Machine Learning Techniques for Diabetes Classification: A Comparative Study

Hiri Mustafa[1], Chrayah Mohamed[2], Ourdani Nabil[3], Aknin Noura[4]
FS, Abdelmalek Essaadi University, TIMS LABORATORY, Tetuan, Morocco[1, 3, 4]
ENSATE, Abdelmalek Essaadi University, TIMS LABORATORY, Tetuan, Morocco[2]

***Abstract*—In light of the growing global diabetes epidemic, there is a pressing need for enhanced diagnostic tools and methods. Enter machine learning, which, with its data-driven predictive capabilities, can serve as a powerful ally in the battle against this chronic condition. This research took advantage of the Pima Indians Diabetes Data Set, which captures diverse patient information, both diabetic and non-diabetic. Leveraging this dataset, we undertook a rigorous comparative assessment of six dominant machine learning algorithms, specifically: Support Vector Machine, Artificial Neural Networks, Decision Tree, Random Forest, Logistic Regression, and Naive Bayes. Aiming for precision, we introduced principal component analysis to the workflow, enabling strategic dimensionality reduction and thus spotlighting the most salient data features. Upon completion of our analysis, it became evident that the Random Forest algorithm stood out, achieving an exemplary accuracy rate of 98.6% when 'BP' and 'SKIN' attributes were set aside. This discovery prompts a crucial discussion: not all data attributes weigh equally in their predictive value, and a discerning approach to feature selection can significantly optimize outcomes. Concluding, this study underscores the potential and efficiency of machine learning in diabetes diagnosis. With Random Forest leading the pack in accuracy, there's a compelling case to further embed such computational techniques in healthcare diagnostics, ushering in an era of enhanced patient care.**

Keywords*—Machine learning; support vector machine; artificial neural networks; decision tree; random forest; logistic regression; Naive Bayes; principal component analysis; classification; diabetes

I. Introduction

In recent times, diabetes has prominently risen as a pervasive and potentially lethal ailment, with its effects resonating across age groups and genders. This condition, fundamentally shaped by the body's compromised insulin production, interferes with carbohydrate metabolism. This interference results in heightened blood sugar levels, precipitating a slew of symptoms such as augmented thirst, hunger, and frequent urination [1]. A concerning facet of this disease is its accentuated and adverse impact on women, as reflected in their lower survival rates and compromised quality of life [2].

The malaise manifests in three main forms: Type 1, Type 2, and gestational diabetes. Type 1 is predominantly an autoimmune disorder seen in children, leading to the annihilation of pancreatic insulin-producing cells. In contrast, Type 2 emerges when there's heightened insulin resistance across various organs, eventually pushing the pancreas beyond its production capacities. An added layer of complexity is gestational diabetes, which particularly afflicts pregnant women owing to their pancreas's insufficient insulin output during pregnancy [2]. Furthermore, the diabetes spectrum has more grim facets, capable of inducing long-term harm and malfunctioning in diverse organs like the eyes, kidneys, heart, blood vessels, and nerves [4].

Given the multifarious nature of this disease, physicians find themselves navigating a diagnostic labyrinth. Early diagnosis becomes paramount, serving as the linchpin in circumventing and mitigating potential complications [5]. Fortunately, recent technological strides, predominantly within the machine learning spectrum, proffer novel solutions. Machine learning, a potent sub-discipline of artificial intelligence, harnesses algorithms and statistical frameworks to parse voluminous datasets, unveiling patterns and correlations that often remain concealed from conventional statistical techniques [3].

Positioned against this backdrop, our study delves into the potential of machine learning as a transformative tool in diabetes diagnostics. Six pivotal machine learning classification paradigms - namely, Support Vector Machine, Artificial Neural Networks, Decision Tree, Random Forest, logistic regression, and Naive Bayes - are meticulously examined using the PIDD dataset. By anchoring our assessment on accuracy, we render a holistic comparison of these algorithms' performance nuances.

The paper is structured to facilitate a coherent reader journey. Post this introduction in Section I, Section II immerses into the expansive realm of related works, detailing classification modalities used previously in diabetes prediction. Section III sheds light on our chosen methodologies and intricacies of the PIDD dataset. The crux of our findings unfolds in Section IV, with Section V diving into discussions and implications of these outcomes. Finally, Section VI encapsulates our conclusions, while also hinting at prospective research trajectories.

