Integrating Causal Inference and Machine Learning for Early Diagnosis and Management of Diabetes

Sahar Echajei, Mohamed Hafdane, Hanane Ferjouchia, Mostafa Rachik
Department of Mathematics and Computer Science, Faculty of Sciences Ben M'sik-Hassan II University of Casablanca, Morocco

***Abstract*—In the context of the increasing prevalence of diabetes, this work focuses on integrating causal inference with Machine Learning (ML) for early diagnosis and effective management of diabetes. We applied a series of advanced techniques to improve model performance, including the use of data preprocessing methods, evaluation of variable importance and causal analysis, Feature Engineering methods, and hyperparameter optimization. The diabetes prediction model is a Stacking ensemble model that combines the predictions of several base models (namely: Random Forest Classifier, XGBClassifier, Gradient Boosting Classifier). Initial results showed a precision of 0.70, a recall of 0.70, an Area Under Curve (AUC) of 0.768, and a Mean Cross Entropy (MCE) of 0.299. After optimization, precision increased to 0.73, recall to 0.73, AUC to 0.798, and MCE improved to 0.271. This approach has demonstrated a significant improvement in diabetes prediction, suggesting that the integration of causal inference and Machine Learning is a promising path for the diagnosis and management of diabetes. The reduction in MCE, alongside improvements in precision, recall, and AUC, underscores the effectiveness of our optimization techniques in enhancing model reliability and performance.**

***Keywords*—Machine learning; classification; causal inference; Bayesian networks; ensemble technique; diabetes diagnosis**

I. Introduction

Diabetes is a chronic disease that is steadily increasing, posing a major challenge to health systems worldwide. Early diagnosis and effective management are essential to reduce complications and improve the quality of life of patients. This disease, characterized by chronic hyperglycemia[1], manifests when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces.

From traditional methods to innovative approaches, the literature reveals a variety of techniques for diagnosing and managing diabetes. However, gaps remain, particularly in terms of predictive accuracy and operational efficiency.

In this context, the advent of Artificial Intelligence (AI), particularly Machine Learning, which allows systems to learn and evolve from data without being explicitly programmed [1], [2], opens new perspectives for diabetes diagnosis. ML is increasingly used to analyze complex datasets and can help identify patterns and trends that are difficult to detect by traditional analytical methods.

In parallel, Causal Inference, a statistical method that allows for inferring cause-and-effect relationships from data, combined with Machine Learning [3], offers prospects for overcoming several limitations of conventional methods by identifying non-obvious relationships between variables and the disease.

[1]World Health Organization, "Diabetes". https://www.who.int/news-room/fact-sheets/detail/diabetes.

This study aims to leverage these advanced technologies to improve early diagnosis and management of diabetes. We present a model integrating causal analysis and various Machine Learning techniques to analyze patient data in innovative ways. The objective is to provide a tool capable of extracting meaningful and actionable knowledge from vast health datasets, thus helping to bridge the gap between traditional diagnostic methods and the individual needs of patients for personalized and proactive diabetes management.

II. Related Work

The integration of causal inference and machine learning in diabetes risk prediction has seen notable advancements, significantly enhancing the precision and reliability of predictive models. Researchers [3]-[5] have extensively explored a range of ML algorithms, such as Neural Networks, Decision Trees, Random Forest, Naïve Bayes, and Support Vector Machines, to predict diabetes effectively. Gradient boosting techniques, especially XGBoost, have proven to be highly effective in classification tasks, delivering superior performance [6], [7]. Ensemble learning methods [8]-[13], particularly stacking models, have gained popularity due to their ability to improve model robustness and accuracy by combining outputs from several base learners using a meta-learner.

The processes of feature engineering and selection are vital in developing high-performance predictive models. Techniques like Random Forest-based feature importance are crucial for enhancing model interpretability and performance [14]. Causal inference has proven to be a powerful technique for identifying cause-and-effect relationships within health data [15], [16], providing more profound insights than traditional correlation methods. The core principles of causal inference have been widely applied in the healthcare sector, particularly to refine treatment strategies and improve patient outcomes. Specifically, in diabetes research, causal inference methods have been employed to model causal relationships within clinical datasets. At the heart of causal inference are Bayesian networks, which offer the possibility of modeling the probabilistic dependency relationships between variables [17].

Formally, a Bayesian network can be described by the following Formula (1):

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \backslash \text{Parents}(X_i)) \quad (1)$$

where, $P(X_1, X_2, \ldots, X_n)$ represents the joint probability of n random variables, and $Parents(X_i)$ are the direct precursors of the variable X_i in the network. This formula encapsulates the essence of Bayesian networks, allowing for the decomposition of the complexity of interactions between diabetes risk factors

and its clinical manifestations into simpler and calculable relationships.

