computation

Article

An Improved Homogeneous Ensemble Technique for Early Accurate Detection of Type 2 Diabetes Mellitus (T2DM)

Umuhire Mucyo Faustin * and Beiji Zou

School of Computer Science and Engineering, Central South University, Changsha 410083, China; bjzou@csu.edu.cn

* Correspondence: faustinu@yahoo.com

Abstract: The objective of the present study is to improve the genetic algorithm (GA) supremacy in selecting the most suitable and relevant features within a highly dimensional dataset. This results in cost reduction and improving classification performance. During text classification, employing terms such as features using vector space representation can result in a high dimensionality of future space. This condition presents some issues, including high computation cost in data analysis and deteriorating classification accuracy performance. Several computational feature selection techniques can be applied in eliminating the least significant features within a dataset, including a genetic algorithm. The present study improved the performance of the classifier in classifying Pima Indian diabetes data. Despite the popularity of GA in the feature selection area, it does not provide the most optimal features due to one of its underlying issues: premature convergence due to insufficient population diversity in the future generations. GA was improved in its crossover operator using two steps: define a variable slice point on the size of the gene to be interchanged for every offspring generation and apply feature frequency scores in deciding the interchanging of genes. The above obtained results to the proposed technique will be better results than the results for standard GA. Our proposed algorithm attained an accuracy of 97.5%, precision of 98, recall of 97% and F1-score of 97%.

Keywords: type 2 diabetes mellitus; machine learning; homogenous ensemble; decision tree; genetic algorithm; Pima Indian Diabetes Dataset

check for updates

Citation: Faustin, U.M.; Zou, B. An Improved Homogeneous Ensemble Technique for Early Accurate Detection of Type 2 Diabetes Mellitus (T2DM). *Computation* **2022**, *10*, 104. https://doi.org/10.3390/computation10070104

Academic Editor: Demos T. Tsahalis

Received: 16 March 2022
Accepted: 10 June 2022
Published: 23 June 2022

1. Introduction

During text classification, such as medical data with machine learning or deep learning techniques, employing terms such as features using vector space representation can cause a high dimensionality of feature space and sparsity [1,2]. This kind of condition introduces some issues, including high computation cost in data analysis and reducing the classification accuracy performance [3]. The present study enhanced the classifier's performance in classifying a diabetes medical condition [4]. To address this concern [5], the majority of studies introduce feature selection techniques [6], which is a feature dimensionality reduction approach [7]. Several evolution techniques were used for feature selection subset search as a sort of optimization issue, which include particle swarm optimization (PSO), ant colony optimization (ACO), and genetic algorithm (GA) [8]. With well-optimized parameter tuning, PSO can locate optimal parts of the complicated search space and effectively traverse the search space to obtain global optima, but PSO comprises several numeral mathematical operations that need user-specified parameters with difficult-to-find optimal values [9]. ACO is inspired by ants' optimization behavior and effectively discovers optimum features by using ants' shortest path; however, it suffers from insufficient pheromone update rules. On the other hand, GA is more ideal for large-scale issues, since it replicates the process of natural selection in producing an optimal feature set. Its effectiveness in feature selection is due to its capacity to search for solutions in a large search area and its high performance in optimization problems.

Running GA as a search function in many applications, on the other hand, has certain acknowledged flaws, such as premature convergence [10]. The population diversity problem in feature selection, where simple GA starts with a varied population but after a few generations converges to a point where each chromosome tends to have similar solutions, is one of the causes that contribute to this difficulty. Premature convergence is one of the challenges in feature selection, which claims that it is one of the drawbacks when applied using basic GA. This constraint makes it difficult for the algorithm to find better answers in succeeding generations. The authors claimed that the aimless searching or simply randomized process in the crossover operator to create offspring might affect the algorithm's convergence rate. Several studies [11] applied GA feature selection and recorded inadequacy when selecting suboptimal features.

In this study, we proposed Enhanced Genetic Algorithm (EGA), which seta variable slice point on the size of the gene to be interchanged for every offspring generated and applied feature frequency scores in deciding the interchanging of genes. Our evaluation proved that our proposed technique obtained better results than standard GA.

Contributions: The main contributions of this study are the following:

(1) Improve genetic algorithm feature selection technique with improved convergence properties that better explore good solutions in the search space.
(2) Divide the dataset randomly into smaller subsets using the average splitting technique and separately model each subdivision using the decision tree classifier.
(3) Improve the performance of the proposed homogenous ensemble technique by using the accuracy ranking technique to retain excellent performing base models and reject worse performing ones.

To effectively evaluate the performance of our proposed scheme technique, the Pima Indian Diabetes Dataset (PIDD) was used. A comparative study is then conducted using well-known ensemble techniques, including XGBoost, Gradient boost and Catboost, as well as some recently published studies. Finally, our proposed scheme performance was evaluated using Accuracy, F1-score, Recall, Precision, and Auroc metrics.

2. Literature Review

In recent years, predictive classification using machine learning algorithms has been vibrant among data science researchers. The accuracy obtained by most researchers is reasonably good and acceptable (70–85%); see, for instance [12–14]. Nevertheless, there is a lot to be understood in this area. This section presents some ensemble learning techniques and genetic algorithm feature selection related to our proposed study. Machine learning techniques can be applied in a variety of disciplines due to their excellent powerful classification capabilities. Therefore, there are many research studies on diabetes prediction continuously formulating novel techniques to improve classification accuracy performance. An example of such a technique is an ensemble learning method. Perveen et al. (2016) in [15] compared the performance of Adaboost, J48 decision tree, and bagging using the Canadian primary care sentinel surveillance network medical dataset to classify diabetic and non-diabetic patients. The results obtained after the experiment using the weka data-mining toolkit show that the Adaboost ensemble technique outperforms the bagging ensemble technique and the J48 decision tree technique. Vijayan and Anjali (2016) in [11] compared decision tree, SVM and naïve Bayes as Adaboost ensemble technique base classifiers. The authors constructed integrated models based on the three mentioned base classifiers for the early prediction of diabetes. The PIDD dataset was used to compare and evaluate the performance of the proposed integrated AdaBoost classifiers. The study revealed that integrating a single machine learning base classier yielded higher performance than a single machine learning classifier. The authors in [16] applied different machine learning ensemble techniques such as bagging, AdaBoost and random forest along with the PSO feature selection method on a heart disease medical condition dataset. The experimental results show that bagging ensemble techniques outperformed other ensemble classifiers.