As we traverse this research landscape, our study is guided by the pressing questions: How do these machine learning paradigms stack against each other for diabetes prediction on the PIDD dataset? Furthermore, can they truly emerge as reliable instruments for diabetes diagnostics?

II. Related Work

In the annals of modern healthcare research, the strategic deployment of machine learning to grapple with the monumental challenge of diabetes classification has unfailingly occupied a spotlight [6]. The intrigue and allure of this intersection between computational prowess and medical insight have galvanized countless researchers to charter previously unexplored terrains.

Vandana Bavkar, with an academic rigor that's now cited extensively, delivered a magnum opus—a systematic review that scrutinized the versatile applications of machine learning, data mining techniques, and tools in the expansive canvas of diabetes research [6]. His explorations weren't just confined to the realms of prediction and diagnosis. They ventured further, diving deep into the intricacies of diabetic complications, the mystique of genetic predispositions juxtaposed against environmental triggers, and the labyrinth of healthcare management. It was in the revelations of Bavkar's investigation that the bedrock importance of prediction and diagnosis was underscored, positioning them as cornerstone applications of machine learning in the diabetes research tapestry [7].

Parallel to Bavkar 's seminal work, Hassan et al. [8] charted their own research trajectory, focusing on the prediction dynamics of diabetes mellitus. Armed with a range of machine-learning classifiers, their study served as a testing ground for techniques such as K-nearest neighbors, Support Vector Machine, and Decision Tree. The metrics they employed—precision, accuracy, sensitivity, and specificity—offered a comprehensive lens through which to evaluate the performance of these classifiers.

Further enriching this research milieu, Kaur et al. delineated a study wherein a quintet of predictive models was brought to the fore [9]. These included stalwarts like Decision Tree, Support Vector Machine, and Naive Bayes. The Pima Indian Diabetes dataset and the R Data Manipulation Tool became their canvas. In a different vein, Zhang et al. concocted a rather innovative approach, introducing a hybrid model that synergized random K-means with Decision Tree, specifically tailored to forecast diabetes risk [10]. Other scholarly forays in this domain have seen the inception of predictive architectures grounded on the Weighted Feature Selection of Random Forest and the XGBoost Ensemble Classifier [11]. Yet another groundbreaking initiative leaned into a logistic regression model, ingeniously augmented by the feature transformation capabilities of XGBoost [12].

Each of these studies, while diverse in methodology and focus, echoes a singular sentiment: the paramount importance of machine learning's role in not just predicting and classifying diabetes, but also in unearthing the intricate dance of genetics and environment, and in revolutionizing healthcare delivery for diabetic patients.

But herein lies an undeniable truth. Despite the richness of insights and the plethora of methodologies that have emerged from these academic odysseys, the horizon of diabetes classification using machine learning still holds vast expanses yet to be charted. The quest for impeccable prediction accuracy continues, as does the endeavor to spotlight risk factors in their nascent stages. With this study, our ambition is lucidly clear: to augment the extant knowledge reservoir by meticulously assessing the efficacy of a spectrum of machine learning algorithms, all in the context of the revered Pima Indians Diabetes Data Set [13].

To punctuate our intentions and situate our efforts in the grander scheme of academic pursuits, it's paramount to acknowledge the teeming body of studies that have previously addressed this challenge. This dense and rich academic tapestry underscores both the significance and the complexity of the diabetes classification conundrum.

III. Dataset and Methods

A. Dataset Description

In this study, we utilized the Pima Indians Diabetes Data Set [14], which is a widely used dataset in diabetes research. This dataset was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases and is available for public use from the UCI Machine Learning Repository. The dataset consists of 768 instances, each containing information about female patients of Pima Indian heritage. The dataset includes various attributes such as age, BMI, blood pressure, skin thickness, insulin level, and diabetes pedigree function, along with the target variable indicating whether the patient has diabetes or not.

The clinical descriptors for these attributes are presented in Table I.

TABLE I. The Clinical Descriptors Of The Variables

Number	Attribute	Description	Type
1	Npreg	Number of pregnancies	Numeric
2	Glu	Plasma glucose concentration	Numeric
3	BP	Diastolic blood pressure (mm Hg)	Numeric
4	SKIN	Triceps skinfold thickness, (mm)	Numeric
5	Insulin	Insulin dose, (mu U/ml)	Numeric
6	BMI	Body Mass Index (weight in kg/ (size m)2)	Numeric
7	PED	Diabetes pedigree function (heredity)	Numeric
8	Age	Age (Year).	Numeric
9	class	Target variable (0 or 1)	Numeric

B. Methods

The intellectual pursuit of understanding diabetes through the prism of machine learning necessitates the application of a robust and multifaceted methodology. In light of this, our investigation unfurled in a series of calibrated steps, each meticulously designed to serve a specific purpose within the broader research framework.

