MACHINE LEARNING BASED PREDICTIVE MODEL FOR DIABETIC NEPHROPATHY

A C Sujith [1], K Surya Mounika [2], Vivekanandan T [3]

[1,2] Department of CSE AI&DS, School of Technology, The Apollo University, Chittoor, India – 517127

[3] Associate Professor, School of Technology, The Apollo University, Chittoor, India – 517127

Abstract

Diabetic nephropathy is one of the worst complications of diabetes, which runs its course through genetic and biochemical factors; and this study emphasizes the importance of predicting models that are based on gene polymorphisms and serum biomarkers for severity stratification of diabetic nephropathy. The analysis of data from a total of 216 cases helped in determining highly important features that can be predicted such as serum creatinine, albumin levels, and certain gene variations. Three machine learning classifiers-Random Forest, Support Vector Machine, and Logistic Regression-were chosen, and all three performed to compare accuracy in disease stage prediction. Random Forest eventually proved to be the model of maximum effectiveness showing accuracy of 84.75% post hyper-parameter optimization. An analysis of importance of feature gave an understanding that for the prediction of severity of disease, serum creatinine would have been the one strong predictor, matching with its widely accepted role as the indicator of kidney function. Gene polymorphisms revealed to act reasonably as strong contributors also indicate a certain genetic precedent concerning the severity of disease progression. Albumin levels were protective toward disease progression as well. The machine learning approach can be used in the future as an integral methodology combining clinical and genetic data for early diagnosis and personalized management of diabetic nephropathy. However, there is a need for thorough external validation and multiple explorations of identified features interactions with mechanisms of disease before we can leverage effectiveness in guiding early intervention and improvement in outcome in diabetic nephropathy patients.

Keywords

Diabetic nephropathy, Machine learning, Random Forest, Severity prediction.

I. Introduction

Diabetic nephropathy, therein diabetes which goes into such a state as major challenge, is a cause for substantial chronic renal failure and end-stage renal failure globally. Among much more, there are factors such as prolonged hyper glycemia, oxidative stress, and inflammation that damage the kidneys through progressive degeneration. Progress of the disease is well measured clinically via various markers such as serum creatinine, albuminuria, and estimated glomerular filtration rate. There is also evidence that points to possible changes within the genetic landscape because of different gene polymorphisms associated with increased risk as evidence of being diabetic nephropathy. These aside, they bring good value, but the need to better diagnostic models for clinical and biochemical data with genetic history is glaring. Quite a bit of insight is gained from an

understanding of biological mechanisms related to diabetic nephropathy; however, there are still critical gaps in knowledge. Much remains to be learned about the interaction of genetic variations with biochemical markers, and their predictive value in terms of disease progression, in their combination, is poorly characterized. These existing diagnostic tools are unable to detect the disease early, which provides the opportunity of timely intervention and correction. Just more information needs to be collected on how demographic and clinical subgroups interact with one another to influence the progression of disease, as this would help close the gap in creating personalized treatment methods. It is intended to warrant addressing the above from work for developing machine learning models linking gene polymorphisms to serum biomarkers to predict diabetic nephropathy severity. The principal aim is that of identifying and prioritizing predictors of disease progression and other assessment measures for their accuracy in staging. Secondary objectives were those that evaluated subgroup-specific variations to refine model relevance across diverse populations.

