A Prediction Model for Type 2 Diabetes Using a Genetic Algorithm Feature Selection Method and XGBoost Classifier

Danah Albaaj

Abstract

Type 2 diabetes is a chronic illness with both a high prevalence and high incidence rate worldwide. Individuals who are left undiagnosed or untreated face a high risk of severe health complications, which can eventually lead to death. It is thus vital that proper methods for identifying these high-risk individuals are created. Recently, the application of predictive modeling for identifying patients at risk of type 2 diabetes has shown great promise as access to electronic health records continues to increase. However, the performance and generalizability of current predictive models for type 2 diabetes still require improvements. This paper proposes a hybrid risk prediction model based on an evolutionary computation feature selection and ensemble learning method given little literature exists regarding the use of evolutionary computation methods in creating a type 2 diabetes prediction model. In the proposed model, a genetic algorithm is wrapped with an extreme gradient-boosting (XGBoost) classifier for optimal feature selection before being passed into an XGBoost classifier. The proposed model was tested on the UCI Pima Indians Diabetes Dataset (PIDD) and then evaluated by comparing the results of the various performance metrics against previous models that had used the data set. Although the results did not show a significant increase in predictive performance over previous methods, the proposed model's results compared to baseline preprocessed model indicate future research is necessary to understand the full potential of this method.

Acknowledgements

I would like to express my thanks to the people who have helped me most throughout my project. A special thanks of mine goes to my mentor, Alan Garfinkel, who helped me in completing the project by sharing his thoughts, offering critical feedback and always pushing me to look beyond the surface.

I also wish to thank my brother for his personal support and attention, his continued willingness to be my second set of eyes, and for inspiring me to go my own way.

Finally, I wish to thank my mother. I wouldn't be where I am today without her endless encouragement and support. Thank you, Mom.

Contents

1 Introduction

Type 2 diabetes is a chronic disease characterized by high blood glucose levels stemming from the body's resistance to insulin produced. According to the World Health Organization's [14] recent figures, approximately 383 million individuals have Type 2 diabetes, with an expected increase of near 50% prevalence within the next 25 years. Early prevention and intervention are crucial to lowering the risk of complications such as heart disease, kidney disease and failure, neuropathy, and diabetic retinopathy. In light of this, identifying methods for predicting these high-risk patients is a matter of global public health.

The use of data mining has proven useful in predicting patient outcomes and diagnosis, with a plethora studies producing risk prediction models for type 2 diabetes over the past two decades. These models play an important role in the clinical setting as they provide clinicians the ability to extract hidden patterns and useful knowledge from patient data that may have gone unnoticed otherwise. Moreover, recent research and the growth of type 2 diabetes prevalence indicates a continued need for new models with greater performance accuracy. Fortunately, increased access to electronic patient electronic health records provides researchers with an opportunity to create new models and revise the current prediction methods in use. [8] Much of the literature's current focus is on applying various feature selection methods with classification algorithms (most often SVM, ANN, DT) to determine which methods provide an increase from the current threshold of 80%.[8]

Evolutionary algorithms are at the forefront of techniques to show promise for medical data mining. Evolutionary algorithms are a class of technique inspired by the process of evolution; thus its mechanisms apply the concepts of natural selection and survival-of-the-fittest to solve real problems. Genetic algorithms (GA) are one popular approach within evolutionary programming that researchers Tan [17] and Karegowda et al. [5], [6] have shown as a viable feature selection method in the preprocessing stage. In their work, Karegowda et al. [5] demonstrates how using GAs for feature selection improves the accuracy of classifiers. When applied to the PIDD, the GA improved the classification accuracy of Naive Bayes and C4.8 to 86.47% and 85.71% respectively. Karegowda et al.'s [6] subsequent work combined a GA with a back propagation network (BPN) to the same data set, producing 84.71% predictive performance, an improvement from the 79.5% of the BPN alone. Here, the GAs utilize the concepts of natural selection to include randomness when solving for the optimal feature subset. This makes it possible for the models to incorporate a truer understanding of how diseases and the human population interact with one another.

