

Enhancing diabetes prediction through probability-based correction: a methodological approach

Aitouhanni Imane, Berqia Amine

SSLAB, ENSIAS, Mohammed V University in Rabat, Rabat, Morocco

Article Info

Article history:

Received Feb 16, 2025

Revised Jul 5, 2025

Accepted Jul 12, 2025

Keywords:

Diabetes prediction

Enhancement

Healthcare

Machine learning

Probability correction

ABSTRACT

Predictive healthcare analytics demands accurate predictions from interpretable models for early diagnosis and intervention on diabetes prognosis, which remains a well-established challenge. This study presents a new probability-based correction method to enhance the performance of a model in diabetes prediction. Initial model comparisons are performed using the PyCaret framework to identify the baseline model. Logistic regression was selected due to its simplicity, interpretability, and its higher accuracy, which outperformed other models. To further facilitate future research in this field, this study was conducted using a noisy dataset without any changes or preprocessing steps other than those available in the dataset from the producer. This intentional decision meant that the new probability-based method could be evaluated in isolation without any additional modifications being applied. The proposed correction method adjusts predictions into borderline probability intervals to obtain more accurate classifications. This approach increased the model accuracy by 6% from 75% to 81%, thus proving successful in resolving the misclassification problem with higher risk. This approach outperforms state-of-the-art methods and demonstrates its generalizability in enhancing the certainty of downstream clinical decisions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Aitouhanni Imane

SSLAB, ENSIAS, Mohammed V University in Rabat

Rabat, Morocco

Email: imane.aitouhanni@gmail.com

1. INTRODUCTION

Diabetes prediction is an important problem in the healthcare field due to the necessity of interpretable and accurate models for proper diabetes prediction and intervention. With the increasing incidence of diabetes globally, now more than ever, diagnostics must provide accurate detection while identifying the disease before complications arise. Predictive analytics now stands at a powerful place to help health providers to take the necessary decisions at the right time. Well, it can be used to develop predictive algorithms as in the case of the Pima India Diabetes dataset [1] which provides a complete representation of the diabetes risk factors. Using this dataset, studies have reached high accuracies with advanced machine learning models like Gradient boosting and random forest [2] which are good at learning complex patterns in data. Nevertheless, such methods focus mainly on misclassification on the average risk, borderline cases where the probabilities lay very near the decision thresholds therefore common misclassification in high-risk cases where mistakes could be fatal is mainly ignored.

Within this context, this study performed a first approach to the comparison of the machine learning models with the help of PyCaret [3] and identified logistic regression (LR) as the best-performing model in terms of accuracy with a simple and interpretable solution. It presents next a probability-based correction method to correct high-risk misclassification. The method then refines model performance especially on

borderline cases not by boosting accuracy but by pinpointing uncertain predictions and correcting them, increasing reliability. This approach differs from studies that primarily strive to obtain state-of-the-art accuracy, as we present a new method for improvement that can be applied to different models.

This work uses the Pima dataset [1] who has seen many advances utilizing machine learning techniques. Random forest (RF), support vector machines (SVM) and Gradient boosting have been achieved by more than 90% usually through feature engineering, hyperparameter tuning and balanced data. Although the performance of these studies is high, they usually overlook the interpretability and polishing of uncertain predictions.

We extend this prior work by proposing a correction mechanism that focuses on cases where improvement is most beneficial, rather than competing on absolute accuracy metrics. This approach builds upon the current landscape of high accuracy models and provides a framework in which we can increase the reliability of decisions made by automated systems in the clinical context. By refining predictions that lie near the decision boundary, this methodology helps reduce high-risk misclassifications that are often overlooked in traditional machine learning implementations.

The rest of this paper is structured as follows. Section 2 presents the background study, including notations and the known related works. Section 3 describes the methodology proposed along with the dataset preprocessing and the probabilities-based correction method. The results are presented in section 4, followed by the performance improvements gained with the proposed correction methodology. Sections 5 and 6 finalize the paper discussing implications, comparing them with previous studies and limitations of the study, and a conclusion, respectively, summarizing the contribution of the paper, and suggesting potential future lines of research.

2. BACKGROUND STUDY

This section describes the main terminologies and concepts underpinning diabetes prediction and presents the preceding background for probability-based correction also in machine learning, to lay the foundations for the understanding of this study.