Maclin (1999) in [17] compares the performance of decision tree, neural network, bagging and boosting ensemble techniques.

The study's experimental results demonstrated that bagging and boosting performed better than single classifiers such as decision tree and artificial neural network. The study also revealed that bagging outperformed boosting ensemble technique. Kala et al. (2011) in [18] proposed a new, improved technique for diagnosing breast cancer using Artificial Neural Networks (ANN). This study's proposed model was used to solve breast cancer prediction using a genetic algorithm to select the best features in a dataset and Artificial Neural Network for classification. In conclusion, the above study was shown to be both efficient and scalable. In what follows, Paul and Choubey (2017) in [5] proposed a new hybrid algorithm using a genetic algorithm (GA) for selecting the most suitable features in the PIDD dataset, and in [19] the Radial Basis Function Neural Network (RBFN) was used for classifying patient with diabetes and non-diabetes. The authors concluded that the hybrid method was better than the RBFN alone. Next, the authors applied in [20] a genetic algorithm feature selection technique to find the most relevant features and eliminate the redundant features using different medical datasets. The SVM classification technique was applied. The authors reported a significant classification improvement across the datasets when GA feature selection was applied compared to when all features were used. In our current study, we aim to build and develop on what has been previously accomplished and come up with a well-performing ensemble learning technique for T2DM prediction.

3. Proposed Scheme

3.1. Dataset Description

The Pima Indian Diabetes Dataset (PIDD) was used to conduct our study, which is publicly available on the Kaggle dataset repository in CSV format [20]. This dataset contains young females of at least 21 years of age of Pima Indian heritage living around Phoenix, Arizona, in the USA. The dataset contains 768 records of patients—268 (34.9%) patients tested positive for diabetes, 500 (65.1%) patients tested negative for diabetes—and nine attributes, including class variables. The dataset attributes and statistics are presented in Table 1. This dataset reports a class imbalance that will occur when there is a great difference between minor classes and major classes in classes with binary values (0 or 1). This study's primary purpose is to predict whether a person would test positive using diabetes medical test results that are provided in the dataset. We solve a binary class problem with a class value "1" being interpreted as a patient who tested positive for diabetes. The class value "0" is interpreted as a patient who tested negative for diabetes. Our proposed scheme is shown in Figure 1.

Table 1. PIDD Dataset Description and Statistics.

Attribute Number	Attribute Name	Attribute Description	Standard Deviation	Mean	Type
1	Pregnancies	Occurrences of pregnancy	3.4	3.8	Numeric
2	Glucose	Plasma glucose concentration in a 2 h oral glucose tolerance test	32.0	120.9	Numeric
3	Blood Pressure	Diastolic blood pressure (mmHg)	19.4	69.1	Numeric
4	Skin Thickness	Triceps skin fold thickness (mm)	16.0	20.5	Numeric
5	Insulin	2 h serum insulin (mu U/mL)	115.2	79.8	Numeric
6	BMI	Weight in kg/(height in m^2)	7.9	32.0	Numeric
7	Diabetes Pedigree Function (DPF)	Diabetes pedigree function	0.3	0.5	Numeric
8	Age	Age of person (years)	11.8	33.2	Numeric
9	Outcome	Class variable (0 for non-diabetic and 1 for a diabetic)			

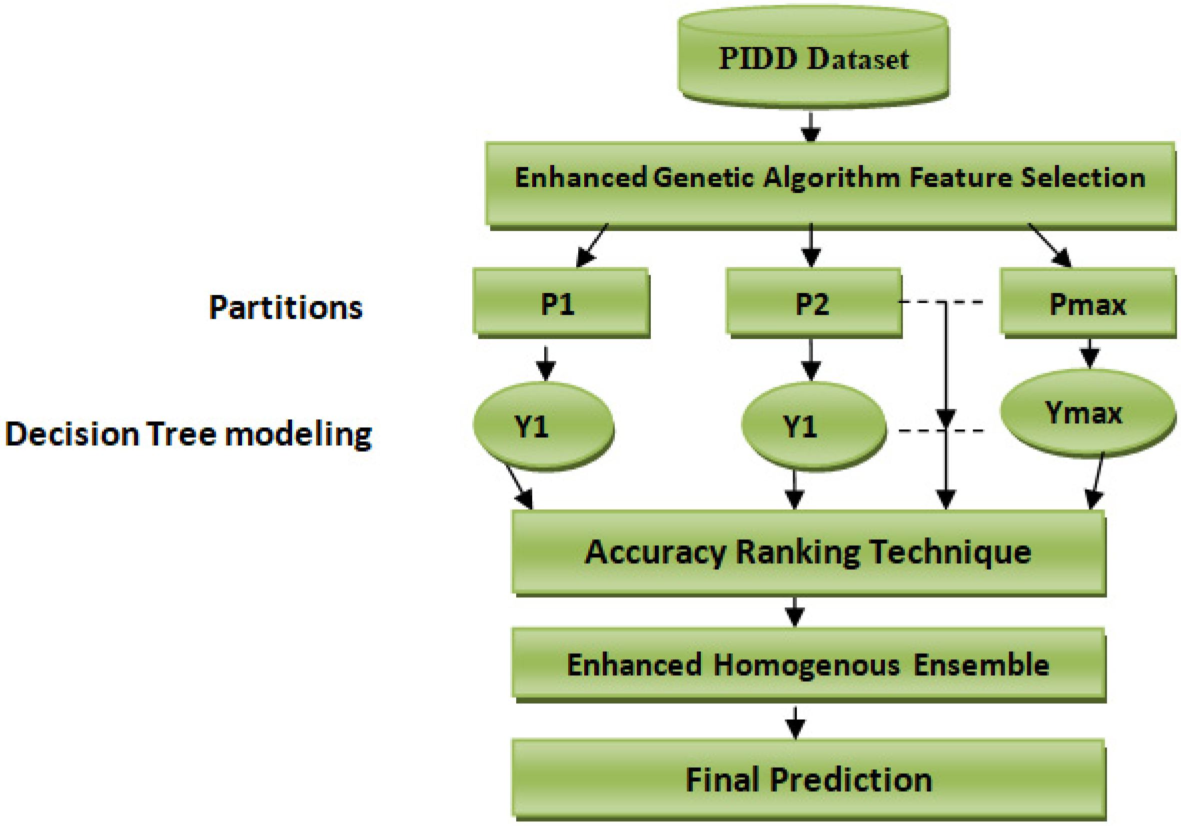
PIDD Dataset
Enhanced Genetic Algorithm Feature Selection
Partitions
P1
P2
Pmax
Decision Tree modeling
Y1
Y1
Ymax
Accuracy Ranking Technique
Enhanced Homogenous Ensemble
Final Prediction

Figure 1. Proposed Scheme.