III. METHODOLOGY

The methodology of this research is outlined in this section, which includes: (1) description of data selection and data processing, (2) description of causal analysis and ML models, and (3) description of the techniques used to validate and evaluate the models. The process of developing these models, executed using Python for scripting and analysis, is shown in Fig. 1.

A. Data Selection and Processing

In our study, we utilized the Health Facts database from Cerner Corporation [18], which compiles de-identified and detailed clinical records from a broad spectrum of healthcare facilities, including 130 hospitals and integrated delivery networks across the United States, spanning a decade from 1999 to 2008. Health Facts, a voluntary initiative for organizations employing the Cerner electronic medical record system, provided a rich and comprehensive foundation for our research. The dataset comprises essential demographic information including gender, age, and ethnicity; clinical metrics such as diagnoses and blood glucose levels; and various clinical measurements including renal function, creatinine levels, and heart rate. Additionally, it contains other pertinent health data such as BMI, height, and weight. The variables are classified into numeric and nominal types.

To ensure a focused and relevant dataset, we established specific inclusion criteria targeting patients who had undergone blood glucose measurements, aiming to identify those at potential risk or already diagnosed with diabetes. Through this selective process, we identified 34.367 unique patients and proceeded to analyze 88 different variables. This methodological approach enabled us to accurately identify factors associated with an increased risk of diabetes, leveraging a database that is both exhaustive and representative of the national population.

The cleaning and preparation of medical data marked the beginning of our exploration. We started by removing irrelevant variables, such as patient identifiers and hospital codes, to reduce noise and focus our analysis on information directly impacting outcomes. To enhance the robustness of our database, records with more than 40% missing data were deleted, ensuring data reliability for analysis. Missing values were imputed using the mode for categorical variables and the Iterative Imputer method for numerical variables, preserving relevant information without introducing significant bias. The Iterative Imputer algorithm employs a round-robin approach [19], [20], utilizing regression models to estimate the missing values within a feature by leveraging the remaining features as predictors.

