An Overview of Diabetes Prediction through Ensemble Learning

Lokendra Argal[1] and Dr. Pinaki Ghosh[2]

[1]Research Scholar, Sanjeev Agrawal Global Educational (SAGE) University, Bhopal, India, 462022
[2]Professor, Sanjeev Agrawal Global Educational (SAGE) University, Bhopal, India, 462022
[1]lokendra.argal@gmail.com, [2]pinaki.g@sageuniversity.edu.in

Abstract. *Diabetes is a chronic condition marked by elevated blood sugar levels and presents a growing global health concern. Early and accurate prediction is essential for effective management, enabling timely intervention and reducing strain on healthcare systems. This study investigates the use of ensemble learning techniques to enhance diabetes prediction. Ensemble learning, which combines the strengths of multiple machine learning models, is employed to improve predictive accuracy and reliability. The paper provides a comprehensive overview of key ensemble methods, including bagging, boosting, and stacking, and evaluates their effectiveness in comparison to individual models. Results demonstrate that ensemble approaches consistently outperform standalone algorithms, offering a more robust and dependable framework for diabetes prediction. These findings highlight the significant potential of ensemble learning in medical diagnostics and its critical role in advancing predictive healthcare technologies.*

Keywords: - Diabetes prediction, Ensemble Learning, Deep Learning, Classifiers, Bagging, Stacking

INTRODUCTION

Medical diagnosis is one of the challenging and crucial tasks in medical science. To predict the diabetes disease, data are taken from patients like plasma glucose concentration, diastolic blood stress, and triceps skin fold thickness, serum insulin, body mass, age etc. Then the patient consults to a specialist doctor. The physician takes the decision using his/her knowledge and experience based on these factors. The process of taking the decision is very lengthy and sometimes takes a few weeks or months that make the physician work's very difficult [1]. Nowadays, a huge number of medical datasets are easily available that are useful for research in different sectors in medical science. So, it is hard or sometimes become impossible to handle the massive data by a human. Therefore, effective computer-based approaches are taken place over the traditional modalities [2]. The computer-based systems increase the correctness and save time as well as money. Diabetes is one of the frequent diseases that targets the elderly population worldwide.

Diabetes is considered as a chronic disease associated with an abnormal state of the human body where the level of blood glucose is inconsistent due to some pancreas dysfunction that leads to the production of little or no insulin at all, causing diabetes of type 1 or cells to become resistant to insulin, causing diabetes of type 2 [3]. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes. Even though it's incurable, it can be managed by treatment and medication. Individuals with diabetes face a risk of developing some secondary health issues such as heart diseases and nerve damage. Thus, early detection and treatment of diabetes can prevent complications and assist in reducing the risk of severe health problems. Many researchers in the bioinformatics field have attempted to address this disease and tried to create systems and tools that will help in diabetes prediction. They either built prediction models using different types of machine learning algorithms such as classification or association algorithms. Decision Trees, Support Vector Machine (SVM), and Linear Regression were the most common algorithms [4].

Over the years, it has been found that people with the following health characteristics face a greater risk against diabetes:

- A Body Mass Index value greater than 25
- Members of the family suffering from diabetes
- People with HDL cholesterol concentration in the body less than 40 mg/dL
- Prolonged hypertension having gestational diabetes
- People who have suffered from polycystic ovary disorder in the past
- People belonging to ethnic groups like African American, or Native American, or Latin American, or Asian-pacific aged over 45 years
- Having an inactive lifestyle

When a doctor diagnoses that an individual has pre-diabetes, they suggest the individual better their lifestyle. Adopting a fitness regime and a good diet plan can help prevent diabetes [5].

