## **ENSEMBLE LEARNING APPROACH FOR EARLY DETECTION OF DIABETES USING MACHINE LEARNING AND DEEP LEARNING**

V P Shrinikhaythan1, Aditri S Madhavan2, Omana J3

1,2,3School of Computer Science and Engineering, VIT University, Chennai

1[shrinikhaythan.vp2024@vitstudent.ac.in](mailto:shrinikhaythan.vp2024@vitstudent.ac.in), 2[aditri.smadhavan2024@vitstudent.ac.in](mailto:aditri.smadhavan2024@vitstudent.ac.in), 3[omana.j@vit.ac.in](mailto:omana.j@vit.ac.in)

## **Abstract:**

Diabetes mellitus is a chronic and potentially life-threatening disease which requires early detection. Prevalent since ancient times, it is a disease with multiple forms of diagnosis, physical and chemical, but with no definitive [cure. It](http://cure.it) is an entirely manageable disease to live with when given adequate treatment, but if left untreated, it can have a long-lasting impact on the human body and, in worst cases, can even lead to death. Therefore, the need for a timely diagnosis becomes increasingly necessary, and Machine Learning (ML) is a judicious method that can be deployed to predict diabetes at an early stage. Based on previous studies of the high performing models - performing Artificial Neural Network (ANN) and Catboost, this study aims to identify the most suitable ensemble model, utilising renowned data balancing techniques such as SMOTE-ENN and SMOTETOMK-ENN. Tests with both balancing techniques revealed SMOTETOMK-ENN to be the more effective technique when combined with the models mentioned earlier. Therefore, it was used along with Bayesian Optimisation of fine-tuning hyperparameters to create an ensemble model through a weighted voting system. The proposed ensemble model surpasses the evaluation metrics with flying colours, achieving a resounding accuracy of 98.94%, which is superior to other individual models of the same models known and thus makes it a better choice for use in risk-sensitive fields like healthcare, providing more accurate detection.

**Keywords:** Diabetes Mellitus (DM), Artificial Neural Networks (ANNs), Catboost, Synthetic Minority Oversampling Technique (SMOTE), Edited Nearest Neighbours (ENN), SMOTETOMK, Weighted voting system, Bayesian Optimization, one hot encoding, KNNImputer, Isolation Forest, Local Outlier Factor (LOF), Z score Normalisation, Matthew Correlation Coefficient (MCC), Negative Predicted Value (NPV)

**I INTRODUCTION**

Diabetes mellitus (DM) is scientifically defined as a metabolic disease characterised by symptoms resulting from the excessive amount of glucose present in the blood. This excess amount is believed to be present due to insufficient insulin secretion or insulin resistance [1]. According to studies done on the topic, it is expected that in the coming years, a large section of people will be affected by diabetes as compared to before. Such a considerable increase prompts people to take more care in obtaining an accurate diabetes prediction and determining the best ways to manage it [2]. While this disease is not life-threatening, it still has a fatality rate of 85%. It is mainly because accurate clinical detection of diabetes takes so long that people start developing other diabetes induced symptoms such as heart attacks, other organ failures, etc, which are life-threatening [3]. All these symptoms typically appear 7-12 years after the onset of the disease, during which time it is often unnoticed. Therefore, the degree of manifestation and associated complications is closely related to the detection period of DM[4]. Despite all of this, it is worth noting that diabetes has no proper cure, only treatments that can help control it [5]. Therefore, effective management, timely medical intervention, and avoidance of long-term health ailments associated with DM all depend on the early detection of DM. As the management of the disease primarily focuses on addressing its symptoms, it involves controlling blood pressure and maintaining the health of the eyes and feet [6]. For this purpose, laboratory methods such as the oral glucose tolerance test (OGTT), fasting blood sugar (FBS), and HbA1c tests are typically employed in a traditional diagnostic process. These techniques are known to be both expensive and time-consuming.

On the other hand, emerging machine learning-based models appear to be non-invasive, economical, and equally adequate substitutes [7]. Machine learning (ML), a branch of artificial intelligence, enables the production of results from existing data without the explicit need for programming [8]. It analyses and mines large-scale data (big data), allowing the systems to adapt and learn from experiences [9]. It computes models to solve all-around problems and understand things like a human would. As a result, Machine learning (ML) models are commonly used for binary classification, often employing feature selection and other models. Another subset of Artificial Intelligence is Deep Learning (DL), a model that demonstrates the ability to learn image features automatically [10]. DL produces results of high accuracy, as it has large learning structures that allow it to delve deep into data traits and make inferences. It utilises neural networks, inspired by the human brain's functioning [11].

