Tec Empresarial
P-ISSN: 1659-2395; E-ISSN: 1659-3359

MACHINE LEARNING BASED DIABETES DISEASE PREDICTION

Prof. Raghuvir Joshi
Department of Computer Science, Shree Om College of Computer Science
Email: raghuvirjoshi87@gmail.com

Dr. Ishaan Tamhankar
Department of Information Technology, Surendranagar University
Email: prof.ishaantamhankar@gmail.com

ABSTRACT
Machine Learning is a field of computer science to let computers learn like human beings. It is a set of algorithms through which we want computers to learn and take decisions like us. We want computers to learn from the past data, build a model for learnt concepts and use that model for future predictions. Machine Learning is being used for the predictions in every field of real world applications. From business to education, entertainment to healthcare, the diverse algorithms of Machine Learning are improving accuracies of real world applications significantly. Even though Machine Learning has been proven to be acceptably accurate, it is still challenging to use it for critical applications such as for health care where inaccurate results may cause severe negative impact. Designing Machine Learning based solution is indeed a very challenging task if developed model is going to be used for patients. This paper presents progress of Machine Learning based Diabetes Disease Prediction along with result analysis of our implementation with Orange tool.
Keywords – Diabetes Disease, Machine Learning, Performance Matrix

I. INTRODUCTION
Machine learning lets computers learn from the past data and build models for future predictions. Due to high accuracy and efficiency of Machine Learning based applications, almost every field of real world application is working towards adopting intelligent decision making. ML has also been widely implemented and accepted with healthcare based applications. From symptom analysis to detection of disease, ML is applicable from every small task to major task in healthcare.

According to WHO, The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014. Prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. Proper detection and proactive maintenance surely help the patients in longer living of lives. Diabetes is one of those few diseases those are difficult to diagnose at initial state due to their progressive symptoms. Very often, patients don't notice the initial symptoms and suffer once they are diagnosed with high risk stage diabetes. Our research

work is based on identifying possible stage of diabetes based on results of some observations. We focus on very easy to note observations which are easy to arrange by patients with comfortable lab. Tests. This work focuses on predicting whether a patient has a possibility of diabetes or not. This will help patients to decide whether they should consult medical professional for better diagnosis or not.

II. MACHINE LEARNING

Machine learning is a computer technology that enables machines to learn from data and improve their performance on specific tasks over time. Instead of being explicitly programmed to perform a task, machines use algorithms to learn patterns from examples and make predictions or decisions based on that knowledge. This helps them tackle complex problems and make better choices as they encounter new data [1,2,3].

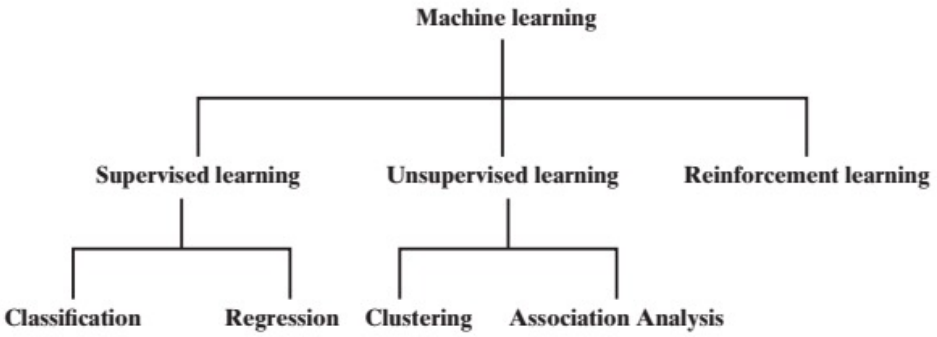


Figure – 1 – Machine Learning Types [1]

In supervised learning, algorithms are trained on a dataset consisting of input-output pairs, where the inputs are typically features or attributes, and the outputs are corresponding labels or target values. The algorithm learns to map inputs to outputs by minimizing a predefined loss function, adjusting its parameters iteratively through techniques like gradient descent. Unsupervised learning, on the other hand, deals with data lacking explicit labels or target outputs. Instead, the algorithm seeks to uncover underlying patterns, structures, or relationships within the data through techniques such as clustering, dimensionality reduction, or density estimation. Reinforcement learning involves an agent interacting with an environment, where it learns to make sequential decisions to maximize cumulative rewards. The agent receives feedback in the form of rewards or penalties based on its actions, and through exploration and exploitation strategies, it aims to learn an optimal policy for decision-making [1,2,3].