3.2. Data Cleansing

Missing Values

Missing values imputation is a critical preprocessing approach. The Pima Indian Diabetes Dataset contains several missing and impossible figures for a normal living human being, such as 0 glucose, 0 blood pressures, 0 skin thickness, 0 insulin, and 0 BMI, indicating that the real value is missing. Missing values affect the performance of the classifier severely. So, it is essential to handle the missing values carefully. We applied K-mean clustering with a based distance data imputation technique [21].

3.3. Data Subdivision

In this section, the proposed data subdivision technique will be presented. Let dataset (PIDD) be denoted by DT, where DT = $\{(a_i + b_i), 1, 2, 3, 4 \ldots\ldots\ldots N\}$. The dataset's independent variables are denoted as:

$$a_i = \{x_{i1,}x_{i2,}x_{i3,}\ \ x_{i4} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots x_{in}\} \quad (1)$$

b_i denotes the outcome variable of the dataset where $y_i\, \epsilon \{0, 1\}$. For example, let one independent dataset variable of the different instance be:

$$a_j = \{a_{1j,}a_{2j,}a_{3j,}\ \ xa_{4j} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots a_{Nj}\} \quad (2)$$

The weighted average of a_j is computed as:

$$a_j = \frac{1}{N}\sum_{i=1}^{N} \omega\}t_i a_{ij}, \omega\}t \geq= 0,\ i = \sum_{i=1}^{N} \omega\}t_i = 1 \quad (3)$$

Data with additional weight contribute more to the weighted average than those with low weight. From Formula (3), weight can never be negative, although it can be 0.

To develop a proposed data subdivision algorithm that randomly subdivides the dataset into reduced subsets, i.e., each subset, the algorithm will randomly pick independent variable sets. The randomization technique can yield any amount of trees, and this will reduce variance and improve performance.

Let the original dataset be DT, which is the root node, and it represents the whole population, which can be divided into distinct homogenous sets. The algorithm will randomly pick a data point from variable sets (features). The data point is substituted in the set for further selection. Dataset DT is divided into two subsets using average based partitioning rules:

$$DT = \begin{cases} DT_{1,} & if\left(x_{ij} \ < x_j\right) \\ DT_{2,} & if\left(x_{ij} \ \geq x_j\right) \end{cases} \tag{4}$$

Each child is considered separately, i.e., DT_1 and DT_2 as the root node; as in (4), DT_1 generate DT_{11} and DT_{12}, while DT_2 gives DT_{13} and DT_{14}. This partitioning process will continue until the termination rule is reached. To ensure that the dataset is not over partitioned, we proposed a maximum tree height stopping technique, where H_{max} terminates the algorithm from infinite. The tree halts growing when $H = H_{max}$; in the root node, the tree height is H = 0 and H1 = 1 for DT_1 and DT_2.

After the dataset has been divided into smaller subsets, the decision tree classifier is utilized to model each partition independently. As a result, decision tree classification is characterized by a high degree of robustness and interpretability. In addition, the decision tree utilizes Gini impurity to determine the likelihood of a wrongly classified variable when randomly selected.

$$Gini = 1 - \sum_{i=1}^{J} p_1^2 \tag{5}$$

P_i is the probability of an object being classified into a particular class, and we now have a forest of T_{max} trees all fittedwith the decision tree algorithm.

3.4. Feature Selection Using Enhanced Genetic Algorithm (EGA)

GA imitates the natural selection process in obtaining the optimal feature set in a given dataset. It can search for the solution in an immense search space, and its fantastic performance in the optimization problem is due to its success in the feature selection technique. GA is well known to have a premature convergence flaw. Simple traditional GA starts with a diverse population after a small number of generations converged to a certain point where each chromosome will contain a similar solution [22].

Figure 2 displays the workflow of the GA feature selection technique. Premature convergence is the biggest challenge in feature selection, primarily when implemented using simple GA. The will hamper the algorithm from looking for a better solution in each subsequent generation. In this study, EGA was improved in crossover operator by:

- Allocating placeholder slice point on the size of the gene to be interchanged for every offspring generated.
- Applying features occurrence scores in determining the swapping of genes.

EGA is deployed as a feature subset optimizer, therefore increasing the performance of the machine learning classifiers.

EGA Procedure

Chromosome Encoding: The binary encoding of chromosomes is widely adapted. In our study, the chromosome is a bit string with the value 0 and value 1. Value 1 indicates the feature is included, while 0 means the feature is not included in the feature set with the length equal to the total number of features.

Population Initialization: A whole number from 0 and 1 is randomly populated to each chromosome with a length equal to the total number of features in the dataset.

Fitness Function: In the current population, each chromosome fitness is computed as follows:

$$fitness = (BF * acc(F)) + (1 - BF) * \left(\frac{1}{F}\right) \tag{6}$$

where F represents active or selected dataset features, and BF is the balancing factor between the features subset size and machine learning classification accuracy. Acc (F) is the measured accuracy score of the ML algorithm on the already selected feature subset.

Selection: Two parents' chromosomes selection is based on the ranking selection approach. This will ensure that chromosomes with superior fitness will not be lost in the forthcoming generation; this will give more opportunity to the supreme chromosome to mate in the pool to generate better offspring.

Crossover: The genetic algorithm is enhanced on this operator as follows:

- The crossover point was improved by adjusting a placeholder slice point for the genes to be interchanged when generating offspring.
- The crossover operator is managed and controlled by a collective feature occurrence score to decide on swapping the selected genes.

The feature subset showed by a particular cut point was calculated to obtain the cumulative frequency score. This procedure is similar to other feature subsets. Afterwards, the subset of both parents was compared. The feature subset with a more excellent score is stored as child1. The different feature subset is accumulated as child2. The variable-slicing multi-point crossover is illustrated in Figure 3.

Mutation: The offspring undergo a mutation process using mutation probability such as bit flipping.

Population Update: When a certain number of chromosomes have been reached in the population, a new population will be generated and passed to the next generation.

GA Parameters: Population size—300, Number of generation—10, Crossover operator —2-point crossover, Crossover probability—0.6, Mutation operator—bit flip, Mutation probability—0.333, Selection—Roulette wheel.