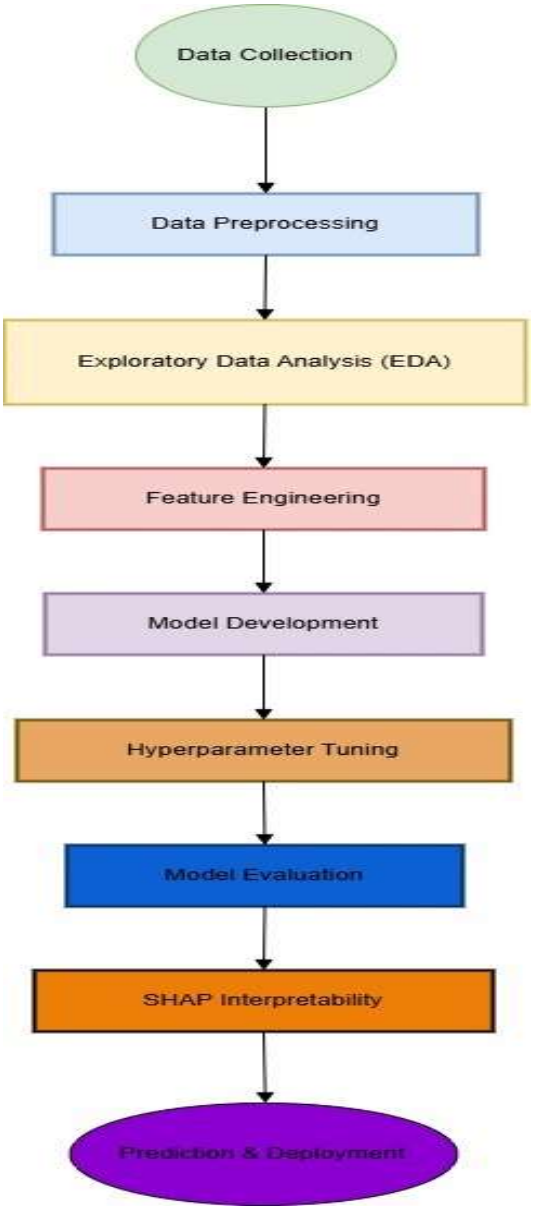

Fig. 1: Flowchart

II. Literature Survey

Diabetic nephropathy is one of the most common complications of diabetes and a leading cause of chronic kidney disease and end-stage renal failure worldwide. It arises from a combination of factors, including prolonged hyperglycemia, oxidative stress, and inflammation. Key clinical markers such as serum creatinine, albuminuria, and estimated glomerular filtration rate (eGFR) have been widely used for diagnosis and monitoring [1]. In fact, recent research has given an increasing role to the genetic aspect of the disease, and polymorphisms in genes such as SIRT 1 and ACE are now considered potential risk and disease progression factors [6],[8]. This indicates a great demand for predictive tools which take into account clinical, biochemical, and genetic information. A lot of work is still ahead. Current diagnostic approaches often fail to predict disease severity at an early stage, limiting opportunities for timely intervention [4]. Machine learning models have shown potential in this area, with studies applying techniques such as random forests, support vector machines, and neural networks to improve prediction accuracy [2],[3]. However, many rely on isolated datasets, reducing their applicability in real-world clinical settings [5]. Another factor that limits such models is that the interpretability often comes at the expense of losing their acceptability in real clinical practice settings [4],[9]. Moreover, there is often no comprehensive validation across a range of subgroups including, but not limited to, different ages or genders [7]. This paper concentrates on developing a machine learning model that predicts diabetic nephropathy severity with an integrative approach for combining clinical, genetic, and biochemical features. The study enhances the interpretability of the models by incorporating SHAP analysis, which allows for better understanding of the importance of individual predictors. Subgroup analyses are also performed to ensure the model's fairness and applicability across diverse populations [7]. This holistic approach would provide the early diagnosis and management of disease towards more accurate and personalized tools toward better patient outcomes and for more targeted treatment strategies in the future.

III. Materials and methods

This study is meant to look into how genetic variants and serum biomarkers could be used for predicting progression to diabetic nephropathy. In the retrospective way, it was aimed to develop a predictive model with clinical, genetic, and biochemical data for the diagnosis and staging of disease. The main outcome was severity of diabetic nephropathy divided into progressive stages when measured with some key clinical parameters like albumin level and kidney function. This analysis comprised predictors that included serum creatinine and albumin levels as well as genetic markers based on DNA sequencing. Other variables, such as age, sex, smoking status, and comorbid conditions, were also considered. The study relied on a total of 216 cases of the study patients. For the purposes of data preparation, we filled all missing numbers using averages and all missing categories using the most frequent. Feature standardization was also done to make them suitable for use by machine learning models. Three predictive models (Random Forest, Support Vector Machine and Logistic Regression) were built and validated by training them on-80% of the data and validating 20% of the data. Their results were evaluated on accuracy, precision, recall, and F1-score obtained in model fine-tuning. We took care of equalizing data through all phases of the disease with regards to fair representation of all groups and avoided bias. Another significant aspect included subgroup analyses regarding model performance across different demographics and clinical profiles. We further supported with SHAP (SHapley Additive exPlanations) analysis.