As our work is focused on analyzing past data of patients to build a model that can predict possibility of diabetes for new patients, we will use supervised learning for implementation.

III. SUPERVISED LEARNING

As explained in Section-II, Supervised Learning algorithms are trained on input-output pairs – that forms the dataset of past records. Various algorithms are developed which are discussed in brief here.

Decision Tree: A decision tree is a hierarchical structure used in machine learning for classification and regression tasks. It comprises internal nodes representing decisions based on features and leaves representing outcomes. Decision trees partition the feature space into subsets based on informative features to facilitate straightforward decision-making. They're interpretable and useful for visualizing decision processes but prone to overfitting, mitigated by pruning or ensemble methods like Random Forests.

Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming feature independence. It calculates class probabilities given input features by multiplying conditional probabilities. Despite its simplicity, Naive Bayes often performs well, especially for text classification tasks. It's computationally efficient and suitable for real-time applications but may produce suboptimal results with correlated features and cannot handle missing values effectively.

K-Nearest Neighbor (KNN): K-Nearest Neighbor (KNN) is a straightforward algorithm for classification and regression tasks. It classifies data points by majority voting among their nearest neighbors in feature space. KNN is non-parametric, versatile, and suitable for multi-class classification, regression, and outlier detection. However, its performance can be sensitive to the choice of K and the distance metric, and it may struggle with irrelevant or noisy features.

Support Vector Machine (SVM): Support Vector Machine (SVM) is a powerful supervised learning algorithm for classification, regression, and outlier detection. It finds the optimal hyperplane that separates different classes while maximizing the margin. SVM is effective in high-dimensional feature spaces and robust to local optima, but its performance depends on the choice of kernel function and parameters. SVM's interpretability may be limited compared to simpler models, and it can be computationally expensive with large datasets.

Neural Network: A neural network is a computational model inspired by the brain's interconnected neurons. It consists of layers of artificial neurons that learn to map input data to output predictions through training. Neural networks automatically learn hierarchical representations of data, capturing complex patterns, but may lack interpretability. They require large amounts of labeled data and computational resources but achieve state-of-the-art performance in various domains.

IV. LITERATURE REVIEW

As ML has started getting wide acceptance in healthcare, over the years many researchers have proposed solutions for health care related predictions. Some of the recent developments are explained in this section.

A paper titled "Identifying Ethical Considerations for Machine Learning Healthcare Applications" proposes a systematic framework for identifying ethical concerns in Machine Learning - Health Care Applications (ML-HCAs), aiming to facilitate ethical evaluations throughout development and implementation stages [4]. Another paper, "Ethical Machine Learning in Healthcare," addresses how machine learning models may amplify existing health inequities and emphasizes the importance of framing ML ethics through social justice [5]. "Secure and Robust Machine Learning for Healthcare: A Survey" surveys various application areas in healthcare, focusing on security and privacy aspects, and discusses methods to ensure secure and privacy-preserving machine learning [6]. "Automated Machine Learning: Review of the State-of-the-Art and Opportunities for Healthcare" reviews literature on automated machine learning (AutoML) to assist healthcare professionals with limited data science expertise, identifying opportunities and barriers in its utilization [7]. "Artificial Intelligence with Multi-functional Machine Learning Platform Development for Better Healthcare and Precision Medicine" explores how AI and machine learning contribute to healthcare discovery and precision medicine, emphasizing the integration of electronic health records and AI for real-time decision support [8]. "Synthetic Data in Machine Learning for Medicine and Healthcare" discusses the proliferation of synthetic data in AI for medicine and healthcare, raising concerns about software vulnerabilities and policy challenges [9]. "How to Develop Machine Learning Models for Healthcare" outlines guidelines for developing machine learning models in healthcare, with a focus on improving clinical decision support and enhancing patient care [10].

Specific to diabetics predictions, many researchers have proposed their solutions which are discussed as below.