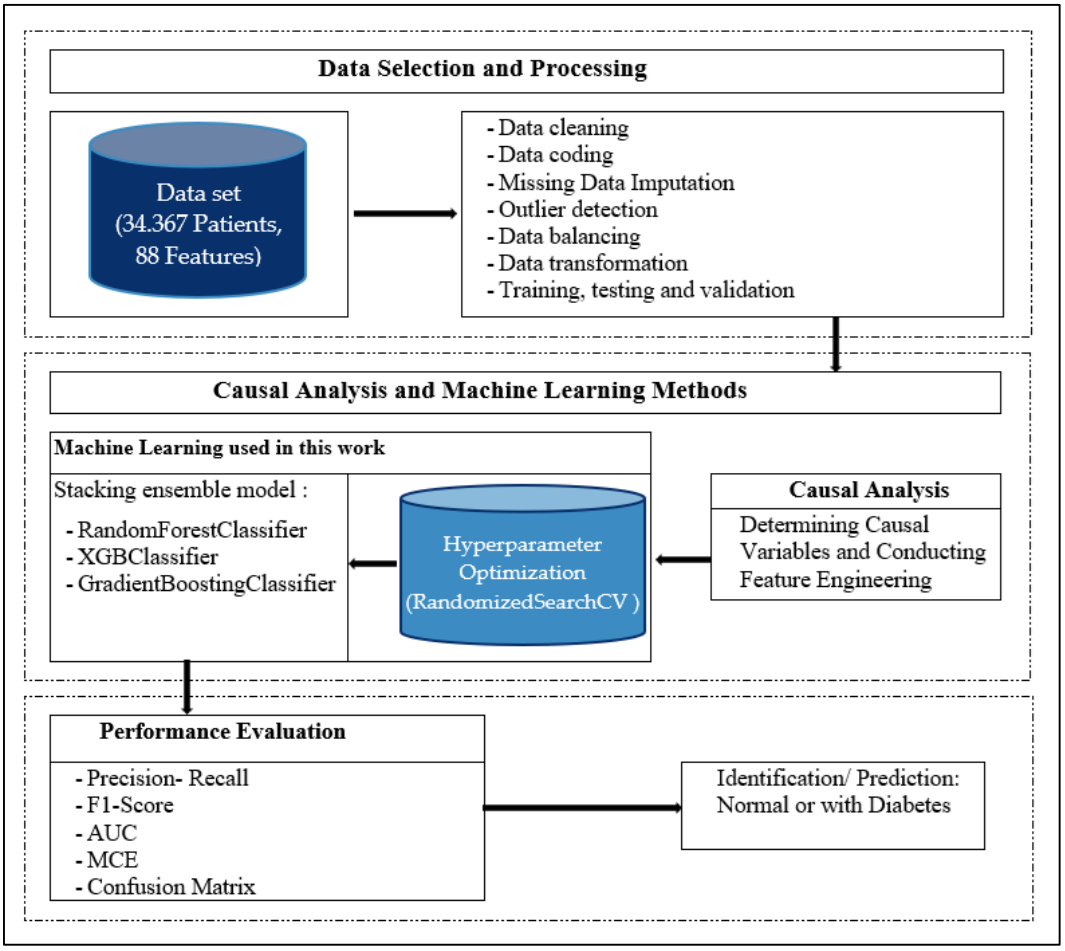

Fig. 1. Data preparation and analytical framework for diabetes prediction.

Furthermore, the application of the Isolation Forest method for outlier identification and removal bolstered our dataset's integrity by eliminating extreme values that might distort outcomes. The Isolation Forest algorithm constructs a binary tree to directly analyze outlier data instances [21]-[23]. To address the common problem of class imbalance in medical datasets, which can bias model performance in favor of the majority class, the Synthetic Minority Over-sampling Technique method (SMOTE) was employed to balance the training data [24]. Subsequent data normalization was essential for aligning the scales of different variables [25], enabling ML models to converge more rapidly and stably.

B. Causal Analysis and Machine Learning Methods

The XGBClassifier, an XGBoost model for classification [26], was utilized for initial predictions, taking advantage of its renowned performance and speed in classification tasks. XGB operates through the successive, iterative creation of an ensemble of simple models, like decision trees, building them one after another. Each new model in the sequence aims to address the inaccuracies of its predecessors. This process enhances the predictive accuracy of the ensemble and mitigates overfitting by minimizing a specified loss function [27]. XGBoost can be succinctly described as an ensemble learning methodology grounded in decision trees, as seen in Fig. 2, employing Gradient Descent as its fundamental objective function. This framework offers considerable versatility and efficiently leverages computational resources to achieve the expected outcomes.

The integration of causal analysis with Machine Learning constitutes the core of our methodological approach. Initially, variable analysis was conducted using the Random Forest Classifier to evaluate their importance [28]-[30], selected for its ability to efficiently handle large datasets and provide a reliable estimate of variable importance without making prior assumptions about data distribution.

Causal analyses were then conducted in two main stages, employing Bayesian network-based inference techniques [31], [32], to identify variables with potential causal relationships to diabetes:

- An initial causal analysis with all variables was conducted using the Hill Climbing Search and Bayesian Network algorithms to explore the dataset comprehensively. The Hill Climbing Search algorithm, an optimization tool, searches for the most effective network structures by maximizing a score function that evaluates each structure's quality in relation to the observed data. In conjunction with the Bayesian Network algorithm, which constructs a probabilistic model to represent causal relationships among variables, this phase enabled the identification of complex interactions and the main precursors of diabetes within our dataset.
- Subsequently, a more refined causal analysis targeted a restricted set of variables, focusing on those identified as impactful on diabetes by the initial Bayesian analysis and those deemed most influential by the Random Forest Classifier. This iterative use of the Hill Climbing Search and Bayesian Network algorithms in the second analysis phase confirmed and validated the initial findings, concentrating on a narrower subset of variables. This methodological approach strengthens the reliability of our conclusions, ensuring our model is robustly grounded for effective diabetes diagnosis and management.

After identifying and validating key causal variables, such as age and gender, we adjusted them based on specific weightings and established interactions between some variables, like creating ratios and products. Integrating these adjusted and interactive variables into our Machine Learning database aimed to refine our prediction accuracy, leveraging the causal relationships between variables and the incidence of diabetes.

To further refine our model, we implemented two key techniques aimed at hyperparameter optimization: (1) using Randomized Search CV to fine-tune the hyperparameters of three base models: RandomForestClassifier, GradientBoostingClassifier, and XGBClassifier, and (2) implementing a Stacking ensemble model further enhanced predictive performance by combining multiple base models [33], [34], producing more accurate and robust predictions.

