

Comparative Analysis of Machine Learning Models for Predicting Diabetes: Unveiling the Superiority of Advanced Ensemble Methods

Gregorius Airlangga^{1✉}

¹ Information System Study Program, Engineering Faculty, Atma Jaya Catholic University of Indonesia, Indonesia

Informasi Artikel

Riwayat Artikel

Received : April 15, 2024

Revised : April 21, 2024

Accepted : April 23, 2024

Keywords:

Extra Trees; LightGBM; DNN; Diabetes; Ensemble Learning

Kata Kunci:

Extra Trees; LightGBM; DNN; Diabetes; Ensemble Learning

ABSTRACT

This study presents a comprehensive analysis of various machine learning models to predict diabetes. The research evaluates and compares the predictive performance of advanced ensemble techniques Extra Trees Classifier and LightGBM with traditional machine learning algorithms and simpler deep neural network (DNN) architectures. The dataset comprises numerous features pertinent to diabetes diagnosis, such as glucose concentration, BMI, and insulin levels, among others. A methodology, including polynomial feature transformation and ten-fold cross-validation, was employed to ensure the study's reliability and the models' capability to generalize. The advanced ensemble models, Extra Trees and LightGBM, achieved stellar predictive metrics, with the former attaining a near-perfect ROC AUC, accuracy, precision, and an F1 score close to 1. LightGBM followed closely, demonstrating the high efficacy of ensemble methods in complex data settings. These results were contrasted with significantly lower performance metrics from DNNs and respectable, albeit lower, scores from traditional models like Decision Trees, Random Forest, KNN, and XGBoost.

ABSTRAK

Penelitian ini menyajikan analisis komprehensif dari berbagai model pembelajaran mesin untuk memprediksi penyakit diabetes. Penelitian ini mengevaluasi dan membandingkan kinerja prediktif dari teknik ensemble seperti Extra Trees Classifier dan LightGBM dengan algoritma pembelajaran mesin tradisional dan arsitektur jaringan saraf tiruan (DNN) yang sederhana. Dataset mencakup berbagai fitur yang relevan untuk diagnosis diabetes, seperti konsentrasi glukosa, BMI, dan tingkat insulin, di antara lainnya. Metodologi yang digunakan, termasuk transformasi fitur polinomial dan validasi silang sepuluh kali lipat, yang mana digunakan untuk memastikan keandalan studi dan kemampuan model untuk digeneralisasikan. Model ensemble: Extra Trees dan LightGBM, mencapai metrik prediktif ROC AUC, akurasi yang hampir sempurna, serta presisi, dan skor F1 yang mendekati 1. LightGBM juga menghasilkan nilai yang kompetitif, menunjukkan efikasi tinggi dari metode ensemble dalam pengaturan data yang kompleks. Hasil-hasil ini dikontraskan dengan metrik kinerja yang signifikan lebih rendah dari DNN, meskipun lebih rendah, dari model tradisional seperti Decision Trees, Random Forest, KNN, dan XGBoost.

Corresponding Author :

Gregorius Airlangga

Information System Study Program, Engineering Faculty, Atma Jaya Catholic University of Indonesia, Indonesia

Jl. Jend. Sudirman No.51 5, RT.004/RW.4, Karet Semanggi, Kecamatan Setiabudi, Kota Jakarta

Selatan, Daerah Khusus Ibukota Jakarta 12930

E-mail: gregorius.airlangga@atmajaya.ac.id

INTRODUCTION

Diabetes mellitus stands as a formidable challenge in global health, affecting millions worldwide with its chronic implications and serving as a precursor to a range of potentially fatal complications (Lin et al., 2023; Subramaniyan et al., 2023; Zeinelabdeen et al., 2024). The persistent increase in diabetes prevalence, as highlighted by the World Health Organization's reports, underscores the urgency for enhanced diagnostic strategies that can tackle the early detection and management of this disease (Francis et al., 2024; Jayaprabha & Priya, 2024; Motala et al., 2022). The prevalence of diabetes and its burden on health systems makes it imperative to leverage new technological advancements to improve outcomes and reduce the associated healthcare costs (Agarwal et al., 2022; Bode et al., 2021; Glennie et al., 2022). The traditional approach to diabetes prediction has predominantly revolved around basic statistical methods and classic machine learning techniques (Das et al., 2021; Modak & Jha, 2023; Suryanarayana et al., 2024). Initial models were largely dependent on logistic regression and decision trees, which, while foundational, often oversimplify the complex interdependencies of medical data variables (Buaroiu & Bâra, 2024; Carlucci et al., 2020; Gobi et al., 2023). With the advent of more sophisticated machine learning techniques, there has been a notable shift towards models that integrate multiple algorithms to improve prediction accuracy and model robustness (Chekroud et al., 2021).