1) Data preprocessing: Central to the fabric of any data-driven study is the sanctity of the data itself. Recognizing this, our first port of call was to refine and purify the data landscape. We embarked on a rigorous journey of data preprocessing, which, at its core, was about ensuring the

reliability and accuracy of the outcomes. Recognizing the potential pitfalls of missing values, these were diligently identified and addressed with a strategic blend of imputation or outright deletion, depending on the context.

2) Feature selection: Beyond just raw data, the richness of features often dictates the nuances of the results. With this philosophy in mind, we ventured into the realm of feature selection. The objective was straightforward yet critical: to streamline the dataset by spotlighting the most consequential attributes for diabetes classification. From the vast repertoire of available techniques, we leaned on the classical Principal Component Analysis (PCA). It's a tool that elegantly navigates the dimensions of data, projecting from a higher-dimensional space to a lower one, all while retaining features pivotal to dataset variance.

3) Machine learning algorithms: With the data landscape prepped, the stage was set to deploy the titans of machine learning. Six algorithms, each renowned for its distinctive virtues and latent challenges, were chosen for the diabetes classification task:

a) Support Vector Machine (SVM): The SVM stands tall as a supervised classifier, renowned for its prowess in both regression and classification tasks [15]. Fig. 1 shows SVM algorithm. Originated by Vapnik [16], SVM's genius lies in its capacity to delineate data into classes, both linearly and non-linearly. At its core, SVM conjures hyperplanes in a high-dimensional milieu. The ultimate aspiration? A hyperplane that segregates data classes with the widest possible margin. Non-linear classification gets a boost through a bouquet of kernel functions, each striving to maximize hyperplane margins [17].

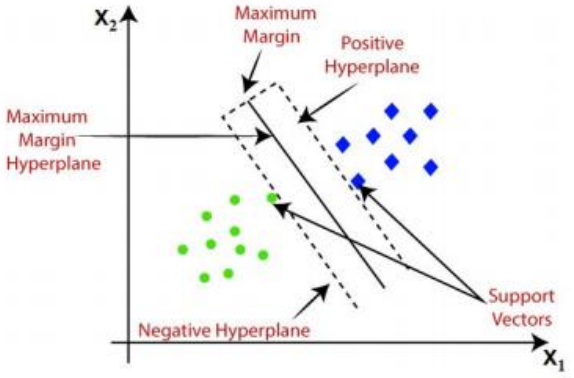

Fig. 1. Support vector machine algorithm [18].

b) Artificial Neural Networks (ANN): Channeling inspirations from the intricate mesh of human neural architecture, ANNs exemplify the confluence of biology and computation [19]. Introduced in the 1950s, ANNs mirror the workings of the human brain's myriad neurons, with artificial neurons and weighted interconnections taking center stage [20]. There are three essential layers in a neural network: input layer, hidden layer, and output layer. The input layer is in charge of accepting data from the user. Fig. 2 shows an example of MLP network with two inputs, five neurons in the hidden layer. The output layer will provide us with the results. The hidden layer is the layer that sits between the input and output layers. On the same layer, there is no interaction between neurons [21]. If the input vector is $\vec{x}$, the weight vector is $\vec{w}$, and the activation function is a sigmoid function, the output is as follows:

$$y = \text{sigmoid}\,(\vec{x}.\vec{w}) \quad (1)$$

and the sigmoid is as follows:

$$\text{sigmoid}\,(x) = \frac{1}{1+e^{-x}} \quad (2)$$

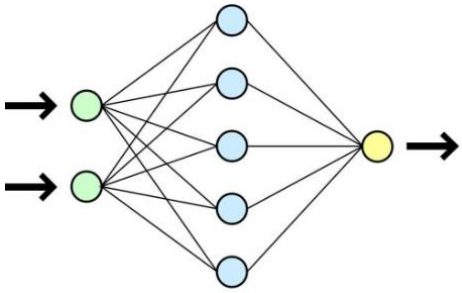

Fig. 2. Example of an MLP network with a hidden layer with two inputs, five neurons in the hidden layer, and one output.

c) Decision Tree: Decision Trees, both elegant and insightful, offer a flowchart-like structure to visualize and make decisions. Whether for classification or regression, they rely on a series of attribute tests, guiding data from root to leaf, ultimately culminating in a class prediction [22]. The algorithmic underpinnings encompass three operations: determining terminal nodes, associating non-terminal nodes with tests, and assigning a class to terminal nodes (see Fig. 3) A plethora of algorithms, from ID3 to CTREE, have been proposed for decision tree formulation [23].