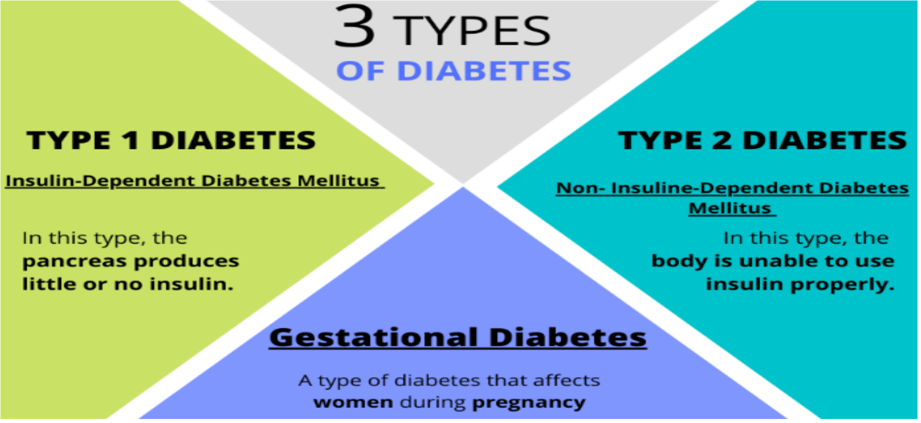
3 TYPES
OF DIABETES
TYPE 1 DIABETES
Insulin-Dependent Diabetes Mellitus
In this type, the pancreas produces little or no insulin.
TYPE 2 DIABETES
Non- Insuline-Dependent Diabetes Mellitus
In this type, the body is unable to use insulin properly.
Gestational Diabetes
A type of diabetes that affects women during pregnancy

Figure 1: Types of Diabetes.

Diabetic illness is the most prevalent disease that affects humans. It is caused by an inadequate synthesis of insulin and excessive blood sugar levels in the blood. Before doing a clinical examination, it is necessary to look for many signs and symptoms of diabetes. Although the newly discovered symptoms are basic to detect in a handbook, the accuracy of diabetes prediction continues to be a serious challenge. Although newly discovered symptoms are easy to access in a handbook, the accuracy of diabetes prediction continues to be a significant challenge [6]. There are several researchers that are devoted to this topic, with the goal of accurately diagnosing diabetic condition via the collection of large amounts of data. The standard stages for recognizing diabetic illness make use of the smallest amount of processes possible, but they fall short of achieving the highest possible detection accuracy.

The three main types of diabetes are shown in figure 1 and explained below:

- Type 1 diabetes
- Type 2 diabetes

- Gestational diabetes (diabetes during pregnancy)

Type-1 Diabetes- In type-1 diabetes, the body loses the ability to make insulin. Insulin is a vital hormone made by the pancreas, a gland in the body near the stomach. It is needed to change glucose, the body's primary energy source [7], into energy. Type 1 diabetes usually occurs in children or adults under age 30, but can develop at any age. Environmental factors, such as viral infections, some diseases, chemicals and stressful situations may play a role, but the specific role of these factors still is not clear.

Type 2 Diabetes- Type 2 is the most common form of diabetes. In type 2 diabetes, the onset occurs slowly and over time when the pancreas cannot produce enough insulin. The pancreas eventually begins to tire. Insulin production levels off, and the body cannot keep up with the amount of glucose in the blood, triggering type 2 diabetes. The condition may not be diagnosed right away, however, because often there are no visible symptoms. Some people with type 2 diabetes need to take insulin or medication to help their bodies use insulin better [8].

Gestational Diabetes- Gestational diabetes occurs when a woman's body cannot produce enough insulin during pregnancy. There are usually no symptoms. Pregnant women should be tested for diabetes between the 24th and 28th week of pregnancy. A type of Diabetes Sometimes, gestational diabetes reveals undiagnosed type 2 diabetes. If this is the case, diabetes remains after pregnancy, and the blood glucose will become high if diabetes is not treated.

II. Steps for Diabetes Prediction

General steps for diabetes prediction [9] are shown in figure 2 and explained below.