A substantial body of research has been dedicated to exploring various ensemble techniques such as Random Forests, AdaBoost, and Gradient Boosting Machines (Alam et al., 2023; Li, T., 2024; Vamvakas et al., 2022). These methods have been recognized for their superior ability to mitigate overfitting, a common pitfall in predictive modelling by aggregating the predictions of several base estimators to improve generalizability over unseen data (Nguyen et al., 2023). However, the application of newer, more efficient ensemble methods like LightGBM and Extra Trees in diabetes prediction is still not extensively documented or explored in academic literature, particularly in direct comparison with traditional models (Afsaneh et al., 2022). The emergence of ensemble methods marked a significant advancement in predictive modeling, addressing several limitations of earlier approaches by enhancing both performance and predictive accuracy (Mahbub et al., 2023). Extra Trees, for instance, extends the Random Forest algorithm by altering the way splits are chosen in the construction of the trees (Livieris et al., 2021). This method introduces additional randomness into the model, which can help in discovering more robust patterns in the data, potentially leading to better generalization on new, unseen datasets (González et al., 2020).

Similarly, LightGBM, a gradient boosting framework that uses tree-based learning algorithms, has gained popularity due to its efficiency and effectiveness, especially in large datasets (Gamel et al., 2024). Its novel approach to handling large data sets through gradient-based one-sided sampling and exclusive feature bundling allows it to perform with high efficiency, making it particularly suitable for medical datasets, which often contain a vast array of parameters and records (Shehadeh et al., 2021). This research aims to delve deeper into the capabilities of LightGBM and Extra Trees classifiers, examining their efficacy in handling complex, nonlinear relationships inherent in diabetes-related data. The primary objective is to evaluate whether these advanced ensemble techniques can outperform traditional models in predicting diabetes outcomes, thus providing a more accurate, reliable, and timely diagnosis. Despite the advancements in ensemble methods, there is a notable research gap in their direct comparison across diverse medical datasets, particularly those related to chronic diseases like diabetes (Li, M., 2023). Most comparative studies have focused on either non-medical datasets or have not fully exploited the potential of these advanced algorithms in a structured, comparative manner (Ahmad et al., 2021; Alinsaif, 2024; Ge et al., 2023).

This study seeks to fill the existing research gaps by providing a thorough comparative analysis of the performance of advanced ensemble models against traditional predictive models on a standardized diabetes dataset. By focusing on key performance metrics such as accuracy, precision, recall, and F1 scores, this research will offer significant insights into the strengths and

limitations of each model, thereby contributing to the broader field of medical predictive analytics. The article is structured to first outline the methodologies used in preparing the dataset, including the implementation of polynomial feature transformations to enrich the model's input features. Following this, the configuration and optimization of both the LightGBM and Extra Trees models are discussed. The subsequent section presents a detailed analysis of the results obtained from these models, comparing them against benchmarks set by traditional predictive models. The discussion section then interprets these results in the context of their practical implications for diabetes prediction, followed by a conclusion that summarizes the findings and outlines potential directions for future research.

RESEARCH METHODS

This research adopts a quantitative approach to evaluate the efficacy of advanced ensemble machine learning models in predicting the occurrence of diabetes. The study utilizes a well-documented Diabetes Dataset that comprises demographic, medical history, and clinical measurement data for individuals diagnosed with diabetes. The dataset is representative of a diverse population, encompassing various age groups, gender, and ethnic backgrounds, making it ideal for developing and testing predictive models that could be generalized to a broader population. The summary of our research activity is presented in the figure 1.