**1.1 Problem Statement**

As explained above, the early detection of diabetes is crucial. Several studies have undertaken this task, resulting in multiple powerful machine learning models for early and accurate prognosis. However, all these models suffer from variations of limitations, such as overfitting, sensitivity to data imbalance, and poor generalisation of datasets. This study thus aims to create a comprehensive ensemble model free from such limitations by combining two other standalone models, namely ANN and Catboost, both tested with prominent data balancing techniques, SMOTE-ENN and SMOTETOMK-ENN, to gauge the better model among them and combine them under a weighted voting system to get the best out of them. This ensures the study's aim to improve diagnostic accuracy, thereby providing healthcare practitioners with a more accurate and reliable diagnostic tool.

**1.2 Research objectives**

The primary objective of this paper is to determine the optimal combination for creating a highly reliable and robust ensemble with high diagnostic accuracy for the early detection of diabetes. The key objectives are:

* To design a novel ensemble framework for the early detection of diabetes by combining the benefits of ANN and Catboost models for improved diagnostic performance.
* To incorporate advanced data balancing techniques like SMOTE-ENN AND SMOTETOMK-ENN with both the individual models to figure out their best combination.
* To implement a weighted voting system to Ensemble the ANN and Catboost models to create an optimised ensemble model for better performance.
* To assess the model's performance using comprehensive metrics, such as accuracy, precision, recall, F1 score, MCC, and NPV. The use of MCC is because it provides a very balanced view of model performance by considering all four components of the confusion matrix. And NPV is used to measure the reliability of pessimistic predictions in the model, which is of utmost importance, especially in critical fields like healthcare.

Section II discusses various other related works. Section III discusses the research methodology used to achieve the proposed model. Section IV analyses the model's results, demonstrating its superiority. Section V concludes the paper by summarising the key findings and suggesting avenues for future improvement. Section VI contains the reference papers referred to in this paper.

**II LITERATURE REVIEW**

Diabetes Mellitus (DM), a disease plaguing a varied multitude of people around the world with increasing incidence rates, is no joke to take lightly. If such a disease remains unchecked, its associated symptoms will run rampant on one's body, affecting the organs. Following this analogy, many researchers have analysed this topic using multiple techniques, resulting in numerous innovative approaches and methods, even when limiting the scope to machine learning and deep learning models. Some of the more innovative ones include the following:

In [12], Artificial neural networks (ANN) and CatBoost were used for the early detection of diabetes mellitus. Based on clinical and demographic characteristics, both models demonstrated significant promise in distinguishing between cases with and without diabetes. CatBoost offers excellent accuracy and interpretability due to its strong generalisation capabilities and reliable handling of categorical data. Conversely, the ANN model contributed to competitive performance metrics, achieving an accuracy of 95% above by effectively capturing intricate nonlinear interactions in the data. Although both models performed well, the comparative study revealed that specific project needs, including interpretability, Training time, and computational resources, can influence the decision between them. The results underscore the significance of machine learning methods in the medical field, enabling early diagnosis and improving patient outcomes. Future research may focus on enhancing model performance by incorporating more diverse datasets, refining feature engineering, and optimising hyperparameters. Furthermore, the use of these models in clinical decision-support systems can be crucial for optimising resources and promoting preventive healthcare practices.

In [13], the dataset was obtained from Pima India, which contained approximately 768 records. Preprocessing methods, including the removal of null values and outliers, were applied to the dataset. On the clean dataset, several machine learning models were applied, including Logistic Regression, Random Forest, Decision Tree, and SVM, among others. Hyperparameter fine-tuning was performed using grid search. The results show that the Logistic Regression model is the most accurate, with an accuracy of 82.8%. But for further development, it needs to be tested on larger datasets. Focusing on the multiplicity of data sources and their correlation to the prognosis of DM in patients, according to the results of [14], the most significant risk factors of diabetes are the body mass index (BMI), glucose levels and age. Results showed 76.30% accuracy for the Random Forest (RF) method and 78.57% accuracy for the Neural Network Algorithm. It has been noted that the study utilised a smaller dataset, and the methods employed may not apply to other categories of the population or similar situations. Additionally, factors such as lifestyle and genetics, which are known as potential influences, were not considered in the study.