In their proposed model, Xu and Wang used a weighted features selection method based on random forest (called RF-WFS) to select the optimal feature subset of the PIDD before passing that subset into an extreme-gradient boosting (XGBoost) classifier. Prior to the feature selection method, preprocessing and XGBoost had a 88.28% predicitive accuracy, while the complete model had a 93.75% accuracy. Both stages outperformed the previous studies with popular classification algorithms. [19] Given these results, it is possible that the combination of a GA feature selection method with the XGBoost classifier will provide an increase in performance. This work presents a hybrid prediction model using the GA wrapper method proposed by Karegowda [5] to identify the optimal features subset of the given Pima Indians Diabetes Dataset (PIDD) and XGBoost classifier.

2 Methods

In this study, we proposed a new diabetes risk prediction model. Firstly, k-nearest neighbors imputation was used to replace missing values of the original diabetes data set. Then, the optimal feature subset was selected from the updated data set by a genetic algorithm wrapper method which utilized the XGBoost classification accuracy to determine fitness scores. Once the optimal feature subset was identified, an XGBoost classifier was constructed for diabetes risk prediction. The model was then evaluated using 10-fold cross validation. The flow chart for the model building is in Figure 1.

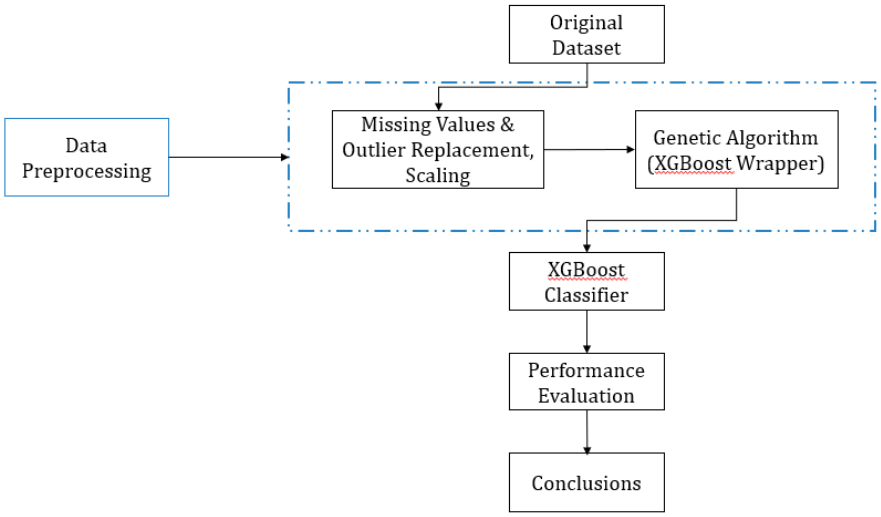


Figure 1: Flowchart of Model Building

2.1 Data Preprocessing

This experiment was performed using the Pima Indians Diabetes Data (PIDD) set, a commonly used publicly available data set from the University of California, Irvine (UCI) machine learning repository [1]. The PIDD set contains information of 768 females of Pima Indian heritage. The outcomes of the data set represent whether the patient is diabetic or not, with an imbalanced distribution of 268 "tested_positive" cases and 500 "tested_negative" cases. The data set contains a total of eight numeric attributes related to the diagnosis of Type II diabetes. Detailed information on the attributes found in the data set can be found in Table 1 and Figure 2.

