International Journal of

INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Predictive Recommendations for Diabetes using Ensemble Techniques with Model Explanations

Jayshree Ghorpade*[1], Balwant Sonkamble[2]

Submitted: 06/05/2023 **Revised**: 14/07/2023 **Accepted**: 06/08/2023

Abstract*:* The modern life-style and post corona impressions have impacted the public health with unending medical issues. The facts-figures of International Diabetes Federation (IDF) Atlas depicted that approximately 11% of the adults along with the teenagers are having one of the major non-communicable diseases called diabetes while few of them are still ignorant about the health disorders. The precautionary measures must be taken to minimize the effect of diabetes by furnishing premature diagnosis and thus assist the people to evade the medical complications. The proposed study delves into the medical data with electronic health records and various learning models with advanced algorithmic techniques. The research outcome based on predictive recommendations helps to portray the safety measures to avoid health hazards. The proposed algorithm with ensemble approach that aims to combine different algorithmic techniques and weighted probability for the heterogeneous data performs better with optimized estimates and model explanations using SHAP. The state-of-art-study depicts that a good sense of disease understanding can help to improve almost every facet of health sustainability and motivate the society to know the at-risk health alerts in advance. The study focuses on ensemble model with gradient boost, random forest, extreme gradient boosting, etc. to depict the reliability of proposed technique and demonstrate the comparative analysis with various benchmark datasets.

Keywords: *Machine Learning, Predictive recommendations, Ensemble, SHAP, Diabetes mellitus*

1. Introduction

It is projected that the rise in diabetes with an intensification of nearly 46% is taken forward due to the climatic, genetic, demographics, stress, socio-economical, etc. influences. The significant factors that have raised the diabetes mellitus type-2 (DMT2) embrace the modern life-style, the spread of age group including the teenagers, obesity with rising overweight, obliviousness to physical activities, etc. The proportion of diabetic complication evolution fluctuates across people with the understanding factors that amend the medical interventions [1]. The circulatory system, the urinary tract, nerves, and eyesight all experience fatal complications as a consequence of diabetes. It is vital to identify the elements that impede or encourage the emergence of these difficulties. Finding the elements that halt or stimulate the progression of these issues is of the utmost importance. The initial stages of the condition [2], particularly the variables that increase the risk of diabetes DMT2 development must be explored to avoid the health challenges.

Researchers explore different classifiers to expect the useful insights by analyzing the data. The current data handling methodologies have problems due to algorithmic stability. The complications arise when the imbalanced data-sets has the under representations of the class that refers to the concept of interest. An efficient feature engineering conduct elects a subset of unique features and assesses the list of potential features to determine its applicability. It regulates the feasibility of the model and reduces the feature extraction costs that leads to better generalization. The multiobjective Harris Hawk Optimization (HHO) optimal feature selection technique is suggested in [3] for the binary classification problem. Also, validating and defending model predictions can be difficult due to the trade-off among interpretability and accuracy, especially in high-stakes fields like medical care, law enforcement, stock analysis, etc. [4]. By providing a flexible way to comprehend complicated models while retaining a significant level of accuracy, SHAP values help in addressing this challenge. The ensemble learning having the knowledge of certain limitations of the individual models, tries to supervise the strengths and short comings of these models; thus, resulting in best possible generalized estimations [5]. Ensemble learning technique minimizes the risk in decision making by combining the various base learners and adapting to the situations or choices to improve the predictability of the merged model. The hypothesis with enhancements in performance metrics of the model is investigated to understand the statistical behavior of the ensemble group candidates and gain knowledge [6]. Experimental analysis with the publicly available data from UCI Machine Learning (ML) repository suggest that these proposed algorithmic ensemble learning techniques show

[1] *Research Scholar, P.I.C.T., S.P.P.U., Asst. Prof., M.I.T.W.P.U., Pune, India*
ORCID ID : 0000-0002-2131-3618
[2]*Department of Computer Engineering, P.I.C.T., S.P.P.U., Pune, India*
ORCID ID : 0009-0001-8681-6347
** Corresponding Author Email: jayshree.aherghorpade19@gmail.com*

some enrichments in the model outcome. The remarkable accomplishment of algorithmic ensemble techniques validates that they manage the trade off to yield better outcome [7]. A hypothesized technique showed the advantages of boosting technique known as AdaBoost and Wagging that represented a variant of Bagging [8].

The fundamental differentiator of the proposed model is the interpretation, which categorizes the precise patient metrics from multi-source data that adds to each distinctive role in the progression of health complication. In general, the proposed model along with the clinical interpretation identifies in advance the patients at risk for health problems. The model understanding with SHapley Additive exPlanations (SHAP) [9] explored the interpreters to assist the diagnosis for complications based on the projected study of predictive recommendations. The model's interpretability helps to understand the cause of a decision suggested based on the predictions [10]. The predictive recommendations based on the interpretability of the proposed model makes it simple for the user to comprehend the necessities of certain forecasts and choices with health sustainability. It assists in wisely using the extracted features with their relationships for the predictive performance. The prevalence of chronic conditions such as cardiovascular diseases, endocrine diseases, obesity, age, etc. are the comorbidities that intricate the fitness issues and these people are more susceptible to diabetes mellitus type-2 disease [11]. These issues lead to a necessity for having the automatized patient center methods. The recommendations for health sustainability emphasizes the safeguard monitoring practices of the comprehensive life with rational, spiritual & corporal actions. Thus, it will normalize the medical treatments and minimize the allied cost as the interpreted model clarifies the clinically pertinent elucidation of each contributing individual in the sample towards the prognosis.