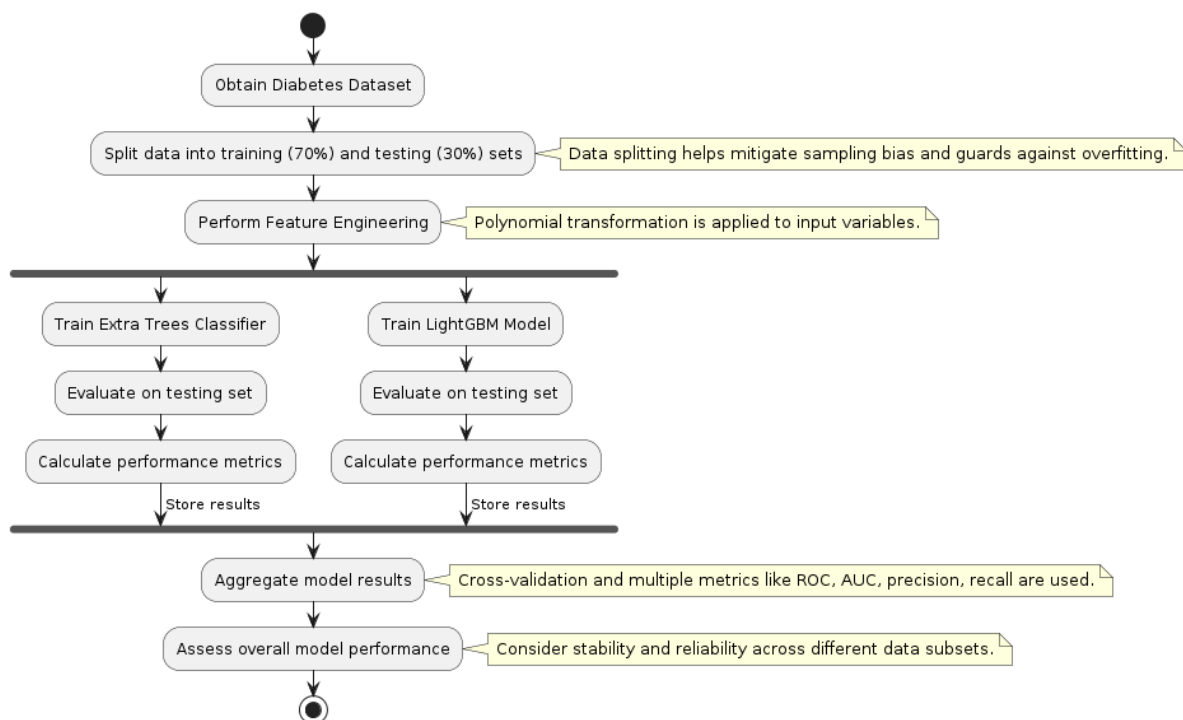


Figure 1. Activity diagram of Research Methods

Dataset

The dataset used in this study encompasses comprehensive entries for several hundred individuals, capturing a broad demographic and medical spectrum typical of the general population at risk for diabetes. Each record includes critical information across various features, such as the number of pregnancies, fasting plasma glucose levels, blood pressure measurements, skin thickness, serum insulin levels, body mass index (BMI), a synthetic estimate known as the diabetes pedigree function, and age. These features are scientifically validated markers associated with the diagnosis of diabetes. The outcome variable is binary, delineating the presence (1) or absence (0) of diabetes, which facilitates a clear clinical endpoint for the classification task (Aboelnaga, 2023).

To support the robustness and validity of the predictive models developed in this study, the dataset underwent a rigorous data splitting process. The data was randomly divided into a training set, which comprises 70% of the total data, and a testing set, accounting for the remaining 30%. This random stratification is crucial as it helps mitigate sampling bias, ensuring that both subsets are representative of the overall population. This partitioning serves multiple essential purposes. Firstly, the training set provides a substantial and rich source of examples from which the models can learn. This segment of the dataset allows the models to detect and internalize complex patterns and relationships among the features without being tailored to peculiarities that might only appear in the testing set. Secondly, the testing set, isolated from the model during the training phase, acts as a new and unseen data source for evaluating the model's performance. This separation is vital for assessing the model's ability to generalize its learned patterns to new data, thereby providing a realistic measure of how the model would perform in practical, real-world scenarios. Moreover, the separation of data into training and testing sets safeguards against overfitting. Overfitting occurs when a model is excessively complex, capturing noise rather than representing the underlying distribution of data. An independent testing set is crucial for detecting this phenomenon, as an overfitted model would perform well on the training data but poorly on unobserved data.