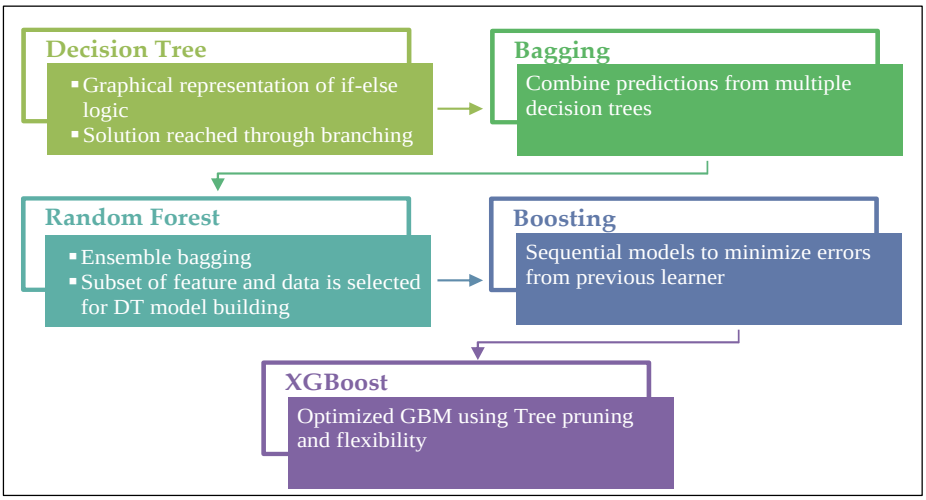

Fig. 2. The evolution of XGBoost from tree-based models [7].

Hyperparameter optimization through Randomized Search CV is an efficient technique to explore a vast parameter space and identify the optimal configuration for our model [35]. This technique offers an optimal balance between computational efficiency and the ability to discover hyperparameter combinations that maximize model performance. Using Randomized Search CV ensures that our model is not only robustly adapted to our dataset's specifics but also refined to achieve the best possible performance in terms of precision, recall, F1 score, AUC, MCE, etc. Simultaneously, adopting a Stacking ensemble model is based on the principle that diversity in prediction methods contributes to a significant improvement in the overall model performance. This synergy exploits each model's unique strengths while mitigating their individual weaknesses, leading to superior generalization capacity and prediction accuracy.

In summary, our approach to hyperparameter optimization, coupled with strategic adjustment of key causal variables, constitutes a methodical approach aimed at maximizing the diagnostic efficacy of our model in the early prediction of diabetes. This combination of precise adjustments and thorough optimization seeks to raise the standard of accuracy and reliability necessary for clinical applications in diabetes diagnosis.

C. Performance Evaluation

The model's performance post-optimization was assessed using several key metrics [25]:

- Precision: This measures the proportion of correct predictions (true positives, TP) among the predicted positive cases (true positives, TP and false positives, FP).

$$Precision=TP/(TP+FP)$$

- Recall: This evaluates how many actual positive cases were correctly identified by the model, compared to the total actual positive cases (true positives, TP and false negatives, FN).

$$Recall=TP/(TP+FN)$$

- F1 Score: As a harmonic mean of precision and recall, this metric evaluates the balance between these two metrics.

$$F1\ Score=2 \times (Precision \times Recall)/(Precision+Recall)$$

- AUC: This provides an overall measure of model performance, indicating its ability to distinguish between classes (diabetic and non-diabetic in our case). Specifically, it quantifies the model's overall performance by calculating the area under the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at different decision thresholds (2). An AUC close to 1 indicates superior model performance, with better distinction between positive and negative classes.

$$AUC = \int_0^1 TPR(\text{FPR})d\text{FPR} \quad (2)$$

where, TPR is the true positive rate and FPR is the false positive rate.

- MCE: This metric assesses the model's prediction accuracy from a probabilistic perspective, providing insight into the confidence of its predictions (3). Lower MCE values indicate higher confidence and accuracy in the predicted probabilities.

$$MCE = -\frac{1}{N}\sum_{k=1}^{N}[y_i\ log(\hat{y}_i) + (1+y_i)\ log(1-\hat{y}_i)] \quad (3)$$

where, y_i is the actual label, $\hat{y}_i$ is the predicted probability for the i-th observation, and N is the total number of observations.

- The Confusion Matrix is an essential tool for calculating these metrics, enabling a detailed analysis of model performance by identifying not only successes but also the types of errors made, as shown in Fig. 3.

		Predicted Class	
		Positive	***Negative***
Actual Class	***Positive***	TP (The number of positive cases correctly identified)	FN (The number of positive cases incorrectly identified as negative)
	Negative	FP (The number of negative cases incorrectly identified as positive)	TN (The number of negative cases correctly identified)

Fig. 3. Confusion matrix for model performance evaluation.

We compared these metrics before and after optimization to assess the improvements made by the Stacking ensemble model and hyperparameter optimization, demonstrating the efficacy of our approach in enhancing diagnostic accuracy for early diabetes prediction.