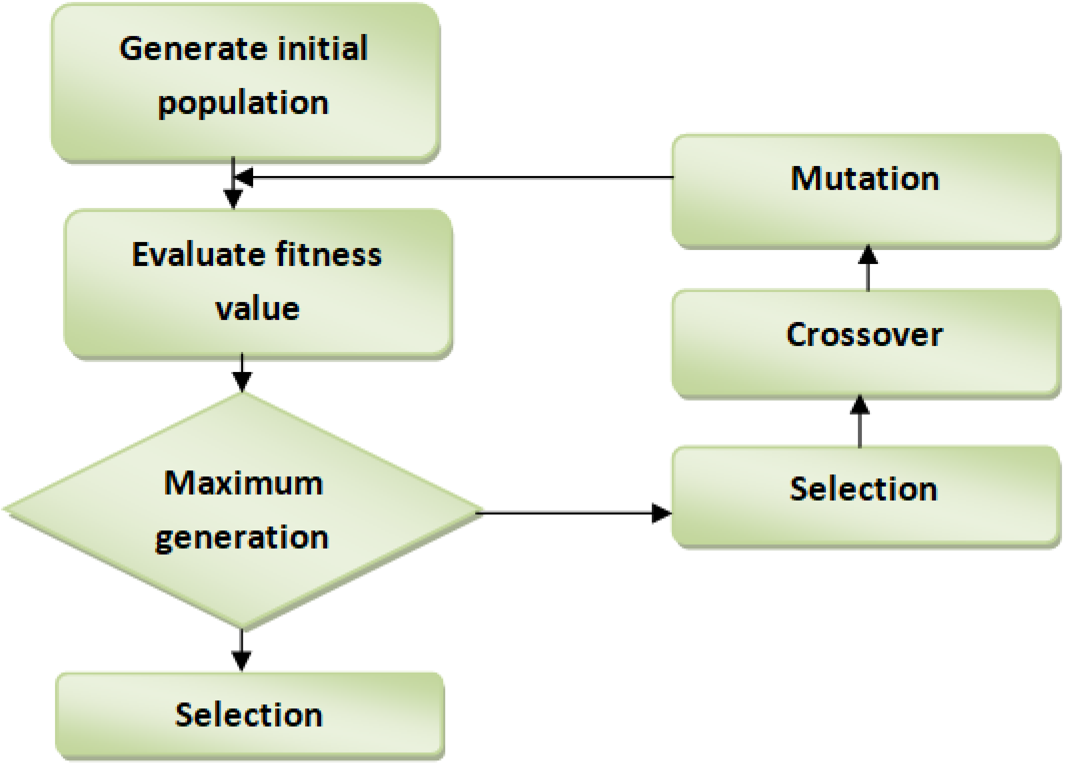

Figure 2. Workflow of feature selection using GA.

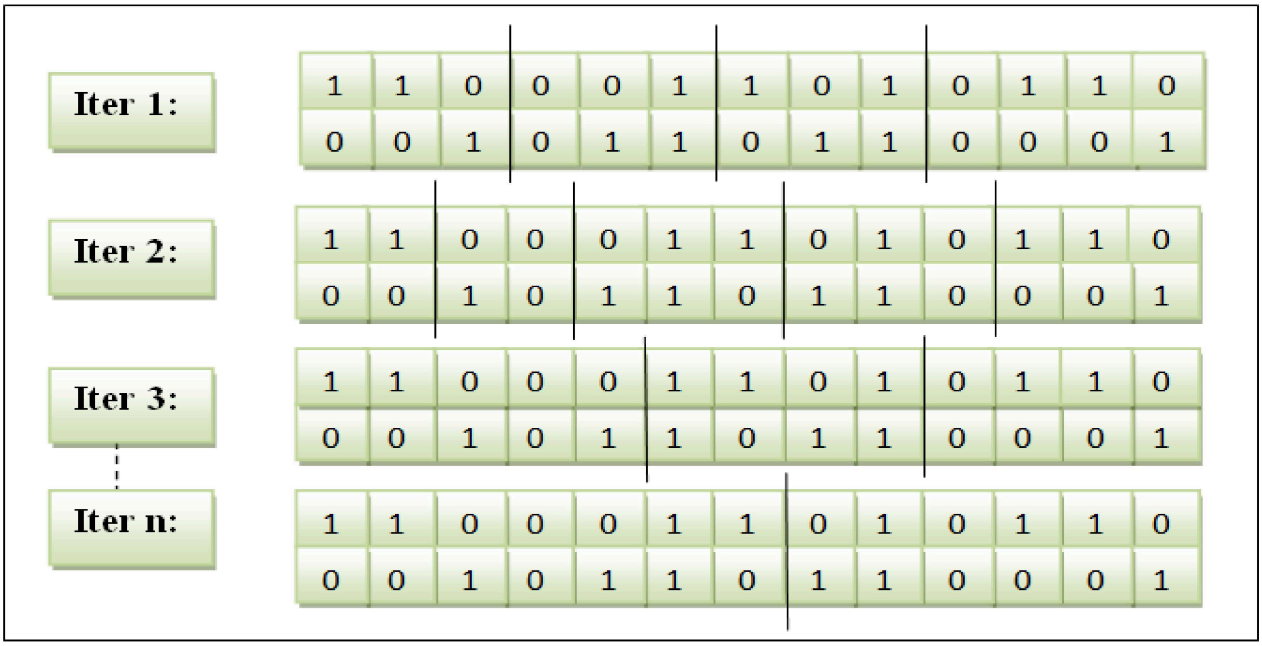

Figure 3. Proposed placeholder slice point for genes interchanging in the offspring generation cycle.

3.5. Enhanced Ensemble Algorithm

To compute a homogenous ensemble from multiple base decision tree models classifiers, first, we identified and retained the best-performing base classifiers and removed the worst-performing base classifiers. Then, a homogenous accurateness weighted aging classifier was then utilized to compute ensemble from only the best-performing decision tree models.

