diagnostics

MDPI

Article

Machine Learning-Based Diagnosis and Ranking of Risk Factors for Diabetic Retinopathy in Population-Based Studies from South India

Abhishek Vyas [1], Sundaresan Raman [1], Sagnik Sen [2,3], Kim Ramasamy [2], Ramachandran Rajalakshmi [4], Viswanathan Mohan [4] and Rajiv Raman [5,*]

[1] Birla Institute of Technology & Science, Pilani 333031, India
[2] Aravind Eye Hospital, Madurai 625020, India
[3] Moorfields Eye Hospital, London EC1V 2PD, UK
[4] Dr. Mohans Diabetes Specialities Centre, Chennai 600086, India
[5] Shri Bhagwan Mahavir Vitreoretinal Services, Sankara Nethralaya, Chennai 600006, India
* Correspondence: rajivpgraman@gmail.com

Abstract: This paper discusses the importance of investigating DR using machine learning and a computational method to rank DR risk factors by importance using different machine learning models. The dataset was collected from four large population-based studies conducted in India between 2001 and 2010 on the prevalence of DR and its risk factors. We deployed different machine learning models on the dataset to rank the importance of the variables (risk factors). The study uses a *t*-test and Shapely additive explanations (SHAP) to rank the risk factors. Then, it uses five machine learning models (K-Nearest Neighbor, Decision Tree, Support Vector Machines, Logistic Regression, and Naive Bayes) to identify the unimportant risk factors based on the area under the curve criterion to predict DR. To determine the overall significance of risk variables, a weighted average of each classifier's importance is used. The ranking of risk variables is provided to machine learning models. To construct a model for DR prediction, the combination of risk factors with the highest AUC is chosen. The results show that the risk factors glycosylated hemoglobin and systolic blood pressure were present in the top three risk factors for DR in all five machine learning models when the *t*-test was used for ranking. Furthermore, the risk factors, namely, systolic blood pressure and history of hypertension, were present in the top five risk factors for DR in all the machine learning models when SHAP was used for ranking. Finally, when an ensemble of the five machine learning models was employed, independently with both the *t*-test and SHAP, systolic blood pressure and diabetes mellitus duration were present in the top four risk factors for diabetic retinopathy. Decision Tree and K-Nearest Neighbor resulted in the highest AUCs of 0.79 (*t*-test) and 0.77 (SHAP). Moreover, K-Nearest Neighbor predicted DR with 82.6% (*t*-test) and 78.3% (SHAP) accuracy.

Keywords: diabetic retinopathy; ranking; risk factors; machine learning

Citation: Vyas, A.; Raman, S.; Sen, S.; Ramasamy, K.; Rajalakshmi, R.; Mohan, V.; Raman, R. Machine Learning-Based Diagnosis and Ranking of Risk Factors for Diabetic Retinopathy in Population-Based Studies from South India. *Diagnostics* **2023**, *13*, 2084. https://doi.org/10.3390/diagnostics13122084

Academic Editor: Jae-Ho Han

Received: 18 May 2023
Revised: 3 June 2023
Accepted: 5 June 2023
Published: 16 June 2023

1. Introduction

Diabetes mellitus (DM) is a metabolic syndrome with an increasing prevalence and high mortality rate [1]. The prevalence of diabetes in people aged 20–79 years has increased from 61.3 million in 2011 to 77 million today, and a further 77 million are pre-diabetic, raising significant concerns about the public health burden of this condition [2,3]. By 2030, it is estimated that approximately 101 million people in India will have diabetes [4–6].

Diabetic retinopathy (DR) is a common ocular complication of DM and is considered one of the leading causes of vision loss and impairment in adults in the working-age group [7,8]. According to a cross-sectional survey in England in 1990–1, the leading cause of blindness was macular degeneration, which accounted for 49% of blind registrations, and glaucoma was at 12%, and diabetes was at 4%. However, in the working-age group of

16–64 years, diabetic retinopathy was attributed to 12% of blindness, while diabetes was the most critical cause of blindness [9].

The number of people affected by diabetes-related retinal disease is 382 million worldwide, and by 2025, that number is anticipated to rise to 592 million [10]. The estimated prevalence of DR is around 34.6% (approximately 93 million individuals), and 10.2% have an advanced stage of the disease [11]. According to the National Diabetes and Diabetic Retinopathy Survey report 2015–2019, India has a DR prevalence of 11.8% in the population aged above 50, and 10.6% of patients are at risk of losing vision [12].

As DR typically does not manifest symptoms until the disease has progressed, only screening methods, such as a routine eye exam or retinal photography, can detect the disease in its early stages. However, because of the increasing number of people diagnosed with diabetes, systematic screening of all people with diabetes may be a big challenge.

DR is also predicted in the literature using deep learning systems. Developing deep learning systems requires standardized grading of retinal images, which is often a challenge. Though screening through retinal photographs is the gold standard, it is important to identify the groups with the highest risk of developing DR so that photographic screening can be prioritized, especially in populations with a high prevalence of diabetes. Therefore, there is a need to look for systemic factors related to DR, which can play a key role as a prescreening tool for DR.

Several factors, such as high blood pressure, postprandial hyperglycemia, albuminuria, serum creatinine, glycosylated hemoglobin, and plasma glucose levels, are significantly associated with the risk of DR [13–19]. Therefore, understanding the role of risk factors is important for developing a strategy to improve global eye health. A previous study has determined that diabetes patients older than 50 years with diabetes duration > 5 years and systolic blood pressure > 140 mm Hg could be targeted to achieve optimal detection of vision-threatening diabetic retinopathy [20].

Risk factors for DR have been identified using statistical techniques in the literature [21–27]. Moreover, the ranking of various risk factors for DR has not received much attention in the literature. Ranking risk factors aims to streamline screening programs and focus on the most important ones.

Clinical data on risk factors has been used in the literature to predict DR. In this context, Cichosz et al. [28] used a linear classification model to predict which individuals had diabetic retinopathy based on data obtained from the National Health and Nutritional Examination Survey (NHANES, 2005–2008) [29] on the oral glucose tolerance test (OGTT), FPG, or HbA1C, and retinal imaging. Using information regarding HbA1c, BMI, waist circumference, age, SBP, urinary albumin, and urinary creatinine, they constructed a model that predicts the presence of retinopathy with a negative predictive value of 99% and a positive predictive value of 22%. Ogunyemi et al. [30] used clinical data from urban safety-net clinics and public health data from the Centers for Disease Control and Prevention (CDC) National Health and Nutrition Examination Survey to learn RUSBoost [31] and AdaBoost [32] ensemble classifiers for predicting retinopathy. The results show that the clinical dataset was not very good at predicting diabetic retinopathy. The best RUSBoost ensemble had an accuracy of 73.5%, a sensitivity of 69.2%, a specificity of 55.9%, and an AUC of 0.72 on cases that had never been seen before (the test data). Tsao et al. [33] built a prediction model for the DR in type 2 diabetes mellitus using data mining techniques, including Decision Trees, Support Vector Machines, Logistic Regressions, and Artificial Neural Networks. The performance of Support Vector Machines was better than that of the other machine learning algorithms. It achieved an accuracy of 79.5% and an AUC of 0.839 using a percentage split (i.e., the data set was divided into 80% as training and 20% as a test).