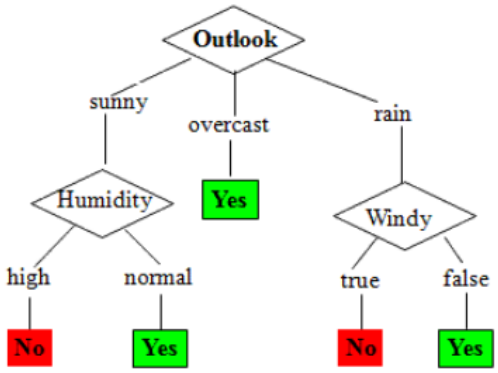

Fig. 3. Example of a decision tree.

d) Random Forest: Emerging from the shadows of decision trees is the Random Forest—a brainchild of Breiman [24]. It's an ensemble approach, creating a 'forest' of decision trees from randomly chosen data subsets (see Fig. 4) The collective wisdom of this forest then votes or averages, producing classifications or regressions, respectively [25][26][27].

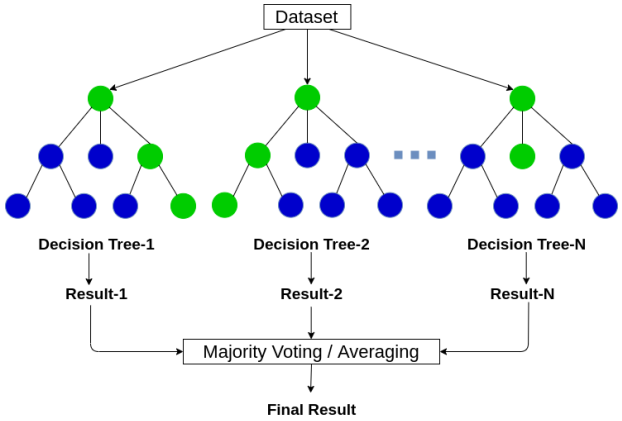

Fig. 4. Random forest.

e) Logistic Regression:

A stalwart in the classification domain, Logistic Regression evaluates probabilities through the sigmoid function, discerning relationships between binary dependent and independent variables. The sigmoid's magic rests in its ability to produce binary outputs based on weighted inputs. If the sigmoid output surpasses 0.5, the prediction is 1; otherwise, it's 0. The sigmoid/logistic function is calculated as follows:

$$y(x) = \frac{1}{1+e^{-x}} \quad (3)$$

where, y is the output which is the result of the weighted sum of the input variables x.

f) Naive Bayes: Grounded in the probabilistic paradigm, the Naive Bayes classifier champions the Bayes Theorem [1]. It presumes that each class feature exists in isolation—hence the "naive" tag. The algorithm computes the posterior probability,

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (4)$$

where, $P(C|X)$ is the posterior probability of the target class.

- $P(X|C)$ is the probability of the predictor type.
- $P(C)$ is the probability that class C is correct.
- $P(X)$ is the prior probability of the predictor.

In many intricate real-world scenarios, Naive Bayes has showcased exceptional classification prowess.

IV. Experimental Results

The selection of the Pima Indians Diabetes Data Set for this study was a deliberate choice. This dataset, which has garnered significant attention in the data science community, offers intricate nuances and a wealth of attributes that allow for an exhaustive evaluation of machine learning algorithm performances.

- 1st Experiment: Comprehensive Approach with All Variables.

Our first experiment was anchored in a holistic approach, wherein all available features from the dataset were utilized. This comprehensive method was designed to create a baseline performance, which future models in our study would either strive to match or surpass. The accuracy metrics corresponding to this experiment, for various algorithms, are tabulated in Table II.

TABLE II. Accuracy when using all variables.

Methods	Accuracy validation
Random Forest	0.982
Decision Tree	0.966
SVM	0.954
Logistic Regression	0.794
ANN	0.948
Naïve Bayes Classifier	0.788

As evinced from the results in Table II, the Random Forest classifier emerged as the frontrunner, delivering an impressive accuracy of 0.982, thereby setting a solid benchmark for subsequent experiments.