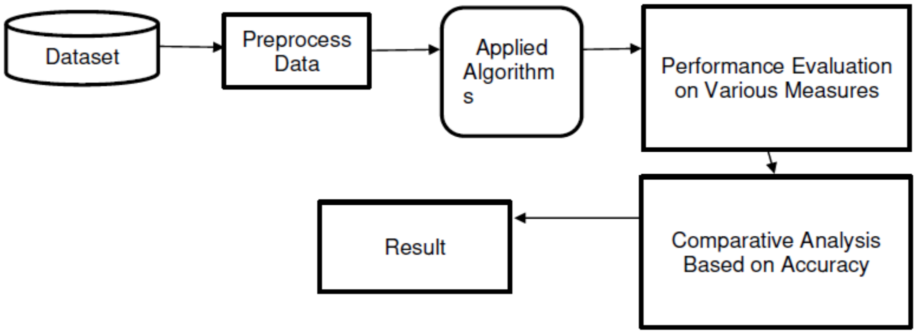
Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

Figure 2: Steps for Diabetes Prediction.

Dataset- The main Objective of using the dataset was to predict through diagnosis whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification [10].

Data preprocessing- In real-world data there can be missing values and/or noisy and inconsistent data. If data quality is low then no quality results may be found. It is necessary to pre-process the data to achieve quality results. Cleaning, integration, transformation, reduction, and Discretization of data are applied to pre-process the data. It is important to make the data more appropriate for data mining and analysis with respect to time, cost, and quality.

Data cleaning consists of filling the missing values and removing noisy data. Noisy data contains outliers which are removed to resolve inconsistencies [11].

Applied Algorithms- Choose an appropriate machine learning algorithm(s) for diabetes prediction. Commonly used algorithms include logistic regression, decision trees, random forests, support vector machines, and neural networks. The choice of model depends on factors such as the size and complexity of the dataset, interpretability requirements, and desired predictive accuracy. Split the dataset into training and testing subsets. Use the training data to train the selected model(s) on the features and corresponding labels (i.e., whether a patient has diabetes or not) [12]. Data mining techniques are also used to extract useful information to generate rules. Association rule mining is an important branch to determine the patterns and frequent items used in the dataset.

Performance Evaluation- Assess the performance of the trained model(s) using the testing data. Common evaluation metrics for binary classification tasks like diabetes prediction include accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve analysis. Fine-tune the model hyperparameters to optimize its performance. This process may involve techniques such as grid search or random search [13].

Comparative Analysis- There are various algorithms available for building predictive models, each with its own strengths and weaknesses. Comparative analysis helps in selecting the most suitable model(s) for the specific task at hand. By comparing the performance of different models using relevant evaluation metrics, we can identify which model(s) offer the best predictive accuracy and generalization capabilities [14]. Comparative analysis provides insights into the strengths and weaknesses of different models, which can guide further optimization efforts.

III. Frequently Used Methods for Diabetes Prediction

Early detection of the diabetes is the need of current epidemiology of diabetes because it can cause more and severe complication with time [15].There is an urgent need to predict diabetes in the population so that proper precautions and treatment can be started to avoid its further escalation. In recent past the scientific community has changed its focus towards early and accurate prediction of diabetes using robust computational methods. Artificial intelligence and soft computing techniques provide an important role in the implementation of human ideas. These systems locate a place in the medical diagnosis and are also involved in human health related fields of application. The computational intensive methods should have high precision and must be validated on multiple datasets from different population being global disease as shown in figure 3. In the cur-rent report, different diabetes prediction computational methods were discussed and the possible suggestions are provided to make them more practical.

Machine learning- Machine learning is a growing branch of computational algorithms that is intended to copy the human intelligence utilizing knowledge from surrounding environment [16]. It is an area of computer science to learn patterns from data to make the sense of previously unknown inputs [17]. Generally, there are two types of machine learning first one is deductive learning and second is inductive learning. The deductive learning predicts new knowledge from existing data and knowledge whereas the inductive learning takes examples and simplifies it instead of beginning with existing knowledge. There exist three types of learning depending upon the input and feedbacks these are unsupervised learning, supervised learning, and reinforcement learning. There are vast range of methods related to classification and diagnosis of diabetes in the literature.