The paper includes a brief overview of ensemble approach and model interpretations as discussed in Section 2. The mathematical significance of proposed research is depicted in Section 3. Section 4 portrays the experimental results with its discussion. To finish, section 5 states the comparative analysis and the last section gives the concluding remarks.

2. Previous Work

Ensemble learning is an all-purpose multiple classifier approach that is strategically produced with various Machine Learning techniques to seek and solve the computational problems. The fundamentals of this novel meta-strategy are to improvise the function approximations, classifications, predictive power, etc. of the model. It tries to minimize the likelihood of selecting the irrelevant data and thus assigns a confidence to the decision by considering the effective features of the model. The data-fusion and error reduction technique with function approximation focuses on incremental learning which is opted to solve the classification problems [12, 19]. The recent advancements in Machine Learning algorithms and computational techniques with ensemble approach has led to multiple investigations in interdisciplinary research with efficient data processing mechanisms along with model interpretations to generate appropriate recommendations. The electronic health record (EHR), which includes several forms of data about patients such as transcript signs, diagnosis, medication, laboratory, etc. is one invaluable source that is generally unexplored. The multi-source data depicts many facets of an individual's health. The literature study states that across the world there is a rise in mortality rate by 13% [13], which is mostly seen in working-class of a nation. The challenges to process the medical data, fraud detection, image processing, etc. requires more prominent algorithms while dealing with real time imbalanced datasets.

The reviews take the attention of the researchers on ensemble learning and its computational approach to help address the healthcare issues by offering sensible recommendations based on accurate predictions. The modern statistical studies and analysis have posed the requirement to convert the biomedical input data from unstructured form to structured form with valuable information to perform the high-level clinical research and have enhancements in algorithmic techniques to solve the medical issues [14]. One of such data is the need to analysis the risk factors for the diabetic disease. Identifying the most appropriate classifier to solve a classification problem is one of the crucial tasks. It is observed that the classifiers with the low error on training data can give false impressions for the unseen data while building the classification model and acquiring its performance. The classifiers having similar pseudo behaviour while training phase and depicting nearly same evaluations on the validation data may be selected at random. But the research challenge is to choose an appropriate classifier model that may be selected at random amongst the others having same performance while training; but with this ambiguity there is a possible risk of electing an unfortunate model with lower performance. Thus, there's a research gap that demands for the ensemble strategy which will implement an algorithm encouraging the performance metrics and exhibiting some level of diversity of the base learners. The important aspects to ensure the diversity is through data resampling or by changing the structure of the individual learners [15]. While solving the classification problems the diversity can be implemented with various training parameters that will help to achieve unlike decision boundaries. Finally, the strategic merging of the multiple classifiers will lessen the total error.

Machine Learning models are interpreted using SHAP (SHapley Additive exPlanations) values. They give a mechanism to comprehend the role played by every attribute or parameter in the forecasting decisions made by the model. The SHAP values enable the understanding of the extent to which each feature/parameter affects the outcome or prediction generated by the model by offering each attribute a specific numerical value. The goal of SHAP is to determine each contributor's contribution to the interpretation of the estimate of an instance [16]. Using the explainability approach, the Shapley values from constructive coalitional theory of games were determined. The feature values are often thought of as coalition members. In accordance with the Shapley values, the payout and prediction should fairly be distributed across the characteristics. The cooperative game theory, notably acts as the foundation for SHAP value analysis where the importance of each player's contribution towards the final result is assessed by captivating into account all feasible player alliances. Such a perception is extended to specific features in a Machine Learning model by using SHAP values and producing proper analysis [17]. The relevance of features in the model as a whole can be resolute using SHAP values. The factors that have the most effects on outcome decisions may be built by studying the average magnitude of SHAP values throughout a dataset. One of the main advantages for the use of SHAP is the fact that it meets a number of desired qualities, such as local correctness, fair distribution, consistent result, etc. The summative measure of the SHAP values represents the variance between the output of the model and the average output throughout the dataset, which quantifies the well-adjusted attribution of feature significance. Thus, thoughtful direction and size of the feature's impact on predictions is made easier to grasp by the visualization assistance of SHAP values [18]. Every attribute on the output produced by SHAP is explicated in terms of its importance.