Let $P_a\ (\Psi_i)$ indicate the occurrence of the correct prediction of classifier Ψ_i and let Ψ_i denote the total number of iterations that the formula has been used for in the ensemble. The classifier's weight is denoted by Ψ_i and is represented as:

$$P_a\ (\Psi_i) > P_a^{\Pi}\ then\ \omega(\Psi_i) = P_a = (\Psi_i) \tag{7}$$

$$else\ \omega(\Psi_i) = \frac{P_a\ (\Psi_i)}{\sqrt{itter(\Psi_i)}} \tag{8}$$

where P_a^{Π} represents the mean accurateness of the classifiers in the ensemble technique. The final prediction of the homogeneous ensemble technique Ψ is obtained as:

$$\Psi(x) = i\ \ if \tag{9}$$

$$\sum_{t=1}^{T_{max}} \omega(\Psi_t)F_t^i(x) = max_{j\{1,2....j\}} \sum_{t=1}^{T_{max}} \omega(\Psi_t)F_t^j(x) \tag{10}$$

The proposed ensemble technique functions by assigning weights to the numerous classifier algorithms, depending on the classifier algorithm accuracy and time spent in the ensemble. If the classifier algorithm's weight drops below a determined threshold, the classifier algorithm is removed from the ensemble. By utilizing accuracy as the optimization criterion, this ensures that the ensemble technique attains the optimal results.

3.6. Model Validation

There are two main validation techniques to validate the machine learning model performance: the hold-out technique and the K-fold cross-validation technique. Each method's choice primarily depends on the magnitude of the dataset and each classification problem's goal. The hold-out technique splits data into two: that is, training data and testing data. The K-fold cross-validation technique is the technique that has been applied in this study; this is a preferred and widely applied validation technique in many other studies. In this technique, the data are split into K equivalent size of folds. K-1 groups are used for the model's training, while the rest is used for the classifier. This procedure will be iterated until each fold of 10 folds has been used as a testing set and the same case as to each k. The accuracy of the classifier is computed [23]. The final evaluation is computed based on the accuracy mean. Our study applied k = 10, representing 90% of the training data and 10% used for testing. This validation method has the following merits:

- Decreases the variance in the prediction errors.
- Reduces overfitting and overlapping of data between training and testing.

4. Experimental Result

4.1. Experimental Setup

This section explains the experimental results obtained after evaluating our proposed enhanced genetic algorithm for optimal feature selection and improved homogenous ensemble learning for classification in detail. The experiments were conducted on a 3.3 GHz Intel dual-core i3 processor with 8 GB of RAM running the Windows 10 operating system to evaluate the proposed scheme's performance. A Jupyer notebook (3.7.6) was used for implementation, and Python Programming language (3.8) was used to analyze the dataset and classify T2DM. Various libraries, such as genetic algorithm, numpy, pandas, matplotlib, and scikit-learn, were used to perform different tasks. The simulation was performed to evaluate our proposed scheme against other ensemble techniques such as XGBoost, Gradient boost and CatBoost.

4.2. Performance Evaluation Metrics

This section discusses the performance evaluation of the improved proposed homogenous ensemble technique for effectiveness and efficiency in the early diagnosis and prediction of T2DM. The performance evaluation of any machine learning classifiers is measured in terms of accuracy metrics; however, relying on only accuracy metrics could be misleading from time to time. Therefore, we used other metrics, such as precision, recall, F1-score, and area under the curve (AUC), and a 10-fold cross-validation technique to evaluate our proposed binary classification model's performance.

The efficiency of the machine learning classifier algorithms is evaluated using parameters that are obtained from the confusion matrix, which is True Positive (TP), which means T2DM is detected as T2DM, True Negative (TP), which means Normal predicted as Normal, False Positive (FP) which means Normal is predicted as T2DM, and finally False Negative (FN), which means T2DM predicted as Normal.

Accuracy is defined as the metrics determining the number of correctly classified classes from the total samples in the testing dataset. The accuracy metric is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

Precision is the total number of correct positive results obtained divided by the total number of positive results predicted by the classifier algorithm. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

Recall: The total number of correct positive results is obtained divided by the total number of all relevant samples. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

F1-Score: It is used to evaluate a test's accuracy. The F1-score is the combination Mean between Precision and Recall metrics. The range for the F1-score is [0, 1]. It tells you how accurate and robust your classifier is. It is calculated as follows:

$$F1 = 2* = \frac{1}{\left(\frac{1}{Precision}\right) + \left(\frac{1}{Recall}\right)} \tag{14}$$

AUC (Area Under Curve) is the performance metrics founded on the varying threshold values for binary classification problems. The AUC metric shows the model's ability to distinguish the binary classes. The higher the AUC, the better the model. It is calculated by plotting the TPR (True Positive Rate), i.e., sensitivity or recall vs. FPR (False Positive Rate), i.e., specificity, at various threshold values.

4.3. GA Convergence Properties

This section gives a detailed analysis of EGA and GAMP convergence properties given the same GA parameter settings, as shown in Table 2. In Figure 4, it can be observed that the proposed enhanced genetic algorithm made a consistent improvement over GAMP, especially after the 40th generation. GA-MP tried to compete with EGA on several generations, and after reaching the 20th generation, it started experiencing the exhibition. As a result of the exhibition, loss of diversity was encountered, which eventually led to not exploring the global search space; therefore, they cannot produce offspring better than their parents, resulting in an early convergence.

Table 2. Best parameters settings that attained highest Accuracy in EGA.

Parameter	**Value**
Generation	100
Population	200
Mutation probability	0.05
A trade-off between the number of features and classification accuracy	0.85
Crossover type	EGA—placeholder slice point GAMP-Multi-Point

EGA produced consistent improvement throughput; therefore, it was able to make better offspring, which are passed to the succeeding generation, and it explored the search space better. EGA achieved a better convergence property than standard GA feature selection, thus choosing the most optimal features within a dataset, which improves classification performance.

Table 3 displays selected features using EGA and GAMP techniques where EGA picked the five most optimal features, while GA-MP chose the six features.

Table 3. Best parameters settings that attained highest Accuracy in EGA.