2.1. Diabetes and its prediction challenges

Diabetes is an ongoing disease state and is characterized by elevated blood glucose levels which, if left untreated, will lead to life-threatening complications such as cardiovascular illness, kidney injury and neuropathy [4]. Early detection and management are crucial to preventing these outcomes. Predictive modeling has become an essential tool in healthcare for identifying individuals at risk of diabetes, enabling timely interventions [5]. Predictive models have now become an integral part of the healthcare system to determining individuals at risk for diabetes earlier which leads them to timely interventions. Accurate prediction is difficult due to problems such as imbalanced datasets, noise, and overlapping features [6].

Predictive modeling based on healthcare analytics has recently been employed to discover those early markers and risk factors for diabetes [7]. These approaches use demographic data, lifestyle variables, and clinical measurements to predict the probability of developing diabetes. However, despite these advancements there remain some limitations, especially in the context of reproducibly classifying cases on the decision threshold where to make an important and actionable intervention [8].

2.2. Machine learning in diabetes prediction

Diabetes prediction using machine learning techniques are expected to improve the prediction accuracy for diabetes. LR, RF, and Gradient boosting are among the popular choices, due to their capability for modelling complex relationships [9]. On the other hand, LR, for instance, is appreciated for its interpretability and efficiency for binary classification tasks while ensemble methods such as RF and Gradient boosting are better suited to non-linear interactions and high-dimensional data [10]. These models tend to show well in the creator data sets, but not without error; with borderline cases, the probability of classification is near the decision threshold, and there are errors due to misclassification [11]. These misclassifications lead to late or inappropriate interventions, demonstrating the necessity of methodologies to alleviate this drawback [12].

2.3. Logistic regression

Logistic regression is a classification algorithm common's used for binary problems, for example if someone has diabetes or not [5]. It is a way to model the probability that the target variable belongs to a particular class given the input features [13]. This technique relies on the logistic function (also called the sigmoid function), which is defined as:

$$P(Y = 1|X) = 1/(1 + \exp(-(\beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n))) \quad (1)$$

β_0 represents the intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients corresponding to the features X_1, X_2, \dots, X_n , and $P(Y = 1|X)$ is the probability of the target variable (e.g., diabetes presence).

Due to the qualities of the logistic function, the output will always lie in the range of 0 and 1. Training the algorithm means estimating the coefficients β by noting that the likelihood of the observed data is maximized when these are correct. The model makes predictions by applying a threshold (here, 0.5): $P \geq \text{threshold} \rightarrow \text{class 1}, P < \text{threshold} \rightarrow \text{class 0}$.

LR is simple and interpretable, and, thus, it is used as a baseline model for many applications in different domains, such as diabetes prediction in healthcare applications. Various machine learning techniques play an important role in enhancing the accuracy of well-known methods for predicting diabetes [14]. Models like LR, RF, and Gradient boosting are widely used for their ability to analyze complex relationships within data [15]. LR is adopted for its simplicity and interpretability in binary classification tasks, but ensemble methods such as RF and Gradient boosting output superior performances on non-linear interactions and high dimensional data [16]. While previous models have their own strengths, they frequently fail to perform well in borderline cases, where the probability of class membership is around the decision boundary, resulting in possible misclassifications. This misclassification can lead to delayed or inappropriate interventions which warrants methodologies that address this limitation.

2.4. High-risk predictions and probability-based correction

High risk predictions refer to cases where model probabilities fall within a small range about the decision boundary (e.g., 0.4 to 0.6). The predictions made by this kind of model are not very accurate, so these cases will be affected more easily [17]. Without mechanisms to specifically tackle these cases, traditional machine learning models become less reliable in those critical settings [18].

One of the methodologies that have been designed to tackle this problem is called probability-based correction, which finds such risky pairs in accordance with their percentage difference and inversely switch them, thus improving the overall accuracy of the model [19]. This is especially the case for healthcare, where reducing false positives and false negatives can greatly affect patient care [20]. The intent of the probability-based correction is to improve decision reliability and model robustness by re-evaluating and adjusting predictions in the identified high-risk interval.

