Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach

Minyechil Alehegn, Rahul Raghvendra Joshi, Preeti Mulay

Abstract: Now-a-days there is increase in people suffering from DM (Diabetes mellitus) and this number is growing continuously. So, it is a considerable chronic disease. MLTs (Machine Learning Techniques) can act as a savior for early diagnosis and prediction of DM. ML is another side of Artificial Intelligence so that be used for prediction, recommendation and recovery from disease in early stages. The system proposed in this paper makes use of two datasets viz. PIDD (Pima Indian Diabetes Dataset) and 130_US hospital diabetes data sets. Techniques used for datasets analysis are Random Forest, KNN, Naïve Bayes, and J48. Ensemble approach facilitates in achieving better results. The accuracy of proposed ensemble approach is 93.62% for PIDD and 98.56% for 130_US hospital dataset.

Index Terms: J48, Diabetes, Ensemble KNN, Naïve Bayes, Random forest,

1 INTRODUCTION

Diabetes usually known as DM. It is a kind of metabolic diseases in which patients suffer from blood glucose problems due to abnormal production and release of insulin. As per WHO report on 14th November 2016, i.e. on World Diabetes Day, 422 million adults are living with diabetes, and 1.6 million people who lost their life due to DM [1]. In 2016, 1.6 million deaths were directly caused by diabetes [1]. So, it is one of the seriously need to be considered kind of chronic disease around the world. DM can cause damage to different body parts viz., nerves, eyes, heart, to name a few. Every year millions of people got affected by this life threatening disease in both civilized and non-civilized parts of the world. CDCP (Center for Disease Control and Prevention) project that during 2001 to 2009 there is 23% increase in Type II diabetes in US [2]. Many countries, Organization, and different health sectors are also worried about this chronic disease for achieving control and prevention in order to mitigate it in early stages, so that person life can be saved. Different variants of DM are there viz. Type I, Type II, Juvenile and Gestational. Type I is insulin dependent, Type II is insulin independent, Gestational can happen during pregnancy and Juvenile diabetes after birth of a baby. According to Canadian Diabetes Association (CDA) in coming 10 years that is during 2010 to 2020, there will be a predictable growth from 2.5 to 3.7 million for people suffering from chronic diseases [3]. So, by looking at these statistics diabetes and other chronic diseases analysis plays a vital role in saving patients life. Moloud et al. [4] discussed ML algorithms used for analysis and prediction purpose will have different processing powers. Meng, Xue-Hui, et al. [5] showed that ML methods/tactics are helpful in getting better insights, patterns from input health data. Bashir, Saba, et al. [6] confirmed that single machine techniques are less effective as compared distributed implementation as well doesn’t work well for single dataset. So, in this paper proposed approach makes use of two datasets i.e. PIDD and 130_US hospital diabetes datasets. Also, ML techniques used here are Random Forest, KNN, Naïve Bayes, and J48 etc. This paper spans over five sections viz., related work, methodology, proposed prediction and classification along with results, conclusions. References used in this paper are consolidated at the last.

2 RELATED WORK

Song et al. [7] explained and described using various factors such as Age, Glucose, BP, BMI, Skin Thickness etc. Diabetes Pedigree function, insulin, and pregnancy parameters not included [7]. Small sample data used in this study for the prediction of DM. Algorithms used were EM, LR, GMM, SVM and ANN. ANN showed high accuracy and performance [7]. Loannis et al. [8] used Naïve Bayes, SVM, and LR. 10 fold cross validation evaluation method was applied. These individual methods compared based on their accuracy value. Among these three algorithms SVM provides visible result than that of other prediction. Data origin, its kind, and dimensionality are some of the factors related to accuracy. Nilashi et al. [9] for noise removal and pre-processing used two clustering techniques viz., PCA and EM. The medical datasets used were Diabetes, Heart, etc. Removal of noise from instances helped to get better accuracy. Yunsheng et al. [1] removed parameters having less influence which in a way can increase or improve performance and gives better outcome .Francesco et al.[10] to escalate the accuracy of algorithm used feature selection mechanism which plays a great role in improvisation of both accuracy as well as performance. HT, DT, BN, MP, Jri, and RF methods were used for analysis and prediction. Best first (BF) and greedy used for feature selection. Hoeffding Tree was shown good performance and better accuracy, Pradeep et al. [11] not used any cross validation. J48, NBs, Cv Parameter selection (CPS), Simple Cart (SC), ANN, ZeroR, filtered classifier, and KNN techniques were applied. Naïve Bayes provide better accuracy in case of DM than other techniques. ANN and KNN gave accurate result for other datasets than that of other prediction algorithms. Sajida et al. [12] used three technique viz. Bagging, J48 and