Table 1: Dataset Description

Number	Attribute	Description
1	Pregnancies	No. of times pregnant
2	Glucose	Plasma glucose conc. 2 hrs in an oral glucose tolerance test
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-hour serum insulin
6	BMI	Body mass index (kg/m^2)
7	Pedigree	Diabetes Pedigree Function
8	Age	Years

	Tested Negative (N=500)	Tested Postive (N=268)	Overall (N=768)
Pregnancies			
Mean (SD)	3.30 (3.02)	4.87 (3.74)	3.85 (3.37)
Median [Min, Max]	2.00 [0, 13.0]	4.00 [0, 17.0]	3.00 [0, 17.0]
Glucose			
Mean (SD)	110 (26.1)	141 (31.9)	121 (32.0)
Median [Min, Max]	107 [0, 197]	140 [0, 199]	117 [0, 199]
BloodPressure			
Mean (SD)	68.2 (18.1)	70.8 (21.5)	69.1 (19.4)
Median [Min, Max]	70.0 [0, 122]	74.0 [0, 114]	72.0 [0, 122]
SkinThickness			
Mean (SD)	19.7 (14.9)	22.2 (17.7)	20.5 (16.0)
Median [Min, Max]	21.0 [0, 60.0]	27.0 [0, 99.0]	23.0 [0, 99.0]
Insulin			
Mean (SD)	68.8 (98.9)	100 (139)	79.8 (115)
Median [Min, Max]	39.0 [0, 744]	0 [0, 846]	30.5 [0, 846]
BMI			
Mean (SD)	30.3 (7.69)	35.1 (7.26)	32.0 (7.88)
Median [Min, Max]	30.1 [0, 57.3]	34.3 [0, 67.1]	32.0 [0, 67.1]
DiabetesPedigreeFunction			
Mean (SD)	0.430 (0.299)	0.551 (0.372)	0.472 (0.331)
Median [Min, Max]	0.336 [0.0780, 2.33]	0.449 [0.0880, 2.42]	0.373 [0.0780, 2.42]
Age			
Mean (SD)	31.2 (11.7)	37.1 (11.0)	33.2 (11.8)
Median [Min, Max]	27.0 [21.0, 81.0]	36.0 [21.0, 70.0]	29.0 [21.0, 81.0]

Figure 2: Summary Table of Initial Data

An initial investigation into the data revealed five attributes Glucose, BloodPressure, Insulin, BMI, and SkinThickness contain 5, 35, 374, 11, and 227 zeroes, respectively. These values are physically impossible and must be removed before the zeros can influence classification tree size to create biased results. [15] Then, we will make use of k-nearest neighbors (kNN) imputation methods which identify the 'k' best values for imputation based on the overall similarity the 'k' samples have with the missing sample. Nearest neighbor imputation provides improvements in predictive performance due to its ability to maintain the data structure. [2] It is important to note that kNN imputation requires the data be scaled since the ranges of the given data are not the same. However, using standard scaling methods would cause a significant loss of our data since many of the features contain outliers relevant for diagnosis. Given the mean is heavily influenced by outliers in the data, we must use a method such as RobustScaler() in Python's sci-kit learn package. This method allows us to identify outliers in a specific quantile range of the data before imputing them with the median. [13] Finally, given the numerical value of the pregnancies attribute holds little connection to type 2 diabetes, the attribute was binarized so 0 represents patients who have not been pregnant and 1 represents those who have been pregnant. [18]

2.2 Genetic Algorithm and XGBoost Wrapper Feature Selection

The next step of the model was to determine which attributes were key in classifying diabetes. Feature selection has been demonstrated to have numerous advantages such as removing

redundant data, improving data quality, and performance improvement in speed and accuracy. [9] Feature selection methods are classified into three categories: filters, wrappers, and hybrid. Filter methods are performed independent of any machine learning algorithms. Filter methods are also considered suitable for high dimensional data sets due to their low computational cost and speed but are unreliable with classification tasks. Wrapper methods are performed in conjunction with machine learning methods. This makes the wrapper methods slower than filter methods but better suited for classification tasks due to their use of classifiers during the feature selection process. Hybrid methods attempt to combine the positive aspects of both filter and wrapper methods. Given that the data set is low-dimensional, we will see better classification accuracy using a wrapper method. However, for higher dimensional data sets one should consider the use of a filter or hybrid method.