Researchers have designed multiple experimentations on the Pima Indian diabetes dataset which was provided by the National institute of diabetes and digestive and kidney diseases [UCI]. The analytical study for Diabetes Mellitus (DM) in United States during 2016-2021 revealed the significant outpouring of more than 30% in mortality. The grownups with age 25-45 years were more susceptible to diabetes [20]. The proposed ensemble algorithm emphasis on the learning strategies with multiple techniques to improvise the working of the Machine Learning techniques and models. The pancreas in abdomen produces a hormone called insulin, which works as an essential component to allow blood sugar to enter the cells of the body to be utilized as fuel and produce energy. The insulin resistance occurs when cells don't react to insulin as they should in people with DMT2 type of diabetes. Accordingly, the blood sugar ultimately raises as a result of the pancreas' inability to keep up and thus it can lead to diabetes mellitus type 2 and prediabetes. The medical forecast by World Health Organization (WHO) states that DMT2 diabetes will double worldwide by 2040 if proper corrective measures are not taken in time. Signs and symptoms of DMT2 often show up gradually over a period and might be ignored for an extended duration where in certain cases, there may be no indications at all. Since these indications might be problematic to recognize, it's vital to be aware of potential risk factors [21] and visit the respective medical professionals if it happens to experience any of these conditions. Thus, research gap demands for more study to determine the most effective way to promote lifestyle modifications, particularly among those at highest risk of developing diabetes and suffering undesirable health effects. The people need to understand the ideal causes for DMT2, so more learning-modeling studies are essential with the predictive recommendations.

3. The Proposed Technique

The research study demonstrates the progressive and unconventional tactics using ensemble of univariate and multivariate Machine Learning algorithms with well-trained classifiers. Furthermore, it is anticipated that the base learners will have accurate performance measures rather than random assumptions. For the procedure to be effective, the particular learners must have specialized knowledge of the task being learned along with a range of fault determinations. The ensemble of classifiers helps to produce a progressive performance in a committee classifier where weak leaners are applied to diverse data from the training set's distributions and further these outputs from every ensemble associate are combined aptly for classification. Further, in order to offer a correlation among the input data and output predictions, the model should be interpretable, which will assist to intimate the appropriate conclusions. The claim for explainable or interpretable models has been particularly marked in the medical work. Predicting an individual's therapeutic result is crucial but it's also important to take into consideration the individual's traits in a quantitative and comprehensible way, such as vitals, medication usage, etc. In addition, excellent models should offer practical guidance for prevention that can be bought into actions. Simply putting a patient into one of several health categories is not particularly useful [22]. The benefit of a model is determined by the explanations of what has to be done in order to change an undesired condition or the detection of early risks.

3.1. The Model

By translating the data into an efficient and insightful format that the Machine Learning algorithms can use, the processing of input enables the model to investigate the beneficial aspects in the system. The most prevalent approaches for building ensemble classifiers are bagging and boosting. The basic learner used in typical algorithms for Machine Learning is employed on the data sample to create the classifiers. The bootstrap samples from the training data are processed by randomly selecting 'n' cases with replacement and evaluated by base algorithms in Bagging. The central objective of ensemble learning is acknowledging the limitations of single Machine Learning techniques and harnessing the performance of multiple base learners to minimize the variance and bias trade off that affects the overall predictive power of the model. The generalized ensemble formations can produce the classifiers that will limit the variance and bias errors associated with single Machine Learning techniques. Bagging is a method that reduces the error of variation without raising the bias trade off; whereas, boosting reduces bias. Ensemble classifiers are robust with capabilities to perform more than the individual learners. Few of the boosting algorithmic techniques are Gradient Boost (GB), extreme Gradient Boost (XGBoost), AdaBoost, etc. The Random Forest (RF) algorithm is the bootstrap aggregation technique with Decision Tree (DT) as its base learner.

The boosting strategy operates by recurrently applying a weak Machine Learning technique on varied distributions over the training data, by constructing a composite classifier by combining the produced classifiers. The layout of the classifier developed in the previous iteration determines how the experimental data are distributed. For computing the overall distribution of the experimental data and combining the projections from each individual classifier, each boosting algorithms have their own set of criteria. During the execution of each iteration, in boosting by resampling, the probability distribution is used for resampling the training subsets. Whereas in boosting by reweighting, the weights allocated to each example are utilized in each repetition to train the base classifier. The original boosting techniques by Robert Schapire and Yoav Freund were not adaptive. Schapire and Freund built an adaptive boosting algorithm that worked better for solving the binary classification problems. The idea of stochastic gradient boosting has motivated to build an algorithmic technique that will adapt to resolve the classification problems. Precisely, the data is resampled and in each iteration a subset of data is generated with a size lesser than that of the original experimental data with weighted probability distribution.

The proposed technique seeks to identify and select the dependable important aspects. These explanatory characteristics features or factors explain how the response variable is forecasted.