Feature Engineering: Enhancing Predictive Capacity

The process of feature engineering is pivotal in molding raw data into a format that is more suitable for modeling, especially for a complex medical condition like diabetes which involves multifaceted biological interactions. To address this, polynomial feature transformation was employed on the dataset. This technique transforms the input variables into polynomial features up to the second degree. By including not only the original features but also their square terms and interaction terms, this method allows the predictive models to capture both the linear and the more intricate non-linear relationships among the variables. For example, the interaction between glucose levels and BMI could be crucial for predicting diabetes, a relationship that might not be as effectively captured by linear models. The introduction of these polynomial features aims to enrich the feature set, thereby enhancing the potential predictive power of the models. This step is fundamental in providing a deeper insight into the complex patterns that underlie the development and progression of diabetes.

In implementing the predictive models, two advanced ensemble techniques were selected based on their robust performance in handling large datasets and their demonstrated efficacy in various machine learning challenges. The first model, the Extra Trees Classifier, builds on the foundation of conventional Random Forest models but introduces additional randomness in the way splits are made within the trees. This variation stems from using random thresholds for each feature rather than searching for the best possible thresholds (as is typical in Random Forests). The introduction of this randomness helps in creating an ensemble of uncorrelated trees, which often results in a more accurate and stable prediction aggregate. The second model, LightGBM, stands out due to its gradient-based learning algorithm optimized for speed and efficiency. LightGBM accelerates the training process without sacrificing accuracy by using histogram-based algorithms, which bucket continuous feature values into discrete bins. This approach reduces the computational overhead significantly, allowing the model to handle large volumes of data swiftly. Moreover, LightGBM improves on traditional gradient boosting methods by focusing on leaf-wise growth rather than level-wise growth, which often leads to better performance on complex datasets. Together, these models represent the cutting edge in ensemble machine learning techniques, chosen specifically for their ability to process large and complex datasets efficiently while maintaining high levels of accuracy. By integrating these advanced methods with enriched feature sets from polynomial transformations, this study leverages both innovative data preparation techniques and the latest advancements in machine learning technology. This combination is tailored to explore the nuances of diabetes prediction, aiming to provide a model that is not only highly predictive but also capable of interpreting the intricate biological and environmental interactions that contribute to the onset of diabetes.

Assessing model performance

In the realm of predictive modeling, particularly for a condition as complex and variable as diabetes, it is imperative to thoroughly evaluate the performance of the models. To achieve this, a series of metrics were employed, each contributing uniquely to the understanding of the model's effectiveness. The Receiver Operating Characteristic (ROC) curve and the corresponding Area Under the ROC Curve (AUC) provide a powerful visual and numerical measure of the model's ability to discriminate between the two classes—those with and without diabetes. A higher AUC indicates better model performance, with a value of 1.0 representing perfect discrimination and 0.5 denoting no discriminative power. Further, accuracy, precision, recall, and F1 score were also calculated. Accuracy measures the proportion of true results (both true positives and true negatives) in the population, providing an overall effectiveness rate of the model. Precision focuses specifically on the proportion of positive identifications that were actually correct, a crucial metric in medical diagnostics where false positives can lead to unnecessary treatment or anxiety. Recall, or sensitivity, measures the model's ability to capture all available positive cases, which is vital for conditions like diabetes where early detection can significantly alter patient outcomes. The F1 score, a harmonic mean of precision and recall, balances these metrics, providing a single score that gauges the model's reliability in identifying positive cases without an overwhelming number of false positives.

To ensure that the models are both stable and reliable, cross-validation techniques were implemented. The dataset was subjected to ten-fold cross-validation, a robust method that involves partitioning the data into ten equally sized subsets. Each subset serves as a test set at some point, while the remaining nine subsets are used for training. This iterative process is repeated ten times, with each of the ten subsets used exactly once as the test data. Such a method significantly reduces the variability in model assessment and guards against overfitting—ensuring the model tested has not just memorized the data, but has learned to generalize from it. The deliberate selection of advanced ensemble techniques alongside sophisticated feature engineering is crucial for navigating the complexities inherent in diabetes prediction. The incorporation of polynomial features enhances the models' capacity to decipher complex interactions between variables—interactions that simpler models might miss, but which are often key to understanding disease progression. Furthermore, the robust validation technique of cross-validation ensures that the findings are not artifacts of the particular sample of data used but are instead reproducible and reliable across different data subsets. This methodological rigor is vital for ensuring that the predictive models developed can be effectively applied in real-world settings, providing reliable, actionable insights into diabetes risk.