IV. Results

Before optimization, our model demonstrated a precision of 0.70, a recall of 0.70, an AUC of 0.768, and a Mean Cross Entropy of 0.299. The initial confusion matrix indicated a balance between classes but hinted at potential for improvement, particularly in reducing false positives and negatives.

Significant improvement was observed after optimization: Precision increased to 0.73, enhancing the model's ability to correctly identify diabetes cases and thereby reduce the number of false positives. Recall also improved to 0.73, highlighting better detection of actual diabetes cases, crucial for early patient management. An AUC of 0.798 indicated a clearer distinction between diabetic and non-diabetic patients, showing increased model sensitivity and specificity. Moreover, the MCE improved to 0.271, reflecting higher confidence and accuracy in the model's probabilistic predictions.

The post-optimization confusion matrix revealed better class distinction, with a notable reduction in classification errors, essential for avoiding incorrect diagnoses and ensuring appropriate patient treatment.

Fig. 4 illustrates the comparison of model performance before and after optimization in terms of precision, recall, F1-Score, AUC, and MCE. As can be seen, each metric demonstrated significant improvement following optimization, underscoring the effectiveness of our approach.

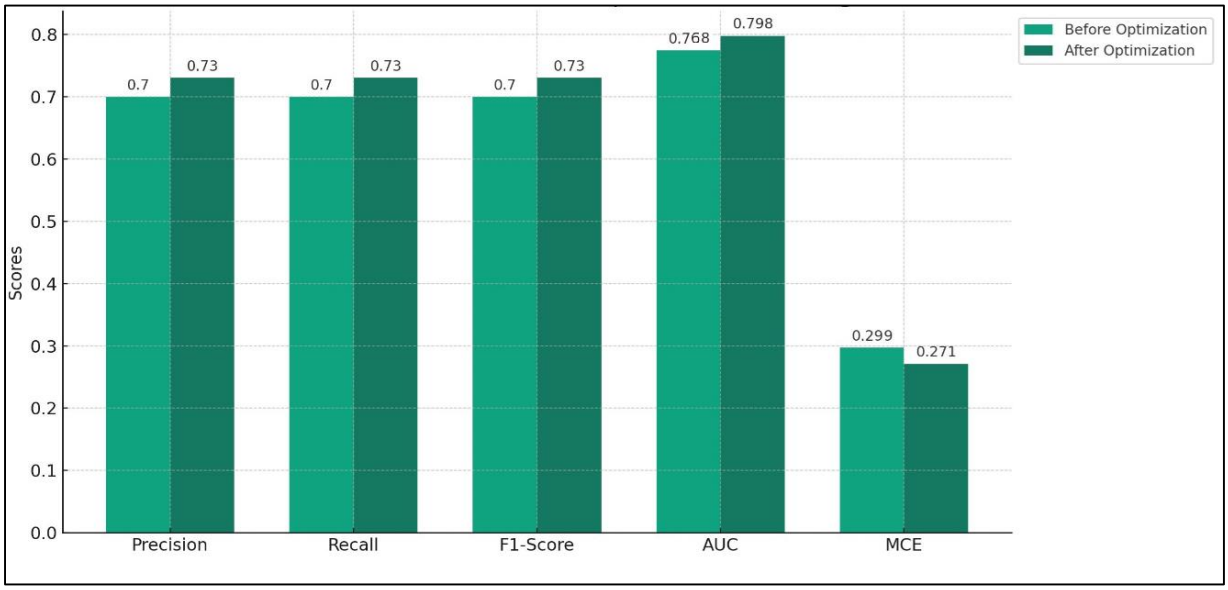

Fig. 4. Comparison of model performance before and after optimization.