This experiment will be investigating the effectiveness of a genetic algorithm (XGBoost) wrapper method in extracting the optimal feature subset. The genetic algorithm is a random search algorithm that generates weighted feature subsets to reduce redundant data, it's process is represented in Figure 3. The following steps were then used to create the wrapper method:

1. Create random initial population using Boolean arrays to determine the features used,
2. Enumerate over fixed number of generations,
3. Determine the fitness of each individual in this generation using the classification accuracy for each individual in the population,
4. Select n fittest individuals to pass their 'genes' to next generation,
5. Randomly select how each set of parents will exchange their genes to create new individual,
6. Determine if new individual will have a mutation with probability = m,
7. Loop over list of selected parents and generate children, add them to current population.

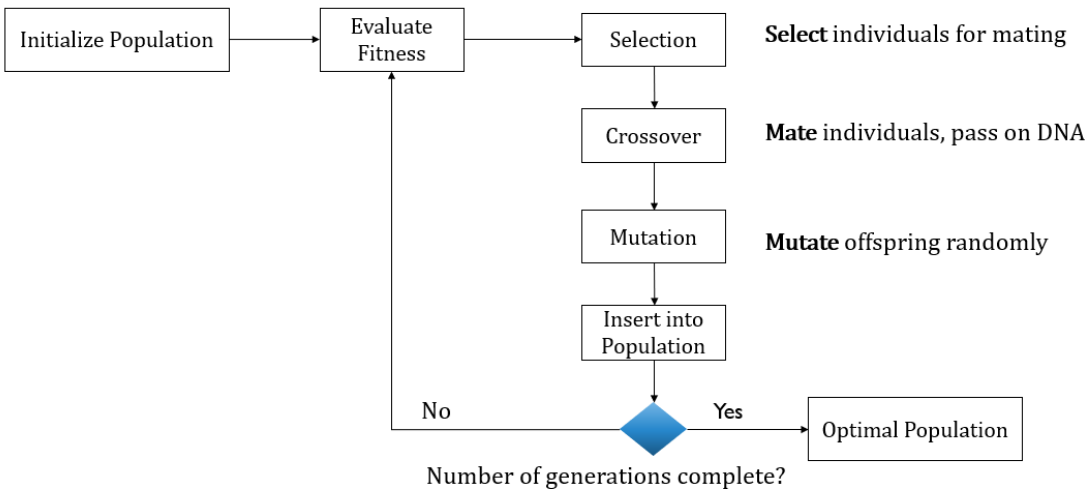


Figure 3: Visual Representation of Genetic Algorithm

2.3 XGBoost Classifier

XGBoost is an ensemble learning algorithm known for being an improved version of the gradient boosting algorithm. Boosting combines the power of weak learners with high bias and weaker predictive power to generate a stronger learner that reduces the bias and variance. The trees made through boosting contain fewer splits, making them highly interpretable. In XGBoost, classification and regression trees (CART) are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals. Parameters like the number of trees or iterations, the rate at which the gradient boosting learns, and the depth of the tree, could be optimally selected through validation techniques like k-fold cross validation. XGBoost is unique in its scalability, speed, and ability to handle sparse data, parallel learning, and option to penalize models for complexity using L1 and L2 regularization to prevent overfitting. [4]