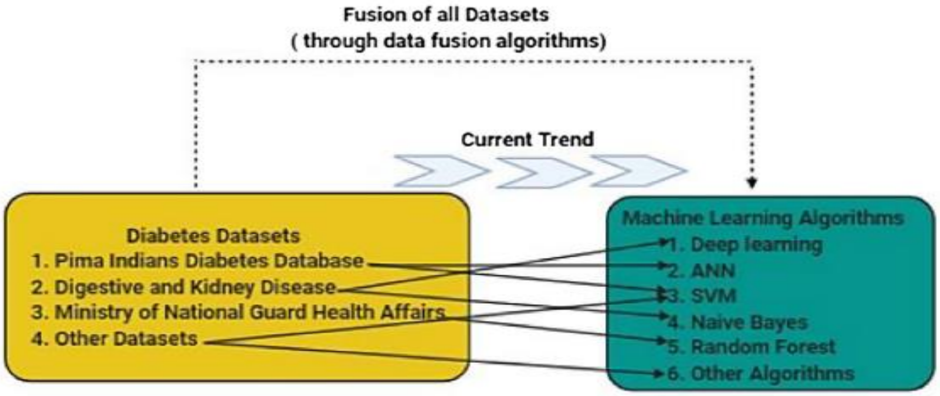
Fusion of all Datasets
(through data fusion algorithms)
Current Trend
Diabetes Datasets
1. Pima Indians Diabetes Database
2. Digestive and Kidney Disease
3. Ministry of National Guard Health Affairs
4. Other Datasets
Machine Learning Algorithms
1. Deep learning
2. ANN
3. SVM
4. Naive Bayes
5. Random Forest
6. Other Algorithms

Figure 3: Diabetes datasets with fusion algorithms.

Artificial neural network- It is a model that is motivated by the bio-logical neural network of the human brain and is used to guess functions that depend on a huge number of unknown inputs [18].It is the collection of interconnected neurons that interchange information among each other and connections have weighted which can be adjusted to get appropriate results [19]. It contains mainly three layers: first is input layer: in this layer neurons accept the inputs and their probability from external world for processing in the model. Second is hidden layer: in this layer neurons accept the input from input layer and forward the output to the output layer. This is the layer where weights are allotted to various probabilities of inputs. The neuron with larger weight is assigned to an input. Third layer is output layer: neurons of output layer are represented with expected attribute values to the external word as output.

Support vector machine (SVM) - It is a supervised learning algorithm which is used for regression analysis and classification. It gets hold of the number of examples and allocates them to one or two categories as stated by the condition it belongs. Then this training algorithm constructs a model that allots the new examples to one of the categories. The foundations of SVM were developed by Vapnik and it is widely held due to many attractive properties [20]. In parallel it reduces the empirical classification error and increases the geometrical margin, that's why it is called as maximum Margin classifiers [21] SVM model is an illustration as point in space; plotted so that the illustrations of the disconnected group are divided by a clear gap that is as wide as possible.

Bayesian network- In Bayesian network set of random variables and their conditional dependencies are symbolized using directed acyclic graph (DAG). This is a supervised learning technique [1]. It is a graphical model that conceals relationship between variables. When this method is used in combination with statistical techniques, then this model has several data analysis advantages. First, the model conceals dependencies between all the variables so it quickly handles the conditions such as missing data entries. Second is that this model can be used to learn the casual relationships so it can be used to get the reorganization of the problem area and expect the consequence of interference.

Back propagation algorithm- Paul Werbos has developed the backpropagation algorithm in 1974 and rediscovered by Rumelhartand Parker [2]. It is a technique of training the artificial neural network to execute a given task. It is used in layered feed-forward artificial neural networks in which the artificial neurons are arranged in layers and send forward signals and subsequently propagate the errors backward.