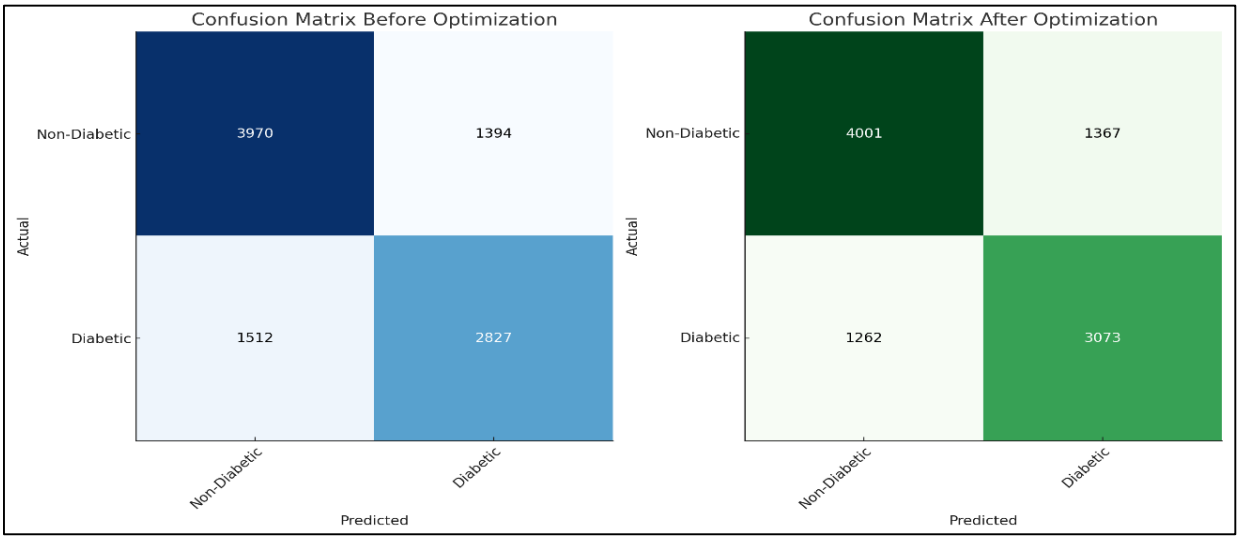

Fig. 5. Visualization of confusion matrices before and after optimization.

Fig. 5 provides a visualization of the confusion matrices before and after optimization to better understand the improvement in classification. The matrix on the left shows the initial distribution of predictions, while the one on the right demonstrates a better distinction between diabetic and non-diabetic classes after optimization. An increase in true negatives (from 3970 to 4001) and true positives (from 2827 to 3073) is observed, indicating an increased ability of the model to correctly identify diabetes cases, as well as a general improvement in precision and recall.

The comparison of performance metrics before and after optimization underscores the effectiveness of our methodological approach, demonstrating the enhancements brought about by integrating causal analysis techniques and hyperparameter optimization into the modeling process.

V. Discussion

The results obtained in this study highlight the effectiveness of integrating causal inference and machine learning (ML) in significantly improving diabetes diagnosis and management. Our model's enhanced ability to distinguish between diabetic and non-diabetic patients after optimization, as evidenced by improved performance metrics, suggests that this method can be particularly useful in clinical settings for anticipating high-risk cases and personalizing treatments.

In a comparative analysis with other studies that utilized the Health Facts database from Cerner Corporation, our model demonstrates superior performance metrics. For instance, a study [36] investigated the use of machine learning models and achieved an AUC of 0.686 for the Random Forest model, which is lower than the AUC of 0.798 achieved by our optimized model. Another study [37] focused on predictive modeling for diabetes using machine learning techniques. Their KNN model

achieved precision and recall values below 0.68, specifically reporting a precision of 0.68, recall of 0.61, and an F1-score of 0.64. These results further underscore the advancements achieved in our study, where post-optimization metrics for precision, recall, and F1-score all improved to 0.73. This improvement reflects our model's enhanced ability to correctly identify both diabetic and non-diabetic patients, reducing the number of false positives and negatives.

This research confirms that using advanced predictive models allows for better utilization of medical resources by targeting the most necessary interventions and potentially reducing the costs associated with late-stage complication treatments. It could also aid in the formulation of more effective, data-driven strategies for public health.

Although the results are promising, it is important to recognize certain limitations. The dependence of our model on the quality and diversity of the data is a major consideration. Our study relies on data from a single dataset, Health Facts from Cerner, which, although comprehensive, may not capture all clinical nuances present in a larger or global population. Additionally, the model could benefit from integrating additional variables not considered in this study, such as genetic data, certain biomedical markers, or lifestyle data, which could potentially improve the accuracy of the predictions.

For future research, several directions can be envisaged, including: (1) applying the model to other datasets to evaluate and enhance its robustness, (2) integrating additional variables that could influence diabetes diagnosis, such as genetic or environmental factors, and (3) interdisciplinary collaboration with experts in diabetology, epidemiology, and behavioral sciences to enrich the analysis and provide a more holistic understanding of diabetes dynamics. This approach represents a significant advancement in diabetes diagnosis and management, paving the way for more effective and personalized treatments for other chronic diseases where early detection and personalized treatment are paramount.

VI. Conclusion

This study explored the application of causal inference combined with advanced Machine Learning techniques to improve early diagnosis and management of diabetes, a growing global public health challenge. The results obtained not only demonstrate the viability of this approach but also its potential to significantly transform current clinical practices by providing more precise and effective tools for diabetes management.

Through a rigorous process of optimization and analysis, we significantly improved the model's performance, as evidenced by enhanced metrics. This improvement underscores the importance of understanding causal relationships not only to predict health events but also to positively influence clinical outcomes through targeted and personalized interventions.