RESULTS AND DISCUSSION

The comparative analysis of machine learning models presented in the table 1 provides a revealing insight into their performance metrics in predicting diabetes. The table highlights the scores across five critical metrics: Receiver Operating Characteristic Area Under the Curve (ROC AUC), Accuracy, Precision, Recall, and F1-Score. The Extra Trees Classifier demonstrates exceptional performance with a perfect ROC AUC of 1.0, illustrating its impeccable ability to discriminate between the classes. With a near-perfect accuracy of 99.76% and precision of 99.37%, the model shows a superior ability to correctly identify cases of diabetes and a minimal rate of false positives. The recall is also at 100%, indicating that the model successfully identifies all positive instances, which, combined with its high precision, yields a remarkable F1-Score of 99.68%. LightGBM shows similarly impressive results with a ROC AUC of 99.88%. Its accuracy stands at 99.69%, with precision slightly lower than Extra Trees at 99.15%, yet still remarkably high. LightGBM also achieves a perfect recall score and an F1-Score of 99.57%, substantiating its efficacy in the detection and prediction of diabetes within the study population.

CNN and RNN, representing deep learning approaches, fall short in comparison to the advanced ensemble models. The CNN's ROC AUC of 56.14% and RNN's 47.99% suggest a limited ability to differentiate between patients with and without diabetes. The accuracy rates

hover around the 50% mark, which is only marginally better than random guessing. Both models show a moderate recall rate, with RNN particularly high at 87.38%, which indicates a tendency to identify a high number of true positive rates but at the cost of precision, leading to lower overall accuracy. Decision Trees and Random Forest provide moderate results, with ROC AUCs of 68% and 84% respectively. These models offer a balanced trade-off between false positives and false negatives, as seen in their precision and recall scores. However, their accuracy and F1 scores indicate there is room for improvement in model performance. KNN shows limited effectiveness with a ROC AUC of 64%, accuracy of 72%, and precision and recall both at 66%. These scores suggest that while KNN is somewhat capable of identifying diabetic cases, it does so inconsistently and with a considerable error rate. XGBoost stands out among the traditional models with a higher ROC AUC of 87%, accuracy of 85%, and a balanced precision and recall of 81% and 82% respectively, culminating in an F1 score of 82%. These results show XGBoost's competency in predicting diabetes, though it still falls short of the advanced ensemble methods.

The advanced ensemble models, Extra Trees and LightGBM, set a high benchmark in predictive performance. Their high precision and recall scores indicate they can be trusted to minimize both false positives and false negatives—essential qualities for a predictive tool in a clinical setting. The significant gap in performance between these models and traditional algorithms underscores the value of choosing advanced machine learning techniques for complex health-related datasets. The disparity in performance between ensemble methods and deep learning models (CNN and RNN) can be attributed to the nature of the data and the models' characteristics. Deep learning models often require vast amounts of data to perform at their best, which may not have been available in this study. On the other hand, ensemble methods are known for their ability to handle smaller, more complex datasets efficiently, making them better suited for this particular task. The moderate performance of traditional machine learning models suggests they may be more suitable for datasets that are less complex or when interpretability is a priority. However, for critical healthcare applications where the cost of misdiagnosis is high, the advanced ensemble methods prove to be the most reliable.

Tabel 1. Comparison of Machine Learning Performance

Model	ROC AUC	Accuracy	Precision	Recall	F1-Score
Extra Trees	1.0	0.9976	0.9937	1.0	0.9968
LightGBM	0.9988	0.9969	0.9915	1.0	0.9957
CNN	0.5614	0.53	0.5338	0.6893	0.6017
RNN	0.4799	0.49	0.5046	0.5340	0.5189
Decision Tree	0.68	0.76	0.71	0.67	0.69
Random Forest	0.84	0.83	0.80	0.80	0.80
KNN	0.64	0.72	0.66	0.66	0.66
XGBoost	0.87	0.85	0.81	0.82	0.82

CONCLUSION

The exploration of various machine learning models in predicting diabetes has culminated in several key insights. Advanced ensemble models, particularly the Extra Trees Classifier and LightGBM, have demonstrated superior performance, substantiating their potential as robust predictive tools in the healthcare domain. These models exhibited remarkable accuracy, precision, recall, and F1 scores, indicating their efficacy in correctly identifying diabetic cases with high reliability and minimal error.