As risk factors have been used to predict DR, it is important to screen significant risk factors from the many presented in the dataset. There is no method to identify the top risk factors for DR. The paper gives a novel approach to ranking risk factors to identify the most significant ones. These risk factors could aid in developing a risk factor-based

algorithm that can aid in the prescreening of DR. The algorithm can help as a prescreening tool for detecting the need for using fundus photographs for identifying referable and non-referable DR.

With a myriad of risk factors for most diseases, it is essential to identify the most important ones, which can be fed to machine learning models for improved classification. We have developed an algorithm for ranking the risk factors for diabetic retinopathy. We have incorporated two techniques for ranking: first, statistical methods (using *p*-values) and second, Shapley Additive Explanations (SHAP). Each of these two methods serves as a validation for the other. Our proposed algorithm can potentially rank risk factors for any other disease.

Our study also predicts DR using five machine-learning classification models, Decision Tree (DT), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), and Naive Bayes (NB). Finally, we have employed an ensemble of machine learning models to predict DR.

The paper is organized as follows: Section 2 discusses materials and methods. Section 3 gives the results, followed by a Discussion in Section 4 and a Conclusion in Section 5.

2. Material and Methods

2.1. Samples and Data Preprocessing

The sample dataset was collected from four large population-based studies conducted in India between 2001 and 2010 on the prevalence of DR and its risk factors [34–37]. All methods were performed following the relevant guidelines and regulations. The study was approved by the Institutional Review Boards of Madras Diabetic Research Foundation, Chennai, India; Vision Research Foundation, Chennai, India; and Aravind Eye Care System, Madurai, India. Informed consent was obtained from the participants according to the Declaration of Helsinki before collecting the data. These studies had patient-level data and included previously diagnosed and newly diagnosed diabetics. In this study, we included data on people aged 40 and older to obtain uniform data for analysis. In the current study, the diagnosis of new diabetes was defined as FBS > 7 mmol/L or >126 mg/dL at the time of initial screening. Age at presentation, duration of diabetes (for known individuals with diabetes), gender, history of hypertension, obesity, cardiovascular disease (CVD), and smoking history were among the sociodemographic and clinical parameters shared by all studies. The prevalence of the stages of DR is 1% proliferative DR, 12.4% mild/moderate non-proliferative DR, 1.4% severe non-proliferative DR, and 3.7% diabetic macular edema.

Data from all four studies was entered into a Microsoft Excel spreadsheet. The total number of people with DR was 857, and those without DR were 3133. An ophthalmologist provided a group of features contributing to the disease directly and indirectly. We call these features risk factors, and our primary objective in this study is to rank these risk factors, which include the history of hypertension status, insulin treatment status, systolic blood pressure status, glycosylated hemoglobin (HBA1c) value, duration of diabetes mellitus, fasting blood glucose, gender, body mass index, and age. Ordinal encoding is applied to categorical risk factors to convert them into continuous risk factors as machine learning models only understand numbers in data. Normalization was applied to continuous risk factors so that the deviation of the variables did not affect classification or model interpretation. Like previous medical data studies, we replaced the missing values with the mode for binary data and the median for numerical data [38].

Ranking features can be done using Random Forests and Logistic Regression. If Random Forests are used, equal importance is given to correlated features. Furthermore, they give preference to features with high cardinality. Logistic regression assumes linearity between the dependent variable and the independent variables. Moreover, it requires no multicollinearity between independent variables. An independent two-sample *t*-test ranks the risk factors according to their *p* values. A lower *p*-value denotes more importance. Furthermore, Shapely additive explanations [39] are used, giving Shapely values a suitable

measure of feature importance. The higher the Shapely value, the higher is the importance of the feature. The coding was performed in Python using a Google Colaboratory notebook with a CPU frequency of 2.30 GHz, 2 CPU cores, the Haswell CPU family, and 12 GB of available RAM. The Python libraries used were sklearn, imblearn, numpy, pandas, matplotlib, collections, and scipy. The purpose of using machine learning classification models is to validate the *t*-test and SHAP-based rankings.

2.2. t-Test

A *t*-test (also known as the student's *t*-test) is a tool for evaluating the means of one or two populations using hypothesis testing. A *t*-test may be used to evaluate whether a single group differs from a known value (a one-sample *t*-test), whether two groups differ from each other (an independent two-sample *t*-test), or whether there is a significant difference in paired measurements (a paired or dependent samples *t*-test). An independent two-sample *t*-test ranks the risk factors according to their *p* values. A lower *p*-value denotes more importance.

The *t*-test assumes that the independent samples of two populations have the same variance and are normally distributed. As there are two samples from a population with unequal variances, the *t*-test is reasonably robust to the violation of its first assumption. A *t*-test repeated measure design yields small effects due to the small sample error. It also results in the effective management of individual differences. One group is available for testing, which may result in less data noise.

2.3. Shapley Values

Lloyd Shapely, in 1953 [40], proposed the concept of Shapley values, which numerically evaluate the value of playing a game. It is important to interpret a model's prediction correctly. It provides an insight into how a model may be improved, engenders user trust, and supports understanding the modeled process. The model itself is the best explanation of a simple model. A simple explanation model is used for complex models, such as ensembles or deep networks, as an interpretable approximation of the original model. In multicollinearity, Shapley regression values are feature importances for linear models. The method requires the model to be retrained on all feature subsets $S \subseteq F$, where F is the set of all features. It assigns to each feature an importance value that represents the effect on the model prediction, including that feature. To compute this effect, a model $f_{S\cup\{i\}}$ is trained with that feature present, and another model f_S is trained with the feature withheld. Then, from the two models, predictions are compared with the current input $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, where x_S represents the values of the input features in the set S. The preceding differences are computed for all possible subsets $S \subseteq F\backslash\{i\}$ because the effect of withholding a feature depends on other features in the model. The Shapley values are then computed and used as feature attributions. They are a weighted average of all possible differences:

$$\varphi_i = \Sigma \frac{|S|!(|F| - |S| - 1)!}{|F|!} \quad [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{1}$$

$$S \subseteq F\backslash\{i\}$$

Shapely values are a unified measure of feature importance. These are solutions to Equation (1) as they are the Shapley values of a conditional expectation function of the original model. The higher the Shapely value, the higher is the importance of the feature.

