Diabetes mellitus early stage risk prediction using machine learning algorithms

Sarra Samet
Laboratory of Mathematics, Informatics and Systems (LAMIS)
University of Larbi Tebessi
Tebessa, Algeria
sarra.samet@univ-tebessa.dz

Mohamed Ridda Laouar
Laboratory of Mathematics, Informatics and Systems (LAMIS)
University of Larbi Tebessi
Tebessa, Algeria
ridda_laouar@yahoo.fr

Issam Bendib
Laboratory of Mathematics, Informatics and Systems (LAMIS)
University of Larbi Tebessi
Tebessa, Algeria
bendibissam@yahoo.fr

2021 International Conference on Networking and Advanced Systems (ICNAS) | 978-1-6654-2821-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICNAS53565.2021.9628955

Abstract— **Diabetic patients are on the rise. Diabetes is one of the most debilitating diseases. Undiagnosed and untreated diabetes can lead to a number of health issues, including heart disease and stroke. It is necessary for the patient to visit a diagnostic institution and contact a doctor. With the advent of machine learning, this important problem has been overcome. A primary objective of this work is to build a model that can reliably predict a person's probability of developing diabetes. To detect diabetes at an early stage, six supervised machine learning classification methods and a hybrid model based on the top three findings are employed. UCI's machine learning repository provides access to the Pima Indians Diabetes Database, which is used in the experiments. All of them are evaluated based on a variety of measures. It is highlighted that the hybrid model which got an accuracy of 90,62% performs better than other state-of-the-art methods.**

Keywords— Diabetes prediction, Data analysis, Hybrid algorithm, Data mining, Machine learning.

I. INTRODUCTION

There are a variety of chronic diseases in the globe today, which are found in both developed and developing countries. Diabetes mellitus, or diabetes as it is now known, is one of these chronic diseases that is responsible for premature death in humans. Increased appetite, thirst, weight loss, weariness, impaired eyesight, and frequent urination are all symptoms. This occurs when the patient's pancreas is unable to create enough insulin, resulting in the body's inability to regulate sugar and glucose levels in the blood. Another cause of diabetes is insulin resistance in cells found in the liver, muscle, and fat [1].

Diabetes has been shown to be on the increase in Asian nations such as India. Various health sectors are working together to anticipate the emergence of chronic illnesses in the future, possibly saving lives. Diabetes can damage numerous human organs, including the eye, heart, nerves, and kidneys [2].

Currently, doctors take a blood sample from their patients and evaluate the sugar concentration in their blood to diagnose diabetes. This procedure used by doctors is time-consuming, and it includes features such as blood pressure, insulin levels, the patient's age, and body mass index (BMI) [2].

If a patient's ancestor has diabetes, there is a possibility that the patient will develop diabetes in the future. Type 1 and Type 2 diabetes are the two forms of diabetes. There are currently no intrusive techniques to predict whether a patient has type 1 diabetes. Type 2 diabetes is unrelated to insulin levels, and it has been found to be more widespread in Pima Indians than in any other population on the planet. If type 2 diabetes is identified early enough, it can be cured [3], [4].

Data mining, machine learning (ML), and statistical approaches are all part of predictive analysis. Through predictive analysis of healthcare data, important judgements and projections may be formed. Predictive analytics can make use of machine learning and regression technologies. To improve patient care, optimize resources and improve clinical results, predictive analysis aims at diagnosing the illness as correctly possible [5], [6].

In medical applications, machine learning has proven to be a viable support tool. It can detect trends in medical records that people are unable to detect. It can aid in analyzing and making future predictions. These predictions can be used to help people avoid contracting certain diseases. Diabetes mellitus is one such ailment. This disease has no cure, although it can be avoided if necessary precautions are taken [7], [8], [9].

To determine if a patient is diabetic or not, Random Forest (RF), Support Vector Machines (SVM), Bayesian Network (BN), and K-Nearest Neighbor (KNN) models, Decision Tree (DT), Artificial Neural Networks (ANN), and Logistic Regression (LR) models can be utilized well.