- 2nd Experiment: Exploring the Power of PCA for Dimensionality Reduction.

Principal Component Analysis (PCA) stands as a testament to the efforts of countless researchers aiming to refine large data volumes into their most significant components. With the aspiration to condense the dataset into its primary four components, representing 71% of its inherent variance, there was an optimistic expectation for data efficiency without sacrificing critical information.

The performance outcomes derived from this approach are detailed in Table III.

TABLE III. Accuracy when using the PCA.

Methods	Accuracy validation
Random Forest	0.97
Decision Tree	0.962
SVM	0.888
Logistic Regression	0.728
ANN	0.826
Naïve Bayes Classifier	0.744

A glance at Table III reveals a pivotal observation: while the Random Forest algorithm continued to exhibit stellar accuracy at 0.97, it was clear that the unmodified data carried nuanced intricacies not entirely captured by PCA. It's a gentle reminder of the delicate balance between data reduction and the preservation of intricate patterns.

- 3rd Experiment: A Deep Dive into Correlation Dynamics.

One of the guiding principles of this experiment was to unearth the relationships and patterns present among the dataset's variables. As machine learning models continue to advance in complexity, a nuanced comprehension of how variables interact and influence each other is paramount.

Fig. 5 and 6 graphically depict the interactions between the class variable and other attributes, as well as the overarching correlation matrix respectively.

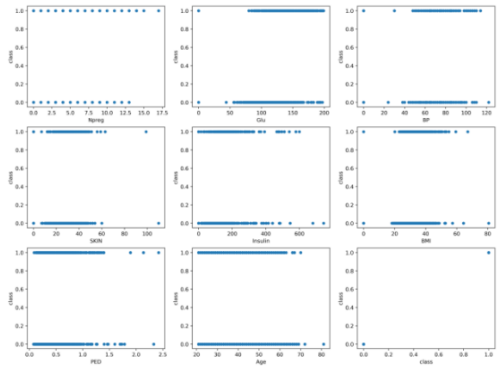

Fig. 5. The class variable as a function of the other variables.

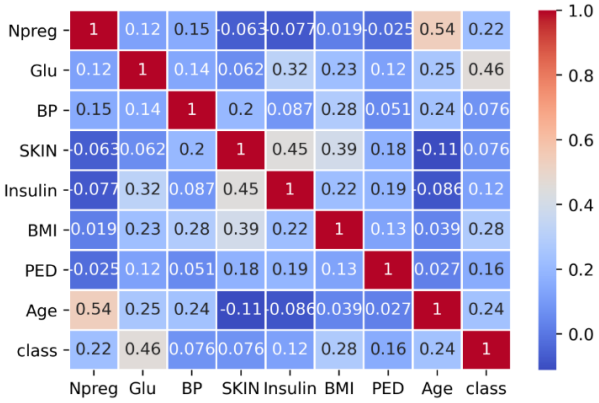

Fig. 6. Correlation matrix.

The metrics arising from this correlation analysis, especially when excluding the 'BP' and 'SKIN' attributes are enumerated in Table IV.

TABLE IV. ACCURACY WITHOUT THE USE OF 'BP' AND 'SKIN'.

Methods	Accuracy validation
Random Forest	0.986
Decision Tree	0.974
SVM	0.964
Logistic Regression	0.792
ANN	0.806
Naïve Bayes Classifier	0.784

An illuminating discovery from this analysis was the marginal contribution of the 'BP' and 'SKIN' variables. By sidelining these variables, the Random Forest algorithm, known for its dynamic adaptability, achieved an apex accuracy of 0.986, highlighting the value of informed feature selection in machine learning.

- 4th Experiment: Spotlight on Prime Features.

The emphasis of this experiment was on identifying and evaluating the predictive power of four critical attributes: number of pregnancies, plasma glucose concentration, body mass index, and age. These features, singled out for their perceived significance, were put to the test to determine their collective predictive prowess.

The outcomes, with focus solely on these attributes are presented in Table V.

TABLE V. ACCURACY WHEN USING THE NUMBER OF PREGNANCIES, PLASMA GLUCOSE CONCENTRATION, BODY MASS INDEX, AND AGE.