GAMP Selected Features	**EGA Selected Features**
Glucose	Glucose
Age	Insulin
BMI	BMI
Pregnancies	Diabetes Pedigree Function
Diabetes Pedigree Function	Age
Blood Pressure	

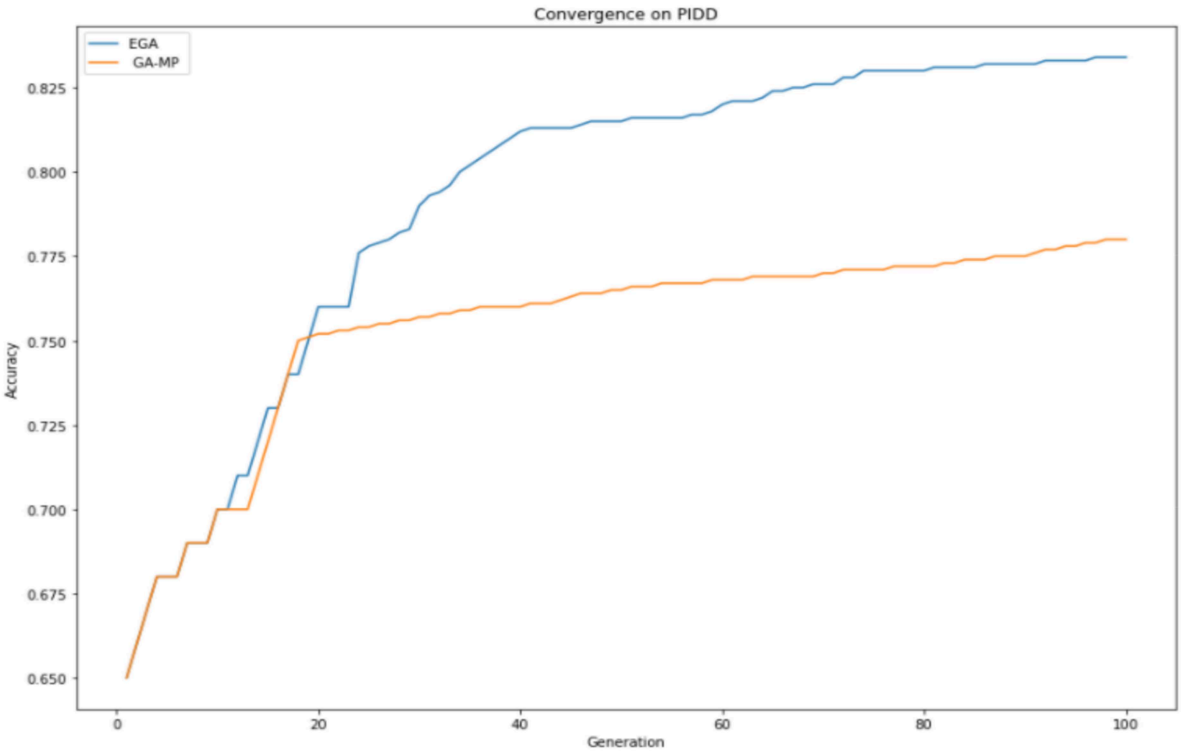

Figure 4. Convergence on PIDD.

4.4. Classification Results

The primary purpose of this study is to present an effective and efficient scheme for the early detection and prediction of T2DM using the machine learning technique. To validate our proposed enhanced homogenous ensemble scheme effectiveness and efficiency, a comparative study was conducted with other well-known ensemble techniques such as XGBoost, Gradient Boost and CatBoost.

From Table 4, it can be observed that our enhanced homogenous ensemble classifier outperformed other conventional ensemble techniques, with an accuracy of 94.47%, Precision of 94%, Recall of 94% and F1-score of 95%, while Gradient Boost classifier registered the lowest classification accuracy. The high performance was achieved due to the application of the accuracy ranking technique, which eliminates the least performing base classifiers, hence improving classification performance.

Table 4. Evaluation performance of classifiers algorithms models before EGA feature selection.

Algorithm	Accuracy	Precision	Recall	F1-Score
XGBoost	90.36	89	88	89
Gradient Boost	86.90	86	86	87
Cat Boost	88.53	88	88	89
Enhanced Homogenous Ensemble	94.47	94	94	95

Table 5 indicates that the performance of the classifier algorithms has been improved when the feature selection techniques have been applied as compared to when using all features provided by the dataset. EGA found an optimal features subset that recorded the highest classification accuracy compared to GA-MP; this is because EGA has improved convergence properties, which will maintain population diversity, proving the ability to find a better solution in search space and end up obtaining the most optimal feature selection.

Table 5. Evaluation performance of classifiers algorithms models after EGA feature selection.

Algorithm	EGA Feature Selection				GA-MP Feature Selection			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
XGBoost	93.25	92	93	92	90.89	91	90	91
Gradient Boost	91.22	90	90	91	88.16	88	87	88
Cat Boost	92.46	92	92	93	89.76	89	89	90
Enhanced Homogenous Ensemble	97.5	98	97	97	95.27	94	95	95

Our proposed scheme, which combines the improved homogenous ensemble technique and improved genetic algorithm feature selection, outperformed other well-known classification algorithms with an Accuracy of 98.75%, Precision of 98%, Recall of 98% and F1-score of 99%.

Figure 5 further evaluated the performance of our proposed algorithm using AUROC curve. Our approach outperformed other algorithms by scoring 98% while gradient boost was the last ranked algorithm with 91.6%.