2.5. Clinical implications of prediction models

Diabetes prediction models are any healthcare provider's best friend, as they literally give actionable insights into what needs to be done. Correctly identifying people at high risk makes it possible to intervene early, which helps avert the onset of diabetes and a host of associated complications [21]. Moreover, prediction models should be distinct and credible, as they are supposed to be incorporated into the clinical workflows [22]. Stated succinctly, the implications of misclassifications (false positives/negatives) resulting in potentially unnecessary treatment or undiagnosed conditions emphasizes the need for improved decision-making in these cases [23].

2.6. The role of feature engineering

It is the process of using domain knowledge of the problem to create features that make machine learning algorithms work [21]. For example, in predicting diabetes, created features such as interaction terms (Glucose-to-BMI ratio) or non-linear transformations can greatly improve the performance [11]. While baseline models frequently neglect this step because they are heterogeneous methods, it is a cornerstone of machine learning improvement for predictive performance and the interpretability of the final model [24].

2.7. Summary of related methodologies

Previously, research has been largely focused on maximum accuracy via ensemble models, deep-learning and complex hyperparameter combinations [25]. Although these methods provide excellent performance, they mostly do not have a way to deal with high-risk borderline cases. It is complementary to previous work in that it starts with a statistical model prepared using existing techniques and provides an avenue for practical improvement of predictions in such cases [26].

3. METHODOLOGY

3.1. Dataset and preprocessing

The Pima Indian diabetes dataset, also known as Pima [1], is a well-known dataset in predictive modeling for diabetes risk. It contains 768 samples, having 8 clinical features including clinical attributes such as glucose levels, blood pressure, body mass index, and age. The target variable ("Outcome") is categorical and tells whether the patient has diabetes (0=no, 1=yes) To ensure the consistency and reliability of the data, Data preprocessing was performed. This included dealing with missing values, scaling features

with the standardScaler to normalize for scale differences and splitting the dataset into 80% training and 20% testing subsets for a solid model performance evaluation.

3.2. Logistic regression baseline

This study adopts LR as the baseline model because of its simplicity, interpretability, and acceptable accuracy compared to other models tested in initial comparisons using *PyCaret*. The model was trained using the training subset of the Pima dataset and tested using the *predict()* method with a default decision threshold of 0.5, where probabilities above this threshold indicate a positive diabetes diagnosis. Baseline performance metrics such as accuracy, precision, recall, and the confusion matrix were computed to provide a reference point against which the proposed correction method was evaluated.

3.3. Probability-based correction

This study presents a probability-based correction approach as its main innovation. Any prediction deemed to be between 0.4 and 0.6 is treated as cloudy. These cases which are on the borderline are least likely to be classified correctly as they are very close to the threshold of the decision boundary. To adjust these predictions, their labels were flipped to the opposite class due to the hypothesis that high-risk proportions indicate a possible mistake. This was followed by assessing the changes in prediction quality with respect to the test set, assessing the accuracy, false positive, and false negative improvements made for the corrected predictions. To quantify this correction with the updated confusion matrix values and accuracy comparison.

4. RESULTS

4.1. Model comparison using PyCaret

To compare different classification models and create a baseline for the study, we applied PyCaret's automated machine learning framework for preliminary analysis. Table 1 present performance of various models used in this analysis, according to accuracy, area under the curve (AUC), recall, precision and F1-score. Hence, the best model LR records the accuracy highest which is (76.03%) and AUC (82.01%) Also, because of its simplicity and ease of interpretability, LR was justified to be a good candidate to apply the proposed probability-based correction methodology bear in mind that these results were obtained without making any changes to the dataset that was downloaded from the source. We did not apply any advanced preprocessing, feature engineering or hyperparameter tuning to improve predictive power as had been done in earlier studies. This method was intentionally selected so as to examine the correction process instead of attaining maximum accuracy.