The study's comparison revealed a notable divergence in the effectiveness of different modeling approaches. While traditional machine learning algorithms such as Decision Trees, Random Forest, KNN, and XGBoost offered varying degrees of success, they were outperformed by the advanced ensemble methods. Deep learning models like CNN and RNN, despite their potential in many applications, did not match the ensemble techniques' performance, potentially

due to the complex nature of the dataset and the requirement of extensive data to learn effectively. This investigation highlights the critical role of model selection in predictive analytics for medical applications. In situations where accuracy is paramount, and the cost of misdiagnosis is high, advanced ensemble methods stand out as the preferred choice. They provide a compelling balance between high predictive power and the ability to manage complex datasets, making them highly suitable for tasks such as diabetes prediction.

In light of the findings, the healthcare sector can look toward advanced machine learning techniques as a means to enhance diagnostic processes and patient care. These methods have the potential to facilitate early detection and intervention, which are crucial in managing diabetes effectively. The integration of these models into clinical practice could lead to more personalized and timely treatment strategies, ultimately improving health outcomes for individuals with diabetes. As the field of machine learning continues to evolve, it will be important to replicate this research with newer models and larger, more diverse datasets. Continuous refinement and testing across various populations will help in fine-tuning these models to cater to the nuances of diabetes manifestations across different demographics.

ACKNOWLEDGEMENTS

We express our deepest gratitude to Atma Jaya Catholic University of Indonesia for the unwavering support and resources provided throughout the duration of this research project. Our sincere thanks also extend to the Information System Study Program, whose rigorous academic environment and collaborative ethos have been instrumental in facilitating our work.

REFERENCES

- Aboelnaga, E. (2023). Diabetes Dataset. Kaggle. <https://www.kaggle.com/datasets/ehababoelnaga/diabetes-dataset>
- Afsaneh, E., Sharifdini, A., Ghazzaghi, H. & Ghobadi, M. Z. (2022). Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review. *Diabetology & Metabolic Syndrome*, 14(1), 196.
- Agarwal, S., Simmonds, I. & Myers, A. K. (2022). The use of diabetes technology to address inequity in health outcomes: limitations and opportunities. *Current Diabetes Reports*, 22(7), 275–281.
- Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M. & Alhumam, A. (2021). Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Applied Sciences*, 11(3), 1173.
- Alam, M. S., Al-Ismail, F. S., Hossain, M. S. & Rahman, S. M. (2023). Ensemble machine-learning models for accurate prediction of solar irradiation in Bangladesh. *Processes*, 11(3), 908.
- Alinsaif, S. (2024). COVID-19 Image Classification: A Comparative Performance Analysis of Hand-Crafted vs. Deep Features. *Computation*, 12(4), 66.
- Buaroiu, A.-C. & Bâra, A. (2024). A Descriptive-Predictive-Prescriptive Framework for the Social-Media--Cryptocurrencies Relationship. *Electronics*, 13(7), 1277.
- Bode, B., King, A., Russell-Jones, D. & Billings, L. K. (2021). Leveraging advances in diabetes technologies in primary care: a narrative review. *Annals of Medicine*, 53(1), 805–816.
- Carlucci, S., De Simone, M., Firth, S. K., Kjærgaard, M. B., Markovic, R., Rahaman, M. S., Annaqeeb, M. K., Biandrate, S., Das, A., Dziedzic, J. W. & others. (2020). Modeling occupant behavior in buildings. *Building and Environment*, 174, 106768.
- Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z.,

