is observed that SVM has got the highest accuracy among all classifiers. They have also discussed on macrovascular and microvascular diabetic complications, for these researchers used temporal data mining and machine learning algorithms for risk stratification.

In [5], the authors derived a set of predictive models of type2 diabetes complications based on electronic health record, and model validation is done. To deal with missing values and class imbalance in RF, Stepwise feature selection is made with the logistic regression. Various Classification models are used like Logistic regression, Naïve Bayes, Support vector machine and random forest. A risk score is scored based on the temporal threshold, complications, and onset date registered. Clinical Historical dataset for more than ten years has taken from the hospital of Pavia, Italy is considered. The final model taken has got an accuracy of 83%.

Methods used are center, profiling, predictive models training, predictive models construction and predictive models validation. In center profiling optimize features are selected to do an initial analysis. In predictive models training, focused on microvascular complications like nephropathy, neuropathy and retinopathy. In predictive models construction, after the first visit, the patient is predicted whether he will develop microvascular complications or not ?, got the best results for retinopathy and neuropathy cases.

In [6], the authors predicted incident of diabetes using medical records of cardiorespiratory fitness. Methodologies like Data preprocessing, features selection, multiple linear regression,
information gain ranking is done using a Decision tree. Naïve Bayes, Logistic regression and random forest on the Henry ford fit dataset (Patients who underwent treadmill stress). To handle imbalanced datasets, the Synthetic minority oversampling technique (SMOTE) was used. Combined three classifiers i.e. RF, NB, and LMT, and has got an accuracy of 92% achieved higher accuracy of 3.04% compared to other researcher’s prediction model. Efficiently predicted cardiorespiratory fitness data using ensemble machine learning and SMOTE methods.

In [7], authors did an enhancement in prediction model and have got an accuracy of 95% on PIDD and other two datasets as Donated by Dr. Schorling from the Department of Medicine of the University of Virginia School of Medicine and collected an online questionnaire. Improved K – means algorithm is used to remove incorrectly clustered data and to get an optimized dataset where preprocessing is done, and Logistic Regression is used to classify remaining data i.e., whether a person has diabetes or not. Data mining toolkit is used where preprocessing, classifying, ranking algorithms, and the visual interface is done. 10 fold cross-validation is used so that it reduces the bias associated with the random sampling method. The model is evaluated by the confusion matrix. The Mathews correlation coefficient (MCC) is used to measure the accuracy of binary classification. Kappa statistics is used to test the consistency of the model.

In [8], Dimensionality is reduced using Principal Component Analysis and minimum redundancy maximum relevance (mRMR) to avoid redundant features. Decision Tree – C4.5, Random forest, and the Neural Network used as classifiers, 5 fold and 10-fold cross validation method is used on PIDD and physical examination clinical database received in Luzhou, China. The Random forest has got the highest accuracy of 80%. RF is a multifunctional machine learning method and plays a significant role in the ensemble machine learning method. Compare to PCA, mRMR has got the best efficiency.

In [9], the authors developed a deep learning model for retinopathy detection. Various color retina images dataset was taken from the kaggle website. Feature extraction is done using the Convolutional Neural network, and SVM, and KNN classifiers are used for detection. The proposed CNN model achieved good results in discovery. The proposed model used regression activation mapping (RAM) to get more accurate results.

In [10], the authors designed a prediction model. They selected significant attributes by giving sequences to each attribute and used classifiers like J48, Random Forest, Naïve Bayes, MLP, KNN, and Neural network. PIDD dataset is used and based on the best attribute selection, the result of classification techniques are improved. Naïve Bayes has shown the average accuracy which is the highest among all of 82.30%. Features are mapped effectively from low to high dimensions.

### B. Datasets, Evaluation matrix and features

The evaluation metrics used for Diabetes detection in Table 1. In this table, various datasets used for diabetes detection using machine learning techniques and various preprocessing algorithms are given. Most of the predictions are done on PIDD dataset which is publicly available. Also, accuracy comparison is done for various machine learning algorithms.