By continuing to develop and refine this approach, we can hope to enhance care for diabetic patients while also offering proactive strategies for managing this complex disease and its multiple complications. The ultimate goal is to contribute to a more predictive, preventive, and personalized healthcare paradigm, where clinical decisions are informed by deep data insights and rigorous causal analyses.

Acknowledgment

Diabetes, World Health Organization, Geneva, 2023. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [April 5, 2023].

References

[1] R. J. Woodman and A. A. Mangoni, "A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future," Aging Clin. Exp. Res., vol. 35, pp. 2363-2397, 2023. doi: 10.1007/s40520-023-02552-2.

[2] K. Menon, "Different types of machine learning: Exploring AI's core," Simplilearn.com, 2023. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/types-of-machine-learning. [Accessed: Nov. 29, 2023].

[3] S. Echajei, Y. Chemlal, H. Ferjouchia, M. Rachik, N. E. Haraj, and A. Chadli, "Exploring the intersection of machine learning and causality in advanced diabetes management: New insight and opportunities," in Engineering applications of artificial intelligence, 1st ed., A. Chakir, J. F. Andry, A. Ullah, R. Bansal, and M. Ghazouani, Eds. Berlin: Springer, 2024, pp. 237-262. doi: 10.1007/978-3-031-50300-9_13.

[4] M. Prosperi et al., "Causal inference and counterfactual prediction in machine learning for actionable healthcare," Nat. Mach. Intell., vol. 2, pp. 369-375, Jul. 2020. doi: 10.1038/s42256-020-0197-y.

[5] V. Asvatourian, "Contributions of causal modeling in the evaluation of immunotherapies from observational data," Ph.D. dissertation, Université Paris-Sud, Université Paris-Saclay, Villejuif, France, 2018.

[6] S. Ramchand, M. Ploszajski, A. Virdi, D. Cole, and X. Xie, "Explainable machine learning to uncover distinct phenotypic signatures of hypertrophic cardiomyopathy and Fabry disease," Research Square, Jan. 2024. doi: 10.21203/rs.3.rs-3864850/v1.

[7] Shiksha Online, "XGBoost algorithm in machine learning," Shiksha, 2023. [Online]. Available: https://www.shiksha.com/online-courses/articles/xgboost-algorithm-in-machine-learning/. [Accessed: Aug. 9, 2023].

[8] M. Hiri, M. Chrayah, N. Ourdani, and N. Aknin, "Machine learning techniques for diabetes classification: A comparative study," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 9, 2023. doi: 10.14569/IJACSA.2023.0140982.

[9] J. P. Anderson et al., "Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: An application of machine learning using electronic health records," J. Diabetes Sci. Technol., vol. 10, no. 1, pp. 6-18, Dec. 2015. doi: 10.1177/1932296815620200. PMID: 26685993; PMCID: PMC4738229.

[10] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," J. Biomed. Inform., vol. 59, pp. 185-200, 2016. doi: 10.1016/j.jbi.2015.12.001.

[11] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," Comput. Methods Programs Biomed., vol. 104, no. 3, pp. 443-451, Dec. 2011. doi: 10.1016/j.cmpb.2011.03.018. Epub 2011 Apr. 30. PMID: 21531475.

[12] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: An application for diagnosis of diabetes," IEEE J. Biomed. Health Inform., vol. 19, no. 2, pp. 728-734, Mar. 2015. doi: 10.1109/JBHI.2014.2325615. Epub 2014 May 19. PMID: 24860043.

[13] A. A. Alzubaidi, S. M. Halawani, and M. Jarrah, "Towards a stacking ensemble model for predicting diabetes mellitus using combination of machine learning techniques," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 12, 2023. doi: 10.14569/IJACSA.2023.0141236.

[14] S. DuBrava, J. Mardekian, A. Sadosky, E. J. Bienen, B. Parsons, M. Hopps, and J. Markman, "Using Random Forest Models to Identify Correlates of a Diabetic Peripheral Neuropathy Diagnosis from Electronic Health Record Data," Pain Medicine, vol. 18, no. 1, pp. 107-115, mai 2016, doi: 10.1093/pm/pnw096.

[15] J. Pearl, "Causal inference in statistics: An overview," Statist. Surv., vol. 3, pp. 96-146, 2009. doi: 10.1214/09-SS057.

[16] M. A. Hernan and J. M. Robins, Causal Inference: What If. Chapman & Hall/CRC Monographs on Statistics & Applied Probab, CRC Press, 2024. ISBN: 9781420076165.