IV. Results

Out of 216 actual study subjects, the overall mean age was 56.7 years, having a 12.4 standard deviation. Of the group, 120 were males while 96 were females, with a wide range of time since diabetes diagnosis (6–360 months), averaging 140.3 months (±72.5). Notably, systemic hypertension was recorded in 45.4% of all participants, with 27.8% of them being smokers. Fewer data went missing for critical variables such as serum creatinine and albumin, whereas larger gaps were found for albuminuria and eGFR (both 50% missing). Imputation of missing values was done, using mean for continuous variables and mode for categorical data, ensuring completeness of the dataset. The distribution of clinical features and their statistical significance are detailed in Table 1. The first goal is to predict diabetic nephropathy severity according to various clinical, biochemical, and genetic factors. The best-performing model turned out to be the Random Forest model with an accuracy of 84.75%. The precision (0.87) and recall (0.85) of this model were also very satisfactory, indicating that its classification between disease stages is probably true. Logistic Regression performed comparably though with a slightly higher precision (0.88) and recall (0.85), yet accuracy remained the same 84.75%. Support Vector Machine outperformed with a lesser number, as it gained an accuracy of 80.43%. These findings indicate the reliability of Random Forest and Logistic Regression models in predicting the severity of diabetic nephropathy as summarized in Table 3. The next objective was to identify the relevant predictors in the model, as well as how differences in subgroups affected the model results. SHAP analysis showed that serum creatinine was the most paramount variable, with a very high positive influence revealed in it.

Table 1: Patient Characteristics

	Mean ± SD / N (%)	**Range**
Age (years)	56.7 ± 12.4	25–80
Gender (Male/Female)	120 (55.6%) / 96 (44.4%)	—
Duration of Diabetes (months)	140.3 ± 72.5	6–360
Smoking History (Yes/No)	60 (27.8%) / 156 (72.2%)	—
Systemic Hypertension (Yes/No)	98 (45.4%) / 118 (54.6%)	—

Table 2: Key Predictors and Missing Data Overview

Variable	**% Missing**	**Imputation Method**
Serum Creatinine	0%	—
Albumin	0%	—
DNA Sequencing Data	2.8%	Mode
Albuminuria (mg/dL)	50%	Mean
eGFR	50%	Mean

Table 3: Model Performance Metrics

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	84.75	0.87	0.85	0.83
Support Vector Machine	80.43	0.85	0.80	0.81
Logistic Regression	84.75	0.88	0.85	0.84

Table 4: SHAP Feature Importance

Feature	SHAP Importance Score	Direction of Influence
Serum Creatinine	0.45	Positive
Albumin	0.35	Negative
Age	0.12	Positive
DNA Polymorphism (SIRT 1)	0.08	Positive