Such shortcomings were addressed in the study by [15], which examined the following variables: age, gender, family history of diabetes, family history of hypertension, hypertension measurement, blood sugar level, obesity, and central obesity, all with a P < 0.05. Upon analysis of family history with DM using multiple variables, the highest OR value is found to be 15.101. That is, in other words, people with a family history of SM2 have 15.101 times a greater risk of developing said disease. The blood sugar level displays the lowest odds ratio, with an OR of 0.016 (protective). The Random Forest algorithm achieves a model accuracy of 84%, which is higher than the anticipated 80%.

Similar to the principle of healthcare professionals, [16] employs an AWOD-based method that focuses on individual patient health conditions for final diagnosis. It proposes a method for prioritising information that involves calculating the average values of the expected and acceptable values to determine the appropriate weighing factors. These differentiate the significance of factors for each patient, aiming to display the reality of diverse individual health conditions that affect prediction. Two datasets were primarily used for testing, namely the Pima Indians Diabetes dataset and the Mendeley Data for Diabetes, each with 392 records. The evaluation metrics used were precision, specificity and accuracy. The results of the proposed AWOD-based method revealed 93.22% and 98.95% accuracy for both respective datasets, proving to be more accurate than other ML-based learning methods, such as K-NN, SVM, RF, and DL.

Following [17] and [18], these also address feature selection techniques and data processing methods when applying machine learning methods. [17] uses Random Forest algorithm and gains an accuracy of 98% while [18] works on fuzzy logic decision level fusion and achieves an accuracy of 94.87%. Both methods yield significantly higher results than other works in the same techniques and demonstrate high detection accuracy.

The next problem to be addressed was the high class imbalance in the datasets. [19] uses the DPD dataset, known for its high class imbalance. Using the ProWSyn method to mitigate the potential biases towards the majority class. It also employs two different models, Highway and LeNet, for its ensemble model, combining them using two distinct approaches: blending and hybridisation. It yields an accuracy of 94% and an F1 Score of 96%. These results far exceed those of other DL models, thus proving the effectiveness of its hybrid strategy. Similarly, combining the highway and LeNet models using the same Ensemble approaches results in a 94% accuracy and 94% F1 Score. Such a high value result proves the importance of the HiTCLe and Hi-Le models. It is also worth noting that a 10-fold cross-validation (FCV) strategy was employed in the evaluation process to enhance the reliability of the models. The interpretability was also enhanced using the SHAP method, which guided the model's fundamentals. For further development, the ablation research evaluates each element of the model and provides pointers for improvement.

Another study addressing the class imbalance problem was [20], which proposed an automatic detection and decision-making system that can diagnose diabetes and its symptoms based on risk factors. Using the BRFSS dataset and various machine learning methods, the model was evaluated using comprehensive metrics, including accuracy, precision, sensitivity, specificity, f-measure, and ROC/AUC score. The results revealed that the KNN outperformed all other comparisons, mainly due to the use of SMOTE-ENN to address the dataset imbalance. The study also attempted to use the SMOTE method alone, but it did not yield results that varied from those of the existing approaches. The model's efficiency can be increased by trying out differing datasets and any other upcoming advanced methods.

Another study [21] uses a Deep Belief Network (DBN) enhanced with attention mechanisms, GAN-based data augmentation, and a hybrid loss function. This study aims to address several existing machine learning-based problems, including class imbalance, feature relevance, and classification accuracy. The dataset imbalance problem is addressed through the use of GANs for generating synthetic data, thereby enabling accurate classification of minority cases. This allowed it to prioritise the more important feature, enabling key indicators such as polyuria and polydipsia to be the primary focus. Furthermore, by combining cross-entropy and focal loss, a hybrid function stabilised the model's high performance, even in more challenging classification cases.

A similar hybrid approach was also employed in the study [22], where a DT-DL-based hybrid method was applied to the PIMA dataset, yielding a higher result than other DL-based methods on the same dataset with an accuracy of 96.62%. The dataset can be made more detailed to generalise the proposed machine learning model.