It is a supervised learning algorithm that takes examples of input and outputs and then error is calculated. The idea of this algorithm is to decrease error till the artificial neural network becomes skilled at the data training.

Apriori algorithm- It is used to find out the relationship among the different set of data. Each set constitutes a number of items called transaction. This algorithm output is the set of rules that inform us how often the data sets are combined into one set. The Apriori algorithm is used for association rule learning. Association rule has two parts first is antecedent. These are subsets of items found in the set of data; second is consequent, which is found along with the antecedent. The association rule is described in two terms, first is confidence which reveals about the percentage of datasets with antecedent and second is the support which tells about the percentage of datasets with antecedent along with the consequent [20].

Deep learning- In current scenario deep learning (DL) is considered as one of the most essential machine learning techniques, DL has succeeded to achieve accurate and efficient model in several applications which include image and video analysis, speech recognition and hand writing prediction. DL can be used in both unsupervised and supervised machine learning problem.

IV. Ensemble Learning

Ensemble learning is a technique used to combine two or more ML algorithms to obtain superior performance compared to when the constituent algorithms are used individually. Instead of relying on a single model, the predictions from the individual learners are combined using a combination rule to obtain a single prediction that is more accurate. Generally, ensemble methods can be classified into parallel and sequential ensembles. The parallel methods train different base classifiers independently and combine their predictions using a combiner. A popular parallel ensemble method is bagging and its extension, the random forest algorithm. Parallel ensemble algorithms use the parallel generation of base learners to encourage diversity in the ensemble members. Meanwhile, sequential ensembles do not fit the base models independently. They are trained iteratively so that the models at every iteration learn to correct the errors made by the previous model. A popular type of sequential ensemble is the boosting algorithm. Figure 4 shows block diagrams, which define parallel and sequential ensemble learning. Furthermore, parallel ensembles can be classified into homogeneous or heterogeneous, depending on the base learners' homogeneity.

Homogeneous ensembles consist of models built using the same ML algorithm, while heterogeneous ensembles comprise models from different algorithms. The success of ensemble learning techniques mainly relies on the accuracy and diversity of the base learners. A machine learning model is considered accurate if it has good generalization ability on unseen instances. In contrast, ML models are diverse if their errors on unseen instances are not the same.

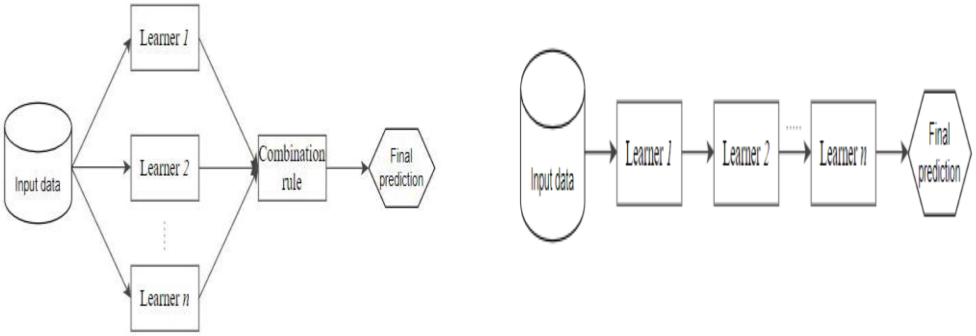

Figure 4: Block diagram of parallel ensemble learning and sequential ensemble learning.

Therefore, diversity is seen as the difference between base learners in an ensemble. Unlike accuracy, there is no general rule of thumb in measuring diversity. Meanwhile, it is challenging to have diversity in the base models when implementing ensemble classifiers. In most ensembles, the base learners are trained using subsets of the same training data, making the models correlated and difficult to achieve diversity. Different ensemble techniques try to achieve diversity heuristically or implicitly. For instance, bagging achieves diversity by subsampling the training data while boosting achieves diversity by reweighting the training data. Furthermore, different techniques are used to achieve diversity among base learners in homogeneous and heterogeneous ensembles.