A paper “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers” is focused on building a framework with a set of machine learning based methods with weighted ensembling concept [11]. "Diabetes Prediction using Machine Learning Algorithms" work is based on analyzing external factors along with regular factors for classification purposed [12]. "A review on current advances in machine learning based diabetes" summarizes the progress of diabetes prediction so far and associated challenges. [13]. “A comparison of machine learning algorithms for diabetes prediction” discusses how different algorithms provide accurate results for a given dataset [14]. “Diabetes prediction using machine learning techniques” provide a discussion for how random forest technique performs better as compared to other techniques [15]. There is much more in literature to explore the development in this field.

V. PROPOSED WORK

DataSet: Our work is focused on diabetics prediction based on some easy to arrange parameters. These parameters can be known to patients or can be arranged with simple laboratory tests. Our focus is on early detection so patients can visit to medical professional for better diagnosis and

treatment. Such early detection can also be done by patients themselves without visiting medical professional. Selection of right dataset suitable for achieving our goal was indeed a challenging task. We have observed multiple datasets made available by researchers and identified most suitable one is "Diabetes prediction dataset" which is a Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data [16]. This dataset suits the best for our work as it includes both – demographic data and medical data. Where demographic data is about gender, age, habits etc., the medical data is about lab. results etc. the various fields of our dataset are listed below.

No.	Field Name	Value Type	Category
1	Age	Numerical	Demographic Feature
2	Gender	Categorical {Male, Female}	Demographic Feature
3	Hypertension	Categorical {Yes, No}	Demographic Feature
4	Heart Disease	Categorical {Yes, No}	Demographic Feature
5	Smoking	Categorical {Never, Former, Ever, Current, Not Current, No Info}	Demographic Feature
6	BMI	Numerical	Demographic Feature
7	HbA1c level	Numerical	Medical Feature
8	Blood glucose level	Numerical	Medical Feature
9	Diabetes	Categorical {Yes, No}	Target

To develop an accurate model, significant number of records are needed. The selected dataset has records of 1,00,000 patients without any invalid or missing values.

Implementation:

We have evaluated our work with three approaches. The algorithm which we used is decision tree and implementation was done with Orange tool.

No.	Features to be analyzed for DT-Decision Tree
1	DT-1: All features of dataset are used.
2	DT-2: Only Medical data related features are used.
3	DT-3: Only Demographic data related features are used.