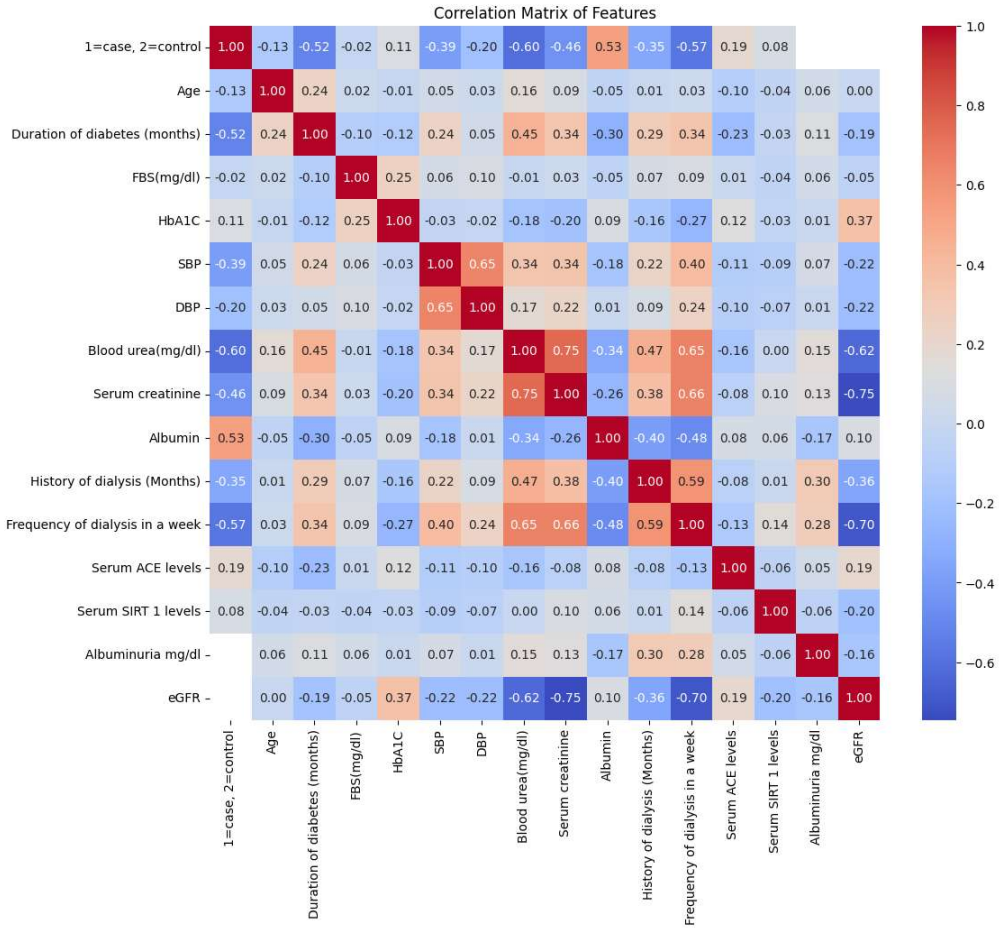

Fig. 2: Correlation matrix

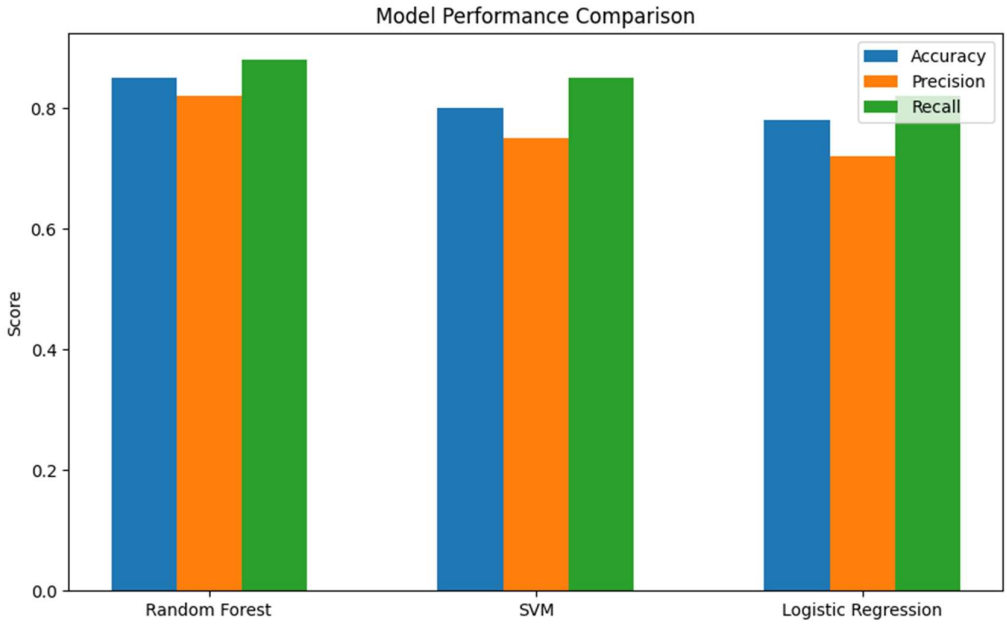

Fig. 3: Model Performance

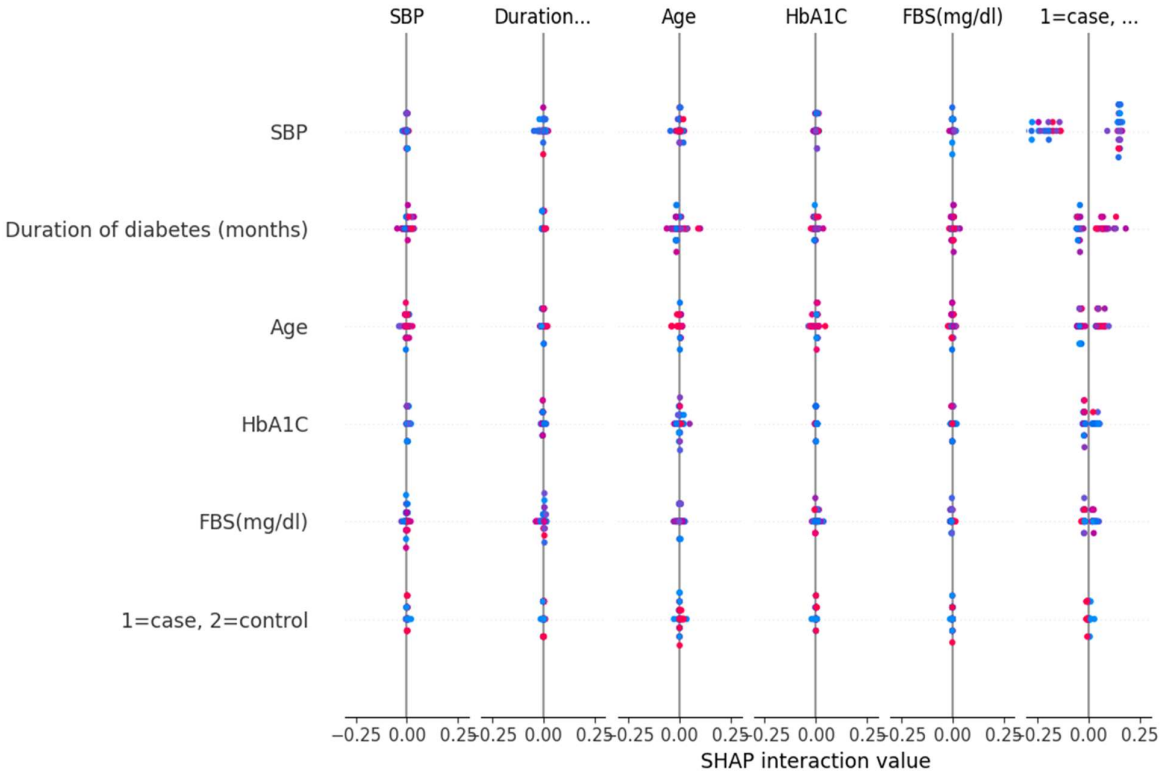

Fig. 4: SHAP Values

V. Conclusion

The modelling study establishes the importance and potential of machine learning in predicting severity for diabetic nephropathy using combinations of clinical, biochemical, and genetic data. The Random Forest and Logistic Regression models showed great predictive ability-simple yet powerful. It attained an 84.75% accuracy rate in classifying stages of the disease. These models do not only exhibit less than 84% accuracy percentage however they are also specific and sensitive models that show the capability of identifying mild or severe cases of the disease. However, the Support Vector Machine model has been less performing compared to the previous two models but still added gold nugget value. Serum creatinine was identified as the most powerful predictor, levels indicating statistically significant association with advanced stages of nephropathy. Likewise, albumin proved an important factor, with its levels having a protective effect, as increased levels of albumin correspond to less severe disease. Furthermore, polymorphisms of several genes like that of the SIRT 1 gene were further empowering and increasing the accuracy of the model as it showcased probable genetic predisposition towards severity of the condition. The results of this study not only prove how important the combination of genetic data and the age-old traditional clinical biomarker would be to forge better early detection and personalized treatment