Methods	Accuracy validation
Random Forest	0.984
Decision Tree	0.97
SVM	0.964
Logistic Regression	0.788
ANN	0.78
Naïve Bayes Classifier	0.78

While these select attributes showcased substantial predictive capability, the Random Forest algorithm highlighted a noteworthy point: focusing exclusively on them, albeit impactful, didn't outperform its previous benchmarks. The model's accuracy, in this context, peaked at 0.984, subtly reminding us of the intricate dynamics within data.

V. DISCUSSION

At the confluence of scientific inquiry, we find an unyielding drive towards understanding, clarity, and the quest for tangible insights. Embedded within the heart of this exploration, our study not only aligns with previous findings but also brings forth novel perspectives in the realm of diabetes research [1,15].

One of the standout revelations was the prowess of the Random Forest classifier. Consistent with the observations by Breiman [24] and further corroborated by Liaw and Wiener [27], the Random Forest's consistent performance with the PIMA dataset reaffirms its position of prominence in machine learning applications.

While our experiments were rooted in rigorous methodologies, they were not without their illuminating moments of introspection. Notably, the outcomes from our dimensionality reduction experiment with PCA deviated from what one might expect from theoretical postulations. Such moments, humbling as they are, serve to underline the subtle yet critical chasm that can exist between abstract mathematical formulations and their tangible manifestations in real-world datasets. This deviation nudges us to approach data science with a blend of both rigor and adaptability, being open to unexpected insights.

Diving further into the dataset's granular details, the number of pregnancies, plasma glucose concentration, body mass index, and age have revealed themselves as potential linchpins in diabetes prediction, much in line with previous research findings [5,8]. Yet, the more subtle role of the 'BP' and 'SKIN' attributes reminds us of the broader landscape of attribute interplay and the importance of not viewing any single attribute in isolation.

Conclusively, this exploration has been an enlightening journey, one that reiterates the power of machine learning but equally underscores the necessity for nuanced, iterative data analysis. As the realms of medical diagnostics and data science continue to intersect, it is these intricate dances between data, theory, and application that will pave the way for transformative insights.

VI. Conclusion and Forward Pathways

Throughout our research, we rigorously applied various machine learning methodologies to the PIMA dataset. A consistent standout was the Random Forest classifier, not merely for its algorithmic prowess but for its adaptability and robustness when pitted against intricate datasets like PIMA. The nuanced roles of attributes, especially 'BP' and 'SKIN', underscore the layered complexity within the dataset and the intricacies of diabetes as a medical condition.

Upon deeper examination, it became evident that while some attributes such as the number of pregnancies, plasma glucose concentration, body mass index, and age played pivotal roles in diabetes prediction, others demanded a more careful evaluation. This balance between attribute importance and the broader attribute interplay deepens our understanding and offers a refined perspective on the dataset's potentials and pitfalls.

Looking ahead, there's a wealth of opportunity. The idea of melding deep learning techniques, such as convolutional and recurrent neural networks, with traditional machine learning offers a promising avenue. As medical datasets continue to expand, they will benefit from architectures designed to handle vast amounts of data and extract intricate patterns. This integration could redefine the landscape of medical predictive modeling, particularly for conditions as multifaceted as diabetes. To encapsulate, our findings have been both affirming and enlightening, and the journey ahead in the realms of medical diagnostics and data science is full of promise. Each step we take is more than just academic progression; it is a stride towards enhancing medical prediction and, ultimately, patient outcomes.

References

[1] S. Deepti and S. D. Singh. "Prediction of Diabetes using Classification Algorithms". Procedia Computer Science, 132(), 1578–1585, 2018, doi:10.1016/j.procs.2018.05.122.

[2] I. Aiswarya, J. S and S. Ronak. "Diagnosis of diabetes using classification mining techniques". International Journal of Data Mining & Knowledge Management Process (IJDKP). Feb. 2015, doi : 10.5121/ijdkp.2015.5101.

[3] W. Emanuel, D. L. Silvia, C. Eleonora, B. Paola and F. Giovanni. "CamurWeb: a classification software and a large knowledge base for gene expression data of cancer". BMC Bioinformatics, 19(S10), 245–256, Oct. 2018, doi:10.1186/s12859-018-2299-7.

[4] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, no. November, pp. 1–10, 2018, doi: 10.3389/fgene.2018.00515.

[5] V. V. Vijayan and C. Anjali, "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach," 2015 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2015, no. December, pp. 122–127, 2016, doi: 10.1109/RAICS.2015.7488400.