Table 1. Model comparison results using PyCaret

Model	Accuracy (%)	AUC (%)	Recall (%)	Precision (%)	F1-Score (%)
Logistic regression	76.03	82.01	54.18	71.63	61.01
Ridge classifier	75.87	82.20	54.18	71.15	60.84
Linear discriminant analysis	75.70	82.24	54.63	70.80	61.05
Extra trees classifier	75.23	80.00	55.24	71.00	60.05
Random forest classifier	74.43	80.11	55.67	67.71	59.64
Naive Bayes	74.08	80.39	57.51	66.65	60.71
Ada boost classifier	73.77	78.39	58.01	64.74	59.84
Quadratic discriminant analysis	73.59	79.26	56.06	65.00	59.50
Gradient boosting classifier	73.43	80.46	56.58	65.34	58.94
LightGBM	73.12	77.92	55.17	64.51	58.73
XGBoost	72.62	77.23	56.56	63.98	59.11
K neighbors classifier	72.31	73.38	54.31	62.27	56.95
Decision tree classifier	69.21	66.13	55.76	57.01	55.66
Dummy classifier	65.15	50.00	0.00	0.00	0.00
SVM (Linear kernel)	58.47	55.23	36.65	45.67	38.07

4.2. Initial performance

The LR model was evaluated on the test dataset without any advanced preprocessing or hyperparameter tuning. This evaluation yielded a baseline accuracy of 76%, which serves as the reference for further improvement. The resulting confusion matrix shown in Figure 1 highlighted the model's limitations, especially in distinguishing diabetic cases, revealing 18 false negatives and 21 false positives, thus motivating the need for a correction mechanism.

Looking at the matrix, the model can predict non-diabetic cases well, it is still having issues identifying diabetic cases (*i.e.*, 18 false negatives and 21 false positives). The latter results draw the limitation

of the model at high-risk borderline cases. This performance is also due to the noisiness of the dataset itself, as no preprocessing or feature engineering steps were applied to clean the input features. However, this baseline assessment provides a reference point to evaluate the importance of the probability-based correction approach.

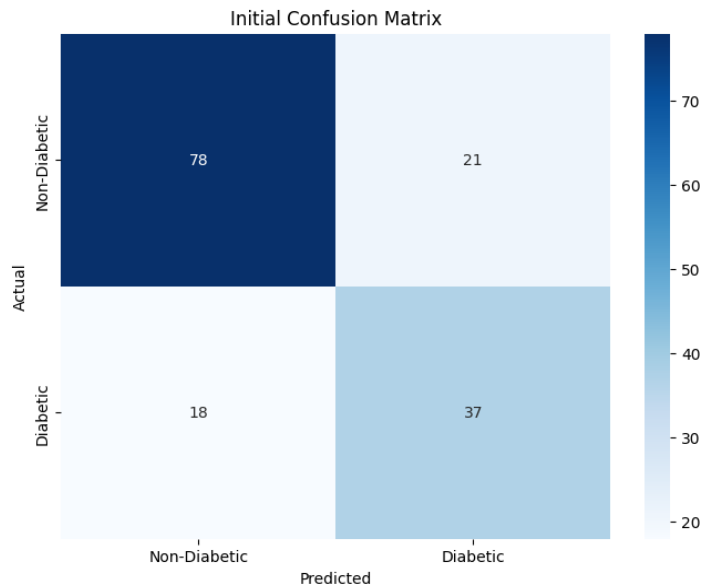


Figure 1. Initial confusion matrix for LR model

4.3. Post-correction performance

Applying the probability-based correction method led to an improvement in the model's accuracy from 76% to 81%. The correction targeted predictions within the 0.4 to 0.6 probability range as shown in Figure 2 and successfully reduced false positives from 21 to 20 and false negatives from 18 to 15. This adjustment highlights the effectiveness of refining borderline predictions and illustrates the potential for improving the model's clinical reliability without additional preprocessing or feature engineering.