2.4 Performance Evaluation

Receiver Operating Characteristic Area Under the Curve (ROC AUC) determines the classification performance of the model on the positive class by examining the false positive rate vs true positive rate at various points. The predictive capability is represented by the area under this curve. This is one of the most common metrics used to determine the performance of a model, making it useful in our comparisons of the proposed model vs previous models. [12] However, due to the class imbalance present in our data and the risks faced due to a false negative rate, it is useful to also examine the model's performance using the precision-recall curve. Note, precision quantifies the number of correct positive predictions made and recall quantifies the number of correct positive predictions made from the total positive predictions available. [3] In addition, we will use stratified k-fold cross validation to get a more accurate measure of the generalizability of the proposed model. Stratified k-fold cross validation does this by dividing the data into k subsets, with k-1 subsets representing our training set and the remainder is the test set. The stratification maintains the class balance of the original data set within each subset of the k folds. The model training and testing is then carried out on these subsets k times, with the final prediction result being the average of the test results for the model.[11]

3 Experimental Results

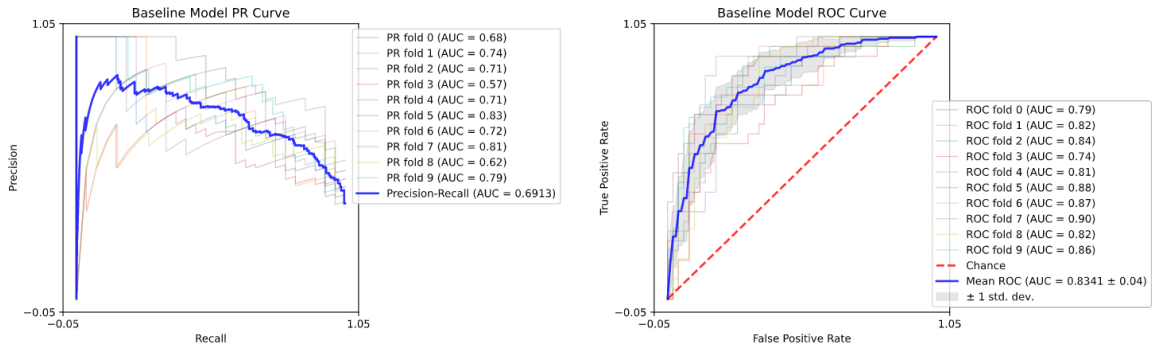
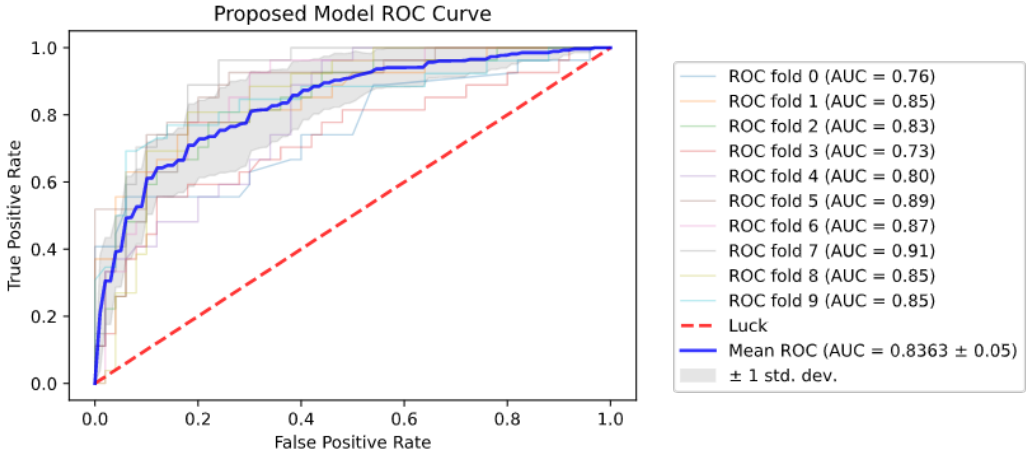