This is why a large amount of data is collected by various organizations. The National Institute of Diabetes and Digestive and Kidney is one such organization that has collected and provided the Pima India Diabetes Dataset. This dataset will be used in this research has 768 instances, each instance having 8 attributes with which the patients were medically tested upon.

II. RELATED WORK

The field of machine learning based prediction on diabetes diagnostic detection has been extensively studied. A Prediction of Diabetes using Classification Algorithms has been suggested in [10]. Focus of this project is to create a model that can better correctly predict diabetes-related risks in patients. There is also a preliminary diagnosis of diabetes in this thesis, which employs three machine learning methods. The outcomes of all three algorithms are computed using a variety of metrics. Precision is assessed in terms of instances correctly categorized and those that are erroneously categorized. Results demonstrate Naive Bayes is superior to other algorithms, with an accuracy of 76.30% rise in the accuracy of the predictions.

978-1-6654-2821-7/21/$31.00 ©2021 IEEE

Diabetes Analysis Using Various Machine Learning Methodologies in [11] shows a more reliable technique for evaluating a patient's diabetic perceived risk is the goal of this research. It uses classification techniques such as Decision Trees, ANNs, and SVMs to classify the patterns. 85 % for NB and 77 % for SVM.

Research on Diabetes Prediction Based on Machine Learning was presented in [12]. Supervised machine learning techniques such as SVM, Naive Bayes classifier and LightGBM are used in this paper. In a comparative examination of classification and identification accuracy, the support vector machine shows the best efficiency.

A survey on current advancements in automated diagnosis of diabetic retinopathy (DR) is proposed in [13]. Deep learning, machine learning and medical image processing are used in this project. DR is diagnosed with a CAD procedure, which is used to diagnose the disease. Medical diagnosis Image Recognition is used to analyze medical pictures.

III. Methodology Used

The proposed approach is as follow:

- Step I: Gather Diabetes Dataset and analyze it.
- Step II: Preprocess the data.
- Step III: Divide the data into two sets: training and testing.
- Step IV: Develop an ensemble model
  - Level 0: Stack kNN, SVM, and Decision Tree in the base layer.
  - Level 1: Create a Logistic Regression Model in the meta layer based on the results of the base layer.
- Step V: Using equations (1) and (2), find the predictive outcome using the testing set on the model.

$$Ensemble\_Model.fit(train\_x\_data, train\_y\_data)(1)$$

$$Predictive\_Outcome = model.predict(test\_x\_data)(2)$$

- Step VI: Determine the model's accuracy.

A. Dataset Specification

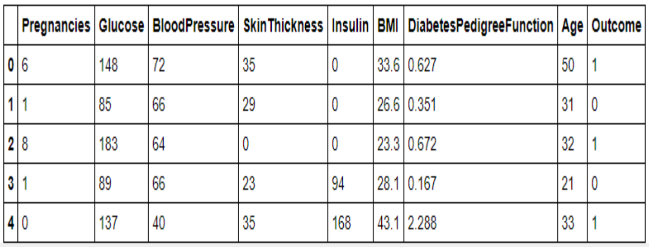
On the Kaggle website, the PIMA diabetes dataset may be found[14]. This dataset, which comes from the National Institute of Diabetes and Digestive and Kidney Diseases, may be used to diagnostically predict whether or not a patient has diabetes based on certain diagnostic metrics provided in the collection.