3.2. Predictive Recommendations

With the right algorithmic strategies and pertinent characteristics, a model's processing may be improved, where the data should be dwelled into to have new cognizance. The comprehensive research study analyses various Machine Learning techniques to improve the outcome of the model and have appropriate recommendations to aid the decisions. The SHAP exhibited the biological links those were learned from the black-box model. The exclusive state-of-art study in the research realm leverages SHAP interpretations to explore the influence of a model's capability to differentiate amongst the groups. The advanced learning algorithms are a shift in paradigm where the traditional programming is intelligently transferred to data-driven criterions and patterns to discover uniqueness in the realm of research studies [23]. The trained model, represented as $\hat{f}(x)$ depicts the predictive recommendations with the interpretability of the research techniques and agnostic methods. The explanatory features within the dataset '*D*' establish a relationship amongst them to deduce the predictor outcome. The observations with an instance obv_j across the feature standards x_i are involved to produce y_i as depicted in (1).

$$y_i \leftarrow \hat{f}(x_i^j)\ \{i \leq |features|,\ j \leq |observations|\}\ \forall i,j \in \mathbb{R}$$
(1)

The interpretability mentioned as intrinsic talks about the simple elementary structure of the model whereas the post hoc one refers with the interpretation approach that deals with the subsequent training of the model such as the permutation feature importance, etc.

The data-driven approach with algorithm transparency helps to gain the model insights based on the trained data that supports the holistic decisions. The model interpretability explains the effect of every instance of the data towards the distribution of the predictor outcome. The impact of feature importance on the average predictions needs to be estimated with the correlation of the parameters that will aid to the ethical beliefs and form the foundation for the predictive recommendations. The long-standing health hazard challenges categorized as cardiovascular, genitourinary, late effects, etc. are associated either with the persons suffering from undiagnosed DMT2 or the ones who are more susceptible to have DMT2 [24].

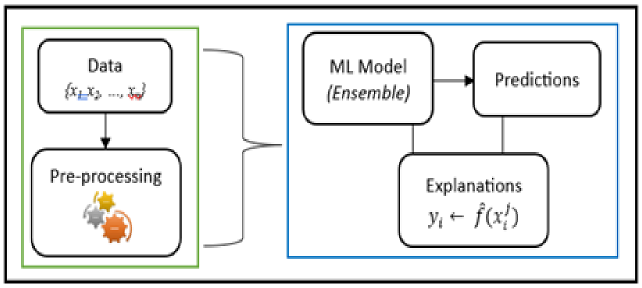

Fig. 1. Proposed Model for Predictive Recommendations

Various factors are responsible for the disparities in various diabetes pervasiveness. The prime risk factors for DMT2 and prediabetes includes being overweight and obese. Encountering a medical diagnosis that anticipates diabetes mellitus (DMT2) in patients enables early management of the ailment, shortens diagnostic duration, along with the savings of the health system cost & respective patient. The research work represents a simple predictive recommendation approach for the same. In particular, if these approaches as shown in Fig.1 are used in sectors that demand the highest degree of responsibility and openness, the study employing predictive recommendations aims to demonstrate the necessity for suitable judgments and predictions provided by Machine Learning techniques and Model Explanations.

4. Results and Discussion

The research study set out a technique to comprehensively investigate and systematically evaluate the conduct and communication of procedures applied to create risk prediction models for estimating the likelihood that persons may either have DMT2 incidentally or undiagnosed in the future.

The research analysis exhibited the correspondence of the extraneous inputs to the proposed model and its impact on the predictions to depict appropriate recommendations to avoid the onset complications. The experimental investigations reveal the proposed research suggestions when applied to the publicly available *Diagnosis, Lab, Transcript, and Medication* data arriving from multiple sources. Evaluation and estimation of each of the multi-source heterogeneous data assisted the predictive recommendations to understand and figure out the efficacy of each data source for the health complications towards DMT2. The *Diagnosis* data showed an accuracy of 80.85% with optimal features. The model interpretation standards are one of the analytical structures which clarifies the complicated Machine Learning models with appropriate forecasts that assist the predictive recommendations. Let '*D*' be the M-dimensional feature vector for a predictive model as shown in (2).

$$\mathrm{D} \leftarrow \{x_{1,} x_{2,} \ldots, x_M\} \quad (2)$$

The study of various specific features that primarily pay for guessing the predictor while interpreting the model helps to acknowledge the fatal consequences in future. The SHAP simply establishes the concern amongst the features and explains how a particular feature is directly involved in determining the outcome (x). Shapley values are allied with coalition game theory where the main objective is to fairly govern the consequence of each performer (e.g. feature) on the complete result of the team/integrated set of features. Recent research study is exploring the assurance for the proposed ensemble Learning algorithm *Ensemble Bootstrap Genetic Algorithm (EnBGA)* [4] for deducing the most relevant crucial optimal features along with the weighted probability model that tries to guarantee the importance of the prediction which will in turn support for the suitable recommendations [25]. The binary classification problem such as Diabetes DMT2 classification in learning paradigm demands for the appropriate framework analysis with concerned features that contribute towards the predictive recommendations.

The SHAP values $\varphi_1, \ldots, \varphi_M$ depending upon the relevance of the features with '$1 < j < M$' as a set of aggregated features along with the characteristic function $v(\ )$ for a predictive model [17] $f(x)$ at $x = x^*$ is depicted in (3). φ (contribution of every feature) is a function that signifies the probable estimations for the features in all subsets $\mathcal{S}$.

$$\varphi_j = \sum_{\mathcal{S} \subseteq M \backslash (j)} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 1)!}{M!} (v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})) \quad (3)$$

where $\varphi_j \leftarrow$predictive variation due to j^{th} observing feature over information of the remaining features.