---

- Minyechil Alehegn is faculty at IT department of Mizan Tepi University, Ethiopia.
- Rahul Raghvendra Joshi and Dr. Preeti Mulay are faculty at CS/IT department of Symbiosis Internal (Deemed University), Pune, Maharashtra, India.

---

IJSTR©2019
www.ijstr.org

1346
Adaboost on CPCSSN dataset to predict the DM at early stage to prevent from early death. Adaboost technique shown effective and improved result compared to other data mining methods. Kamadi et al. [13] data reduction is the problem in case of classification and it also plays a great role in the prediction. To get good accuracy data need to be reduced. PCA does pre-processing. So, for getting better results data need to be reduced. Pradeep & Dr. Naveen [14] showed accuracy of the prediction algorithm is different before and after pre-processing. Decision Tree Technique showed good accuracy before pre-processing the DM dataset. Two algorithms Random Forest and Support Vector Machine does better after pre-processing of DM dataset. Santhanam and Padmavathi [15] dimensionality reduction plays an important role to improve accuracy. K-Means and Genetic Algorithm used to reduce dimensions. 10K folds cross validation mechanism used for evaluating the data. Un-reduced data proved to be less accurate than that of reduced data. Xue-Hui Meng et al. [16] data can be collected by different means like by using distributed questioner also; LR, J48, and ANN. Decision Tree classification technique gave better accuracy than that of ANN and LR. Abdullah et al. [17] OD (Oracle Database) and ODM (Oracle Data Miner) used for storage analysis. Target variables can be identified based on their percentage of necessity. Malgorzata et al. [18] validity of population based forecast was studied using Closed Loop (CL). Juliahippisley et al. [19] used OD score and 5K cross validation methods to get better result. Philippa et al. [20] in his study the most important factors related to Type II diabetes were identified by using Genetics Score (GS) data mining or machine learning algorithm. Robert et al. [90] used CACS and got better results. Pérez et al. [21] ANN, ML or Data Mining prediction algorithms or techniques were applied. ANN replaces human brain with new technology. It is somewhat complex and difficult to for Fresher to work out. Results of this predictor are better in almost all cases. Kazuaki et al. [22] relationship or association of genes was identified using LR method. Muhammad et al. [23] identified risk factors using association rules. E. S. Kilpatrick et al. [24] mean blood glucose risk factor showed best predictor than that of HbA1c in Type I diabetes using Cox Regression (CR).

3 FINDINGS FROM EXISTING LITERATURE RELATED TO DM PREDICTION

ML or Data Mining as well as AI with health care industry are doing well in terms prediction. Findings from existing literature related to DM prediction are as follows:

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Authors</th>
<th>Methodologies</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Welfeng Xu et al.[25]</td>
<td>Adaboost,ID3,NB, and RF</td>
<td>Better accuracy by RF and less accuracy by ID3</td>
</tr>
<tr>
<td>2</td>
<td>Messan et al.[26]</td>
<td>LR,GMMANN,SV, M, ELM</td>
<td>ANN was best with 99% accuracy</td>
</tr>
<tr>
<td>3</td>
<td>Loannis et al.[8]</td>
<td>LR, NB, SVM</td>
<td>84% accuracy by SVM with 10K cross validation</td>
</tr>
<tr>
<td>4</td>
<td>Mehrbaksh et al.[9]</td>
<td>PCA, EM, CART</td>
<td>Noise removal played an important role.</td>
</tr>
<tr>
<td>5</td>
<td>Tao et al.[27]</td>
<td>NBs, KNN, J48, RF, LR, and SVM</td>
<td>Improvisation in cleansing improved system performance.</td>
</tr>
<tr>
<td>6</td>
<td>Yunsheng et al.[7]</td>
<td>KNN, DISKR</td>
<td>Less important attributes and outlier were removed for 78% obtaining better results.</td>
</tr>
<tr>
<td>7</td>
<td>Francesco et al.[10]</td>
<td>Hoeffding Tree, Decision Tree, ANN, Jrip, Bayenet, Greedy Stepwise, Best First, and RF</td>
<td>Hoeffding Tree (HT) showed high accuracy with integration of searching algorithm achieved accuracy of 77.5% than that of other method.</td>
</tr>
<tr>
<td>8</td>
<td>Swarupa et al.[28]</td>
<td>NB, ANN, KNN, J48, zeroR, CPS, SC, and FC</td>
<td>Cross validation mechanism not applied. NB showed relatively good performance of 77.01%.</td>
</tr>
<tr>
<td>9</td>
<td>Sajida et al.[12]</td>
<td>Decision Tree, Bagging, and Adaboost</td>
<td>Adaboost showed better performance than other considered method</td>
</tr>
<tr>
<td>10</td>
<td>Munaza Ramzan [29]</td>
<td>J48, NB, and RF</td>
<td>RF showed improved accuracy than other techniques. 10 fold cross validation was effectual in this study.</td>
</tr>
<tr>
<td>12</td>
<td>Pradeep &amp; Dr.Naveen [29]</td>
<td>J48</td>
<td>Feature selection technique used to improve correctness. Decision tree noted as a better performing algorithm.</td>
</tr>
<tr>
<td>14</td>
<td>Pradeep et al.[11]</td>
<td>RF, KNN, Decision Tree, KNN, RF, and SVM</td>
<td>Before pre-processing DT showed accuracy of 73.82%. RF and KNN showed also good performance after pre-processing.</td>
</tr>
<tr>
<td>16</td>
<td>Sankarana &amp; Dr Pramananda [31]</td>
<td>Association Rule using FP growth and Apriori</td>
<td>Easy and simple decision making helped, acted as a defensive and suggestive medicine.</td>
</tr>
<tr>
<td>17</td>
<td>Xue-Hui Men et al.[5]</td>
<td>KNN, LR, and J48</td>
<td>J48 gave accuracy of 78.27%.</td>
</tr>
<tr>
<td>19</td>
<td>Patil et al.[32]</td>
<td>HPM</td>
<td>It gave high accuracy of 92.38%.</td>
</tr>
<tr>
<td>20</td>
<td>Saba et al.[6]</td>
<td>Adaboost, RF, KNN, LR, NB, SVM, and HVM</td>
<td>Different diseases were studied and output of HMV was attractive.</td>
</tr>
<tr>
<td>21</td>
<td>Amit and Pragati [33]</td>
<td>MLPBayesnet, C4.5, and RF</td>
<td>MLP+Bayesnet showed better performance.</td>
</tr>
<tr>
<td>22</td>
<td>Saba et al.[34]</td>
<td>C4.5, Bagging, ID3, and CART</td>
<td>Bagging classifier showed high accuracy than that of other</td>
</tr>
<tr>
<td>No.</td>
<td>Authors</td>
<td>Method</td>
<td>Description</td>
</tr>
<tr>
<td>-----</td>
<td>----------------------------------</td>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>23</td>
<td>Mounika et al. [35]</td>
<td>NB, ZeroR, and oneR</td>