It consists of numerous medical parameters and one binary-valued dependent (outcome) parameter. This dataset is primarily for females. There are 768 rows and 9 columns in this dataset we have 'Outcome' as the target variable. There are nine attributes in each row, such as:

TABLE I. Description And Type Of Dataset's Attribute

Attribute n°	Attribute Description	Variable type
1	Pregnancies (Number of times pregnant).	Integer
2	Glucose (Plasma glucose concentration within 2 h duration with an oral glucose tolerance test).	Real
3	Blood Pressure (Diastolic/Systolic blood pressure (mm Hg)).	Real
4	The Skin Thickness (Triceps skin fold thickness (mm)).	Real
5	Insulin (2h duration serum insulin (mu U/ml)).	Real
6	BMI (body mass index (weight in kg/(height in m)^2)).	Real
7	Diabetes pedigree function.	Real
8	Age (years).	Integer
9	Outcomes (with diabetes (1) or not (0)).	Binary

These columns denote some specific medical conditions, and the following is a snapshot of the datasets:



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 1. PIMA's snapshot.

B. Data Visualization

Our data is saved in a CSV file, which must be imported into the notebook as a data frame using the Python Pandas module. We can perform numerous data analyses after importing the data. To plot various graphs, we must load the Matplotlib module, which includes all graph-plotting techniques.

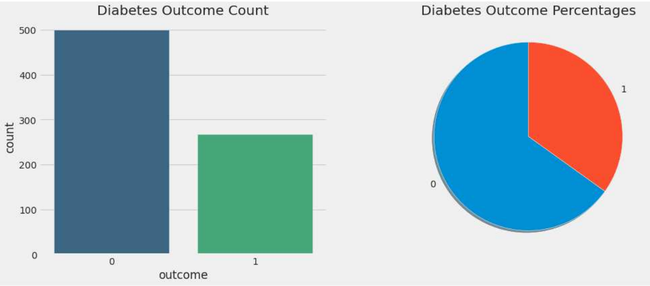
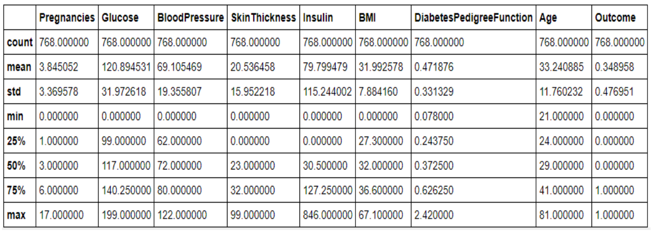


Fig. 2. Diabetes Outcome Count and Percentages.

"fig.2" shows the distribution of the outcome variable there are 500 non-diabetic case and 286 diabetics.

There are a total of 786 cases Non-diabetics are nearly twice as high as diabetic patients.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Fig. 3. Dataset description.

On the figure 3 all of the parameters are calculated using the "df.describe()" method in order to get the central tendency of different fields of the dataset which includes mean, median, mode.

The describe function lists out: total count, mean, standard deviation, minimum value, First quartile(Q1), Median(Q2), Third Quartile(Q3), maximum value.

Count informs us how many NoN-empty rows there are in a feature, the value of std indicates the feature's Standard Deviation. And percentile/quartile for each feature, we can find outliers using them.

It’s normal to find a 0 pregnancies and we notice that there are no zero values in the Age and DiabetesPedigreeFunction.

As we have a very small number of variables, we can have a global comprehension of our variables and their connection. We can see the distribution of the data in form of histograms and we can even draw a line of probability distribution.

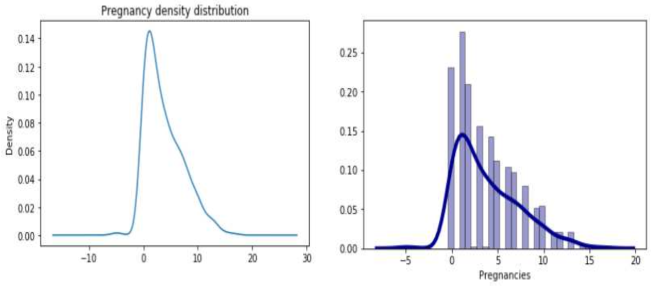


Fig. 4. Pregnancies density distribution and density histogram.