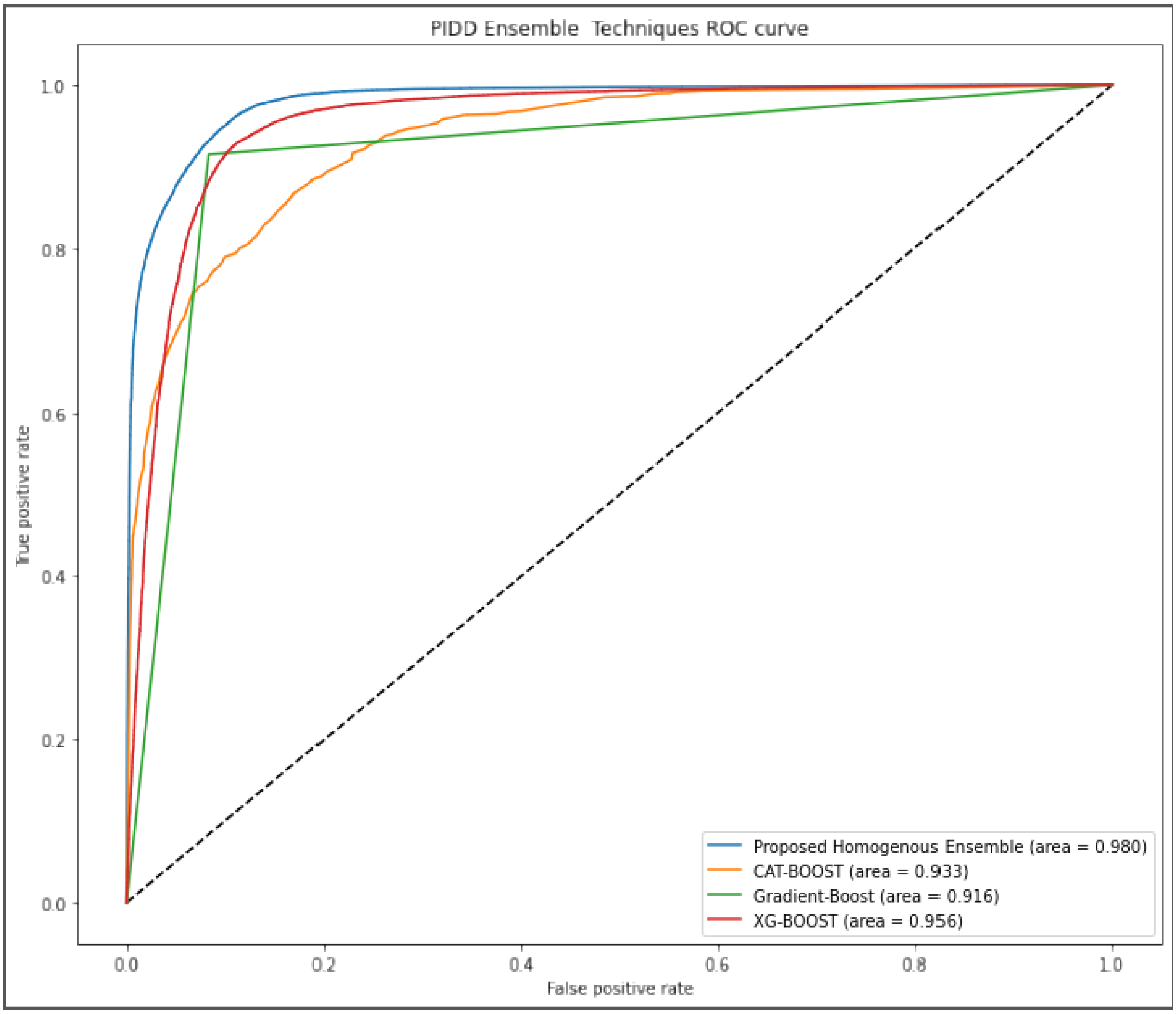

Figure 5. PIDD ROC curve: proposed homogenous against other ensemble techniques.

A Comparative Study with Other Recently Published Studies

We carried out a comparative study between our proposed scheme and some recently published related studies. The experimental results indicate that our proposed scheme outperformed other well-known ensemble techniques. A comparative study was conducted with recently published studies that utilized the PIDD dataset to demonstrate our proposed scheme further. Table 6 displays the comparison; as it can be observed, our proposed scheme technique outperformed other recently published studies.

Table 6. Comparison of our proposed ensemble technique with other recently published studies.

Author/Year	Dataset	Method	Accuracy
(Nnamoko and Korkontzelos, 2020)	PIDD	C4.5 (IQRD +SMOTED) Validation: 10-fold CV	89.5%
(Ramezani et al., 2018)	PIDD	Logistic Adaptive Network based Fuzzy Inference System (LANFIS) Validation: 3-fold CV	88.05%
(Wu et al., 2018)	PIDD	K-means + Logistic regression Validation: 10-fold CV	95.42%
(Mahajan et al., 2017)	PIDD	PCA + ANN Validation: 10-fold CV	92.2%
(Polat et al., 2008)	PIDD	Ensemble of Generalized Discriminant Analysis (GDA) and LS-SVM Validation: 10-fold CV	82.0%
(Alirezaei et al., 2019)	PIDD	KNN + K-means + MOPSO + SVM Validation: 10-fold CV	94.64%
Proposed Approach	PIDD	EGA + Enhanced Homogenous Ensemble Validation: 10-fold CV	97.5%

The experimental results have demonstrated that our proposed scheme has significantly improved the classification performance.

The proposed scheme is a two-step process:

- Enhanced genetic algorithm optimal feature selection, with a better convergence characteristic which can maintain population diversity, is the key to its ability to find the better solution in the search spaces, therefore selecting the most optimal features in a dataset.
- Enhanced homogenous ensemble technique is achieved by randomly portioning dataset into smaller subsets using the average based splitting technique; the subsets are modeled by decision tree classifiers individually.

The homogenous ensemble is generated using decision tree models as base classifiers. We applied the accuracy ranking technique to expel poor ranking classifiers and retain outstanding performing classifiers. These two novel techniques boosted our proposed scheme to achieve the best results as compared to other studies.

The majority of the studies represented in Table 6 did not consider the feature selection approach, and few studies considered the feature selection technique. Their studies did not offer the best classification results because their proposed technique is incapable of selecting the most optimal features within the dataset. The EGA feature selection approach is the reason behind the best classification performance observed by our proposed approach.

Most of the feature selection approaches suffer from premature convergence and cannot select the most optimal features. Our proposed technique involves enhancing genetic algorithm feature selection using two steps: setting a variable slice point on the size of the gene to be interchanged for every offspring generation and applying feature frequency scores in deciding the interchanging of genes. Our experimental results have proven that it is the most efficient in selecting the most optimal features, therefore outperforming other proposed techniques.

4.5. Discussion on the Results

The improved genetic algorithm feature selection with enhanced convergence properties gave an efficient optimal feature subset compared with the standard genetic algorithm. An enhanced homogenous ensemble technique was established to predict T2DM patients accurately. The approach involves randomly subdividing the dataset into smaller subsets using the average-based splitting technique. The smaller subsets are then modeled individually using decision tree classifiers. Homogenous ensemble classifiers then developed from different decision tree models; by applying the accuracy ranking technique, we retained the excellent performing base classifiers and eliminated the least-performing base classifiers. Our experimental results and analysis using the Pima Indian Diabetes Dataset (PIDD) attained an Accuracy of 97.5%, Precision of 98%, Recall of 97%, and F1-score of 97%. In addition, our study also revealed that Glucose, Insulin, BMI, Diabetes Pedigree Function and Age are the leading indicators ofT2DM.

The results show that T2DM high risks can be predicted effectively and efficiently using our proposed scheme. Therefore, our technique can be applied in real diabetes diagnostic centers for the clinical decision-making process.

5. Conclusions

This study proposes a novel method of predicting T2DM involving a two-stage process: enhanced genetic algorithm feature selection (EGA) and enhanced homogenous ensemble classification technique.