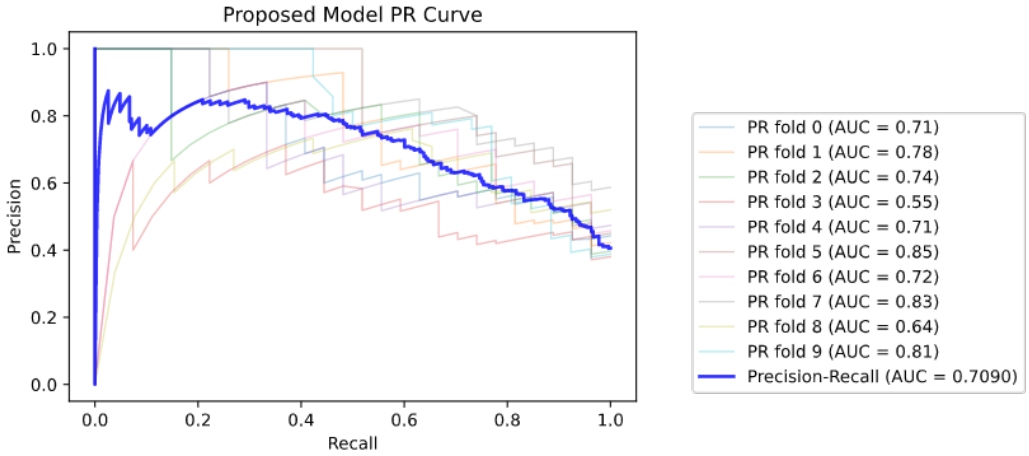
Figure 4: Baseline Model's Cross-Validated PR and ROC Curves

The present study proposes a prediction model based on a GA-XGBoost wrapper feature selection method and XGBoost classifier to accurately predict type 2 diabetes. We first performed kNN imputation to replace disguised missing values and outliers by the median with $k = 10$. The data set with preprocessing and no feature selection is treated as a baseline for comparison of the proposed model's performance. The baseline model's mean ROC AUC is 83.41% and mean PR AUC is 69.13%. Figure 4 illustrates the results for each fold of the 10-fold cross validation.

Next, the data set underwent feature selection via the GA-XGBoost wrapper method. The following values were selected as the genetic algorithm's hyperparameters based on performance. The initial population size (n_pop) was 200, 10 parents were selected each generation (n_parents) for 30 generations (n_gen). Each parent passed on their DNA to create offspring with a crossover rate of 1. These offspring had a mutation rate (r_mut) = 0.03.