#### Table 1: Comparison of Dataset, Algorithms and Metrics Used

<table>
<thead>
<tr>
<th>Authors</th>
<th>Algorithms Used</th>
<th>Dataset used</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>J.Pradeep Kandhasamy, S. Balamurali[1]</td>
<td>J48 Decision tree, K-nearest neighbor, random forest and support vector machine</td>
<td>PIDD</td>
<td>768 patients with eight attributes-number of times pregnant, glucose level, diastolic blood pressure, triceps skinfold thickness, serum insulin, BMI, diabetes pedigree function, age and class</td>
</tr>
<tr>
<td>Neha Sharma, Ashima Singh[4]</td>
<td>Artificial Neural network (ANN)</td>
<td>PIDD</td>
<td>768 patients with 8 attributes-number of times pregnant, glucose level, diastolic blood pressure, triceps skinfold thickness, serum insulin, BMI, diabetes pedigree function, age and class</td>
</tr>
<tr>
<td>Masatoshi Nagata, Koichi Takai et al.[9]</td>
<td>XGBoost algorithm, LSTM algorithm based on RNN and L1R, Yeary health checkup and health insurance dataset from japan</td>
<td></td>
<td>Record of 40,000 people aged 20 to 64 years. – profile information(age, sex), Lab test results (e.g., body mass index, blood pressure, HbA1c), and a health questionnaire (e.g., smoking, alcohol intake, exercise level). overall 33 health checkup items</td>
</tr>
<tr>
<td>Arianna Dagliati, Simone Marini, Lucia Sacchi et al.[35]</td>
<td>Logistic regression, Naïve Bayes, Support vector machine and random forest</td>
<td>Clinical Historical dataset for more than 10 years has taken from hospital of Pavia, Italy is taken</td>
<td>943 records – Demographic(age, gender, time to diagnosis), clinical data(BMI, HbA1c, lipid profile, smoking habit),administrative data(antihypertensive therapy)</td>
</tr>
</tbody>
</table>
### Table I: Comparative Analysis of Various Datasets, Accuracy and Classifiers Used

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Dataset</th>
<th>Accuracy</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting Machine</td>
<td>CPSSN (Canadian</td>
<td>84.7</td>
<td>Hang Lai, Huaxiong Huang et al.[2], 2019</td>
</tr>
<tr>
<td></td>
<td>patients dataset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>CPSSN (Canadian</td>
<td>84.0</td>
<td>Hang Lai, Huaxiong Huang et al.[2], 2019</td>
</tr>
<tr>
<td></td>
<td>patients dataset)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural network with 10-fold cross validation</td>
<td>PIDD</td>
<td>85.24</td>
<td>Raghavendra S, Santosh Kumar J, Raghavendra B. K,[3], 2019</td>
</tr>
</tbody>
</table>
Existential Methods on Diabetes Detection using Machine Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Dataset</th>
<th>Accuracy %</th>
<th>Authors</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>Luzhou</td>
<td>78.52</td>
<td>Quan Zu, Kaiyang ku et al. [38], 2018</td>
<td></td>
</tr>
<tr>
<td>J48</td>
<td>Luzhou</td>
<td>78.06</td>
<td>Quan Zu, Kaiyang ku et al. [38], 2018</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td>PIDD</td>
<td>76.04</td>
<td>Quan Zu, Kaiyang ku et al. [38], 2018</td>
<td></td>
</tr>
<tr>
<td>J48</td>
<td>PIDD</td>
<td>72.75</td>
<td>Quan Zu, Kaiyang ku et al. [38], 2018</td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>PIDD</td>
<td>76.30</td>
<td>Dipti Sisodia, Dileep Sisodia [5], 2018</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>PIDD</td>
<td>65.10</td>
<td>Dipti Sisodia, Dileep Sisodia [5], 2018</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>PIDD</td>
<td>73.82</td>
<td>Dipti Sisodia, Dileep Sisodia [5], 2018</td>
<td></td>
</tr>
<tr>
<td>Support vector machine</td>
<td>PIDD</td>
<td>98%</td>
<td>Gandhi [46], 2014</td>
<td></td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>Tabriz, Iran</td>
<td>97.44</td>
<td>Heydari [47], 2013</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td>Iris Image</td>
<td>89.66</td>
<td>Samant [48], 2017</td>
<td></td>
</tr>
<tr>
<td>Fuzzy Logic</td>
<td>UCI</td>
<td>78.00</td>
<td>Ephizbah [49], 2011</td>
<td></td>
</tr>
</tbody>
</table>

III. PROPOSED SYSTEM DESIGN

![Image of proposed system architecture]

IV. ALGORITHMS

Q- Learning Algorithm

**Input:** inp[1….n] all input parameters which is generated by sensors, Threshold group TMn[1….n] and TMax[1….n] for all sensor.