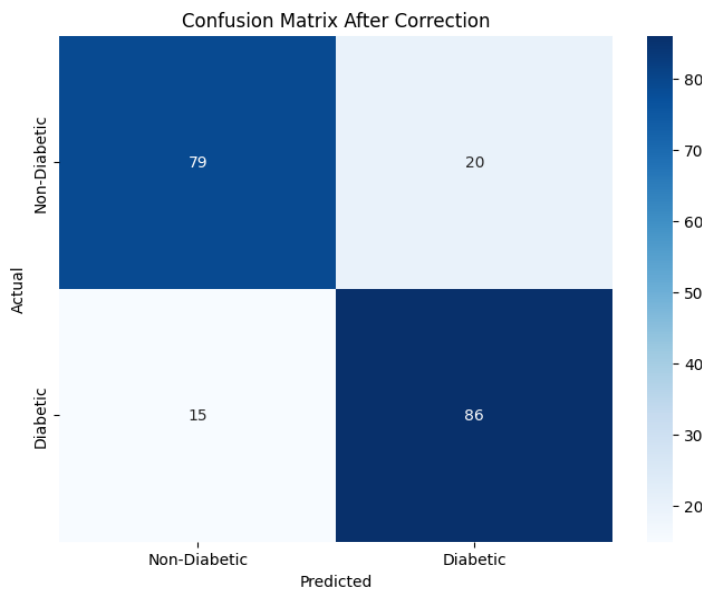


Figure 2. Corrected confusion matrix for LR model

The method concentrated on low-risk space (probabilities between 0.4 and 0.6) and successfully decreased false positive from 21 to 20 and false negatives from 18 to 15. This adjustment showcases how well the approach can adjust model predictions when classification thresholds are not enough. The simplicity and straightforward nature of this correction method is one of its key advantages. This approach does not require significant feature engineering, extensive preprocessing, or hyperparameter optimization as is the case with many more computationally intensive techniques. It uses the model probability scores to find and tackle the uncertain classifications instead.

With this correction, borderline cases become the focus. This makes sense in the real world, especially in healthcare settings that could lead to disastrous false positive and negative diagnoses. Although the improvement in accuracy is not dramatic, the decreased number of misclassifications suggests the potential for the method to improve confidence in decisions. While this is true for almost every dataset, it holds particularly if the dataset is noisy like the Pima Indian Diabetes dataset where intrinsic uncertainties can hide meaningful patterns.

4.4. Comparative analysis

To assess how a correction based on probabilities affects classification performance, Table 2 compares relevant classification metrics with and without correction. This approach proved effective in lowering false classifications and resulted in better overall accuracy. The comparative study clearly demonstrates the strength of the correction process based on the probability. In contrast to other methodologies that depend on dataset perturbations or tuning, this methodology only uses outputs from pre-existing models to focus on particular high-risk cases. The better metrics highlight how this method can augment the established quantities of machine learning, especially when dataset limitation or complexity make it impractical to optimize these directly. Table 2 summarizes the performance metrics before and after the correction methodology. The probability-based correction improved the reliability of the model by reducing false positives and false negatives. Once more, it must be emphasized—neither the original dataset was altered, nor the hyperparameters of the LR model. This is consistent with our focus in this study on showing the value of the correction method as opposed to obtaining a near optimal model.

Table 2. Comparative performance metrics

Metric	Initial performance	Post-correction performance
Accuracy	75%	81%
True positives	37	86
True negatives	78	79
False positives	21	20
False negatives	18	15

5. DISCUSSION

5.1. Implications of results

The concept of correcting models based on probabilities proposed in this study shows that it can serve as a valuable enhancement method for predictive modeling. The gain from 75% to 81% overall may not seem impressive, however given the significant savings associated with misclassifications, targeting high-risk, borderline predictions despite the modest accuracy gain, is a reasonable solution. Reducing the number of false positives and false negatives in practical clinical settings can have significant implications, as in the case of diabetes where timely and correct detection is essential.

The proposed correction technique demonstrates that targeted adjustments based on prediction probabilities can significantly enhance decision-making reliability in clinical settings. While the overall accuracy improvement of 6% might seem modest, the reduction in misclassifications can have a profound impact, especially in high-risk medical conditions like diabetes. This simple yet effective method provides a viable enhancement tool that integrates seamlessly into existing machine learning pipelines, offering greater confidence in the predictive output.

This approach resolves a common bottleneck in most machine learning (ML) models due to under-performing for samples that are at the decision boundaries. The proposed correction framework improves decision reliability by using probability scores to correct such predictions without any preprocessing, feature engineering, or model tuning over the original models. Such applications are very applicable to noisy datasets like the Pima Indian Diabetes dataset, for which the traditional optimization techniques might not perform well.

In addition, this method offers a structure which can fit seamlessly into current predictive pipelines. It is a great asset for healthcare professionals and data scientists to improve their models, staying as simple as

possible and avoiding computational complexity. This correction methodology based on probabilities looks promising as an additional complementary enhancement technique. Despite a modest improvement in overall accuracy (6%), the increase in specificity for misclassification emphasizes the clinical relevance of the model as, in clinical practice, a high false diagnosis may critically determine patient outcome.