(a) ROC Curve for Proposed Model



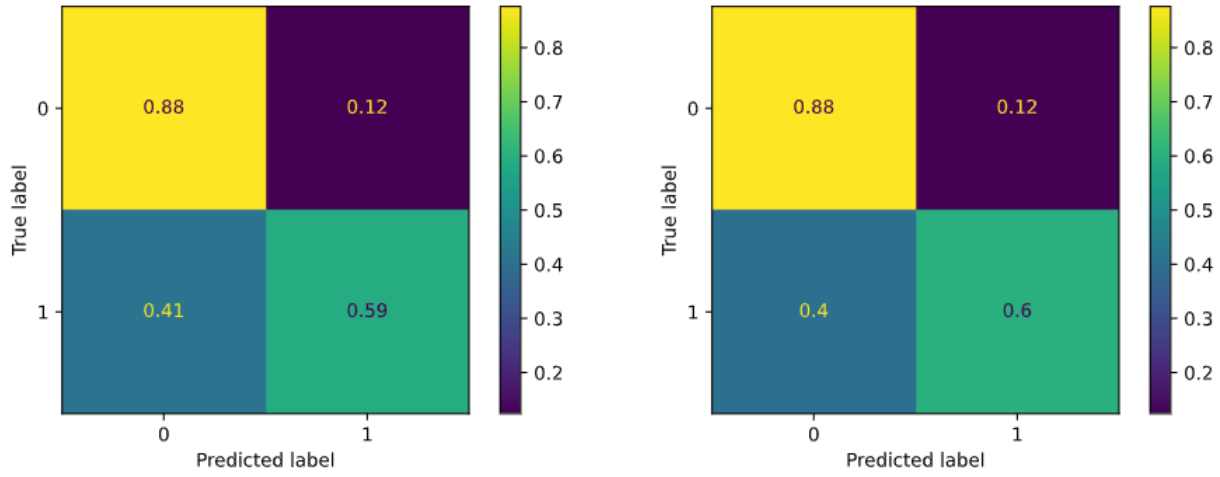
(b) PR Curve for Proposed Model

Figure 5: Comparison of ROC and PR Curve with 10-Fold CV for Proposed Model

Finally, the XGBoost classifier is applied to train the prediction model on the data set after optimal feature selection. Stratified 10-fold cross validation is then used to evaluate the performance of the proposed model. The proposed model's ROC AUC is 83.63%. This is a marginal increase from the 83.41% baseline performance. We also saw an increase from the baseline's precision-recall. The precision-recall AUC of the proposed model is 70.90%, increasing 1.77% from the baseline value. The ROC AUC indicates the model's ability to correctly predict

both the positive and negative outcomes. The precision-recall AUC on the other hand, is an indicator of the model's ability for the positive outcomes only. The results of both curves under 10-fold cross validation are shown in Figure 5. It's worth noting the stochastic nature of the GA feature selection method influenced these results. In some preliminary trials, there were decreases in performance from the proposed to the baseline model. However, these differences were not statistically significant.

The baseline and proposed model's classification performance are shown in more detail in Figure 6, which illustrates both models have similar results. The diagonals of the confusion matrix contain the specificity and sensitivity/recall. These metrics indicate the models' ability to correctly predict the patient outcomes. For this diagnostic test, we want moderate specificity and high sensitivity since the cost of a missed diagnosis is greater than the cost of a wrong diagnosis. The specificity of both the baseline and proposed model is 88.00%. We found the precision of the proposed model to be 83.33%. The baseline model reported a lower precision of 83.10%. Precision, also known as the positive predictive value, is a measure of the predictive performance for the positive class. The reported values of precision indicate both models are predicting true positives at least 83% of the time, which aligns with the low false positive rate. However, the sensitivity of both models do not exhibit the same results. The baseline sensitivity is 59.00% and the sensitivity of the proposed model is 60.00%. Since there is a 1.00% difference between the baseline and the proposed model, we will focus only on the implications for the proposed model. Our proposed model is wrongly classifying individuals of the positive class to be not at risk 40.00% of the time. The contrast between the higher precision and lower recall of the model may result from the class imbalance favoring the negative outcomes. Thus, demonstrating the model is more likely to correctly predict the negative outcomes than the positive outcomes. In addition, such a high false negative rate would prevent the use of our proposed model as a diagnostic tool in any clinical setting.



(a) Confusion Matrix for Baseline Model (b) Confusion Matrix for Proposed Model

Figure 6: Confusion Matrix Comparison for Baseline and Proposed Models

As Table 2 shows, the proposed model provides an increase in performance from the common classification algorithms Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM). However, there is no observed increase in predictive ability when compared to the models proposed by Karegowda et al. or Xu and Wang. Since the GA can harm performance if not tuned properly, it should be noted that Karegowda's GA used a population size of 20, crossover rate of 0.6, mutation rate of 0.033, and 20 generations as the terminating condition which could

account for the significant difference in model performance. Furthermore, the GA requiring further hyperparameter tuning could explain the minimal increase in performance witnessed when transitioning from the baseline model to our proposed model.

