Predictive Analytics in Healthcare using Stacking Ensemble Learning

Chow Wei Lee1, a) and Sook Ling Lew1, b)

1*Faculty of Information Science and Technology, Multimedia University, 75450 Melaka, Malaysia*

2*Centre for Advanced Analytics, CoE for Artificial Intelligence, Faculty of Information Science and Technology, Multimedia University, 75450 Melaka, Malaysia*

*Corresponding author: b) sllew@mmu.edu.my*

*a) dhdlcw0418@gmail.com*

**Abstract.** As the healthcare field continues to evolve, advanced technologies have become essential to address emerging challenges and improve patient outcomes. Predictive analytics in the healthcare sector have been anticipated to revolutionize patient care delivery, resource allocation, and clinical decision-making. However, existing predictive models often struggle with accuracy and robustness due to imbalanced datasets, inadequate handling of missing values, and outliers, leading to suboptimal patient care and intervention strategies. This study aims to develop a robust predictive model for diabetes outcomes, enhance machine learning (ML) model prediction accuracy, and assess the impact of predictive analytics on healthcare decision-making. To address these challenges, a comprehensive preprocessing approach and a robust ensemble learning method are proposed. The Stacking Classifier integrates four base models - Random Forest (RF), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Adaptive Boosting (AdaBoost) - with Extreme Gradient Boosting (XGBoost) as the meta-model. Evaluated using the Pima Indian Diabetes dataset, the proposed model achieved an accuracy of 96.33%, with precision and recall scores of 97.26% and 95.30%, respectively. These results demonstrate significant improvements over traditional models, highlighting the effectiveness of the ensemble approach in predicting diabetes. However, the study is limited by its reliance on a single dataset, suggesting the need for further validation across diverse populations and settings.

# Introduction

As the healthcare field is constantly evolving, advanced technologies have become a necessity to deal with changing challenges and improve patient results [1]. Despite the development of modern healthcare, it is still difficult to precisely predict disease outcomes and as a result any timely intervention or patient care remains subpar [2]. This study focuses on developing and evaluating a predictive model that aims for improving predictions in disease outcome. Systematic evaluation of the effect with predictive analytics has on decision-making processes is currently lacking contemporary healthcare environment [3]. The knowledge of how predictive models affect clinical decisions and patient care is crucial for the success in introducing analytics into healthcare processes [4]. The proposed methodology utilizes ensemble learning techniques, specifically stacking, to combine multiple ML algorithms. This fusion aims to enhance the predictive validity of the model. By leveraging different capabilities of various ML algorithms with an ensemble structure, the study aims at improving the accuracy of the predictions. The Pima Indian Diabetes dataset is utilized to be conducted and tested with the proposed model, given it being widely used upon in diabetes research and predictive analytic experiments. Previous studies have demonstrated the efficacy of ensemble models in achieving higher accuracy compared to individual ML models [5], [6]. Hence, this study is aimed to build upon these findings, providing a detailed analysis of the impact of the model on decision-making processes with healthcare. In conclusion, this study underscores the potential of predictive analytics to transform healthcare by enabling, data-driven approaches to patient care.

# Related WORKS

## Predictive Analytics in Diabetes using Pima Dataset

Predictive analytics in diabetes has gathered significant research attention, especially through the utilization of the Pima Indian Diabetes Dataset. Reza et al. [7] proposed a stacking ensemble model with an accuracy of 95%, which far exceeded conventional ML and deep learning algorithms. The model aims to improve early detection and preventative therapies. Rahim et al. [5] developed a stacking method that achieved 94.17% accuracy. One of the focus areas in this study is early diagnostics with ML models and their combination fused as an SVM-NN model to be used for diabetes containment. Shimpi et al. [6] put forward an analytical model utilizing optimized SVM, KNN and RF with decision-level fusion having the accuracy of 94.27% on the Pima dataset. In 2021, R et al. [8] concentrated on classifying diabetic-positive patients by stacked ensemble modelling which showed improved accuracy in the case of a system (93%) and AUC than other ML models. Malini et al. [9] reviewed ML methods on Pima, out of which XGB Classifier recorded 86% accuracy. Table 1 shows the summary of past studies using the Pima Dataset.