[17] D. Pe'er, "Bayesian network analysis of signaling networks: A primer," Science's STKE: Signal Transduction Knowledge Environment, vol. 2005, p. pl4, 2005. doi: 10.1126/stke.2812005pl4.

[18] B. Strack et al., "Impact of HBA1C measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," BioMed Research International, vol. 2014, pp. 1-11, Jan. 2014. [Online]. Available: https://doi.org/10.1155/2014/781670.

[19] K. Mahalakshmi and P. Sujatha, "The role of exploratory data analysis and pre-processing in the machine learning predictive model for heart disease," in 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8. doi: 10.1109/ACCAI58221.2023.10199714.

[20] O. Noy and R. Shamir, "Time-dependent iterative imputation for multivariate longitudinal clinical data," in The International Conference on Learning Representations (ICLR) 2023, Kigali, Rwanda, 2023.

[21] S. Zhao et al., "An outlier detection based two-stage EEG artifact removal method using empirical wavelet transform and canonical correlation analysis," Biomed. Signal Process. Control, vol. 92, p. 106022, 2024. doi: 10.1016/j.bspc.2024.106022.

[22] S. Nikhitha, S. Shivani, G. Harshavardhan, N. Sworup, and K. Nimrita, "Credit card scam detection using machine learning," in AIP Conference Proceedings, vol. 2742, no. 1, 2024.

[23] K. Naveeda and S. S. M. H. S. Fathima, "Real-time implementation of IoT enabled cyber attack detection system (IoT-E-CADS) in advanced metering infrastructure (AMI) using machine learning technique (MLT)," Research Square, Feb. 2024.

[24] 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, 2023. doi: 10.3390/diagnostics13142383.

[25] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," Multimedia Tools Appl., 2023. doi: 10.1007/s11042-023-16407-5.

[26] A. E. M. Eljialy, M. Y. Uddin, and S. Ahmad, "Novel framework for an intrusion detection system using multiple feature selection methods based on deep learning," Tsinghua Sci. Technol., vol. 29, no. 4, pp. 948-958, Aug. 2024. doi: 10.26599/TST.2023.9010032.

[27] A. Hudon, K. Phraxayavong, S. Potvin, and A. Dumais, "Ensemble methods to optimize automated text classification in avatar therapy," BioMedInformatics, vol. 4, no. 1, pp. 423-436, Feb. 2024. doi: 10.3390/biomedinformatics4010024.

[28] S. J. Rigatti, "Random forest," J. Insur. Med., vol. 47, no. 1, pp. 31-39, Jan. 2017. doi: 10.17849/insm-47-01-31-39.1.

[29] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models," Med. Biol. Eng. Comput., vol. 53, pp. 1305-1318, Mar. 2015.

[30] J. H. Huang, R. H. He, L. Z. Yi, H. L. Xie, D. Cao, and Y. Z. Liang, "Exploring the relationship between 5′AMP-activated protein kinase and markers related to type 2 diabetes mellitus," Talanta, vol. 110, pp. 1-7, Jun. 2013. doi: 10.1016/j.talanta.2013.03.039.

[31] D. Koller and N. Friedman, *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: The MIT Press, 2009.

[32] M. Elakel, "Application of Bayesian structure learning combined with domain knowledge in finding causal networks," M.S. thesis, Tilburg Univ., School of Humanities and Digital Sciences, Dept. of Cognitive Science & Artificial Intelligence, Tilburg, The Netherlands, Jul. 2021.

[33] M. S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, "Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data," Heliyon, vol. 10, no. 2, p. e24536, 2024. doi: 10.1016/j.heliyon.2024.e24536.

[34] R. Sivashankari, M. Sudha, M. K. Hasan, R. A. Saeed, S. A. Alsuhibany, and S. Abdel-Khalek, "An empirical model to predict the diabetic positive using stacked ensemble approach," Front. Public Health, vol. 9, p. 792124, 2021. doi: 10.3389/fpubh.2021.792124. PMCID: PMC8814448. PMID: 35127623.

[35] M. A. Almarzooqi, "Using ML to understand the factors impacting diabetes in diabetic patients," M.S. thesis, Rochester Inst. Technol., Rochester, NY, Sep. 2023.

[36] Y. Shang, K. Jiang, L. Wang, et al., "The 30-days hospital readmission risk in diabetic patients: predictive modeling with machine learning classifiers," *BMC Med. Inform. Decis. Mak.*, vol. 21, Suppl. 2, p. 57, 2021. doi: 10.1186/s12911-021-01423-y.

[37] H. Zouache and I. Bendib, "Classification du diabète avec l'algorithme KNN," Master's thesis, Dept. of Computer Science, Univ. of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria, 2021. Supervised by M. Belazzoug.