The overall value for φ_j raises exponentially with the increase in features for the integrated model. One of the instinctive approaches known as Leave-One-Out (LOO) technique [18] includes the performance of the integrated model with all features verses the one without.t the feature at interest. Accordingly, the significance of each feature is the difference between the complete result with the entire data set and the outcome consequences without the feature at interest. Thus, SHAP explains and identifies the effect of every single contributor on the final model predictions. SHAP values offer local interpretability, which means they may be used to forecast a particular occurrence or observation [26]. The involvement of each attribute to a specific prediction is quantified by SHAP values.

Shapley values provide a tool for assessing a predictor's significance with respect to others. These values reflect the impact of having the knowledge against not knowing about that predictive feature on the model error with loss function while tracking the relationship directions. The graphic makes it possible to see the contribution of multiple features towards the model's projection for a particular observation. The impact is increased with the rise in absolute value. The nearer the feature is to the line separation of red & blue, the greater the degree of its influence on the score, and the size of the bar indicates its impact. As seen in Fig.2 the model's projection for an observation under interest is (x) = *[-2.58]* with higher positive impact of *BMI_avg, Cardiovascular_Disorders* whereas *Nutritional_Disorders* has negative impact. Thus, the predictive recommendation depicts that the observation under interest with these disorders is susceptible to develop DMT2 in future and advised to consult the diabetologists for further medical investigations for healthcare treatments.

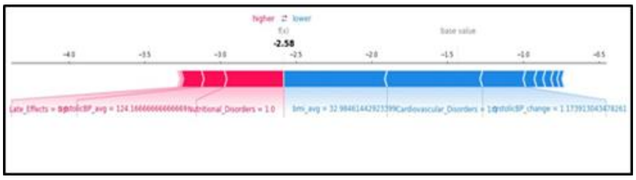

Fig. 2. SHAP with force plot

The beeswarm plot shows a summary of features with their information-dense influence on the model. A single dot indicates each instance of the explanation. The SHAP value determines the dot's x-location. The prominent features represented at the top of the plot in Fig.3 are the at-risk factors which acts as threat for the health to have DMT2 as the forthcoming disease and thus are useful in envisaging inception of DMT2 disease.

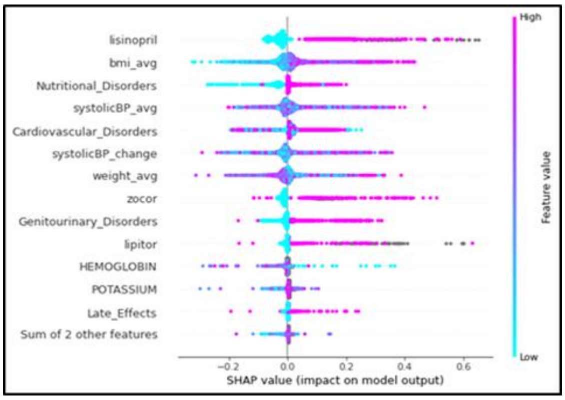

Fig. 3. SHAP with beeswarm plot

The relative significance of various characteristics or features of the multi-source heterogeneous dataset are comprehend to gain substantial visions into how a Machine Learning model produces predictions by utilizing SHAP values. This information portrays the understanding of feature handling, bias consideration, exploring the debugging of model, etc. thus establishing the confidence in the model's guesses and building the trust for predictive recommendations to resolve the challenges. Setting aside preventative measures would be made easier with a more accurate classification of overall risk of T2DM occurrence [27]. Few of the effective recommendations and preventive measures are the resources at the Centers for Disease Control and Prevention (*www.cdc.gov/diabetes/*) and the National Diabetes Prevention Program (*www.cdc.gov/diabetes/prevention/*). The National Institutes of Health assists for the management and avoidance of DMT2 conditions. In order to avoid type 2 diabetes in people at greater risk, the Community Preventive Services Task Force suggests dietary and physical activity intervention initiatives (*www.thecommunityguide.org/findings/diabetes-combined-diet-and-physical-activity-promotion-programs-prevent-type-2-diabetes*).

5. Comparative Analysis

It is challenging to recognize and use the indicators to diagnose diabetes at an early stage wherein the issue is still unresolved. Thus, it has inspired us to tackle the diabetes prediction issue by identifying & analyzing the results of all pertinent solutions with individual investigations. Most of the researchers executed their model on the benchmark PIMA Indian diabetes dataset, which is publicly available at the UCI data repository.