[6] V. C. Bavkar and A. A. Shinde, "Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement". Indian Journal of Science and Technology 14(10): 869-880, 2021, doi: 10.17485/IJST/v14i10.2187.

[7] Y. Liu et al., "Machine Learning For Tuning, Selection, And Ensemble Of Multiple Risk Scores For Predicting Type 2 Diabetes," Risk Management and Healthcare Policy, Volume 12(), 189–198, Nov. 2019, doi:10.2147/rmhp.s225762.

[8] F. Hassan and M. E. Shaheen, "Predicting Diabetes from Health-based Streaming Data using Social Media, Machine Learning and Stream Processing Technologies," International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, pp. 1957-1967, Number 8. 2020.

[9] K. Harleen and K. Vinita, "Predictive modelling and analytics for diabetes using a machine learning approach," Applied Computing and Informatics, Vol. 18 No. 1/2, pp. 90-100, Mar. 2018, doi:10.1016/j.aci.2018.12.004.

[10] Z. Hancui, C. Shuyu, C. Wenqian, and W. Tianshu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 386–390, Nov. 2017, doi:10.1109/ICSESS.2017.8342938.

[11] W. Zhiliang and X. Zhongxian, "A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier," Eleventh International Conference on Advanced Computational Intelligence (ICACI), pp. 278–283, Jun. 2019, doi:10.1109/ICACI.2019.8778622.

[12] Z. B. Xiangyan et al., "Novel binary logistic regression model based on feature transformation of XGBoost for type 2 Diabetes Mellitus prediction in healthcare systems," Future Generation Computer Systems, 129, pp.1-12, Apr. 2022, doi: 10.1016/j.future.2021.11.003.

[13] O. Adigun, F. Okikiola, N. Yekini, and R. Babatunde, "Classification of Diabetes Types using Machine Learning," International Journal of Advanced Computer Science and Applications, Vol. 13, No. 9, 2022.

[14] "Pima Indians Diabetes Dataset | Kaggle," accessed 06 July 2023.

[15] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," Procedia Comput. Sci., vol. 167, pp. 706-716, 2020, doi: 10.1016/j.procs.2020.03.336.

[16] C. CORINNA and V. VAPNIK, "Support-Vector Networks," Mach. Learn., vol. 20, 1995, pp. 273-297.

[17] I. Ahmad et al., "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," IEEE Access, vol. 6, pp. 33789-33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[18] M. M. Abdelsalam and M. A. Zahran, "A Novel Approach of Diabetic Retinopathy Early Detection Based on Multifractal Geometry Analysis for OCTA Macular Images Using Support Vector Machine," IEEE Access, vol. 9, pp. 22844-22858, 2021, doi: 10.1109/ACCESS.2021.3054743.

[19] J. Choi et al., "Convolutional Neural Network Technology in Endoscopic Imaging: Artificial Intelligence for Endoscopy," Clin. Endosc., vol. 53, pp. 117-126, 2020, doi: 10.5946/ce.2020.054.

[20] B. Alic, L. Gurbeta, and A. Badnjevic, "Machine learning techniques for classification of diabetes and cardiovascular diseases," 6th mediterranean conference on embedded computing (MECO) 2017 Jun 11 (pp. 1-4). IEEE. doi:10.1109/MECO.2017.7977152.

[21] Q. Zou et al., "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, pp. 1-10, 2018, doi:10.3389/fgene.2018.00515.

[22] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," Int. J. Sci. Res., vol. 5, pp. 2094-2097, 2016.

[23] S. Singh and P. Gupta, "Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey." International Journal of Advanced Information Science and Technology (IJAIST), Vol.3, No.7, July 2014, doi:10.15693/ijaist/2014.v3i7.47-52.

[24] L. Breiman, "Random Forests," Mach. Learn., vol. 45, pp. 5-32, 2001, doi:10.1023/a:1010933404324.

[25] V. F. Rodriguez-Galiano et al., "An assessment of the effectiveness of a random forest classifier for land-cover classification," ISPRS J. Photogramm. Remote Sens., vol. 67, pp. 93-104, 2012, doi:10.1016/j.isprsjprs.2011.11.002.

[26] V. Svetnik et al., "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," J. Chem. Inf. Comput. Sci., vol. 43, pp. 1947-1958, 2003, doi: 10.1021/ci034160g.

[27] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, 2002, pp. 18-22.