The complete workflow for three decision tree is shown in figure – 2

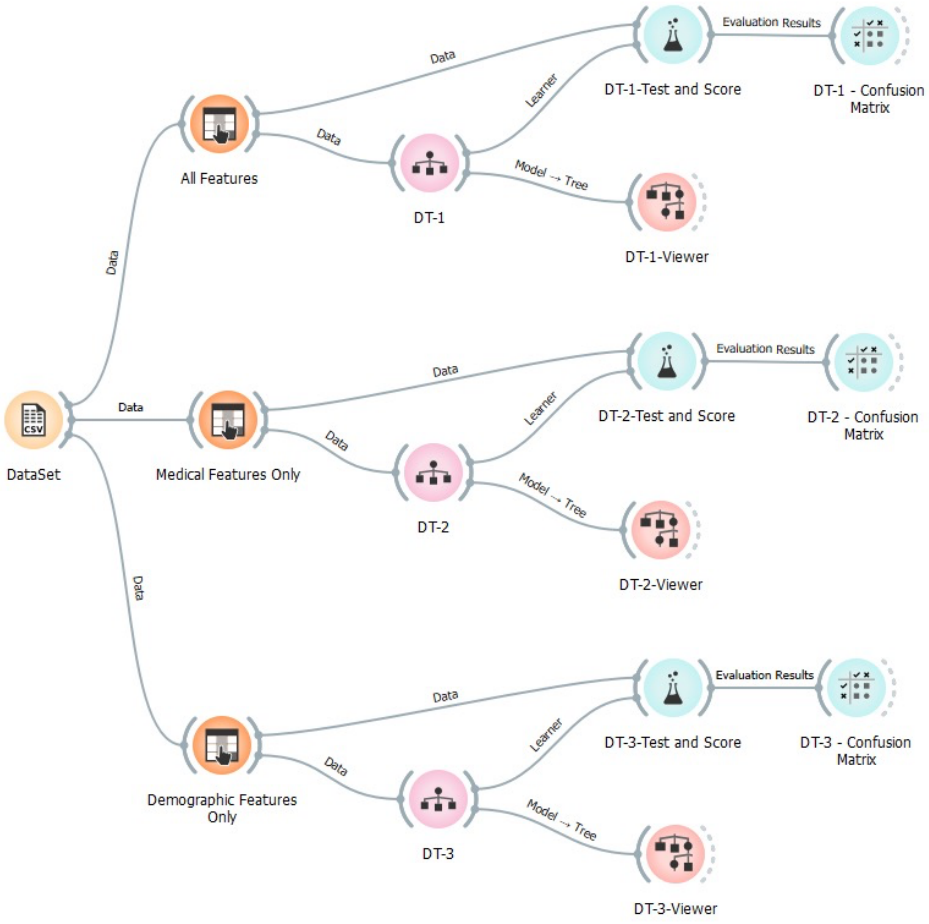


Figure – 2 – Orange Workflow for Decision Trees

VI. PERFORMANCE ANALYSIS

DT-1: All features of dataset are used to build the decision tree. We have demographic data as well as medical data. Specific to medical data – we have two features - HbA1c level and Blood glucose level. HbA1c reflects average blood sugar control over a longer period, while blood glucose levels provide real-time information about current blood sugar levels. We have observed that DT-1

(Decision Tree-1) is only having prediction based on HbA1c value and no other features are used. Figure – 3 shows the model.

Figure – 3 – Decision Tree - 1

DT-2: Only medical data related features are used. As now we had only two features - HbA1c level and Blood glucose level, our DT-2 (Decision Tree-2) is exactly same as modelled in DT-1 (Decision Tree-1) that is influenced only by HbA1c test.

DT-3: Only demographic data related features are used. We have got a model that is visually very complex so figure is not shown here.

Test and Score Results with 5 fold cross validations – DT-1

Model	AUC	CA	F1	Prec	Recall	MCC
DT-1	0.727	0.954	0.946	0.956	0.954	0.661

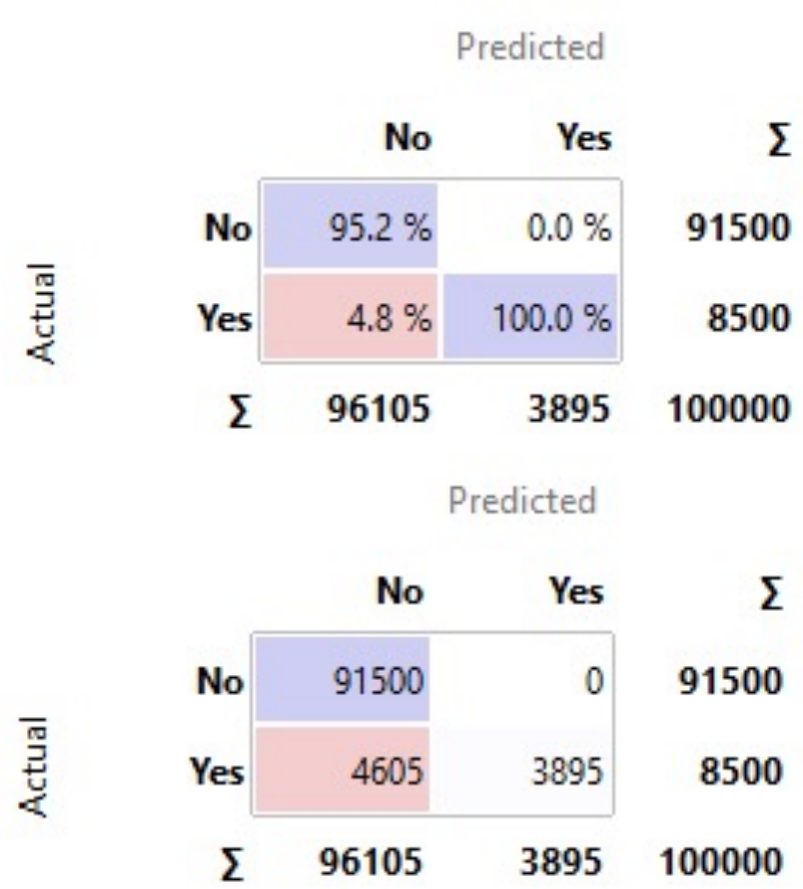
Test and Score Results with 5 fold cross validations – DT-2