SHAP is used because it has a solid theoretical foundation in game theory. Furthermore, among the feature values, the prediction is fairly distributed. Moreover, SHAP has a fast implementation for tree-based models. Although Shapley value computation requires exponential time complexity, machine learning applications employ Shapley value approximation methods, such as Monte Carlo permutation sampling, which approximates Shapley value in linear time [41–43].

2.4. Method Design

The data was divided into an 80% training and 20% test set. Synthetic Minority Oversampling Technique (SMOTE) [44] was applied to the training data as the oversampling method to generate synthetic data in the minority class for solving the class imbalance problem. Following this, the proposed algorithm (Algorithm 1), as shown below, was applied to generate the weights for the risk factors using a *t*-test and SHAP independently.

Algorithm 1: Ranking of Risk Factors.

Begin
Input: Set of risk factors ($R = R_1, R_2, R_3 \ldots, R_N$)
1. **for** each risk factor, R_i **do**
2. Apply independent sample *t*-test to R_i and calculate its *p*-value. A lower *p*-value denotes more importance.
3. Apply SHAP to each model, namely, SVM, DT, KNN, LR, and NB, to calculate the Shapley value of R_i for each model. A higher Shapely value denotes higher importance.
4. **end for**
5. Sort the risk factors in step 2 in increasing order of their *p*-values.
6. Sort the risk factors in step 3 in decreasing order of Shapely values.
7. The sorted risk factors are the risk factors $R_1, R_2, R_3 \ldots, R_N$ ranked independently using a *t*-test and SHAP.
8. The weights of risk factors $R_1, R_2, R_3 \ldots, R_N$ are set based on the ranking results. For a risk factor R_i with rank order *r*, its weight is set as $d - r$, where *d* is the number of risk factors.
9. Set the previous cross_val_score as prev_AUC = 0
10. **for** each i in the *N* (number of risk factors) ranked with the *t*-test and SHAP **do**
11. **for** each model j in M(SVM, DT, KNN, LR, and NB) **do**
12. Calculate the cross_val_score for the risk factor i as curr_AUC
13. **if** prev_AUC $\geq$ curr_AUC
14. Remove the risk factor from the ranked order and set its weight to zero.
15. **else** prev_AUC = curr_AUC
16. **end for**
17. **end for**
18. **return** Weights assigned to Risk factors for all models in the *t*-test and SHAP.
Output: Weights assigned to risk factors for all models in the *t*-test and SHAP.
End

The ensemble weights across all five models, namely, K-Nearest Neighbor (KNN) [45], Decision Tree (DT) [46], Support Vector Machine (SVM) [47], Logistic Regression [48], and Naive Bayes [49] are computed for both *t*-test and SHAP. The above steps are repeated three times to enhance robustness. An average is computed to arrive at the final model weights (for both the individual model and the ensemble case). As can be seen from Figure 1, the weights are sorted in decreasing order of importance for each model in the *t*-test and SHAP to generate the ranking of risk factors. The ranked risk factors are added individually to compute the best combination of risk factors producing the highest AUC metric to predict DR separately using a *t*-test and SHAP.

The *t*-test is used because it is robust to the violation of its assumption that the independent samples of two populations have the same variance and are typically distributed. SHAP is used in the algorithm because it has a solid theoretical foundation and fairly distributes the prediction among the feature values.

Let N be the number of risk factors, and K be the number of machine learning models. The algorithm proposed that employs the *t*-test has a time complexity of $O(N^2 \log N + N \times K)$ and that using a SHAP has a time complexity of $O(K \times N^2 \log N)$, assuming a linear approximation in the Shapley evaluation method.