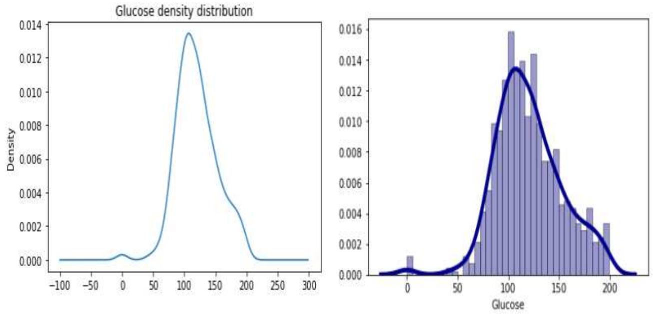


Fig. 5. Glucose density distribution and density histogram.

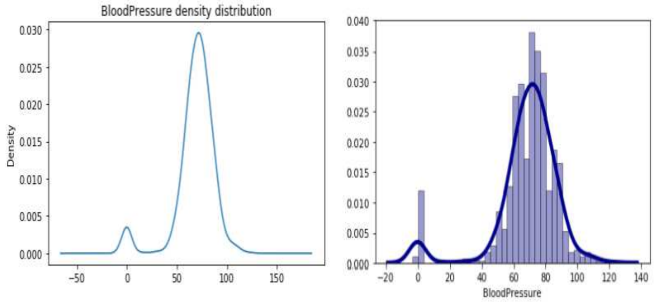


Fig. 6. Blood pressure density distribution and density histogram.

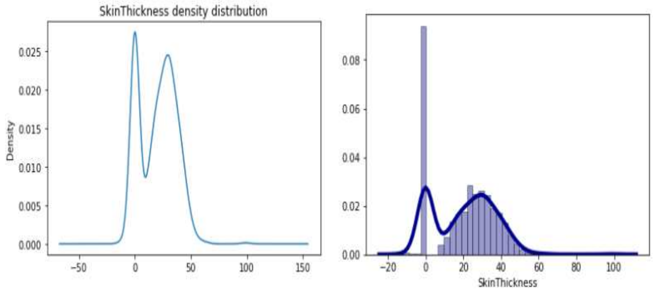


Fig. 7. Skin thickness density distribution and density histogram.

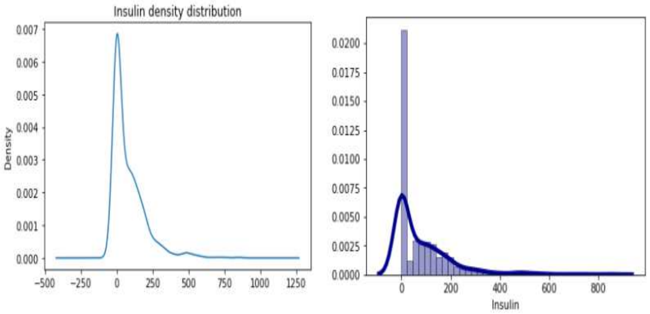


Fig. 8. Insulin density distribution and density histogram.

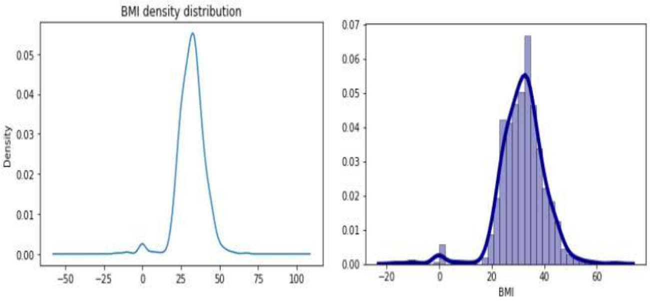


Fig. 9. BMI density distribution and density histogram.