The EGA feature selection technique was achieved to attain the most relevant features and discard the redundant features. Our experimental results and analysis introduced the effectiveness of EGA in dimensionality reduction, therefore raising classification performances compared to the normal standard GA (GA-MP) algorithm; this was attributed to the capacity of EGA to sustainpopulationdiversitywhichisthekeytoitsabilityinexploringbettersolutionsinthesearch. Meanwhile, the enhanced homogenous ensemble classifier was able to improve classifier performance due to its ability to eliminate the worse-performing classifiers.

Future Work

It is crucial to bring in huge real hospital data for the continuous training and optimization of our proposed scheme. In addition, it is also necessary to validate our proposed scheme using other chronic medical condition datasets.

Author Contributions: Conceptualization, U.M.F.: methodology, implementation and manuscript drafting and B.Z. did the supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are freely available from the Kaggle dataset repository at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed on 28 April 2022).

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

References

1. Younus, M.; Munna, T.A.; Alam, M.M.; Allayear, S.M.; Ara, S.J.F. Prediction Model for Prevalence of Type-2 Diabetes Mellitus Complications Using Machine Learning Approach. In *Data Management and Analysis*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 103–116. [CrossRef]
2. Barik, S.; Mohanty, S.; Mohanty, S.; Singh, D. Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques. *Intell. Cloud Comput.* **2020**, *2*, 399–409. [CrossRef]
3. Lixandru-Petre, I.-O. A Fuzzy System Approach for Diabetes Classification. In Proceedings of the 2020 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 29–30 October 2020; pp. 1–4. [CrossRef]
4. Alam, T.M.; Iqbal, M.A.; Ali, Y.; Wahab, A.; Ijaz, S.; Baig, T.I.; Hussain, A.; Malik, A.; Raza, M.M.; Ibrar, S.; et al. A model for early prediction of diabetes. *Inform. Med. Unlocked* **2019**, *16*, 100204. [CrossRef]
5. Sarwar, A.; Ali, M.; Manhas, J.; Sharma, V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *Int. J. Inf. Technol.* **2018**, *12*, 419–428. [CrossRef]
6. Shah, J.; Patel, R. Classification techniques for Disease detection using Big-data. In Proceedings of the 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 13–14 December 2019; pp. 140–145. [CrossRef]
7. Raja, J.B.; Pandian, S.C. PSO-FCM based data mining model to predict diabetic disease. *Comput. Methods Programs Biomed.* **2020**, *196*, 105659. [CrossRef] [PubMed]
8. Faruque, F.; Asaduzzaman; Sarker, I.H. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 7–9 February 2019; pp. 1–4. [CrossRef]
9. Deepika, P.; Sasikala, S. Enhanced Model for Prediction and Classification of Cardiovascular Disease using Decision Tree with Particle Swarm Optimization. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; pp. 1068–1072. [CrossRef]
10. Wu, Y.; Liu, L.; Xie, Z.; Bae, J.; Chow, K.-H.; Wei, W. Promoting High Diversity Ensemble Learning with Ensemble Bench. In Proceedings of the 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI), Atlanta, GA, USA, 28–31 October 2020; pp. 208–217. [CrossRef]
11. Colangelo, P.; Segal, O.; Speicher, A.; Margala, M. Artificial Neural Network and Accelerator Co-design using Evolutionary Algorithms. In Proceedings of the 2019 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 24–26 September 2019; pp. 1–8. [CrossRef]
12. Peng, C.; Wu, X.; Yuan, W.; Zhang, X.; Zhang, Y.; Li, Y. MGRFE: Multilayer Recursive Feature Elimination Based on an Embedded Genetic Algorithm for Cancer Classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *18*, 621–632. [CrossRef] [PubMed]
13. Fazakis, N.; Kocsis, O.; Dritsas, E.; Alexiou, S.; Fakotakis, N.; Moustakas, K. Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction. *IEEE Access* **2021**, *9*, 103737–103757. [CrossRef]
14. Li, Z.; Giorgetti, A.; Kandeepan, S. Multiple Radio Transmitter Localization via UAV-Based Mapping. *IEEE Trans. Veh. Technol.* **2021**, *70*, 8811–8822. [CrossRef]
15. Nguyen, M.H.; Le Nguyen, P.; Nguyen, K.; Le, V.A.; Nguyen, T.-H.; Ji, Y. PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model. *IEEE Access* **2021**, *9*, 57338–57350. [CrossRef]
16. Wong, T.-T.; Yeh, P.-Y. Reliable Accuracy Estimates from *k*-Fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1586–1594. [CrossRef]
17. He, X. Statistical Interpretation and Modeling Analysis of Multidimensional Complicated Computer Data. In Proceedings of the 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 29–31 July 2021; pp. 747–751. [CrossRef]
18. Nnamoko, N.; Korkontzelos, I. Efficient treatment of outliers and class imbalance for diabetes prediction. *Artif. Intell. Med.* **2020**, *104*, 101815. [CrossRef] [PubMed]
19. Driss, K.; Boulila, W.; Batool, A.; Ahmad, J. A Novel Approach for Classifying Diabetes' Patients Based on Imputation and Machine Learning. In Proceedings of the 2020 International Conference on UK-China Emerging Technologies (UCET), Glasgow, UK, 20–21 August 2020; pp. 1–4. [CrossRef]
20. Khan, A.A.; Qayyum, H.; Liaqat, R.; Ahmad, F.; Nawaz, A.; Younis, B. Optimized Prediction Model for Type 2 Diabetes Mellitus Using Gradient Boosting Algorithm. In Proceedings of the 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, 15–17 July 2021; pp. 1–6. [CrossRef]
21. Miazi, Z.A.; Jahan, S.; Niloy, A.K.; Roknuzzaman; Shama, A.; Rahman, Z.; Islam, R.; Badal, F.R.; Das, S.K. A Cloud-based App for Early Detection of Type II Diabetes with the Aid of Deep Learning. In Proceedings of the 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), Rajshahi, Bangladesh, 8–9 July 2021; pp. 1–6. [CrossRef]
22. Mansour, R.F.; El Amraoui, A.; Nouaouri, I.; Diaz, V.G.; Gupta, D.; Kumar, S. Artificial Intelligence and Internet of Things Enabled Disease Diagnosis Model for Smart Healthcare Systems. *IEEE Access* **2021**, *9*, 45137–45146. [CrossRef]
23. Alirezaei, M.; Niaki, S.T.A. A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. *Expert Syst. Appl.* **2019**, *127*, 47–57. [CrossRef]