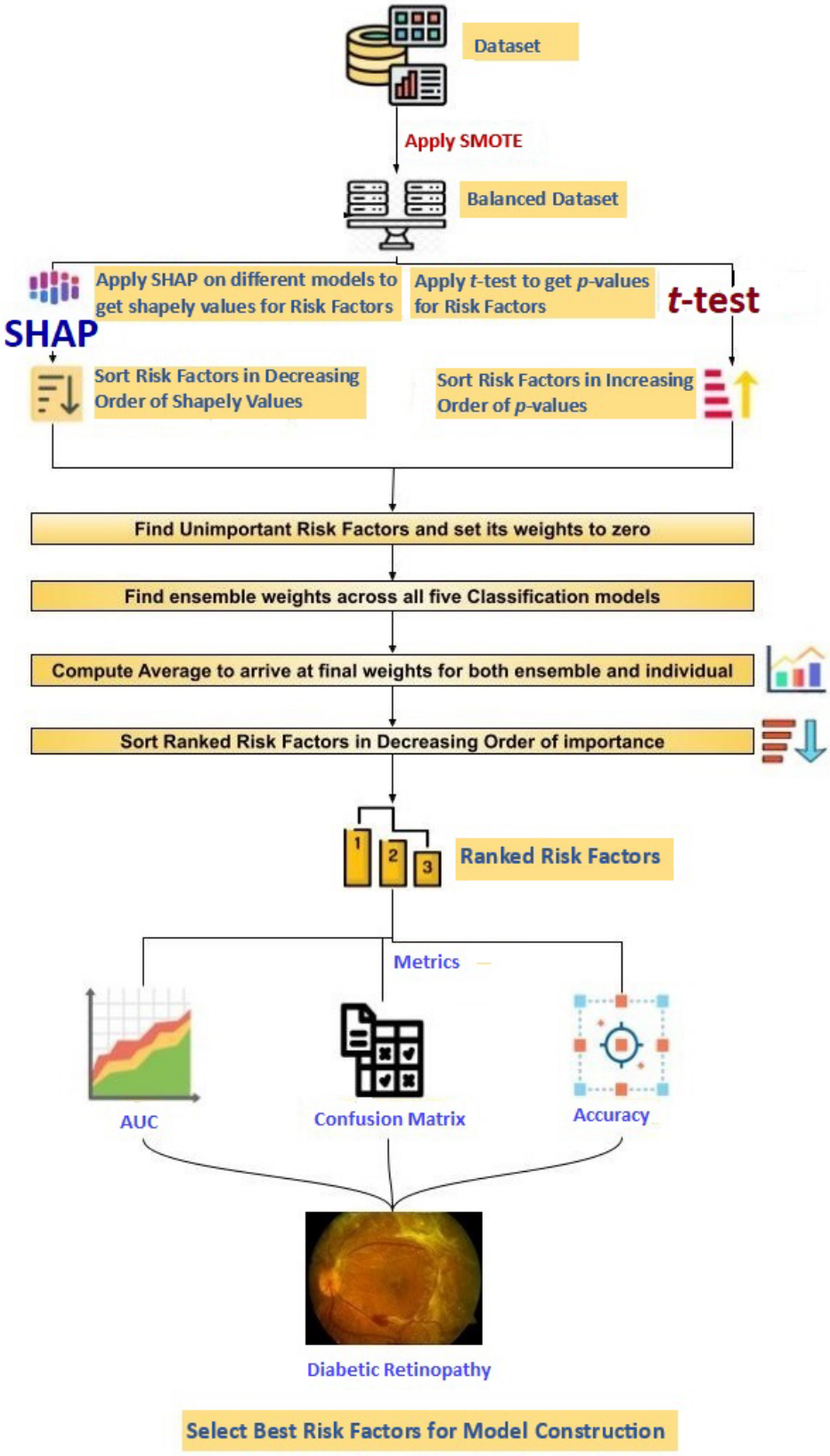

Figure 1. The flow of the method design.

3. Results

3.1. Ranking

Algorithm 1 was applied to the dataset after performing SMOTE, initially using a *t*-test and then SHAP as a ranking measure. It can be inferred from Table 1 that risk factors, such as glycosylated hemoglobin and systolic blood pressure, were found to be the top three risk factors for diabetic retinopathy in all machine learning (ML) models when using a *t*-test for ranking. Furthermore, it can be inferred from Table 2 that risk factors, such as systolic blood pressure and history of hypertension, were found to be among the top five risk factors for diabetic retinopathy in all machine learning models using SHAP for ranking. Table 3 shows that when an ensemble of five models were used, risk factors, such as systolic blood pressure and duration of diabetes mellitus, were found to be in the top four risk factors for DR in both the *t*-test and SHAP-based rankings.

Table 1. Ranking of risk factors using *t*-test and SMOTE with various ML models.

SNo.	Decision Tree	SVM	KNN	Logistic Regression	Naive Bayes
1	glycosylated hemoglobin	glycosylated hemoglobin	glycosylated hemoglobin	glycosylated hemoglobin	glycosylated hemoglobin
2	diabetes mellitus duration	systolic blood pressure	systolic blood pressure	body mass index	body mass index
3	systolic blood pressure	diabetes mellitus duration	fasting plasma glucose	systolic blood pressure	systolic blood pressure
4	fasting plasma glucose	insulin treatment	history of hypertension	gender	gender
5	history of hypertension	fasting plasma glucose	insulin treatment	age	diabetes mellitus duration
6	insulin treatment	history of hypertension	diabetes mellitus duration	insulin treatment	age
7	gender	gender	gender	fasting plasma glucose	insulin treatment
8	body mass index	body mass index	body mass index	history of hypertension	fasting plasma glucose
9	age	age	age	diabetes mellitus duration	history of hypertension

Table 2. Ranking of risk factors using SHAP and SMOTE with various ML models.

SNo.	Decision Tree	SVM	KNN	Logistic Regression	Naive Bayes
1	diabetes mellitus duration	systolic blood pressure	glycosylated hemoglobin	systolic blood pressure	systolic blood pressure
2	glycosylated hemoglobin	fasting plasma glucose	systolic blood pressure	history of hypertension	fasting plasma glucose
3	systolic blood pressure	history of hypertension	fasting plasma glucose	insulin treatment	history of hypertension
4	fasting plasma glucose	insulin treatment	history of hypertension	diabetes mellitus duration	insulin treatment
5	history of hypertension	diabetes mellitus duration	insulin treatment	gender	diabetes mellitus duration
6	insulin treatment	gender	diabetes mellitus duration	glycosylated hemoglobin	gender
7	gender	glycosylated hemoglobin	gender	body mass index	glycosylated hemoglobin
8	body mass index	body mass index	body mass index	fasting plasma glucose	body mass index
9	age	age	age	age	age

3.2. Classification Performance

Tables 4 and 5 show the results for the sensitivity, specificity, AUC, and accuracy of the five individual classifiers using ensemble weights for DR prediction in the *t*-test and SHAP. All the models achieved sensitivity ranging from 0.55 to 0.76, specificity ranging from 0.59 to 0.84, AUCs ranging from 0.71 to 0.79, and accuracy ranging from 64.3% to 82.6%. Out

of the five machine learning classifiers, in terms of sensitivity, Naive Bayes performed the best, with a value of 0.76 in the *t*-test and SHAP. Regarding specificity, KNN performed the best, with a value of 0.84 in the *t*-test and 0.8 in SHAP. In terms of AUC, while using a *t*-test, DT and KNN resulted in the highest AUC value of 0.79 with associated risk factors, such as glycosylated hemoglobin and systolic blood pressure. Similarly, with SHAP, DT and KNN resulted in the highest AUC of 0.77 with associated risk factors, such as systolic blood pressure, history of hypertension, diabetes mellitus duration, insulin treatment, fasting plasma glucose, and glycosylated hemoglobin. KNN achieved the best accuracy of 82.6% in the case of the *t*-test and 78.3% in the case of SHAP. The AUC and accuracy of the DT and KNN of the *t*-test and SHAP are shown in Figure 2a,b,d,e. The receiver operating characteristic (ROC) curve for all models of the *t*-test and SHAP is shown in Figure 2c,f.

Table 3. Ranking of risk factors using an ensemble of ML models.

SNo.	*t*-test + Ensemble	Shapely + Ensemble
1	glycosylated hemoglobin	systolic blood pressure
2	systolic blood pressure	history of hypertension
3	body mass index	diabetes mellitus duration
4	diabetes mellitus duration	insulin treatment
5	gender	fasting plasma glucose
6	age	glycosylated hemoglobin
7	insulin treatment	gender
8	fasting plasma glucose	body mass index
9	history of hypertension	age

Table 4. Metrics for ranking using a *t*-test using ensemble weights (Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Area under the ROC Curve (AUC)).

		***t*-test**		
Model	**Sensitivity**	**Specificity**	**AUC**	**Accuracy**
Decision Tree	0.6	0.83	0.79	0.779
SVM	0.72	0.66	0.75	0.711
KNN	0.58	0.84	0.79	0.826
Logistic Regression	0.68	0.65	0.71	0.657
Naive Bayes	0.76	0.59	0.73	0.668

Table 5. Metrics for ranking using Shapely additive explanations (SHAP) using ensemble weights (Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Area under the ROC Curve (AUC)).