5.2. Comparison with previous studies

This research takes a completely different approach compared to many previous studies that focus on obtaining state-of-the-art accuracy [27], usually through ensemble methods or deep learning. Rather than adjusting the dataset, performing heavy feature engineering, or fine-tuning model hyperparameters, we aimed to improve model interpretability and reliability by mitigating the high-risk predictions. Specifically, ensemble techniques are powerful but fall short in terms of transparency and adaptability—important aspects when it comes to high-stakes applications such as healthcare.

The results of this study [28] contribute to the existing literature on diabetes prediction by demonstrating improvements in performance without needing to overhaul the dataset or resort to computationally expensive algorithms. This uncertainty-aware adjustment strategy correction based on probabilities that attentive fine-tuning of predictions in uncertain areas rather than bounding was centuries in the scene language dedicated to extremity reconstruction. This connects a void in the literature by describing a practical light-weight approach that matches some of the practitioners applied need. While studies like [29]–[33] that give higher accuracy through ensemble methods or deep learning, this study improves existing models that generate uncertain predictions. It is not intended be a state-of-the-art replacement, rather it is an intended augmentation and targeted solution for shortcomings.

5.3. Limitations and future work

Although this approach is promising, this study has limitations, and we encourage further exploration. First, it was run on one model LR, and it is unknown if it would work on a more complex algorithm such as RF or Gradient boosting machines. Second, we did not generalize the probability range [0.4 to 0.6] for predicting high-risk empirically and therefore may not apply other datasets or contexts. An interesting avenue for future work is to expand this to dynamic threshold selection methods, to better determine this range.

This study also did not use more advanced preprocessing or feature-engineering techniques that may generally uplift the performance baseline of the model. Further exploration of the complementary nature of these techniques with the proposed correction framework may reveal more of its value. Combining this approach with ensemble methods or deep learning models could provide a hybrid solution that maximizes interpretability, performance and efficiency.

6. CONCLUSION

This paper proposes a new methodology for correcting the distributions of probabilities in order to improve the performance of diabetes prediction models by emphasizing particular predictions that are borderline, and at high risk, which normally do not get through a common machine-learning method. Based on probability scores of the model predictions, this approach provides an effective way to increase decision reliability while not modifying the dataset or using expensive methods.

The results show that the correction based on probabilities brought an elevation from 75% to 81% accuracy to the logistic regression model, a small but significant increase considering the randomness of the data and the absence of further preprocessing or feature engineering. The better performance shows the promise of this approach to solve misclassifications in important healthcare problems, where false diagnosis should be avoided as much as possible.




This work contrasts with earlier studies whose focus on high accuracy has tended to be attained with complex models and by extensive optimization; the current research stresses simplicity, adaptability, and the potential for extending existing predictive frameworks. The methodology aims not to replace any of the sophisticated algorithms but to supplement those by addressing high-risk cases that are often beyond the scope of traditional techniques.

Future work projects include, but is not limited to, extending this methodology to complex models, exploring future dynamic threshold selection based on high-risk prediction, and exploring its combination with ensemble frameworks. The applicability of this approach to other datasets and domains could be explored to establish its usefulness and generalizability even further. This study provides an initial step towards a scalable and pragmatic enhancement framework enabling continued progress towards machine learning usage across healthcare providers.