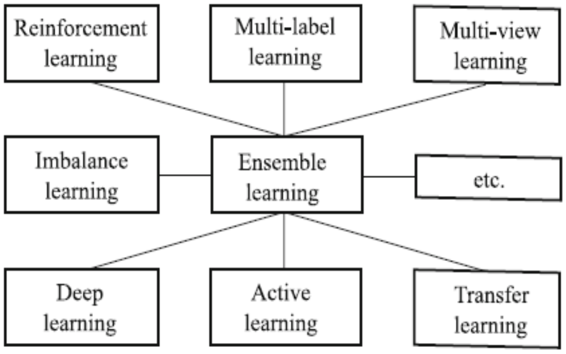

Figure 5: The combination of ensemble learning with other machine learning issues.

For example, heterogeneous ensembles employ different ML algorithms as base learners; therefore, they are essentially diverse. The main challenge in heterogeneous ensembles is obtaining the most effective method to combine the different base learners' predictions. However, the main challenge of homogeneous ensemble methods is ensuring the base learners are diverse even though they use the same ML algorithm. Hence, bootstrap methods such as random forest and boosting methods such as AdaBoost have been developed to achieve diversity in the ensemble.

Ensemble learning is more than a specific algorithm, which makes it easy to combine ensemble method with other machine learning algorithms as shown in figure 5.

V. Characteristics of Ensemble Learning

Overfitting avoidance- When just a small amount of data is available, a learning algorithm is prone to finding many different hypothesis that predict all of the training data perfectly while making poor predictions for unseen instances. Averaging different hypothesis reduces the risk of choosing an incorrect hypothesis and therefore, improves the overall predictive performance.

Computational advantage- Single learners that conduct local searches may get stuck in local optima. By combining several learners, ensemble methods decrease the risk of obtaining a local minimum.

Representation- The optimal hypothesis may be outside the space of any single model. By combining different models, the search space may be extended and hence, a better fit to the data space is achieved. The goal is to induce a classifier that can classify new emails to either spam (sad face) or no spam (smiley face) based on the email-length (x-axis) and number of recipients (y-axis). In this example we use a decision stump as a model. Decision stump is a weak classifier, consisting of a single split.

Class imbalance- There are many machine learning problems for which one class has substantially more examples than other classes. In such cases, machine learning algorithms may develop a preference for the majority class while ignoring minority classes. Ensemble methods may be applied in a way that mitigates the class imbalance problem. One example is to create an ensemble in which each of the inducers is trained using a balanced subsample of the data. Another work showed how combining random under-sampling techniques with ensemble techniques such as bagging or boosting may significantly improve predictive performance for problems with class imbalance. The prediction performance can be further improved by using the EUSBoost method which leverages evolutionary under-sampling to promote diversity among individual inducers.

Concept drift- In many real-time machine learning applications, the distribution of features and the labels tend to change over time. This phenomenon frequently affects the predictive performance of the model over time. Ensemble based approaches often serve as a remedy for this problem. For example, in the dynamic weighted majority (DWM) method, individual decision trees are dynamically created and deleted according to changes in predictive performance. After detecting the drift, new low-diversity and high diversity forests are trained and predictions of new instances are drawn based on a weighted majority voting of the old high-diversity forest, the new low-diversity forest, and the old low-diversity forest.

Curse of dimensionality- Increasing the number of features fed into a machine learning model usually exponentially increases the search space and hence, the probability of fitting models that cannot be generalized. This phenomenon is known as the curse of dimensionality. Certain ensemble learning methods can be used to lessen the impact of the phenomenon.