<td>Effective treatment to both old and young patients. NB showed good performance than that of others.</td>
</tr>
<tr>
<td>24</td>
<td>Nongyao and Rungruttikam [36]</td>
<td>NB, LR, Bagging, ANN, Boosting, and J48.</td>
<td>The concept applied by using boosting or bagging. RF showed accuracy of 85.558%.</td>
</tr>
<tr>
<td>25</td>
<td>Dr Saravana et al. [37]</td>
<td>Analysis algorithm in Hadoop</td>
<td>Focused on treatment in healthcare industry using big data analysis. Low cost for proper treatment.</td>
</tr>
<tr>
<td>26</td>
<td>Veena and Anjali [38]</td>
<td>Decision Stump(DS), SVM, J48, and NB</td>
<td>DS (Decision stump) algorithm provided better accuracy of 80.72%.</td>
</tr>
<tr>
<td>27</td>
<td>Kung et al. [39]</td>
<td>KNN, New EM, and opposite sign test</td>
<td>The hybridization of EM and KNN used in this study.</td>
</tr>
<tr>
<td>28</td>
<td>Saravanathan and velmurugan [40]</td>
<td>SVM, J48, CART, and KNN</td>
<td>J48 prediction showed better output with accuracy of 67.15%.</td>
</tr>
<tr>
<td>29</td>
<td>Seokho et al. [41]</td>
<td>SVM, E2_SVM</td>
<td>It was an ensemble approach. E2_SVM was exposed better prediction than normal support vector machine with 80%.</td>
</tr>
<tr>
<td>30</td>
<td>Rian and Irwansyah [42]</td>
<td>Fuzzy rule</td>
<td>Early detection is done by fuzzy rule.</td>
</tr>
<tr>
<td>31</td>
<td>Yang et al. [43]</td>
<td>NB, BN.</td>
<td>BN showed improved accuracy of 72.3%</td>
</tr>
<tr>
<td>32</td>
<td>Lin [44]</td>
<td>NB, SVM, ANN</td>
<td>The meta classifier used in this study. By applying meta classifier the accuracy of the algorithm can be improved.</td>
</tr>
<tr>
<td>33</td>
<td>Vrushali and Rakhi [45]</td>
<td>CLAT</td>
<td>Estimation of Prediction and severity to diabetes can be accurately analyzed.</td>
</tr>
<tr>
<td>34</td>
<td>Emrana et al. [46]</td>
<td>C4.5 and KNN</td>
<td>C4.5 showed results with accuracy of 90.43%.</td>
</tr>
<tr>
<td>35</td>
<td>Nahla et al. [47]</td>
<td>SVM with rule extraction with SQReX-SVM</td>
<td>Hybridized method.</td>
</tr>
<tr>
<td>36</td>
<td>Kamadi et al. [48]</td>
<td>DT, and Gini index, Gaussian fuzzy function</td>
<td>DT model showed better performance.</td>
</tr>
<tr>
<td>37</td>
<td>Sakorn [49]</td>
<td>Expert system with fuzzy rule</td>
<td>Expert system for better treatment developed.</td>
</tr>
<tr>
<td>38</td>
<td>Ayush and Divya [15]</td>
<td>CART</td>
<td>Study showed 75% of accuracy.</td>
</tr>
<tr>
<td>39</td>
<td>Jae et al. [50]</td>
<td>Linear and Wraper selection</td>
<td>Time complexity was minimized.</td>
</tr>
<tr>
<td>40</td>
<td>Bum et al. [51]</td>
<td>LR and Naïve Bayes, Anthropometry</td>
<td>Study was concentrated on forecast of glucose levels. The results showed 74.1% accuracy.</td>
</tr>
<tr>
<td>41</td>
<td>Asma [52]</td>
<td>Decision Tree(DT)</td>
<td>DT showed results of 78.1768%.</td>
</tr>
<tr>
<td>42</td>
<td>Anjli and Varun [53]</td>
<td>SVM</td>
<td>RF and SVM achieved 72% of accuracy.</td>
</tr>
<tr>
<td>43</td>
<td>Aruna and GA, KNN, and</td>
<td>Association</td>
<td>Association between</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Authors</th>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>Prajwala [55]</td>
<td>Random Forest and Decision Tree</td>
<td>RF prediction showed better accuracy than DT.</td>
</tr>
<tr>
<td>45</td>
<td>Emirhan et al. [43]</td>
<td>Rough Set, ANFIS</td>