The correlation coefficient, which is a useful tool for determining how closely two variables are related, was used by the authors. When the dataset's two properties are closely connected, it is possible to ignore one of them to prevent duplication as shown by Jackins et. al. [28] through the correlogram matrix. As the medical data is complicated, non-normal, and organized by correlation, the process was time consuming. The method with Naïve Bayes (NB) and Random Forest (RF) achieved an accuracy of 74.46%. Mohapatra et. al. projected a schematic process for uncovering the possibilities of diabetes by means of multilayer perceptron [29] with an accuracy of 77.5%. But, the lacuna of the method was the lack of optimal feature selection method. Sisodia et. al. [30] suggested a technique including DT and NB, which gave an accuracy of 76.3%, but the disadvantage was that no proper preprocessing techniques were employed on the dataset; instead, it's a relatively straightforward method used for classification. Tasin et. al. [31] proposed the diabetes prediction system using the XGBoost classifier and explainable AI approach, which represented 74% of accuracy.

Table 1. Comparative analysis for PIMA diabetes dataset

Classifier	Accuracy (%)	F1 Score	Precision	Recall
MLP [29]	77.50	0.83	0.83	0.85
NB, DT [30]	76.30	0.76	0.75	0.76
XGBoost [31]	74.00	0.73	0.73	0.74
SoftVoting [32]	79.10	0.72	0.73	0.72
SVM [33]	75.00	0.73	0.72	0.75
LR [34]	78.26	0.57	0.47	0.71
Proposed	79.87	0.74	0.91	0.81

The designed ensemble soft voting learner by Saloni employs a combination of random forest, logistic regression, and naive bayes to provide binary classification for the PIMA dataset with an accuracy of 79.04% [32]. Saiteja et.al. [33] trains a system to forecast a patient's high blood pressure and diabetes condition using the Machine Learning classification technique called Support Vector Machine that produced an accuracy of 75%.

With the aid of cutting-edge techniques and base classifiers, an empirical evaluation of the predictive recommendation approach has been carried out with the benchmark dataset and it has been observed that the proposed ‘EnBGA’ ensemble algorithm produces the best results with an accuracy of 79.87%. The comparative analysis for the various research contributions with performance evaluation is depicted in Table.1 and Fig.4 for PIMA dataset.

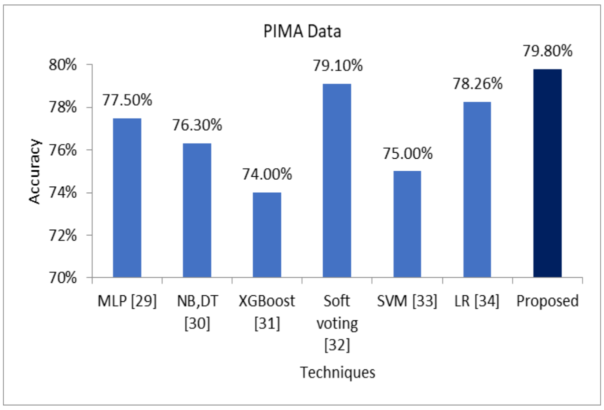

Fig. 4. Comparative analysis with Proposed Technique for PIMA data

Further, the effectiveness of the proposed algorithm was also established with the RTML dataset of diabetes mellitus collected from the employees of Rownak Textile Mills Ltd, Dhaka, Bangladesh [31]. The execution of the proposed ensemble technique as depicted in Fig.5, represented a better performance of accuracy as 97.56% as compared to the existing research initiatives with 96% and 81%.

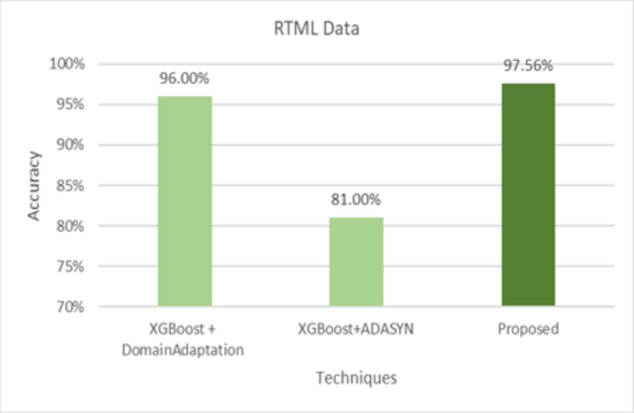

Fig. 5. Comparative analysis with Proposed Technique for RTML data

Breast cancer is another health hazards which is mostly observed in the females and it ranks first among females worldwide in terms of leading causes of death. Here, the breast's cells proliferate out of proportion.

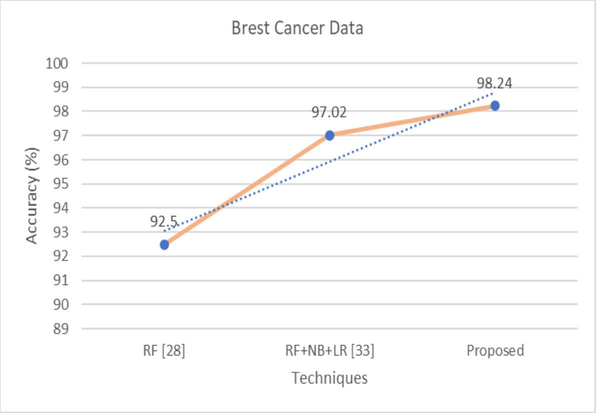

Fig. 6. Comparative analysis with Proposed Technique for Brest-Cancer data

To confirm the robustness of the ensemble technique, the suggested research technique and methodology has also been contrasted with the breast cancer dataset producing an improved performance with an accuracy of 98.24% and AUC of 98%, as compared to earlier results of 92.5% [28] and 97.02% [33] as shown in Fig.6.