**TABLE 1.** Summary of past studies using Pima dataset

|  |  |  |
| --- | --- | --- |
| **References** | **Methods** | **Accuracy** |
| Reza et al. [5] | Stacking Ensemble Learning  Base Models:  RF + Decision Tree (DT) + SVM  Meta Model:  Logistic Regression (LR) | 95.9% |
| Rahim et al. [6] | Stacking Ensemble Learning  Base Models:  RF + KNN + SVM + Naïve Bayes (NB)  Meta Model:  LR | 94.17% |
| Shimpi et al. [7] | Particle Swarm Optimization  Models: RF+SVM+KNN | 94.27% |
| R et al. [8] | Stacking Ensemble Learning  Base Models:  RF + KNN + DT + GB + SVM + NB  Meta Model:  LR | 93.1% |
| Malini et al. [9] | XGBoost | 82% |

## Predictive Analytics in Other Healthcare Settings

The reviewed studies contribute together to show that the use of predictive analytics is widespread in varied settings within healthcare. AR Rao et al. [10] showed that DT regression outperformed sparse regression in healthcare cost prediction, achieving an R2 value of 76%. Aloitabi et al. [11] evaluated models for predicting stroke patient transfers to the ICU, finding similar accuracies (0.96 for DT, SVM, LR; 0.94 for Artificial Neural Network (ANN)) and emphasized the need for standardized model deployment for early risk prevention. Nang Maik et al. [12] addressed "no shows" in outpatient clinics using LR and DT models, highlighting the effectiveness of SMS reminders in reducing non-attendance rates. These studies collectively highlight the potential of predictive analytics for decision-making and targeted interventions to optimize healthcare outcomes through efficient resource allocation. Table 2 shows the summary of past studies that use predictive analytics in other healthcare settings.

**TABLE 2.** Comparison between past studies and proposed method

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **References** | **Datasets** | **Methods** | **Results** | |
| AR Rao et al. [10] | Patient Records sourced from SPARCS (statewide planning and research cooperative system) | DT  LR | R2 Value: 0.76 |
| Aloitabi et al. [11] | Patient Records sourced from King Fahad Medical Center, National Neuroscience Institute, Riyadh Saudi Arabia | DT  ANN  SVM  LR | Accuracy:  DT: 95%  ANN: 94%  SVM: 96%  LR: 96% | |
| Nang Maik et al. [12] | Outpatient appointment data sourced from public hospital in Singapore (Khoo Tech Puat Hospital) Models: RF+SVM+KNN | DT  LR | Identifying key predictors like language preference and SMS response for improving appointment attendance | |

## Predictive Analytics using Ensemble Machine Learning

Gupta et al. [13], R et al. [8], and Mung et al. [14] demonstrate ensemble learning’s effectiveness in healthcare, achieving high accuracy and robustness across various predictive analytics applications. Table 3 shows the summary of past studies using ensemble ML.

**TABLE 3.** Comparison between past studies and proposed method

|  |  |  |  |
| --- | --- | --- | --- |
| **References** | **Datasets** | **Methods** | **Results (Accuracy)** |
| Gupta et al. [13] | Covid-19 Patient Dataset | Stacking Ensemble Learning  Base Models:  DT + RF + SVM + ANN  Meta Model: LR Boosting (LogitBoost) | 93.11% |
| R et al. [8] | PIMA Diabetes Dataset | Stacking Ensemble Learning  Base Models:  RF + KNN + DT + GB + SVM + NB  Meta Model: LR | 93.1% |
| Mung et al. [14] | Datasets from UCI ML repository | Ensemble Learning combining clustering and classification.  Base Models:  Naïve Bayesian  DT, K-NN | Lung Cancer: 99.93%  Heart Disease: 93.56%  Diabetes-disease: 98.06%  Cervical cancer: 98.82% |

# Methods and Materials

A screenshot of a computer

Description automatically generatedThe proposed framework for this study is based on several studies mainly about stacked ensemble ML [7], [8]. The proposed framework is based on stacking ensemble ML learning has several processes involved, which are dataset preparation, data preprocessing, base models, base model results, meta-model, performance evaluation, and results as shown in Figure 1.