<td>ANFIS was important and showed improved performance than Rough Set.</td>
</tr>
<tr>
<td>46</td>
<td>Krati et al. [56]</td>
<td>KNN</td>
<td>Dependency on dataset size.</td>
</tr>
<tr>
<td>47</td>
<td>Anuja and Chitra [57]</td>
<td>SVM</td>
<td>SVM prediction method provided 78% of accuracy.</td>
</tr>
<tr>
<td>48</td>
<td>Thirumal et al. [58]</td>
<td>C4.5, KNN, NB, SVM</td>
<td>C4.5 provided better accuracy of 78.2552% as compared to others.</td>
</tr>
<tr>
<td>49</td>
<td>E. S. Kilpatrick et al. [59]</td>
<td>MBG, HBAc1</td>
<td>MBG was better prediction algorithm than that of HBAc1.</td>
</tr>
<tr>
<td>50</td>
<td>N. Sattar et al. [60]</td>
<td>Association</td>
<td>Novel factors for DM were identified</td>
</tr>
<tr>
<td>51</td>
<td>C. Törn et al. [61]</td>
<td>GAD</td>
<td>GAD performs better and gives proper output.</td>
</tr>
<tr>
<td>52</td>
<td>M. S. Lewi et al. [62]</td>
<td>IGFBP-1</td>
<td>Glucose level is one of the prominent factor for DM.</td>
</tr>
<tr>
<td>53</td>
<td>H. Elding et al. [63]</td>
<td>SDDS (Standard Deviation Score)</td>
<td>Birth date is also associated with diabetic risk.</td>
</tr>
<tr>
<td>54</td>
<td>K. Chien et al. [64]</td>
<td>Cox Regression Coefficients(CRC)</td>
<td>Prediction model developed 5K folds cross validation used.</td>
</tr>
<tr>
<td>55</td>
<td>R. W. Grant et al. [65]</td>
<td>Genetic test</td>
<td>Hypothesis formulated for Type II DM.</td>
</tr>
<tr>
<td>56</td>
<td>Y. Vergouw et al. [66]</td>
<td>MLR</td>
<td>Rules were generated for Type I diabetes.</td>
</tr>
<tr>
<td>57</td>
<td>M. Ekelu et al. [67]</td>
<td>HOMA-beta, HOMA</td>
<td>Gestational diabetes occurs only at the pregnancy time.</td>
</tr>
<tr>
<td>58</td>
<td>B. Isom et al. [68]</td>
<td>OGTT</td>
<td>Family history is one of the factors related to diabetes.</td>
</tr>
<tr>
<td>59</td>
<td>matthew I et al. [69]</td>
<td>Risk Adjustment Technique, R2</td>
<td>DM specific measures done in this study.</td>
</tr>
<tr>
<td>60</td>
<td>Anti-jussi pyykkonenetal [68]</td>
<td>Association</td>
<td>Life style leads to poor health.</td>
</tr>
<tr>
<td>61</td>
<td>Dan ziegler et al. [70]</td>
<td>Cox Proportional Hazard Models (CPHM)</td>
<td>Predictors of mortality for diabetic &amp; non diabetic population studied.</td>
</tr>
<tr>
<td>62</td>
<td>Kyoko kogawa satoet al. [71]</td>
<td>LR</td>
<td>The combined of A1C and fast plasma glucose studied.</td>
</tr>
<tr>
<td>63</td>
<td>Alison jet al. [72]</td>
<td>CPHM</td>
<td>Self rated health profile combined with EQ VAS was an important factor.</td>
</tr>
<tr>
<td>64</td>
<td>Alehegn M et al. [73]</td>
<td>SVM, NN, DS, and PEM</td>
<td>PEm showed better performance.</td>
</tr>
<tr>
<td>65</td>
<td>NAnnekleefstr a et al. [74]</td>
<td>Association</td>
<td>Better living style and Type II diabetes are related to each other.</td>
</tr>
<tr>
<td>66</td>
<td>Meredith et al. [75]</td>
<td>Logistic Regression Models (LRM)</td>
<td>Waist-to-height ratio is the most predictive measure.</td>
</tr>
<tr>
<td>67</td>
<td>Mikael et al. [76]</td>
<td>Mean Square(X2)</td>
<td>Type I diabetes analysis was more effectual using Mean Square method.</td>
</tr>
<tr>
<td>68</td>
<td>caroline et Regression Model</td>
<td>Uric acid was better</td>
<td>methods.</td>
</tr>
<tr>
<td>69</td>
<td>sylvia et al.[78]</td>
<td>Association</td>