		SHAP		
Model	**Sensitivity**	**Specificity**	**AUC**	**Accuracy**
Decision Tree	0.64	0.78	0.77	0.753
SVM	0.72	0.66	0.75	0.747
KNN	0.55	0.8	0.77	0.783
Logistic Regression	0.68	0.65	0.71	0.66
Naive Bayes	0.76	0.59	0.73	0.643

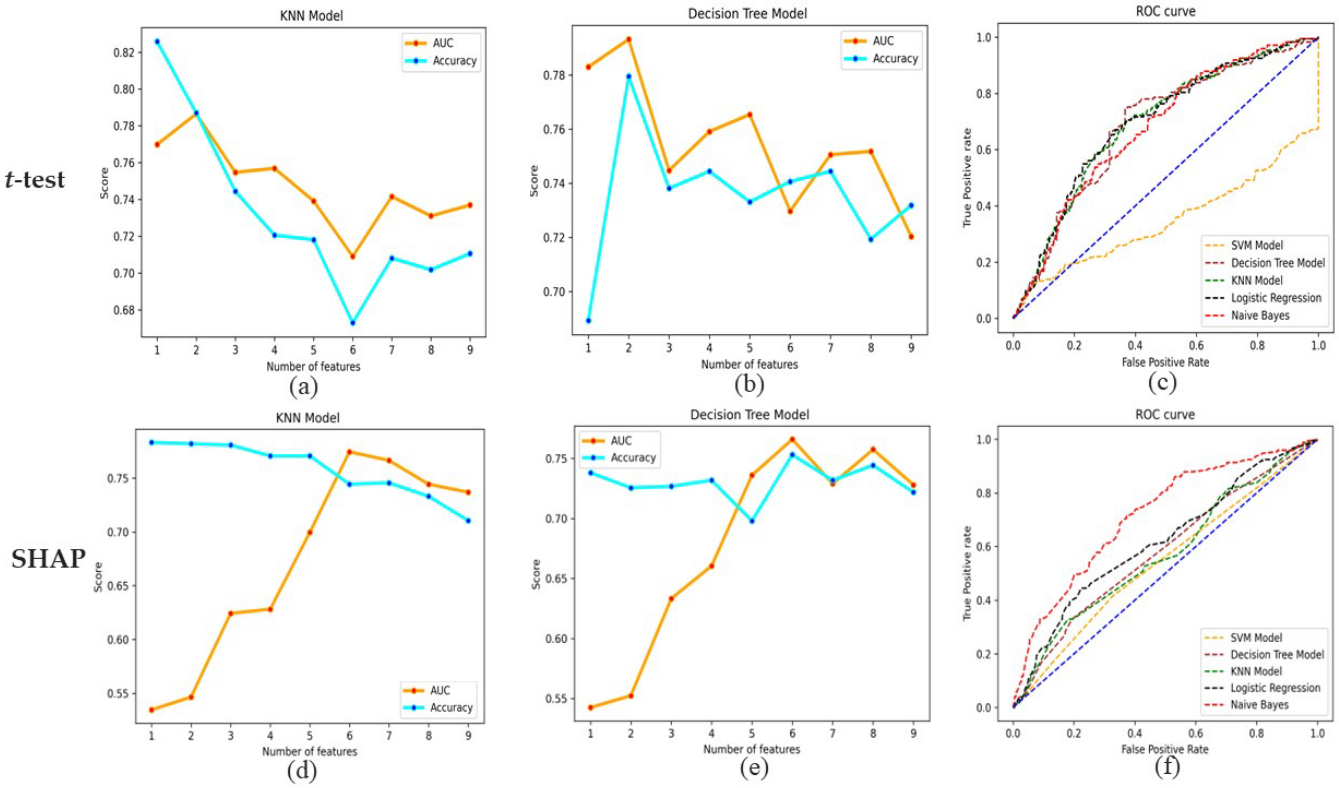

Figure 2. (**a**,**b**,**d**,**e**) AUC and accuracy for the Decision tree and K-Nearest Neighbors in the case of *t*-test and SHAP. (**c**,**f**) ROC curve for all models of *t*-test and SHAP.

4. Discussion

In this study, we propose an algorithm for ranking risk factors to predict DR. Systolic blood pressure was consistently found to be among the top risk factors using the *t*-test, SHAP, and ensemble methods. The sensitivity, specificity, and AUC values for the *t*-test and SHAP are very close to each other, which validates our two methods (using the *t*-test and SHAP) for ranking risk factors. In the case of the *t*-test, DT and KNN resulted in the highest AUC value of 0.79 with associated risk factors, such as glycosylated hemoglobin and systolic blood pressure. Similarly, with SHAP, DT and KNN resulted in the highest AUC value of 0.77 with associated risk factors, such as systolic blood pressure, history of hypertension, diabetes mellitus duration, insulin treatment, fasting plasma glucose, and glycosylated hemoglobin. Comparing the risk factors for AUC, it can be inferred that systolic blood pressure and glycosylated hemoglobin seem to be the critical risk factors for predicting DR.

This study uses two techniques to rank the risk factors: the *t*-test and SHAP. SHAP and *t*-tests take fundamentally different approaches to evaluate the significance of a feature. SHAP values are derived from cooperative game theory and evaluate the contribution of each feature to a specific instance's prediction. In contrast, *t*-tests are a statistical method for testing hypotheses that compares the means of two groups. These distinct methodologies can result in varying conclusions regarding the significance of a feature. SHAP values can detect complex interactions between features that a *t*-test might overlook. In isolation, the *t*-test may indicate that HbA1c is significant, but in the context of other features and their interactions, SHAP values may indicate that HbA1c is not as significant. This may result from confounding variables, multicollinearity, or other intricate feature interactions. A *t*-test's results can be sensitive to sample size and variation. The *t*-test might mistakenly identify HbA1c as a significant factor when it is not due to a lack of statistical power resulting from a small sample size or high variability. In contrast, SHAP values may be more resilient in such circumstances. While age is consistently at the bottom, systolic blood

pressure and diabetes mellitus duration are consistently at the top of the ranking in all the models using the *t*-test and SHAP.

We can infer from Tables 1–3 that age is the least significant risk factor as it was ranked among the last five risk factors by all machine learning models and the ensemble model. The current literature shows the effect of age on the severity of DR is unclear and varies with the population being studied [36]. While Stratton et al. reported that old age impacts the progression of DR [50], other studies identified younger age as a risk factor [51–53]. Therefore, it is likely that the risk of DR may be present irrespective of age; hence, screening should be performed in all age groups.

Although age was not on the list of the top risk factors in this analysis, we did find that duration of diabetes was one of the top four risk factors employing ensemble models of five classifiers. The duration of diabetes is related to the patient's exposure time to other DR-related risk factors. It should, therefore, be of prime importance while targeting screening towards DR. Previous studies have shown that DM duration may be the most important independent risk factor for DR [54–56].