**Figure 1.** Flowchart of proposed framework

**Dataset**

The selected dataset is the Pima Indians Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases and sourced via Kaggle [15]. It comprises 768 samples, with 500 classified as non-diabetic and 268 as diabetic. The dataset includes several clinical attributes used to distinguish diabetic from non-diabetic cases. Table 4 outlines the features and their descriptions.

**TABLE 4.** Features of the dataset with descriptions

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Descriptions** | | |
| Pregnancies | | | Number of times pregnant |
| Glucose | | | Plasma glucose concentration 2 hours in an oral glucose tolerance test |
| SkinThickness | | | Triceps skin fold thickness |
| Insulin | | | 2-hour serum insulin (mmU/ml) |
| BMI | | | Body mass index (weight in kg/(height in m)2 |
| DiabtesPedigreeFunction | | | Diabetes Pedigree Function |
| Age | | | Age (years) |
| Outcome | | | Class Variable (0 to1), 1 = diabetic, 0 = non-diabetic |

## Data Preprocessing

Missing values in BMI, Insulin, Blood Pressure, Skin Thickness, Diabetes Pedigree Function, and Glucose were imputed with median values specific to target classes 0 and 1, preserving data consistency [16]. Next, outliers in features such as Age, BMI, Pregnancies, Glucose, Insulin, Skin Thickness, DBP, and Diabetes Pedigree Function were identified using box plots and domain knowledge. Outliers were replaced with median values relevant to the target classes, ensuring they did not skew the analysis [17].

An Upsampling technique was applied to balance the dataset [18], originally the dataset contained 268 instances of class 1 (diabetes) and 500 instances of class 0 (no diabetes). Figure 2 shows the distribution of outcome.

Using this Upsampling method, class 1 was increased to 500 instances, ensuring that there are equal representations for both classes [18]. This balanced dataset improves the model’s ability to generalize and accurately predict diabetes. Figure 3 shows the distribution of outcome after using the Upsampling technique.

|  |  |
| --- | --- |
| A graph with blue squares  Description automatically generated with medium confidence | A graph with blue rectangular bars  Description automatically generated with medium confidence |
| **Figure 2.** Visualization of the Distribution of Outcome | **Figure 3.** Visualization of the Distribution of Outcome using Upsampling |

The dataset was then split into 70% training and 30% testing sets. Numerical features were standardized using StandardScaler, which adjusted each feature's mean to zero and standard deviation to one [19]. This standardization ensures that no single feature disproportionately influences the learning process [20]. These preprocessing steps were essential for preparing the data for effective model training and testing.

**Base and Meta Models**

This study employed several base models—RF, KNN, SVM, and AdaBoost—each offering unique strengths in classification. Random Forest builds multiple DT and aggregates their results for improved accuracy, while KNN classifies based on proximity in the feature space [21]. SVM identifies the optimal hyperplane separating classes, and AdaBoost improves prediction by emphasizing misclassified instances in successive iterations [22]. As a meta-model, XGBoost was used to combine base model predictions into a final output. XGBoost is a powerful ensemble method offering regularization and internal handling of missing values [9], [22], [23], [24]. The study evaluates model performance using accuracy, precision, recall, and F1-score, all derived from the confusion matrix components: TP, TN, FP, and FN.

**RESULTS**

## Individual Model and Stacking Method Results

Figure 4 compares the performance of various ML algorithms, showing that Random Forest achieved the highest accuracy at 0.94, while SVM recorded the lowest at 0.883. The Stacking Method, which integrates multiple models, delivered superior overall results. Figure 5 presents the confusion matrix of the Stacking Method.