6. Conclusion

The existing Machine Learning models collaborate to produce a solution for a problem by applying the decision fusion strategy of ensemble learning that has shifted the procedure-oriented analysis to data-oriented analysis. The proposed algorithm depicts performance improvement with bootstrap approach and ensemble technique to have an effective method for diabetes diagnosis. With the recent release of electronic health record data and predictive recommendations, there has never been a better chance to use forecasting methods to advance medical care and uncover possibly new risk factors. Establishing confidence in Machine Learning models and guaranteeing accountability in decisions depend on explanations and interpretability. By giving an unbiased and understandable approach to measure the influence of each characteristic feature on model predictions, SHAP values aid in bridging the gap among these ideas. Thus, the predictive recommendations based on suggested methodology with algorithmic techniques along with the SHAP interpretations reveals the risk variables at an early stage. Also, the comparative study for various benchmark datasets assists the effectiveness of the proposed research work.

Author contributions

Jayshree Ghorpade: Methodology, Software & Experimentation.

Balwant Sonkamble: Reviewing & Analysis.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Luis Fregoso-Aparicio, Julieta Noguez, Luis Montesinos and José A. García-García, “Machine learning and deep learning predictive models for type 2 diabetes: a systematic review,” Diabetology & Metabolic Syndrome, 2021, Vol.13. Isuue.148, pp.1-22, doi.org/10.1186/s13098-021-00767-9

[2] P. Colmegna et al., "Evaluation of a Web-Based Simulation Tool for Self-Management Support in Type 1 Diabetes: A Pilot Study," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 1, pp. 515-525, Jan. 2023, doi: 10.1109/JBHI.2022.3209090.

[3] Uzma Ghulam Mohammad,Salma Imtiaz, Manoj Shakya, Ahmad Almadhor and Fareeha Anwar, “An Optimized Feature Selection Method Using Ensemble Classifiers in Software Defect Prediction for Healthcare Systems,” Vol.22, Article ID 1028175 , pp.1-14, Jun’2022.

[4] Ghorpade, J., & Sonkamble, B., “Data-driven based Optimal Feature Selection Algorithm using Ensemble Techniques for Classification,” International Journal on Recent and Innovation Trends in Computing and Communication, Vol.11, Issue.4, pp.33–41.

[5] Agliata A, Giordano D, Bardozzo F, Bottiglieri S, Facchiano A, Tagliaferri R., “Machine Learning as a Support for the Diagnosis of Type 2 Diabetes”, International Journal of Molecular Sciences. 2023; 24(7):6775.

[6] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," in IEEE Access, vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[7] Webb, G. I., & Zheng, Z., “Multistrategy Ensemble Learning: Reducing Error by Combining Ensemble Learning Techniques,” IEEE Transactions on Knowledge and Data Engineering, 16(8), 980-991, 2004.

[8] Dutta, A.; Hasan, M.K.; Ahmad, M.; Awal, M.A.; Islam, M.A.; Masud, M.; Meshref, H. Early, ‘Prediction of Diabetes Using an Ensemble of Machine Learning Models. Int. J. Environ. Res. Public Health 2022, 19, 12378.

[9] E. D. Spyrou and V. Kappatos, "XAI using SHAP for Outdoor-to-Indoor 5G Mid-Band Network," 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2023, pp. 862-866, doi: 10.1109/CSNT57126.2023.10134625.

[10] D. Fryer, I. Strümke and H. Nguyen, "Shapley Values for Feature Selection: The Good, the Bad, and the Axioms," in IEEE Access, vol. 9, pp. 144352-144360, 2021, doi: 10.1109/ACCESS.2021.3119110.

[11] B. Shamreen Ahamed, Meenakshi Arya, Auxilia Osvin Nancy, “Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques”, Front. Comput. Sci., 10 May 2022.

[12] F. Sambo et al., "A Bayesian Network analysis of the probabilistic relations between risk factors in the predisposition to type 2 diabetes," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 2015, pp. 2119-2122.

[13] Diabetes, 5April’2023, World Health Organization Newsroom, https://www.who.int/news-/ news-room/fact-sheets/detail/diabetes

[14] Karina W. Davidson, PhD, MASc, “Screening for Prediabetes and Type 2 Diabetes US Preventive Services Task Force Recommendation Statement”, JAMA. Aug’2021, Vol.326, Issue.8, pp.736-743. doi:10.1001/jama.2021.12531

[15] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag, "Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors", Big Data 2015, Vol3(4), pp.277-287.

[16] N. S. Choudary, V. B. Bommineni, G. Tarun, G. P. Reddy and G. Gopakumar, "Predicting Covid-19 Positive Cases and Analysis on the Relevance of Features using SHAP (SHapley Additive exPlanation)," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), India, 2021, pp. 1892-1896.