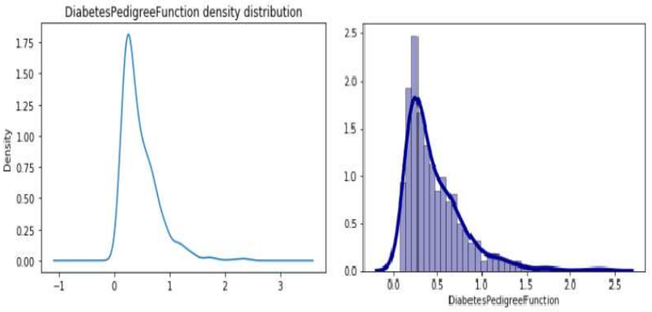


Fig. 10. Diabetes pedigree function density distribution and histogram.

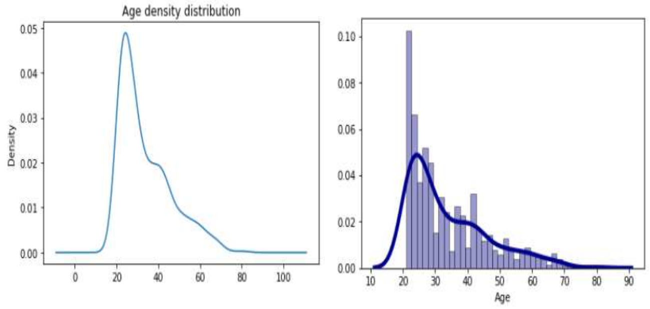


Fig. 11. Age density distribution and density histogram.

To find the correlation of different fields we use corr() and plot it using heatmap() function in seaborn. We can observe the association between the fields in the heatmap below. Correlations are stronger in areas that are lighter, and vice versa in places that are darker or have minimal connection.

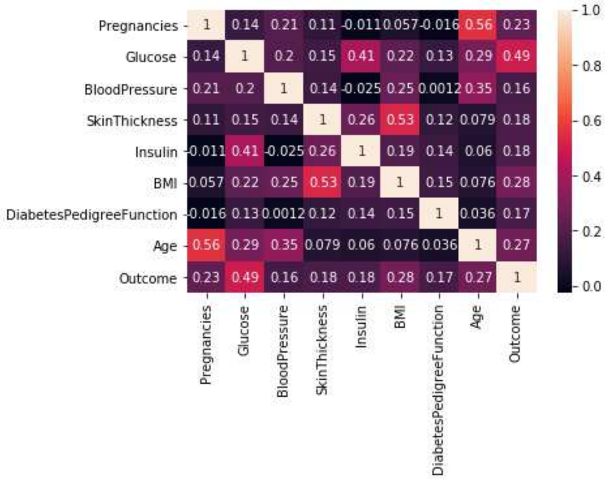
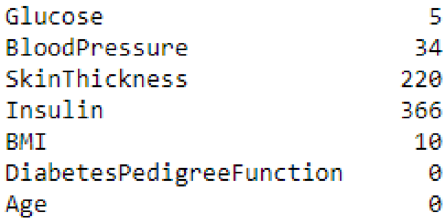


Fig. 12. Correlation between different fields.

From above heatmap we can infer that ‘Glucose’ and ‘Outcome’ have a correlation coefficient of 0.49. We also see a prominent correlation between ‘Age’ and ‘Pregnancies’ i.e. 0.56 which is self explanatory as the age of a woman increases the number of pregnancies she had would tend to increase. We can also see that ‘BMI’ and ‘skin thickness’ have a high correlation with each other. There is also a correlation between ‘insulin’ and ‘glucose’.

C. Data preprocessing

Insulin has the most zeros (366 records), followed by Skinthickness (220 records). The simplest method to deal with these zero-value entries is to delete them, however this removes a large percentage of the records from the dataset and makes it impossible to derive useful insights from the remaining data “Fig.13”.



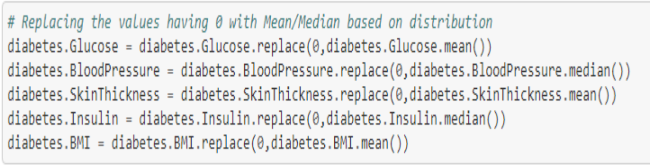
```
Glucose                       5
BloodPressure                34
SkinThickness               220
Insulin                     366
BMI                          10
DiabetesPedigreeFunction      0
Age                           0
```

Fig. 13. Number of records with zero values in each column.