Out of 151 non-diabetic cases, 147 were correctly identified as True Negatives, and out of 149 diabetic cases, 142 were accurately predicted as True Positives. There were 4 False Positives and 7 False Negatives. The Stacking Classifier, combining RF, AdaBoost, SVM, and KNN with XGBoost as the meta-model, achieved a notable accuracy of 96.33%. Precision and recall scores were 97.26% and 95.30%, respectively, underscoring the classifier's effectiveness in minimizing both false positives and false negatives. This highlights the model's robust performance in predicting diabetes onset. Table 5 tabulates the performance of the individual models and Stacking Method.

|  |  |
| --- | --- |
| A graph of different colored bars  Description automatically generated | A blue squares with white text  Description automatically generated |
| **Figure 4.** Accuracy comparison of individual models | **Figure 5.** Confusion matrix of stacking method |

**TABLE 5.** Comparison of individual models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Random Forest | 0.94 | 0.9018 | 0.9866 | 0.9423 |
| AdaBoost | 0.92 | 0.9195 | 0.9195 | 0.9195 |
| SVM | 0.8833 | 0.8750 | 0.8926 | 0.8837 |
| KNN | 0.9 | 0.8521 | 0.9664 | 0.9057 |
| XGBoost | 0.93 | 0.9156 | 0.91 | 0.93 |
| Stacking Classifier | 0.9633 | 0.9726 | 0.9530 | 0.9627 |

## Comparison with Past Studies

To assess the effectiveness of the proposed Stacking Classifier method, it was compared with prior research. In comparison (Table 6), the proposed method in this study—using Random Forest, KNN, SVM, and AdaBoost as base models, with XGBoost as the meta-model—achieved the highest accuracy of 96.33%, highlighting its superior performance. Table 6 presents a detailed comparison of these methods and their respective accuracies.

**TABLE 6.** Comparison between Past Studies and Proposed Method

|  |  |  |
| --- | --- | --- |
| **References** | **Methods** | **Accuracy** |
| R et al. [8] | Stacking Ensemble Learning  Base Models: RF+KNN+DT+GB+SVM+NB  Meta Model: LR | 93.1% |
| Rahim et al. [6] | Stacking Ensemble Learning  Base Models: RF+KNN+SVM+NB  Meta Model: LR | 94.17% |
| Shimpi et al. [7] | Particle Swarm Optimization  Base Models: RF+SVM+KNN | 94.27% |
| Malini et al. [9] | XGBoost | 82% |
| Reza et al. [5] | Stacking Ensemble Learning  Base Models: RF + DT + SVM  Meta Model: LR | 95.9% |
| Proposed Method | Stacking Ensemble Learning  Base Models: RF + KNN + SVM + AdaBoost  Meta Model: XGBoost | 96.33% |

# CONCLUSION

This study highlights the significant potential of ensemble ML techniques in healthcare, particularly for diabetes prediction. Using the Pima Indian Diabetes dataset, four base models—Random Forest, AdaBoost, SVM, and KNN - were integrated within a Stacking Classifier framework, with XGBoost serving as the meta-model. The proposed ensemble model achieved a highest accuracy of 96.33%, with precision of 97.26% and recall of 95.30%, outperforming all individual models. These results demonstrate the effectiveness of ensemble learning in improving predictive accuracy for early disease detection, supporting more timely interventions and better patient outcomes. While the findings are promising, the study is limited by its reliance on a single dataset and the potential for overfitting due to the complexity of the stacking approach. Future work should validate these results using diverse datasets and real-world clinical data to ensure broader applicability and to further optimize model generalization.