<td>LB has direct relationship with diabetes as it was confirmed. So, it is one of the factors related to rise of the diseases.</td>
</tr>
<tr>
<td>70</td>
<td>Wei et al.[79]</td>
<td>SVM</td>
<td>Sex and physical activity are considerable attributes for DM.</td>
</tr>
<tr>
<td>71</td>
<td>Mohammadza et al.[80]</td>
<td>Operating Characteristic Curves</td>
<td>Lipid accumulation product was important predictor for DM.</td>
</tr>
<tr>
<td>72</td>
<td>Farzad et al.[81]</td>
<td>SD</td>
<td>HDLC foretelling exists in young women rather than male.</td>
</tr>
<tr>
<td>73</td>
<td>Michael et al.[82]</td>
<td>Archimedes Model (AM)</td>
<td>AM predicted DM with better accuracy and high level of sensitivity.</td>
</tr>
<tr>
<td>74</td>
<td>Earl et al.[83]</td>
<td>Association</td>
<td>Metabolic syndrome is the strong cause for occurrence of diabetes.</td>
</tr>
<tr>
<td>75</td>
<td>Marion et al.[84]</td>
<td>Logistic Regression (LR)</td>
<td>Females living with Type I diabetes studied.</td>
</tr>
<tr>
<td>76</td>
<td>Jan cederholm et al.[85]</td>
<td>Cox Regression Analysis (CRA)</td>
<td>Type II DM studied only in this study.</td>
</tr>
<tr>
<td>77</td>
<td>Amber et al.[86]</td>
<td>Score, Ukpds, Framingham algorithm</td>
<td>Ukpds and score were useful than Framingham algorithm.</td>
</tr>
<tr>
<td>78</td>
<td>Matthias et al.[87]</td>
<td>Operating Characteristic Curve</td>
<td>Life style is one of the factors for DM occurrence.</td>
</tr>
<tr>
<td>79</td>
<td>Heli et al.[87]</td>
<td>Association</td>
<td>Association of HLA and autoantibody was studied</td>
</tr>
<tr>
<td>80</td>
<td>Makrina et al.[88]</td>
<td>Integrated Discrimination Improvement (IDI)</td>
<td>Lifestyle distinctions may not only reduce diseases possibility for a women who were not pregnant.</td>
</tr>
<tr>
<td>81</td>
<td>Mandy et al.[89]</td>
<td>LR</td>
<td>High risks related to DM studied.</td>
</tr>
<tr>
<td>82</td>
<td>Rachel et al.[90]</td>
<td>Bergman Model</td>
<td>This model gave better outcome.</td>
</tr>
<tr>
<td>83</td>
<td>Daniel et al.[90]</td>
<td>ARX model and model-free ZOH</td>
<td>ARX was better than ZOH in this study.</td>
</tr>
<tr>
<td>84</td>
<td>Malgorzata et al.[91]</td>
<td>Closed Loop (CL)</td>
<td>Validation for population based forecast was studied.</td>
</tr>
<tr>
<td>85</td>
<td>Juliahippisley et al.[91]</td>
<td>QD Score</td>
<td>Five cross validation was used to achieve result.</td>
</tr>
<tr>
<td>86</td>
<td>Philippa et al.[92]</td>
<td>Genetics Score (GS)</td>
<td>Most of the risk factors related to Type II DM were identified.</td>
</tr>
<tr>
<td>87</td>
<td>Robert et al.[93]</td>
<td>CACS</td>
<td>CACS predictor showed better results.</td>
</tr>
<tr>
<td>88</td>
<td>Pérez et al.[94]</td>
<td>ANN</td>
<td>ANN outperformed for DM analysis.</td>
</tr>
<tr>
<td>89</td>
<td>Kazuaki et al.[95]</td>
<td>Logistic Regression (LR)</td>
<td>The genes association was studied and identified in relation to DM.</td>
</tr>
<tr>
<td>90</td>
<td>David G. Clayton on [96]</td>
<td>Association</td>
<td>Type I DM causes studied.</td>
</tr>
<tr>
<td>91</td>
<td>Muhammad et al.[97]</td>
<td>ANOVA</td>
<td>Possibility of Type II DM for future generation was analyzed.</td>
</tr>
<tr>
<td>92</td>
<td>Abu nasar et al.[98]</td>
<td>CLIPS</td>
<td>Expert system for plant diseases developed.</td>
</tr>
<tr>
<td>93</td>
<td>Othman et</td>
<td>ANN</td>
<td>ANN outperformed as</td>
</tr>
</tbody>
</table>