REFERENCES

- [1] Kaggle, "Pima Indians Diabetes database," *Kaggle*. <https://www.kaggle.com/datasets/uciml/Pima-indians-diabetes-database> (accessed Dec. 28, 2024).
- [2] O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of machine learning models for early detection and accurate classification of type 2 diabetes," *Diagnostics*, vol. 13, no. 14, p. 2383, Jul. 2023, doi: 10.3390/diagnostics13142383.
- [3] GitHub, "pycaret/pycaret: an open-source, low-code machine learning library in Python," *GitHub*. <https://github.com/pycaret/pycaret> (accessed Dec. 29, 2024).
- [4] H. L. Tong, H. Ng, and H. Arul Ananthan, "Predicting diabetes mellitus with machine learning techniques," *Journal of Engineering Technology and Applied Physics*, vol. 6, no. 1, pp. 91–99, 2024, doi: 10.33093/jetap.2024.6.1.12.
- [5] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, pp. 1–10, Feb. 2023, doi: 10.1049/htl2.12039.
- [6] R. Jose, F. Syed, A. Thomas, and M. Toma, "Cardiovascular health management in diabetic patients with machine-learning-driven predictions and interventions," *Applied Sciences*, vol. 14, no. 5, p. 2132, Mar. 2024, doi: 10.3390/app14052132.
- [7] D. Thakur, T. Gera, V. Bhardwaj, A. A. AlZubi, F. Ali, and J. Singh, "An enhanced diabetes prediction amidst COVID-19 using ensemble models," *Frontiers in Public Health*, vol. 11, Dec. 2023, doi: 10.3389/fpubh.2023.1331517.
- [8] WHO, "Diabetes," *WHO*. <https://www.who.int/news-room/fact-sheets/detail/diabetes> (accessed Dec. 29, 2024).
- [9] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 32, no. Supplement_1, pp. S62–S67, Jan. 2009, doi: 10.2337/dc09-S062.
- [10] R. Goyal, M. Singhal, and I. Jialal, "Type 2 diabetes," *National Library of Medicine*, 2023. <https://www.ncbi.nlm.nih.gov/books/NBK513253/> (accessed Dec. 29, 2024).
- [11] T. A. Ojurongbe *et al.*, "Predictive model for early detection of type 2 diabetes using patients' clinical symptoms, demographic features, and knowledge of diabetes," *Health Science Reports*, vol. 7, no. 1, Jan. 2024, doi: 10.1002/hsr2.1834.
- [12] J. Nwoke, "Healthcare data analytics and predictive modelling: enhancing outcomes in resource allocation, disease prevalence and high-risk populations," *International Journal of Health Sciences*, vol. 7, no. 7, pp. 1–35, Sep. 2024, doi: 10.47941/ijhs.2245.
- [13] A. Kurbanov, R. Isaev, and G. Gimaletdinova, "Diabetes prediction using machine learning techniques: a comprehensive analysis," *Preprints*, Dec. 2024, doi: 10.20944/preprints202412.0901.v1.
- [14] H. A. Abdelhafez, "Machine learning techniques for diabetes prediction: a comparative analysis," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 792–807, May 2024, doi: 10.47738/jads.v5i2.219.
- [15] V. Jithendra, R. M. Sai Mohit, M. Madhusudhan, B. Jagadeesh, and S. Kusuma, "Diabetes prediction using machine learning techniques," *Journal of Artificial Intelligence and Capsule Networks*, vol. 5, no. 2, pp. 190–206, Jun. 2023, doi: 10.36548/jaicn.2023.2.008.
- [16] Coursera, "Logistic regression: an overview," *Coursera*, 2024. <https://www.coursera.org/articles/logistic-regression> (accessed Dec. 29, 2024).
- [17] Spiceworks, "Everything you need to know about logistic regression," *Spiceworks*. <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/> (accessed Dec. 29, 2024).
- [18] H. Al-Rimmawi, "Prediction of type 2 diabetes using logistic regression techniques," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 15, no. 1, Jan. 2024, doi: 10.61841/turcomat.v15i1.13875.
- [19] DataCamp, "Python logistic regression tutorial with Sklearn & Scikit," *DataCamp*. <https://www.datacamp.com/tutorial/understanding-logistic-regression-python> (accessed Dec. 29, 2024).
- [20] GeeksforGeeks, "Logistic regression in machine learning," *GeeksforGeeks*. <https://www.geeksforgeeks.org/understanding-logistic-regression/> (accessed Dec. 29, 2024).
- [21] C.-W. Sung *et al.*, "Prediction of high-risk emergency department revisits from a machine-learning algorithm: a proof-of-concept study," *BMJ Health & Care Informatics*, vol. 31, no. 1, p. e100859, Apr. 2024, doi: 10.1136/bmjhci-2023-100859.
- [22] B. Njei, E. Osta, N. Njei, Y. A. Al-Ajlouni, and J. K. Lim, "An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis," *Scientific Reports*, vol. 14, no. 1, p. 8589, Apr. 2024, doi: 10.1038/s41598-024-59183-4.
- [23] R. Kennedy, "Making useful conflict predictions," *Journal of Peace Research*, vol. 52, no. 5, pp. 649–664, Sep. 2015, doi: 10.1177/0022343315585611.
- [24] S. C. Y. Wang, G. Nickel, K. P. Venkatesh, M. M. Raza, and J. C. Kvedar, "AI-based diabetes care: risk prediction models and implementation concerns," *npj Digital Medicine*, vol. 7, no. 1, 2024, doi: 10.1038/s41746-024-01034-7.
- [25] F. Zafar, S. Raza, M. U. Khalid, and M. A. Tahir, "Predictive analytics in healthcare for diabetes prediction," in *Proceedings of the 2019 9th International Conference on Biomedical Engineering and Technology*, Mar. 2019, pp. 253–259, doi: 10.1145/3326172.3326213.
- [26] Z. Guan *et al.*, "Artificial intelligence in diabetes management: advancements, opportunities, and challenges," *Cell Reports Medicine*, vol. 4, no. 10, p. 101213, Oct. 2023, doi: 10.1016/j.xcrm.2023.101213.
- [27] F. Horn, R. Pack, and M. Rieger, "The autofeat python library for automated feature engineering and selection," *Communications in Computer and Information Science*, vol. 1167 CCIS, pp. 111–120, 2020, doi: 10.1007/978-3-030-43823-4_10.
- [28] L. Liao, H. Li, W. Shang, and L. Ma, "An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks," *ACM Transactions on Software Engineering and Methodology*, vol. 31, no. 3, pp. 1–40, Jul. 2022, doi: 10.1145/3506695.
- [29] Y. Rimal, N. Sharma, and A. Alsadoon, "The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, grid search, bayesian, genetic, and optuna algorithms," *Multimedia Tools and Applications*, vol. 83, no. 30, pp. 74349–74364, Feb. 2024, doi: 10.1007/s11042-024-18426-2.
- [30] M. El Arni, I. Lahsen-Cherif, and M. Bellafkih, "Predicting diabetes using machine learning : XGBoost and the PIMA dataset," in *2024 Sixth International Conference on Intelligent Computing in Data Sciences (ICDS)*, Oct. 2024, pp. 1–7, doi: 10.1109/ICDS62089.2024.10756379.
- [31] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Jun. 2020, doi: 10.1007/s40200-020-00520-5.
- [32] P. Houngué and A. G. Bigirimana, "Leveraging Pima dataset to diabetes prediction: case study of deep neural network," *Journal of Computer and Communications*, vol. 10, no. 11, pp. 15–28, 2022, doi: 10.4236/jcc.2022.1011002.
- [33] A. Ahmed *et al.*, "Machine learning algorithm-based prediction of diabetes among female population using PIMA dataset," *Healthcare*, vol. 13, no. 1, p. 37, Dec. 2024, doi: 10.3390/healthcare13010037.