The science of data mining has a problem with missing data. It occurs when there are no corresponding values for any given instance, or when the values are irrelevant or improperly collected when the data is provided. The accuracy and performance of a database can be harmed by missing database values. For Pima Indian diabetes, different features have null values in the data, therefore no patient with 0 blood pressure or 0 plasma glucose levels in his body is possible. For good classification results, missing values must be imputed, else inaccurate classification results will ensue [2], [15].

So missing data is a problem that must be carefully handled. There are various strategies for dealing with missing data, such as: removing data points with missing data if we are not data starving, or using the mean/ median of the distribution of the same parameter to fill in missing data for a parameter.

The technique choosed to manage zero values is to substitute them for the mean or median values for the same field.



```
# Replacing the values having 0 with Mean/Median based on distribution
diabetes.Glucose = diabetes.Glucose.replace(0,diabetes.Glucose.mean())
diabetes.BloodPressure = diabetes.BloodPressure.replace(0,diabetes.BloodPressure.median())
diabetes.SkinThickness = diabetes.SkinThickness.replace(0,diabetes.SkinThickness.mean())
diabetes.Insulin = diabetes.Insulin.replace(0,diabetes.Insulin.median())
diabetes.BMI = diabetes.BMI.replace(0,diabetes.BMI.mean())
```

Fig. 14. Replacing missing values.

D. Machine learning approaches

ML is the application of Artificial Intelligence that assists a machine in improving and learning from experiences. The primary goal of ML is to access data and learn from it using computer algorithms. A model is trained on historical data and then used to predict future data. This model may be saved and deployed to create powerful apps [16], [17].

The most generally used methodologies in Machine Learning are supervised learning, which trains algorithms with human-labeled instance input and output data, and unsupervised learning, which gives an algorithm no marked data and asks it to find meaning in its own data [18].

It can be difficult to choose the optimum learning approach for disease prediction because it is dependent on dataset size and user access. In the majority of studies, supervised machine learning (SML) methodologies are used, along with straightforward and simple predictive modeling. Implementing these models in clinical practice can undoubtedly aid in the delivery of better health services and improve specialist decision-making [16], [19]. We picked the supervised one since we already know the outcome of the datasets we have.

The “No Free Lunch” theorem is a principle in machine learning. In a word, nobody's machine-learning algorithm works optimally for every problem and is critically relevant for supervised learning (i.e. predictive modeling). For example, you can't argue that neural networks are always better than decision trees or that decision trees are always better than neural networks. Many factors come into play, including the size and structure of your dataset. As a result, you should attempt a variety of methods for your problem while evaluating performance and selecting the winner using a hold-out “test set” of data [20].

We will use several supervised machine learning techniques for prediction (NB, KNN, SVM, DT, RF, and LR algorithms). In the proposed method, the dataset was divided into two groups.

We are going to strive to make those models more accurate. In order to further enhance precision, a hybrid ensemble model is developed, in which three algorithms are combined with the maximum accuracy and feed into a different model.

E. Hybrid Model

Hybrid Models are a collection of various ML algorithms. They can be used to take advantage of the efficiency of multiple models on certain sets of data, which will increase the overall prediction model's accuracy.

Stacking was employed to create our Hybrid Model. The model has two layers: the base layer (level 0) and the meta layer (level 1). We will use algorithms in the foundation layer that have proven to be accurate in testing.

IV. Experimental Result

This section provides a detailed overview of our experiments for diabetes prediction. We utilize Jupyter Notebook, free software for the conduct of machine learning. We need to import the Sklearn module, which includes the necessary algorithm and functions for machine learning purposes.