4) Proposed Prediction and Classification Method

The dataset from UCI repository i.e. PID and 30-US hospital dataset were considered [54]. PID involves 768 records and 8 characteristics with one target class and 30-US hospital dataset consists of 93743 instances and 48 features. Ensemble or hybrid model with base learner for forecasting applied. Based on literature, four most known data mining or machine learning prediction techniques were considered. They are described as below.

4.1 Random Forest (RF)

It is one of the prediction algorithms in the machine learning area. It is more adaptable to ensemble approach. It can easily tackle large datasets.

4.2 K-Nearest Neighbour (KNN)

It is grouped under the category of lazy prediction technique. It is easy technique helps to group new work based on similarity measure [18]. The training data are sorted in this algorithm. Define k - number of nearby neighbours. Distance between training samples and instance. Estimate inaccessibility of the training sections arranged and the neighbouring neighbour based on the minimum - the remoteness is determined in the subsequent step. Training data for all categories defined. Majority of the class of nearest neighbours have the forecast value of the query record. Euclidean = \sqrt{\sum (x_i - y_i)^2}

4.3 Naive Bayes (NB)

NB is prevalent and fits when the input data is large and need a short computational time. Calculation based on prospect is done by applying Bayes formula [19].

P(D/h) = (P(h/D) * P(D)) / (P(h)) Where P(h) is refers to prior probability of hypothesis, h in this case is true P(D) is refers to prior possibility of training data D P (h/D) is refers to possibility of h given D P (D/h) is refers to possibility of D given h

4.4 J48 (Decision Tree)

It is also called decision tree prediction algorithm. It is the upgraded version of ID3 classification machine learning algorithm. By using this algorithm, it is possible to construct rules which are simple and easy to understand [47]. Check for the above base cases. For each Instance I, find the consistent information gain ratio by splitting on I. Let I better be the feature with the maximum or better normalized information
gain. Make a choice node that splits on $I_{\text{better}}$. Recur or repeat on the sub lists obtained by splitting on $I_{\text{better}}$, and add those bulges as leaf of node.

### 4.5 HYBRID MODEL

By combining 4.1 to 4.4 to one method – accuracy will be maximum [12, 47].