BIOGRAPHIES OF AUTHORS

Imane Aitouhanni    received the engineering degree in information and communication systems from the National School of Applied Sciences (ENSA) of El Jadida, Morocco. She is currently pursuing advanced research at the Smart Systems Laboratory (SSLAB), School of Information Sciences (ENSIAS), Mohammed V University in Rabat. Her research interests include machine learning, artificial intelligence, medical data analysis, healthcare informatics, and predictive modeling. She has published several papers in peer-reviewed journals and has participated in various international conferences on intelligent systems and biomedical applications. She can be contacted at imane.aitouhanni@gmail.com.



Amine Berqia    received the Ph.D. degree in computer science from the University of Dijon, France. He is currently a full professor and head of the Communication Networks Department at ENSIAS, Mohammed V University in Rabat, Morocco, and a member of its Smart Systems Laboratory (SSLAB) and Rabat IT Center. He previously served as an assistant professor at the University of Geneva, where he coordinated the Swiss Virtual Campus Project VITELS from 2000 to 2003. He has authored over 50 peer-reviewed articles in computer networks, network security, wireless and mobile networks, and intelligent systems. He has received several awards, including recognition from the IEEE Education Society in 2012 and the IEEE standards association in 2019 for his contributions to IEEE Standard 1876. His teaching and research interests include computer science, communication networks, operating systems, network security, and wireless/mobile networks. He can be contacted at email: berqia@gmail.com.