There were few investigations on the risk stratification of DR based on ML and risk factors. Azizi-Soleiman et al. [57] reported a model for detecting DR in Iranians using outpatient clinical data. The logit model obtained an AUC of 0.760 by training on the data of 1782 patients (without cross-validation) using backward elimination as a feature selection strategy. Tsao et al. [33] divided the clinical data of 536 Taiwanese patients into training and validation sets (80:20 ratio) and tested how well the four models (Support Vector Machine, Decision Tree, Artificial Neural Network (ANN), and Logistic Regression) could detect DR. They found that the Support Vector Machine performed the best, with an AUC of 0.839. According to Yao et al. [58], an Artificial Neural Network with back propagation outperformed Logistic Regression in DR detection with AUCs of 0.84 and 0.77, respectively. Population-based data are more pertinent to the reality of DR screening programs than hospital-based data [59]. Our study applied machine learning (ML) techniques to population-based data and demonstrated their utility for DR detection. Moreover, we have proposed two techniques to rank the risk factors: the *t*-test and SHAP, which validate each other, obtaining AUCs of 0.79 and 0.77, respectively.

The first limitation of the study is that only a subset of the risk factors suggested by the current literature were considered. There is scope for a larger set of risk factors to be considered to identify the top risk factors that can aid in initial screening for referable DR in populations where ophthalmologists are scarce. Second, it was not possible to evaluate risk factor rankings for each form of retinopathy separately in the present study. The classification of risk factors refers to the risk of diabetic retinopathy, regardless of its severity.

Our study has shown that ML technology successfully ranks important risk factors in large-scale epidemiological studies. Previous studies have demonstrated the vital role of ML in other medical fields, such as T2DM, obesity, and heart failure [60–62]. Our results confirm the excellent performance of ML in predicting diabetic retinopathy. This is the first study that evaluates the importance of risk factors using various ML methods with data from the Indian population and checks the risk factors for diabetic retinopathy.

5. Conclusions

The study aims to reflect the importance of ranking risk factors to find their relevance to DR. We have proposed two techniques to find the relative contributions of risk factors to the presence of DR. In both cases, age contributed the least and systolic blood pressure contributed the most among the nine risk factors considered for the study. Thus, validating both of our proposed techniques, KNN achieved the best accuracy of 82.6% in the case of the *t*-test and 78.3% in the case of SHAP to predict DR. A subset of risk factors given by ophthalmologists was considered in the study. Furthermore, other risk factors, such as demographics, lifestyle, family history, living standards, and ethnicity, need to be explored in further studies as part of the future scope of this work. These risk factors could aid

in developing a risk factor algorithm for DR and aid in the prescreening of DR. The algorithm can be a prescreening tool using fundus photographs to identify referable and non-referable DR. The study can also be extended using images and top risk factors to predict DR. Moreover, the ranking of risk factors for non-proliferative/proliferative DR or diabetic macular edema and whether the ranking would change with the development of macular edema or proliferative DR could be a potential future study.

Author Contributions: Conception or design of the work—S.R., R.R. (Rajiv Raman) and A.V. Data collection—S.S. and K.R. Data analysis and interpretation—A.V., S.R., R.R. (Rajiv Raman) and S.S. Drafting the article—A.V. Critical revision of the article—S.R., R.R. (Rajiv Raman), S.S., A.V., V.M. and R.R. (Ramachandran Rajalakshmi). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Declaration of Helsinki, and approved by the Institutional Review Board of Vision Research Foundation (IRB No 59-2007 P, 23/05/2007).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets analyzed during the current study are not publicly available as it is against the organization/hospital policy. However, they are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. King, H.; Aubert, R.E.; Herman, W.H. Global Burden of Diabetes, 1995–2025 Prevalence, numerical estimates, and projections. *Diabetes Care* **1998**, *21*, 1414–1431. [CrossRef]
2. Anjana, R.M.; Deepa, M.; Pradeepa, R.; Mahanta, J.; Narain, K.; Das, H.K.; Adhikari, P.; Rao, P.V.; Saboo, B.; Kumar, A.; et al. Prevalence of diabetes and prediabetes in 15 states of India: Results from the ICMR--INDIAB population-based cross-sectional study. *Lancet Diabetes Endocrinol.* **2017**, *5*, 585–596. [CrossRef]
3. Saeedi, P.; Petersohn, I.; Salpea, P.; Malanda, B.; Karuranga, S.; Unwin, N.; Colagiuri, S.; Guariguata, L.; Motala, A.A.; Ogurtsova, K.; et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Res. Clin. Pract.* **2019**, *157*, 107843. [CrossRef]
4. Whiting, D.R.; Guariguata, L.; Weil, C.; Shaw, J. IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Res. Clin. Pract.* **2011**, *94*, 311–321. [CrossRef] [PubMed]
5. Anjana, R.M.; Pradeepa, R.; Deepa, M.; Datta, M.; Sudha, V.; Unnikrishnan, R.; Bhansali, A.; Joshi, S.R.; Joshi, P.P.; Yajnik, C.S.; et al. Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: Phase I results of the Indian Council of Medical Research—INdia DIABetes (ICMR—INDIAB) study. *Diabetologia* **2011**, *54*, 3022–3027. [CrossRef]
6. Wild, S.; Roglic, G.; Green, A.; Sicree, R.; King, H. Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. *Diabetes Care* **2004**, *27*, 1047–1053. [CrossRef] [PubMed]
7. Bourne, R.R.; Stevens, G.A.; White, R.A.; Smith, J.L.; Flaxman, S.R.; Price, H.; Jonas, J.B.; Keeffe, J.; Leasher, J.; Naidoo, K.; et al. Causes of vision loss worldwide, 1990-2010: A systematic analysis. *Lancet Glob. Heal.* **2013**, *1*, 339–349. [CrossRef]
8. Klein, B.E.K. Overview of epidemiologic studies of diabetic retinopathy. *Ophthalmic Epidemiol.* **2007**, *14*, 179–183. [CrossRef]
9. Gibson, J.M.; Lavery, J.R.; Rosenthal, A.R. Blindness and partial sight in an elderly population. *Br. J. Ophthalmol.* **1986**, *70*, 700–705. [CrossRef] [PubMed]
10. Pandey, S.K.; Sharma, V. World Diabetes Day 2018: Battling the Emerging Epidemic of Diabetic Retinopathy. *Indian J. Ophthalmol.* **2018**, *16*, 2. [CrossRef] [PubMed]
11. Yau, J.W.Y.; Rogers, S.L.; Kawasaki, R.; Lamoureux, E.L.; Kowalski, J.W.; Bek, T.; Chen, S.-J.; Dekker, J.M.; Fletcher, A.; Grauslund, J.; et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care* **2012**, *35*, 556–564. [CrossRef] [PubMed]
12. Dr. Rajendra Prasad Centre for Ophthalmic Sciences, AIIMS, New Delhi. *National Blindness and Visual Impairment Survey India 2015–2019: A Summary Report. National Programme for Control of Blindness and Visual Impairment*; Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India: New Delhi, India, 2019.
13. Klein, R.; Klein BE, K.; Moss, S.E.; Davis, M.D.; DeMets, D.L. Is blood pressure a predictor of the incidence or progression of diabetic retinopathy? *Arch. Intern. Med.* **1989**, *149*, 2427–2432. [CrossRef]
14. Constable, I.; Knuiman, M.; Welborn, T.; Cooper, R.; Stanton, K.; McCann, V.; Grose, G. Assessing the risk of diabetic retinopathy. *Am J. Ophthalmol.* **1984**, *97*, 53–61. [CrossRef] [PubMed]