approaches for diabetic nephropathy, but promise that further outside validation in independent datasets will strengthen and establish the generalizability of such models. The results, in general, imply that, taking a look at machine learning tools in the direction of patient clinical practice, it could possibly bring interventions much earlier at improving patient outcome in diabetic nephropathic populations towards more targeted and effective treatment.

VI. References

1. Hosseini Sarkhosh, S. M., Esteghamati, A., Hemmatabadi, M., & Daraei, M. (2022). Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms. Journal of Diabetes & Metabolic Disorders, 21(2), 1433-1441. https://link.springer.com/article/10.1007/s40200-022-01076-2

2. Han, H., Chen, Y., Yang, H., Cheng, W., Zhang, S., Liu, Y., ... & Li, K. (2022). Identification and verification of diagnostic biomarkers for glomerular injury in diabetic nephropathy based on machine learning algorithms. *Frontiers in Endocrinology*, *13*, 876960. https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2022.876960/full

3. Maniruzzaman, M., Islam, M. M., Rahman, M. J., Hasan, M. A. M., & Shin, J. (2021). Risk prediction of diabetic nephropathy using machine learning techniques: a pilot study with secondary data. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *15*(5), 102263. https://www.sciencedirect.com/science/article/abs/pii/S1871402121002836

4. Yin, J. M., Li, Y., Xue, J. T., Zong, G. W., Fang, Z. Z., & Zou, L. (2024). Explainable Machine Learning-Based Prediction Model for Diabetic Nephropathy. *Journal of Diabetes Research*, *2024*(1), 8857453. https://onlinelibrary.wiley.com/doi/full/10.1155/2024/8857453