- Belgrave, D., DeRubeis, R., Iniesta, R. & others. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20(2), 154–170.
- Das, S. K., Roy, P. & Mishra, A. K. (2021). Deep learning techniques dealing with diabetes mellitus: a comprehensive study. *Health Informatics: A Computational Perspective in Healthcare*, 295–323.
- Francis, A., Harhay, M. N., Ong, A. C. M., Tummalapalli, S. L., Ortiz, A., Fogo, A. B., Fliser, D., Roy-Chaudhury, P., Fontana, M., Nangaku, M. & others. (2024). Chronic kidney disease and the global public health agenda: an international consensus. *Nature Reviews Nephrology*, 1–13.
- Gamel, S. A., Ghoneim, S. S. M. & Sultan, Y. A. (2024). Improving the accuracy of diagnostic predictions for power transformers by employing a hybrid approach combining SMOTE and DNN. *Computers and Electrical Engineering*, 117, 109232.
- Ge, Y., Guo, Y., Das, S., Al-Garadi, M. A. & Sarker, A. (2023). Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, 104458.
- Glennie, J. L., Berard, L. & Levrat-Guillen, F. (2022). Sensor-based technology: bringing value to people with diabetes and the healthcare system in an evolving world. *ClinicoEconomics and Outcomes Research*, 75–90.
- Gobi, N., Singh, B. P. & Yadav, S. (2023). Machine Learning and Data Mining for Predictive Modeling of Stroke Risk and Post-Stroke Care Management. *2023 IEEE International Conference on ICT in Business Industry \& Government (ICTBIG)*, 1–6.
- González, S., Garc\'ia, S., Del Ser, J., Rokach, L. & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64, 205–237.
- Jayaprabha, M. S. & Priya, V. V. (2024). Early Prediction of Chronic Kidney Disease: A Comprehensive Survey. *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, 45–51.
- Li, M., Jiang, Y., Zhang, Y. & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, 11, 1273253.
- Li, T., Xia, Q., Ouyang, Y., Zeng, R., Liu, Q. & Li, T. (2024). Prospectivity and Uncertainty Analysis of Tungsten Polymetallogenic Mineral Resources in the Nanling Metallogenic Belt, South China: A Comparative Study of AdaBoost, GBDT, and XgBoost Algorithms. *Natural Resources Research*, 1–23.
- Lin, E.-C., Chiang, Y.-C., Lin, H.-Y., Tseng, S.-Y., Hsieh, Y.-T., Shieh, J.-A., Huang, Y.-H., Tsai, H.-T., Feng, S.-W., Peng, T.-Y. & others. (2023). Unraveling the link between periodontitis and coronavirus disease 2019: exploring pathogenic pathways and clinical implications. *Biomedicines*, 11(10), 2789.
- Livieris, I. E., Iliadis, L. & Pintelas, P. (2021). On ensemble techniques of weight-constrained neural networks. *Evolving Systems*, 12, 155–167.
- Mahbub, M. F., Shuvo, W. O. K. & Momen, S. (2023). Prediction of Glycemic Control in Diabetes Mellitus Patients Using Machine Learning. In *Proceedings of the Computational Methods in Systems and Software* (pp. 216–242). Springer.
- Modak, S. K. S. & Jha, V. K. (2023). Diabetes prediction model using machine learning techniques. *Multimedia Tools and Applications*, 1–27.
- Motala, A. A., Mbanya, J. C., Ramaiya, K., Pirie, F. J. & Ekoru, K. (2022). Type 2 diabetes mellitus in sub-Saharan Africa: challenges and opportunities. *Nature Reviews*

Endocrinology, 18(4), 219–229.

- Nguyen, M. H., Trinh, S. H., Ly, H.-B. & others. (2023). Toward improved prediction of recycled brick aggregate concrete compressive strength by designing ensemble machine learning models. *Construction and Building Materials*, 369, 130613.
- Shehadeh, A., Alshboul, O., Al Mamlook, R. E. & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. *Automation in Construction*, 129, 103827.
- Subramaniyan, V., Lubau, N. S. A., Mukerjee, N. & Kumarasamy, V. (2023). Alcohol-induced liver injury in signalling pathways and curcumin's therapeutic potential. *Toxicology Reports*.
- Suryanarayana, G., Kumar, G. B., Anudeep, G., Sai, G. V. & others. (2024). A Comprehensive Survey on Diabetes Type-2 (T2D) Forecast Using Machine Learning. *Archives of Computational Methods in Engineering*, 1–19.
- Vamvakas, A., Tsivaka, D., Logothetis, A., Vassiou, K. & Tsougos, I. (2022). Breast cancer classification on multiparametric MRI--increased performance of boosting ensemble methods. *Technology in Cancer Research & Treatment*, 21, 15330338221087828.
- Zeinelabdeen, Y., Abaza, T., Yasser, M. B., Elemam, N. M. & Youness, R. A. (2024). MIAT LncRNA: A multifunctional key player in non-oncological pathological conditions. *Non-Coding RNA Research*, 9(2), 447–462.