After applying all the algorithms, we evaluated our model to check the accuracy as shown in table 2. It was observed that the exactness of the model is 77.27% after implementation of the Naive Bayes method, and after the Random Forest Algorithm was implemented, it was determined that the model's accuracy is 83.76%. Moreover, it was discovered that the accuracy of the model 78.57% after the use of the Logistic regression technique. After using the KNN method, it has been determined that the best model accuracy is at n = 5, i.e. 88.31%. With regard to Support Vector Machine algorithm, it was found that the accuracy of model is 87.01 % and following application of the algorithm for the Decision Tree i.e. 85.71%.

Once all techniques have been tested, we discovered that kNN, SVM, and Decision Tree provided the best accuracy (Fig.15). Because our problem is classification based, we employed Logistic Regression at the meta layer as shown below:

```
def get_stacking():
    level0 = list()
    level0.append(('knn',
KNeighborsClassifier()))
    level0.append(('svm', svm.SVC()))
    level0.append(('dt',
DecisionTreeClassifier()))
    level1 = LogisticRegression()
    model =
StackingClassifier(estimators=level0,
final_estimator=level1, cv=None)
    return model
```

Then we got an accuracy of 90,62% for the hybrid model.

TABLE II. Model Metrics using Weighted Averaging.

Model	Precision	Recall	F1-score	Accuracy
NB	0.77	0.77	0.77	77.27 %
RF	0.84	0.84	0.84	83.76%
LR	0.78	0.79	0.78	78.57%
KNN (n=5)	0.88	0.88	0.88	88.31%
SVM	0.88	0.87	0.87	87.01%
DT	0.86	0.86	0.86	85.71%
Hybrid model	0.91	0.91	0.90	90.62%

V. Discussion

The findings were compared to previous works, and the results are shown in Table 3. Pima Indians Diabetes Data Set was utilized for all evaluations.

TABLE III. Comparison Of Performance.

Approaches	Models	Accuracy
Ramezani et al. 2018 [15]	Hybrid classifier	88.05%
Harleen Kaur et al. 2019 [17]	KNN	88.00%
Nonso Nnamoko et al. 2020 [20]	C4.5	89.50%
Shekharesh Barik et al. 2021 [21]	XGBoost method	74.10%

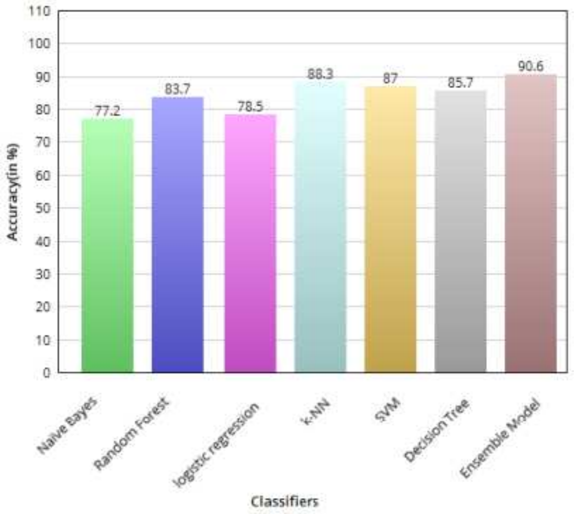


Fig. 15. Experimental results.

VI. Conclusion and Future Scope

In clinical practice, machine learning predictive models can emphasize better guidelines for making decisions about individual patient treatment. Early detection and appropriate therapies are the only ways to lower the death rates caused by chronic diseases. Type 2 diabetes is one of the most common endocrine illnesses in the world today. Utilizing the machine learning prediction models presented in this study, we were able to detect higher accuracy by using a hybrid model that reached an accuracy of 90.62 percent.

The next stage is to apply the numerous strategies outlined in this research study to a different or larger dataset than the one used to increase the accuracy of the model produced for prediction. In addition, we may train the model using attributes such as a person's daily eating habits and exercise routines, as these factors might influence whether or not a person is diabetic. The accuracy of the prediction model may be greatly enhanced by using Deep Neural Networks (DNN) and unsupervised learning techniques on a bigger dataset. Similar prediction models can be developed for a variety of diseases, such as heart disease, cancer, brain tumors, asthma, and so on.