VI. Criteria for Selection of Best Ensemble Method

Despite their high-predictive performance, other models may be preferred over ensemble models for two main reasons. First, predictions usually take a long time for large ensembles as many inducers are applied in order to aggregate a single prediction. This issue appears to be significant in real-time predictive systems. Second, it is almost impossible to interpret ensemble outputs as they consist of the outputs of many inducers. This property usually prevents the use of ensemble models in domains that require a clear and rational explanation for individual decisions.

In addition to predictive performance, there are other factors that may be considered when selecting the best ensemble method for a given problem:

Suitability to a given setting- Methods differ in terms of their capabilities depending on the setting (e.g., class imbalance, high dimensionality, multilabel classification, and noisy data). Therefore, characterizing the learning task and choosing a method accordingly is recommended.

Computational cost- The complexity cost for training the ensemble and the required prediction time for new instances may be considered as important criteria in assessing ensemble models, especially in real-time systems.

Software availability- Different platforms provide their own implementations for machine learning models. The ability to use the same algorithm in several applications may be an important attribute for implementation teams.

Usability- Users prefer to understand how to tune the models they use and hence, they prefer models that provide a clear set of controlling parameters.

VII. Commonly Used Ensemble Learning Algorithms

Bagging- Bagging is a method for generating diverse ensemble for model combination. For unstable classifier like some decision tree and the performance of the classifiers are similar, draw n samples from training data with replacement and update the classifier, run it for some iteration, finally combine all learned classifiers. The random forest has very similar structure, which uses bootstrap sampling. Bagging, which stands for bootstrap aggregating, is one of the earliest, most intuitive and perhaps the simplest ensemble based algorithms, with a surprisingly good performance. Diversity of classifiers in bagging is obtained by using bootstrapped replicas of the training data. That is, different training data subsets are randomly drawn with replacement from the entire training dataset. Each training data subset is used to train a different classifier of the same type. Individual classifiers are then combined by taking a simple majority vote of their decisions.

Boosting- Boosting is a magic-like method, which is also a method for generating diverse ensemble. Boosting combines weak learners to obtain a strong learner. The most widely used boosting algorithm is Adaboost, which stands for Adaptive boost. The Adaboost at each step produces a weak learner and updates the weights of training data (improve the weights of wrong classified data), finally combine these weak learners linearly to form a strong learner. FYI, the weak learner is which could classify better than random in any weights of the training data. Similar to bagging, boosting also creates an ensemble of classifiers by resampling the data, which are then combined by majority voting. However, in boosting, resampling is strategically geared to provide the most informative training data for each consecutive classifier.

Stacking- Stacking is an ensemble of classifiers is first trained using bootstrapped samples of the training data, creating Tier 1 classifiers, whose outputs are then used to train a Tier 2 classifier (meta-classifier). The underlying idea is to learn whether training data have been properly learned. For example, if a particular classifier incorrectly learned a certain region of the feature space, and hence consistently misclassifies instances coming from that region, then the Tier 2 classifier may be able to learn this behavior, and along with the learned behaviors of other classifiers, it can correct such improper training.

Random Forest- A random forest is a special modification of bagging that mixes the bagging approach with a random subsampling method. While bagging works with any algorithm as weak learner, random forests are ensembles of unpruned classification or regression trees. The commonly used growing algorithm for the single decision trees used within the random forest algorithm is CART. Just like bagging, random forests also sample attributes (without replacement) for each tree. The trees are grown to maximal depth (no pruning) and each tree performs an independent classification/regression. Then each tree assigns a vector of attributes or features to a class and the forest chooses the class having most votes over all trees

VIII. Conclusion and Future Scope

This research emphasizes the substantial benefits of applying ensemble learning techniques to the task of diabetes prediction. Ensemble learning, which involves combining the outputs of multiple machine learning models, offers a powerful approach to improving both accuracy and reliability in predictive analytics. By leveraging the diverse strengths of methods such as bagging, boosting, and stacking, we demonstrate that ensemble models can outperform individual algorithms, particularly in managing complex and imbalanced medical datasets. These advanced techniques allow for the aggregation of varied decision boundaries, reducing variance and bias, and thereby improving generalization to unseen data. The findings of this study underscore the potential of ensemble learning to revolutionize medical diagnostics by enabling early and more precise detection of chronic conditions like diabetes. This capability is crucial for timely interventions, more effective treatment planning, and ultimately, better patient outcomes. Furthermore, the improved predictive performance of ensemble models contributes to more efficient healthcare delivery and resource allocation, reducing the overall burden on healthcare systems.