![Fig 1: Proposed Work Flow](image)

### 5 EQUATIONS

#### Table 2 Accuracy of PIDD against considered Algorithms

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Correctly Classified</th>
<th>Incorrectly Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>87.89%</td>
<td>12.11%</td>
</tr>
<tr>
<td>KNN</td>
<td>82.94%</td>
<td>17.06%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>88.41%</td>
<td>11.59%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>89.84%</td>
<td>10.16%</td>
</tr>
<tr>
<td><strong>Proposed Ensemble using stacking</strong></td>
<td><strong>93.62%</strong></td>
<td><strong>6.38%</strong></td>
</tr>
<tr>
<td>Baging-J48</td>
<td>89.97%</td>
<td>10.03%</td>
</tr>
<tr>
<td>Baging-KNN</td>
<td>85.81%</td>
<td>14.19%</td>
</tr>
<tr>
<td>Bagging-Random forest</td>
<td>89.93%</td>
<td>10.07%</td>
</tr>
</tbody>
</table>

#### Table 3 Accuracy of 130 US against considered Algorithms

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Accuracy before fs</th>
<th>Incorrectly classified</th>
<th>Accuracy after fs</th>
<th>Incorrectly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>56.96%</td>
<td>43.04%</td>
<td>84.53%</td>
<td>15.47%</td>
</tr>
<tr>
<td>KNN</td>
<td>46.04%</td>
<td>53.96%</td>
<td>74.59%</td>
<td>25.41%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>56.18%</td>
<td>43.82%</td>
<td>82.26%</td>
<td>17.74%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>48.68%</td>
<td>51.32%</td>
<td>82.80%</td>
<td>17.20%</td>
</tr>
<tr>
<td><strong>Proposed Ensemble using stacking</strong></td>
<td><strong>58.04%</strong></td>
<td><strong>41.96%</strong></td>
<td><strong>88.56%</strong></td>
<td><strong>11.44%</strong></td>
</tr>
<tr>
<td>Baging-J48</td>
<td>57.06%</td>
<td>42.94%</td>
<td>86.96%</td>
<td>13.04%</td>
</tr>
<tr>
<td>Baging-KNN</td>
<td>50.76%</td>
<td>49.24%</td>
<td>77.89%</td>
<td>22.11%</td>
</tr>
<tr>
<td>Bagging-Naïve Bayes</td>
<td>53.76%</td>
<td>46.24%</td>
<td>84.61%</td>
<td>15.39%</td>
</tr>
<tr>
<td>Bagging-Random forest</td>
<td>54.55%</td>
<td>45.5%</td>
<td>85.15%</td>
<td>14.85%</td>
</tr>
</tbody>
</table>

![Fig 2: Accuracy of considered of PIDD for different algorithms](image)

![Fig 3: Accuracy of us-130 hospitals dataset for different algorithms](image)

![Fig 3: Important Features of PIDD](image)

![Fig 4: PIDD Attributes correlation](image)

![Fig 6 Chance of being diagnosed for DM by Glucose](image)
6 Conclusions

Deep Learning or Machine Learning methods have different powers for diverse data sets. In the proposed system two datasets one is large (130_US) and other is small (PIDD) used for analysis. In this work 10K cross validation for evaluation both in single and multiple iterations applied by considering 90% of training and 10 % of testing data. Proposed system used well known and most commonly used machine learning algorithms. Algorithms used in this study are J48, KNN, NB, and Random Forest. The proposed method provides better accuracy of 93.62% in case of PIDD using stacking meta classifier. In case of large dataset 130-us hospital an ensemble method provides better accuracy than single prediction algorithm. Generally in both small and large datasets analysis ensemble method outperformed than a single method. It is also observed that when dataset becomes large the accuracy of the proposed algorithm is not good relatively. NB and J48 prediction algorithm are better for large datasets analysis. KNn technique is not good for large dataset analysis. In this study focus is on DM analysis, in future this hybrid approach or ensemble approach needs to be applied on other diseases for gauging its effectuality.

References


system based ontology for diabetes disease diagnosis. In 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 383-389). IEEE.


[75] MacKay, M. F., Haffner, S. M., Wagenknecht, L. E.,