References

[1] S. Gujral, “Early Diabetes Detection using Machine Learning: A Review,” IJIRST –International J. Innov. Res. Sci. Technol., vol. 3, no. 10, pp. 57–62, 2017, [Online]. Available: http://www.ijirst.org/articles/IJIRSTV3I10027.pdf.

[2] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," IEEE Access, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[3] R. Sehly and M. Mezher, "Comparative Analysis of Classification Models for Pima Dataset," 2020 Int. Conf. Comput. Inf. Technol. ICCIT 2020, vol. 02, pp. 58–62, 2020, doi: 10.1109/ICCIT-144147971.2020.9213821.

[4] G. Luo, "Automatically explaining machine learning prediction results: A demonstration on type 2 diabetes risk prediction," Heal. Inf. Sci. Syst., vol. 4, no. 1, pp. 1–9, 2016, doi: 10.1186/s13755-016-0015-4.

[5] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.

[6] M. Rout and A. Kaur, "Prediction of Diabetes Risk based on Machine Learning Techniques," Proc. Int. Conf. Intell. Eng. Manag. ICIEM 2020, pp. 246–251, 2020, doi: 10.1109/ICIEM48762.2020.9160276.

[7] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, no. xxxx, 2021, doi: 10.1016/j.icte.2021.02.004.

[8] J. Fidalgo, D. Pinho, R. Lima, and M. S. N. Oliveira, "VipIMAGE 2017," vol. 27, 2018, doi: 10.1007/978-3-319-68195-5.

[9] M. Aminul and N. Jahan, "Prediction of Onset Diabetes using Machine Learning Techniques," Int. J. Comput. Appl., vol. 180, no. 5, pp. 7–11, 2017, doi: 10.5120/ijca2017916020.

[10] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[11] P. Sonar and K. Jaya Malini, "Diabetes prediction using different machine learning approaches," Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 367–371, 2019, doi: 10.1109/ICCMC.2019.8819841.

[12] J. Xue, F. Min, and F. Ma, "Research on diabetes prediction method based on machine learning," J. Phys. Conf. Ser., vol. 1684, no. 1, 2020, doi: 10.1088/1742-6596/1684/1/012062.

[13] A. Bilal, G. Sun, and S. Mazhar, "Survey on recent developments in automatic detection of diabetic retinopathy," J. Fr. Ophtalmol., vol. 44, no. 3, pp. 420–440, 2021, doi: 10.1016/j.jfo.2020.08.009.

[14] "pima-indians-diabetes-database." https://www.kaggle.com/uciml/pima-indians-diabetes-database# (accessed Jul. 20, 2021).

[15] R. Ramezani, M. Maadi, and S. M. Khatami, "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis," Alexandria Eng. J., vol. 57, no. 3, pp. 1883–1891, 2018, doi: 10.1016/j.aej.2017.03.043.

[16] N. Jayanthi, B. V. Babu, and N. S. Rao, "Survey on clinical prediction models for diabetes prediction," J. Big Data, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0082-7.

[17] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," Appl. Comput. Informatics, 2019, doi: 10.1016/j.aci.2018.12.004.

[18] T. Chauhan, S. Rawat, S. Malik, and P. Singh, "Supervised and Unsupervised Machine Learning based Review on Diabetes Care," 2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021, pp. 581–585, 2021, doi: 10.1109/ICACCS51430.2021.9442021.

[19] S. L. Cichosz, M. H. Jensen, and O. Hejlesen, "Short-term prediction of future continuous glucose monitoring readings in type 1 diabetes: Development and validation of a neural network regression model," Int. J. Med. Inform., vol. 151, no. December 2020, 2021, doi: 10.1016/j.ijmedinf.2021.104472.

[20] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," Artif. Intell. Med., vol. 104, no. December 2018, p. 101815, 2020, doi: 10.1016/j.artmed.2020.101815.

[21] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques, vol. 153. Springer Singapore, 2021.