**Output:** Trigger executed on appliances, Buzzer execution and GPS message.

**Step 1:** Read all records from database (R into DB)

**Step 2:** Parts [ ]  Split(R)

**Step 3:** CVal = \( \sum_{k=0}^{n} \text{Parts}[k] \)

**Step 4:** check (CVal with Respective threshold of TMn[1….n] and TMax[1….n])

If(true) execute trigger on respective output appliances.

Else Continue;

**Step 5:** T  get current state with timestamp

**Step 6:** if(T.time > Defined Time)

Active GPS for messaging or on buzzer

Else continue.

**Step 7:** end for

Step 8: return DB

Linear Regression Phase

**Input:** User input file data record which contains {symptoms, disease} segment from train database.

**Output:** Projected weight

**Step 1:** Read R {current input from sensor} from current parameters.

**Step 2:** Map with train features with each sample.

**Step 3:** calculate average weight of train DB with same evidences

\[ \text{Weight}_{i….n} = \sum_{k=0}^{n} \langle S_e \rangle \]

**Step 4:** optimized all n instances and select top k instances top[k]

**Step 5:** Return top[k].disease

V. RESULT AND DISCUSSION

The proposed implementation has done with open source environment, in python. Synthetic dataset has used for training as well as testing with cross fold validation with 5 fold, 10 fold and 15 fold respectively.
Various machine learning algorithms have been used to evaluate the performance analysis of entire execution. Below figure 2 shows classification of proposed system with various sample inputs.

![Figure 4: System accuracy of proposed system with false ratio.](image)

The second experiment analysis has done to detect the efficiency of proposed system, the Figure 3 shows the detail description.

![Figure 5: System accuracy of proposed system with false ratio.](image)

VI. CONCLUSION

It is observed that, diabetes detection is done in the most of the areas like retinopathy, neuropathy, nephropathy, cardiovascular patients. Also, prediction can be used for various health claim insurance areas considering historical datasets. Most of the Machine Learning techniques are used on PIDD dataset by extracting various features. Furthermore research is required for detecting and diagnosing this devastating disease.

REFERENCES

17. R. Bruni and G. Bianchi, “Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2349-2361, 2015
Existential Methods on Diabetes Detection using Machine Learning


34. Ioannis Kavakiotis , Olga Tsve c, Athanasios Salifoglou c, Nicos Maglaveras b,d, Ioannis Vlahavas a, Ioanna Chouvarda b,d, “Machine Learning and data mining methods in diabetes research”, ELSEIVER - Computational and structural biotechnology journal”, 8th January 2017


36. Manal Alghamdi, Mouaz Al- Mallah , Steven Keteyian , Clinton Brawner, Jonathan Ehrman “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach : The Henry Ford Exercise (FIT) project “, PLOS one , 24 July 2017

37. Han Wu, Shengqi Yang *, Zhangqin Huang, Jian He, Xiaoai Wang, “Type 2 diabetes mellitus prediction model based on data mining”, ELSEIVER – Informatics in medicines unlocked”, 12th December 2017


AUTHORS PROFILE

Mrs. Vaihali V. Baviskar, is working as an Assistant Professor in G. H. Raisoni Institute of Engineering and Technology, Wagholi, Pune. She has completed her B.E. in Computer Engineering from North Maharashtra University, M.E. in Computer Engineering from Savitribai Phule Pune University and currently pursuing her PhD from Bennett University, Greater Noida. Her Research area is Artificial Intelligence and Machine Learning. She has published 2 National and 6 International Journal Papers. She has an IETE Membership.