5. Xu, Q., Wang, L., & Sansgiry, S. S. (2020). A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning. *Journal of Medical Artificial Intelligence*, *3*. https://jmai.amegroups.org/article/view/5197/html

6. Abedi, M., Marateb, H. R., Mohebian, M. R., Aghaee-Bakhtiari, S. H., Nassiri, S. M., & Gheisari, Y. (2021). Systems biology and machine learning approaches identify drug targets in diabetic nephropathy. *Scientific Reports*, *11*(1), 23452. https://www.nature.com/articles/s41598-021-02282-3

7. Mesquita, F., Bernardino, J., Henriques, J., Raposo, J. F., Ribeiro, R. T., & Paredes, S. (2024). Machine learning techniques to predict the risk of developing diabetic nephropathy: a literature review. *Journal of Diabetes & Metabolic Disorders*, *23*(1), 825-839. https://link.springer.com/article/10.1007/s40200-023-01357-4

8. Su, J., Guo, Y., Hu, J., Peng, J., Dong, Z., Xu, Z., ... & Liu, H. (2024). Identification of diagnostic biomarkers for diabetes nephropathy by multi-chip integrated bioinformatics combining machine-learning strategies and mendelian randomization. https://www.researchsquare.com/article/rs-3936711/v1

9. Liu, M., Li, Z., Zhang, X., & Wei, X. (2024). A nomograph model for predicting the risk of diabetes nephropathy. https://www.researchsquare.com/article/rs-4174033/v1

10. Kanbour, S., Harris, C., Lalani, B., Wolf, R. M., Fitipaldi, H., Gomez, M. F., & Mathioudakis, N. (2024). Machine learning models for prediction of diabetic microvascular complications. *Journal of diabetes science and technology*, *18*(2), 273-286. https://journals.sagepub.com/doi/abs/10.1177/19322968231223726