15. Shiraiwa, T.; Kaneto, H.; Miyatsuka, T.; Kato, K.; Yamamoto, K.; Kawashima, A.; Kanda, T.; Suzuki, M.; Imano, E.; Matsuhisa, M.; et al. Postprandial hyperglycemia is a better predictor of the progression of diabetic retinopathy than HbA1c in Japanese type 2 diabetic patients. *Diabetes Care.* **2005**, *28*, 2806–2807. [CrossRef]
16. Manaviat, M.R.; Afkhami, M.; Shoja, M.R. Retinopathy and microalbuminuria in type II diabetic patients. *BMC Ophthalmol.* **2004**, *4*, 9. [CrossRef] [PubMed]
17. Maghbooli, Z.; Pasalar, P.; Keshtkar, A.; Farzadfar, F.; Larijani, B. Predictive factors of diabetic complications: A possible link between family history of diabetes and diabetic retinopathy. *J. Diabetes Metab. Disord.* **2014**, *13*, 11–15. [CrossRef]
18. Zoppini, G.; Verlato, G.; Targher, G.; Casati, S.; Gusson, E.; Biasi, V.; Perrone, F.; Bonora, E.; Muggeo, M. Is fasting glucose variability a risk factor for retinopathy in people with type 2 diabetes? *Nutr. Metab. Cardiovasc. Dis.* **2009**, *19*, 334–339. [CrossRef]
19. Conway, B.N.; Miller, R.G.; Klein, R.; Orchard, T.J. Prediction of proliferative diabetic retinopathy with hemoglobin level. *Arch. Ophthalmol.* **2009**, *127*, 1494–1499. [CrossRef]
20. Vignesh, T.; Sen, S.; Ramasamy, K.; Kannan, N.; Sivaprasad, S.; Rajalakshmi, R.; Raman, R.; Mohan, V.; Das, T.; Mani, I. Identification of risk factors for targeted diabetic retinopathy screening to urgently decrease the rate of blindness in people with diabetes in India. *Indian J. Ophthalmol.* **2021**, *69*, 3156. [CrossRef]
21. Cohen, O.; Norymberg, K.; Neumann, E.; Dekel, H. Complication-free duration and the risk of development of retinopathy in elderly diabetic patients. *Arch. Intern. Med.* **1998**, *158*, 641–644. [CrossRef]
22. Hietala, K.; Harjutsalo, V.; Forsblom, C.; Summanen, P.; Groop, P.H. Age at onset and the risk of proliferative retinopathy in type 1 diabetes. *Diabetes Care* **2010**, *33*, 1315–1319. [CrossRef]
23. Hu, Y.; Teng, W.; Liu, L.; Chen, K.; Liu, L.; Hua, R.; Chen, J.; Zhou, Y.; Chen, L. Prevalence and risk factors of diabetes and diabetic retinopathy in Liaoning Province, China: A population-based cross-sectional study. *PLoS ONE* **2015**, *10*, e0121477. [CrossRef] [PubMed]
24. Forga, L.; Goñi, M.J.; Ibáñez, B.; Cambra, K.; García-Mouriz, M.; Iriarte, A. Influence of Age at Diagnosis and Time-Dependent Risk Factors on the Development of Diabetic Retinopathy in Patients with Type 1 Diabetes. *J. Diabetes Res.* **2016**, *2016*, 9898309. [CrossRef] [PubMed]
25. Xiao, Y.; Liang, Y.; Lin, Z.; Kong, H.; Du, Z.; Hu, Y.; Ouyang, S. Causes and Risk Factors of Repeated Hospitalization among Patients with Diabetic Retinopathy. *J. Diabetes Res.* **2022**, *2022*, 1–7. [CrossRef] [PubMed]
26. Peters, K.S.; Rivera, E.; Warden, C.; Harlow, P.A.; Mitchell, S.L.; Calcutt, M.W.; Samuels, D.C.; Brantley, M.A., Jr. Plasma Arginine and Citrulline are Elevated in Diabetic Retinopathy. *Am. J. Ophthalmol.* **2022**, *235*, 154–162. [CrossRef]
27. Sabanayagam, C.; Sultana, R.; Banu, R.; Rim, T.; Tham, Y.C.; Mohan, S.; Chee, M.L.; Wang, Y.X.; Nangia, V.; Fujiwara, K.; et al. Association between body mass index and diabetic retinopathy in Asians: The Asian Eye Epidemiology Consortium (AEEC) study. *Br. J. Ophthalmol.* **2022**, *106*, 980–986. [CrossRef]
28. Cichosz, S.L.; Johansen, M.D.; Knudsen, S.T.; Hansen, T.K.; Hejlesen, O. A classification model for predicting eye disease in newly diagnosed people with type 2 diabetes. *Diabetes Res. Clin. Pract.* **2015**, *108*, 210–215. [CrossRef]
29. Centers for Disease Control and Prevention (CDC) (2005–2010). National Center for Health Statistics (NCHS). In *National Health and Nutrition Examination Survey Data*; U.S. Department of Health and Human Services: Hyattsville, MD, USA, 2020.
30. Ogunyemi, O.; Kermah, D. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. In *AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Bethesda, MD, USA, 2015; Volume 2015, p. 983.
31. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2009**, *40*, 185–197. [CrossRef]
32. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
33. Tsao, H.-Y.; Chan, P.-Y.; Su, E.C.-Y. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinform.* **2018**, *19*, 111–121. [CrossRef]
34. Rema, M.; Premkumar, S.; Anitha, B.; Deepa, R.; Pradeepa, R.; Mohan, V. Prevalence of diabetic retinopathy in urban India: The Chennai Urban Rural Epidemiology Study (CURES) Eye Study, I. *Investig. Ophthalmol. Vis. Sci.* **2005**, *46*, 2328–2333. [CrossRef] [PubMed]
35. Raman, R.; Rani, P.K.; Rachepalle, S.R.; Gnanamoorthy, P.; Uthra, S.; Kumaramanickavel, G.; Sharma, T. Prevalence of diabetic retinopathy in India: Sankara Nethralaya diabetic retinopathy epidemiology and molecular genetics study report 2. *Ophthalmology* **2009**, *116*, 311–318. [CrossRef]
36. Namperumalsamy, P.; Kim, R.; Vignesh, T.P.; Nithya, N.; Royes, J.; Gijo, T.; Thulasiraj, R.D.; Vijayakumar, V. Prevalence and risk factors for diabetic retinopathy: A population-based assessment from Theni District, south India. *Br. J. Ophthalmol.* **2009**, *93*, 429–434. [CrossRef] [PubMed]
37. Raman, R.; Ganesan, S.; Pal, S.S.; Kulothungan, V.; Sharma, T. Prevalence and risk factors for diabetic retinopathy in rural India. Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetic Study III (SN-DREAMS III), report no 2. *BMJ Open Diabetes Res. Care* **2014**, *2*, e000005. [CrossRef] [PubMed]
38. Dziura, J.D.; Post, A.L.; Zhao, Q.; Fu, Z.; Peduzzi, P. Strategies for dealing with missing data in clinical trials: From design to analysis. *Yale J. Biol. Med.* **2013**, *86*, 343–358. [PubMed]