Table 2: Comparison of Model Results

Model	**Accuracy**	**Citation**
SVM	75.5%	Kumari and Chitra [10]
LDA	74.4%	Karthikeyani et al. [7]
CART, RF	83.8%	Mira Kania Sabariah et al. [16]
RF-WFS and XGBoost	93.75%	Xu and Wang [19]
GA-NB and NB	86.47%	Karegowda et al. [5]
GA-XGBoost and XGBoost	83.63%	Proposed Model

4 Conclusion

The purpose of this study is to construct an accurate method for diagnosing type 2 diabetes. Previous studies indicated promise in employing evolutionary computation and ensemble methods in the preprocessing stage. Thus, we proposed a model that combined the power of genetic algorithms and XGBoost for optimal feature selection and XGBoost for classification. The model training and testing was performed on the same UCI Pima Indians data set to allow for effective comparisons with previous models in the literature. While the model effectively combines the data preprocessing with the genetic algorithm wrapper method, the feature selection method did not lead to a significant increase in predictive performance. The model was also unable to provide the anticipated improvement in prediction accuracy over the existing models in the literature. However, the performance of the proposed model indicates this method still holds potential and requires further investigation to determine the full scope.

Future research should compare the proposed model's prediction performance against models that employ other ensemble techniques to evaluate fitness scores in the GA-wrapper. In addition, future researchers should investigate the impact of applying more complex genetic algorithms, examining the effect of crossover and mutation type on performance. Finally, the power of the proposed model's techniques on data sets with higher dimensionality and more diversity should be explored.

References

[1] UCI Machine Learning Archive. Pima indians diabetes database, 2015. https://www.kaggle.com/uciml/pima-indians-diabetes-database.

[2] L. Beretta and A. Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics Decision Making*, 2016.

[3] J. Brownlee. Roc curves and precision-recall curves for imbalanced classification, 2020. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/.

[4] XGBoost Developers. Introduction to boosted trees, 2020. https://xgboost.readthedocs.io/en/latest/tutorials/model.html.

[5] A. G. Karegowda, M. A. Jayaram, and A.S. Manjunath. Feature subset selection problem using wrapper approach in supervised learning. *International Journal of Computer Applications*, 1:13–17, 2010.

[6] A. G. Karegowda, M. A. Jayaram, and A.S. Manjunath. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima indians diabetes. *International Journal of Soft Computing*, 2:15–23, 2011.

[7] V. Karthikeyani and I.P. Begum. Comparison a performance of data mining algorithms in prediction of diabetes disease. *International Journal of Computer Science Engineering*, 5:205–210, 2013.

[8] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda. Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15:104–116, 2017.

[9] S. Khalid, T. Khalil, and S. Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *Science and Information Conference 2014*, pages 1–8, 2014.

[10] V.A. Kumari and R. Chitra. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3:1797–1801, 2014.

[11] Scikit Learn. 3.1. cross-validation: evaluating estimator performance, 2020. https://scikit-learn.org/stable/modules/cross_validation.html.

[12] Scikit Learn. 3.3. metrics and scoring: quantifying the quality of predictions, 2020. https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics.

[13] Scikit learn. 6.3. preprocessing data, 2020. https://scikit-learn.org/stable/modules/preprocessing.html.

[14] World Health Organization. Diabetes, 2020. https://www.who.int/en/news-room/fact-sheets/detail/diabetes.

[15] R. Pearson. The problem of disguised missing data. *SIGKDD Explorations*, 8:83–92, 2006.

[16] M.T.M.K. Sabariah, S.T.A. Hanifa, and M.T.S. Sa'Adah. Early detection of type ii diabetes mellitus with random forest and classification and regression tree (cart). *Advanced Informatics: Concept, Theory Application IEEE*, pages 238–242, 2014.

[17] K.C. Tan, C.M. Yu, and T.H. Heng. Evolutionary computing for knowledge discovery in medical diagnosis. *Artificial Intelligence in Medicine*, 27:129–154, 2003.

[18] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang. Type 2 diabetes mellitus prediction model based on data mining. *Informatics in Medicine Unlocked*, pages 100–107, 2018.

[19] Z. Xu, M. A. Jayaram, and A.S. Manjunath. A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier. In *2019 Eleventh Inter. Conf. on Advanced Computational Intelligence*, pages 278–283, 2019.