Looking ahead, future research should focus on integrating heterogeneous datasets across populations and clinical settings to enhance model robustness. Continued refinement of ensemble methods will be key to advancing predictive accuracy, reaffirming the transformative impact of machine learning in the healthcare domain.

References

[1] Ganie, Shahid Mohammad, et al. "An ensemble learning approach for diabetes prediction using boosting techniques." Frontiers in Genetics 14 (2023): 1252159.

[2] Birjais, Roshan, et al. "Prediction and diagnosis of future diabetes risk: a machine learning approach." SN Applied Sciences 1 (2019): 1-8.

[3] Sai, M. Jishnu, et al. "An ensemble of Light Gradient Boosting Machine and adaptive boosting for prediction of type-2 diabetes." International Journal of Computational Intelligence Systems 16.1 (2023): 14.

[4] Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." Procedia Computer Science 167 (2020): 706-716.

[5] Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15.3 (2021): 435-443.

[6] Engineering, Journal Of Healthcare. "Retracted: A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques." Journal of healthcare engineering 2023 (2023): 9872970.

[7] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.

[8] Rani, K. J. "Diabetes prediction using machine learning." International Journal of Scientific Research in Computer Science Engineering and Information Technology 6 (2020): 294-305.

[9] Deberneh, Henock M., and Intaek Kim. "Prediction of type 2 diabetes based on machine learning algorithm." International journal of environmental research and public health 18.6 (2021): 3317.

[10] Joshi, Ram D., and Chandra K. Dhakal. "Predicting type 2 diabetes using logistic regression and machine learning approaches." International journal of environmental research and public health 18.14 (2021): 7346.

[11] Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (IJERT) 9.9 (2020): 921-925.

[12] Ismail, Leila, et al. "Type 2 diabetes with artificial intelligence machine learning: methods and evaluation." Archives of Computational Methods in Engineering 29.1 (2022): 313-333.

[13] Fregoso-Aparicio, Luis, et al. "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review." Diabetology & metabolic syndrome 13.1 (2021): 148.

[14] Maniruzzaman, Md, et al. "Classification and prediction of diabetes disease using machine learning paradigm." Health information science and systems 8 (2020): 1-14.

[15] Kopitar, Leon, et al. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." Scientific reports 10.1 (2020): 11981.

[16] Xue, Jingyu, Fanchao Min, and Fengying Ma. "Research on diabetes prediction method based on machine learning." Journal of Physics: Conference Series. Vol. 1684. No. 1. IOP Publishing, 2020.

[17] Singh, Ashima, et al. "eDiaPredict: an ensemble-based framework for diabetes prediction." ACM Transactions on Multimidia Computing Communications and Applications 17.2s (2021): 1-26.

[18] Geetha, G., and K. Mohana Prasad. "An hybrid ensemble machine learning approach to predict type 2 diabetes mellitus." Webology 18.Special Issue 02 (2021): 311-331.

[19] Fitriyani, Norma Latif, et al. "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension." Ieee Access 7 (2019): 144777-144789.

[20] Rahim, Md Abdur, et al. "Stacked ensemble-based type-2 diabetes prediction using machine learning techniques." Annals of Emerging Technologies in Computing (AETiC) 7.1 (2023): 30-39.

[21] Singh, Sapna, and Sonali Gupta. "Prediction of Diabetes Using Ensemble Learning Model." Machine Intelligence and Soft Computing: Proceedings of ICMISC 2020. Springer Singapore, 2021.