39. Shapley, L.S. *Quota Solutions op n-Person Games1*; Artin, E., Morse, M., Eds.; Princeton University Press: Princeton, NJ, USA, 1953; p. 343.
40. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
41. Williamson, B.D.; Feng, J. Efficient nonparametric statistical inference on population feature importance using Shapley values. In *International Conference on Machine Learning*; PMLR: New York City, NY, USA, 2020; pp. 10282–10291.
42. Tripathi, S.; Hemachandra, N.; Trivedi, P. Interpretable feature subset selection: A Shapley value based approach. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Virtual Event, 10–13 December 2020; pp. 5463–5472.
43. Patel, R.; Garnelo, M.; Gemp, I.; Dyer, C.; Bachrach, Y. Game-theoretic Vocabulary Selection via the Shapley Value and Banzhaf Index. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2789–2798.
44. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
45. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
46. Song, Y.-Y.; Ying, L.U. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130.
47. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
48. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: Berlin/Heidelberg, Germany, 2002.
49. Rish, I. An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop Empir. Methods Artif. Intell.* **2001**, *3*, 41–46.
50. Stratton, I.M.; Kohner, E.M.; Aldington, S.J.; Turner, R.C.; Holman, R.R.; Manley, S.E.; Matthews, D.R. UKPDS 50: Risk factors for incidence and progression of retinopathy in Type II diabetes over 6 years from diagnosis. *Diabetologia* **2001**, *44*, 156–163. [CrossRef] [PubMed]
51. Klein, R.; Klein, B.E.K.; Moss, S.E.; Davis, M.D.; DeMets, D.L. The Wisconsin Epidemiologic Study of Diabetic Retinopathy: III. Prevalence and risk of diabetic retinopathy when age at diagnosis is 30 or more years. *Arch. Ophthalmol.* **1984**, *102*, 527–532. [CrossRef] [PubMed]
52. Klein, R.; Klein, B.E.K.; Moss, S.E.; Cruickshanks, K.J. The Wisconsin epidemiologic study of diabetic retinopathy: XIV. Ten-year incidence and progression of diabetic retinopathy. *Arch. Ophthalmol.* **1994**, *112*, 1217–1228. [CrossRef]
53. Lim, M.C.; Lee, S.Y.; Cheng, B.C.; Wong, D.W.; Ong, S.G.; Ang, C.L.; Yeo, I.Y. Diabetic retinopathy in diabetics referred to a tertiary centre from a nationwide screening programme. *Ann. Acad. Med. Singap.* **2008**, *37*, 753–759. [CrossRef]
54. Abougalambou SS, I.; Abougalambou, A.S. Risk factors associated with diabetic retinopathy among type 2 diabetes patients at teaching hospital in Malaysia. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2015**, *9*, 98–103. [CrossRef]
55. Bamashmus, M.A.; Gunaid, A.A.; Khandekar, R.B. Diabetic retinopathy, visual impairment and ocular status among patients with diabetes mellitus in Yemen: A hospital-based study. *Indian J. Ophthalmol.* **2009**, *57*, 293.
56. Rani, P.K.; Raman, R.; Chandrakantan, A.; Pal, S.S.; Perumal, G.M.; Sharma, T. Risk factors for diabetic retinopathy in self-reported rural population with diabetes. *J. Postgrad. Med.* **2009**, *55*, 92.
57. Azizi-Soleiman, F.; Heidari-Beni, M.; Ambler, G.; Omar, R.; Amini, M.; Hosseini, S.-M. Iranian risk model as a predictive tool for retinopathy in patients with type 2 diabetes. *Can. J. Diabetes* **2015**, *39*, 358–363. [CrossRef] [PubMed]
58. Yao, L.; Zhong, Y.; Wu, J.; Zhang, G.; Chen, L.; Guan, P.; Huang, D.; Liu, L. Multivariable logistic regression and back propagation artificial neural network to predict diabetic retinopathy. *Diabetes Metab. Syndr. Obes. Targets Ther.* **2019**, *12*, 1943–1951. [CrossRef] [PubMed]
59. Taylor-Phillips, S.; Mistry, H.; Leslie, R.; Todkill, D.; Tsertsvadze, A.; Connock, M.; Clarke, A. Extending the diabetic retinopathy screening interval beyond 1 year: Systematic review. *Br. J. Ophthalmol.* **2016**, *100*, 105–114. [CrossRef] [PubMed]
60. Zhang, L.; Wang, Y.; Niu, M.; Wang, C.; Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. *Sci. Rep.* **2020**, *10*, 522. [CrossRef] [PubMed]
61. DeGregory, K.W.; Kuiper, P.; DeSilvio, T.; Pleuss, J.D.; Miller, R.; Roginski, J.W.; Fisher, C.B.; Harness, D.; Viswanath, S.; Heymsfield, S.B.; et al. A review of machine learning in obesity. *Obes. Rev.* **2018**, *19*, 668–685. [CrossRef] [PubMed]
62. Awan, S.E.; Sohel, F.; Sanfilippo, F.M.; Bennamoun, M.; Dwivedi, G. Machine learning in heart failure: Ready for prime time. *Curr. Opin. Cardiol.* **2018**, *33*, 190–195. [CrossRef]