Model	AUC	CA	F1	Prec	Recall	MCC
DT-2	0.727	0.954	0.946	0.956	0.954	0.661

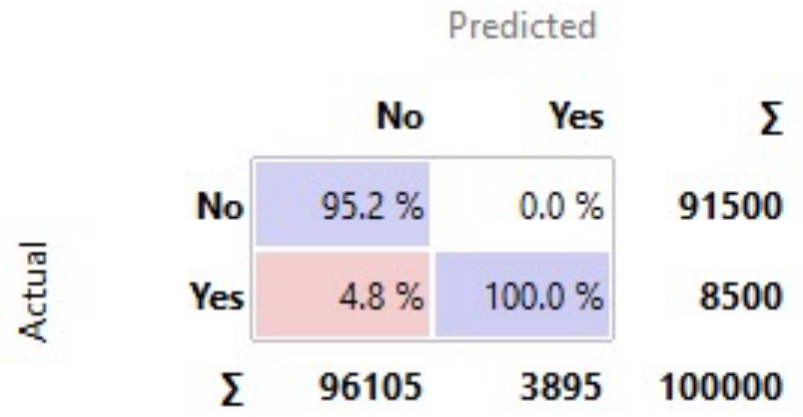
Test and Score Results with 5 fold cross validations – DT-3

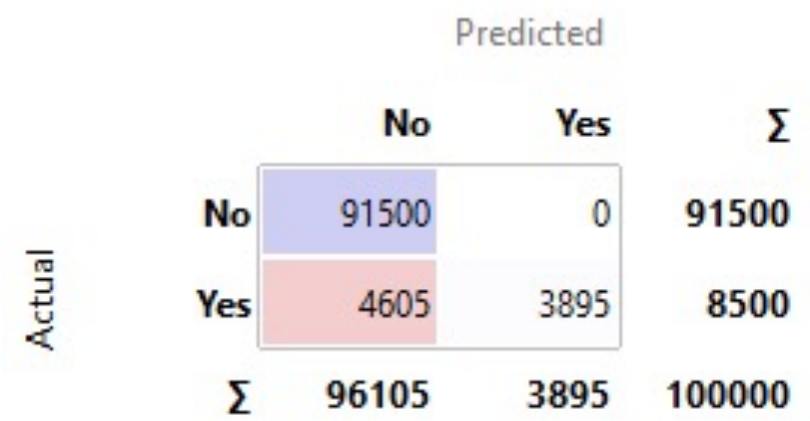
Model	AUC	CA	F1	Prec	Recall	MCC
DT-3	0.533	0.897	0.882	0.872	0.897	0.168

Confusion Matrix – Decision – Tree - 1:



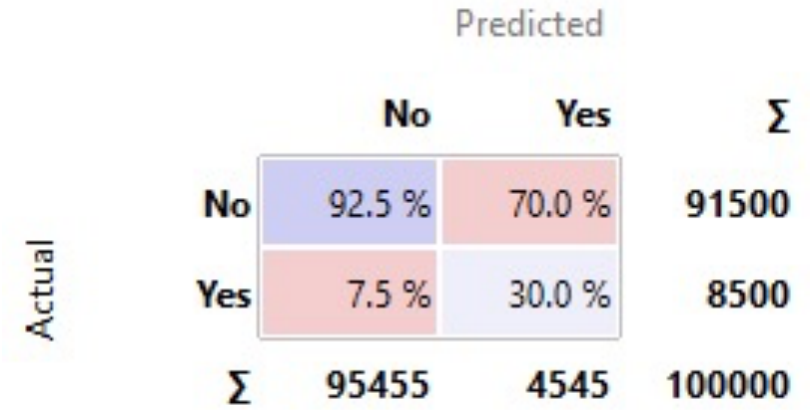
Confusion Matrix – Decision – Tree - 2:



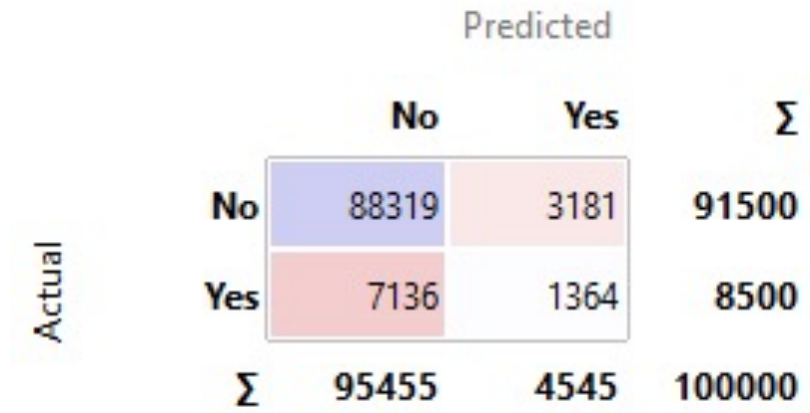


Actual \ Predicted	No	Yes	Σ
No	91500	0	91500
Yes	4605	3895	8500
Σ	96105	3895	100000

Confusion Matrix – Decision – Tree - 3:



Actual \ Predicted	No	Yes	Σ
No	92.5 %	70.0 %	91500
Yes	7.5 %	30.0 %	8500
Σ	95455	4545	100000



Actual \ Predicted	No	Yes	Σ
No	88319	3181	91500
Yes	7136	1364	8500
Σ	95455	4545	100000

Result Analysis: As we discussed the results of DT-1 and DT-2 are exactly same as both of these models are dependent only on 1 feature. Results of DT-3 are less accurate as compared to of DT-1.

CONCLUSION

This research work is based on analyzing a large dataset for diabetic prediction. The dataset has two types of features – medical data and demographic data. We have observed that when both

types of data are analyzed together, the prediction model is highly dependent on medical data and there is no role of demographic data in decision making. Alternate approach to analyze only demographic data would be helpful for people to do prediction without any new medical test. It is obvious that model that is built using medical data will be more accurate as compared to model that is built using demographic data. In our study, we got accuracy of 95.4% for DT-1 (All data) and DT-2 (Medical data). We have got accuracy of 89.7% for DT-3 (Demographic data). Though DT-1 and Dt-2 have higher accuracy, DT-3 is also having satisfactory accuracy that we can trust and explore further.

FUTURE WORK

This research work can be further extended with doing predictions for new instances and verifying the results. Other than the decision tree algorithm, other methods can be evaluated and performances can be measured for improvements. Dataset can also be extended to cover more demographic data that would help in improving corresponding accuracy of model. The dataset can also be arranged for real world medical professionals for better accurate analysis.

REFERENCES

[1] Education, Pearson. Machine Learning, 1e. Pearson Education India., 2018

[2] Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. An introduction to machine learning. Springer, 2019.

[3] Pereira, F. C., and S. S. Borysov. "Machine Learning Fundamentals Mobility Patterns, Big Data and Transport Analytics." (2019): Elsevier 9-29.

[4] Char, Danton S., Michael D. Abràmoff, and Chris Feudtner. "Identifying ethical considerations for machine learning healthcare applications." The American Journal of Bioethics 20.11 (2020): 7-17.

[5] Chen, Irene Y., et al. "Ethical machine learning in healthcare." Annual review of biomedical data science 4 (2021): 123-144.

[6] Qayyum, Adnan, et al. "Secure and robust machine learning for healthcare: A survey." IEEE Reviews in Biomedical Engineering 14 (2020): 156-180.

[7] Waring, Jonathan, Charlotta Lindvall, and Renato Umeton. "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare." Artificial intelligence in medicine 104 (2020): 101822.

[8] Ahmed, Zeeshan, et al. "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine." Database 2020 (2020): baaa010.

[9] Chen, Richard J., et al. "Synthetic data in machine learning for medicine and healthcare." Nature Biomedical Engineering 5.6 (2021): 493-497.

[10] Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. "How to develop machine learning models for healthcare." Nature materials 18.5 (2019): 410-414.

[11] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.

[12] Mujumdar, Aishwarya, and V. Vaidehi. "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.
[13] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15.3 (2021): 435-443.
[14] Khanam, Jobeda Jamal, and Simon Y. Foo. "A comparison of machine learning algorithms for diabetes prediction." Ict Express 7.4 (2021): 432-439.
[15] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (Ijert) Volume 9 (2020).
[16] Diabetes prediction dataset - A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data by Mohammed Mustafa at Kaggle.

Biographies and Photographs

Short biographies (120-150 words) should be provided that detail the authors' education and work histories as well as their research interests. The authors' names are italicized. Small (3.5 X 4.8 cm), black-and-white pictures/digitized images of the authors can be included.