# References

1. S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, “The role of artificial intelligence in healthcare: a structured literature review,” BMC Med Inform Decis Mak **21**(1), 125 (2021).
2. B.K. Sarkar, “Big data for secure healthcare system: a conceptual design,” Complex Intell. Syst. **3**(2), 133–151 (2017).
3. S. Kraus, F. Schiavone, A. Pluzhnikova, and A.C. Invernizzi, “Digital transformation in healthcare: Analyzing the current state-of-research,” Journal of Business Research **123**, 557–567 (2021).
4. E. Vayena, T. Haeusermann, A. Adjekum, and A. Blasimme, “Digital health: meeting the ethical and policy challenges,” Swiss Med Wkly **148**(0304), w14571 (2018).
5. M.A. Rahim, M.A. Hossain, M.N. Hossain, J. Shin, and K.S. Yun, “Stacked Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques,” AETiC **7**(1), 30–39 (2023).
6. J.K. Shimpi, P. Shanmugam, and A.A. Stonier, “Analytical model to predict diabetic patients using an optimized hybrid classifier,” Soft Comput **28**(3), 1883–1892 (2024).
7. M.S. Reza, R. Amin, R. Yasmin, W. Kulsum, and S. Ruhi, “Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data,” Heliyon **10**(2), e24536 (2024).
8. S. R., S. M., M.K. Hasan, R.A. Saeed, S.A. Alsuhibany, and S. Abdel-Khalek, “An Empirical Model to Predict the Diabetic Positive Using Stacked Ensemble Approach,” Front. Public Health **9**, 792124 (2022).
9. M. Malini, B. Gopalakrishnan, and S. Naveena, “Prediction and Detection of Diabetics Mellitus using Different Machine Learning Approaches,” in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, (IEEE, Salem, India, 2022), pp. 323–330.
10. A. Ravishankar Rao, S. Garai, S. Dey, and H. Peng, “Building predictive models of healthcare costs with open healthcare data,” in *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, (IEEE, Oldenburg, Germany, 2020), pp. 1–3.
11. N.N. Alotaibi, and S. Sasi, “Comparison of Predictive Models for Transferring Stroke In-Patients to Intensive Care Unit,” TMLAI **4**(3), (2016).
12. N.L. Ma, S. Khataniar, D. Wu, and S.S.Y. Ng, “Predictive Analytics for Outpatient Appointments,” in *2014 International Conference on Information Science & Applications (ICISA)*, (IEEE, Seoul, South Korea, 2014), pp. 1–4.
13. A. Gupta, V. Jain, and A. Singh, “Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications,” New Gener. Comput. **40**(4), 987–1007 (2022).
14. P.S. Mung, and S. Phyu, “Effective Analytics on Healthcare Big Data Using Ensemble Learning,” in *2020 IEEE Conference on Computer Applications (ICCA)*, (IEEE, Yangon, Myanmar, 2020), pp. 1–4.
15. **UCI Machine Learning Repository,** “Pima Indians Diabetes Database,” Kaggle, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed January 2, 2024).
16. B. Najafi, S. Parsaeefard, and A. Leon-Garcia, “Missing Data Estimation in Temporal Multilayer Position-Aware Graph Neural Network (TMP-GNN),” MAKE **4**(2), 397–417 (2022).
17. L. Huang, W. Chen, Y. Liu, H. Zhang, and H. Qu, “Improving neural machine translation using gated state network and focal adaptive attention networtk,” Neural Comput & Applic **33**(23), 15955–15967 (2021).
18. J.R. Barr, M. Sobel, and T. Thatcher, “Upsampling, a comparative study with new ideas,” in *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, (IEEE, Laguna Hills, CA, USA, 2022), pp. 318–321.
19. Y. Xu, and R. Goodacre, “On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning,” J. Anal. Test. **2**(3), 249–262 (2018).
20. G. Wurdiyanto, H. Candra, H. Holnisar, and V. Pungkun, “The Absolute Standardization Methods of 32P for Calibrate Nuclear Medicine Instruments in Indonesia,” Atom Indo. **45**(3), 183 (2019).
21. S. Carrington, *Supporting Students on the Autism Spectrum in Inclusive Schools: A Practical Guide to Implementing Evidence-Based Approaches* (Taylor & Francis Group, Milton, 2021).
22. O. Rainio, J. Teuho, and R. Klén, “Author Correction: Evaluation metrics and statistical tests for machine learning,” Sci Rep **14**(1), 15724 (2024).
23. A. Mahmoud, and E. Takaoka, “An enhanced machine learning approach with stacking ensemble learner for accurate liver cancer diagnosis using feature selection and gene expression data,” Healthcare Analytics **7**, 100373 (2025).
24. Z. Jinbo, L. Yufu, and M. Haitao, “Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method,” Front. Artif. Intell. **8**, 1553220 (2025).