[17] Martin Jullum, Annabelle Redelmeier and Kjersti Aas, “Efficient and simple prediction explanations with groupShapley: A practical perspective,” XAI.it 21 - Italian Workshop on Explainable Artificial Intelligence, 2021, pp.1-15

[18] Bloch, L., Friedrich, C.M. &amp; for the Alzheimer’s Disease Neuroimaging Initiative, ‘Data analysis with Shapley values for automatic subject selection in Alzheimer’s disease data sets using interpretable machine learning,’ Alz Res Therapy 13, 155 (2021).

[19] Ruchika Malhotra, Anjali Sharma, “Threshold benchmarking for feature ranking techniques”, Bulletin of Electrical Engineering and Informatics, Vol. 10, No. 2, April 2021, ISSN: 2302-9285, pp. 1063-1070.

[20] Lv, Fan & Gao, Xu & Huang, Amy & Zu, Jian & He, Xinyuan & Sun, et. al.,“Excess diabetes mellitus-related deaths during the COVID-19 pandemic in the United States,” eClinical Medicine, Vol.54. 2022.

[21] Nowakowska, M., Zghebi, S.S., Ashcroft, D.M. et al. The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. BMC Med 17, 145 (2019).

[22] Saarela, M., Jauhiainen, S., “Comparison of feature importance measures as explanations for classification models,” SN Appl. Sci. Vol.3, Issue.272, 2021, doi.org/10.1007/s42452-021-04148-9)

[23] Kuhn HW, Tucker AW, ‘Shapley LS. A value for n-person games’, Contributions to the Theory of Games. vol. 2. Princeton, US: Princeton University Press; 1953. p. 307{318.doi:10.1515/9781400881970-018.

[24] 24. Diabetes and Asian Americans. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/library/spotlights/diabetes-asian-americans.html

[25] Wu, C., Wu, J., Luo, C. et al., “Recommendation algorithm based on user score probability and project type,” J Wireless Com Network, 2019, 80.

[26] Glauber H, Vollmer WM, Nichols GA, “A Simple Model for Predicting Two-Year Risk of Diabetes Development in Individuals with Prediabetes”, Perm J. 2018; 22:17-050.

[27] Chukwuebuka Ejiyi, Zhen Qin, Joan Amos, Makuachukwu Ejiyi, Ann Nnani, Thomas Ugochukwu, Victor Kwaku Agbesi, Chidimma D., Chidinma O., "A robust predictive diagnosis model for diabetes mellitus using Shapley-incorporated machine learning algorithms," Healthcare Analytics, Vol.3, 2023, 100166, ISSN 2772-4425.

[28] V. Jackins, S. Vimal, M. Kaliappan, M. Y. Lee, “AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes,” Fe Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, 2021.

[29] Mohapatra, S.K., Swain, J.K., Mohanty, M.N. (2019). “Detection of Diabetes Using Multilayer Perceptron,” In: Bhaskar, M., Dash, S., Das, S., Panigrahi, B. (eds) International Conference on Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 846. Springer, Singapore. https://doi.org/10.1007/978-981-13-2182-5_11.

[30] Sisodia, D., & Sisodia, D.S., ‘Prediction of Diabetes using Classification Algorithms’. Procedia Computer Science, Vol.132, 2018, pp.1578-1585.

[31] Tasin, I., Nabil, T.U., Islam, S., Khan, R.,”Diabetes prediction using machine learning and explainable AI techniques,” Healthc. Technol. Lett. 10, 1–10 (2023). https://doi.org/10.1049/htl2.12039.

[32] Saloni Kumari, Deepika Kumar, Mamta Mittal, “An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,” 'International Journal of Cognitive Computing in Engineering,' Volume 2, 2021, Pages 40-46, ISSN 2666-3074.

[33] Saiteja C., Gahangir H., Ayush Goyal, Anupama B., Sayantan B., Devottam Gaurav, Sanju Mishra, “Smart home health monitoring system for predicting type 2 diabetes and hypertension,” Journal of King Saud University - Computer and Information Sciences, Vol.34(3), 2022, pp.862-870, ISSN 1319-1578.

[34] Joshi RD, Dhakal CK., “Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches,” International Journal of Environmental Research and Public Health. 2021; 18(14):7346.

[35] Ramasamy, J. ., Doshi, R. ., & Hiran, K. K. . (2023). Three Step Authentication of Brain Tumour Segmentation Using Hybrid Active Contour Model and Discrete Wavelet Transform. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 56–64. https://doi.org/10.17762/ijritcc.v11i3s.6155

[36] Robert Roberts, Daniel Taylor, Juan Herrera, Juan Castro, Mette Christensen. Enhancing Collaborative Learning through Machine Learning-based Tools. Kuwait Journal of Machine Learning, 2(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/177

[37] Agrawal, S.A., Umbarkar, A.M., Sherie, N.P., Dharme, A.M., Dhabliya, D. Statistical study of mechanical properties for corn fiber with reinforced of polypropylene fiber matrix composite (2021) Materials Today: Proceedings, .