Considering the issue of classification accuracy, study [23] incorporates KNNImputer during data preprocessing to address missing values and then integrates it into an ensemble model. The proposed stacked ensemble voting classifier (comprising XGBoost, Random Forest, and Extra Trees) achieved an accuracy of 97.49%, indicating the effectiveness of using KNNImputer. This result also far exceeds those of other ML models of the same type and aims to apply DL models in the future.

Based on research on symptoms such as nerve damage caused by diabetes that affect the heart's function, the HRV dataset is utilised in [24] in conjunction with deep learning techniques. A CNN 5-LSTM with an SVM algorithm achieved an accuracy of 95.7%, the highest score obtained when using the HRV dataset for automatic heart rate variability (HRV) detection. It is said to be non-invasive, flexible and a reliable tool for diagnostic purposes. It can be improved in the future by using a larger dataset. With a sufficiently large dataset, DL methods will make the process of detecting anomalies much easier. The inferred information can be considered a warning to take cautionary measures.

In [25], the optimised SVM and the DBN (deep belief neural network) with PSO were compared. PSO is a hyperparameter fine-tuning method based on the movement of swarms and is very efficient compared to traditional methods. The dataset was collected through a survey. As a result, the DBN-PSO achieved higher accuracy as compared to SVM. However, it still needs to be tested on larger datasets to draw meaningful conclusions.

All in all, these studies show a trend of growing effectiveness of machine learning anf deep learning approaches in the early detection of diabetes. Despite its obvious efficacy, it also faces a multitude of problems such as the class imbalance, limited dataset diversity, incomplete consideration of lifestyle and genetics of test subjects. These all provide a significant space for future developments. Dealing with all these limitations can definitely lead to a more accurate diagnostic performance. Taking in to account the varied models and algorithms used, this research aims at picking the best features out of them and combining them using the ensembling technique to produce a more enhanced diagnostic engine.

**2.1 Research Gap**

It is evident from previous studies that both ANN and Catboost are powerful machine learning models in their own right, having shown auspicious results in early diabetic detection. ANN is known for its ability to model complex nonlinear interactions, and Catboost boasts the ability to handle categorical data and produce interpretable results. The power of combining such features remains undiscovered, mainly, creating a significant research gap. One that can be filled to create a more accurate, clinically applicable diagnostic machine model.

The features and challenges of existing ML/DL models for early diabetes detection are provided in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.no** | **Name of research paper** | **Methodology** | **Features** | **Limitations** |
| 1 | **[1]** | SS-HPT-P/pNPMN-PBase | Proposes a model for cell replacement therapy to combat diabetes. | The transplanted cells may not be able to withstand attacks from the body's immune system. |
| 2 | **[2]** | Decision tree (DT)-based random forest (RF)  Support vector machine (SVM) | Provides an evaluation of other research on diabetes prediction that uses the DT-RF and SVM models | Final results are entirely based on the analysed data and cannot be generalised. |
| 3 | **[3]** | K-Nearest Neighbour (KNNs) | Proposal of a Diabetes Classification and Prediction Model (DCPM) | The dataset used in the study has not been handled well and therefore contains outliers and missing values. |
| 4 | **[4]** | AUC  AUPRC  RMSE | Compares novel machine learning methods with standard regression techniques for early prediction of diabetes. | More approaches for ensemble models can be explored. |
| 5 | **[5]** | Decision TreesSupport-Vector Machine (SVM)Naïve BayesJ48Logistic RegressionRandom ForestConvolutional Neural Networks (CNNs)Restricted BoltzmannLSTM and CNNGradient BoostingC4.5Linear Discriminant Analysis | Gives a comprehensive report of the various ML methods used for diagnosing diabetes at an early stage. | No framework has shown a 100% accuracy, with models KNN and RF showing the highest values of 98%. |
| 6 | **[6]** | Artificial Neural Network (ANNs) | Proposed a novel machine learning model for diagnosing diabetes | The model yields better results when it has previously worked on the same task. |
| 7 | **[7]** | Decision Tree  Random Forest  Neural Network | Proposes a model using multiple ML methods for the early detection of diabetes. | The dataset used concentrates on a single fasting glucose level as the index, and therefore requires a diverse selection of indices for generalisation. |
| 8 | **[8]** | Random Forest (RF)  Decision Tree (DT)  K-Nearest Neighbour (KNN)  Extreme Gradient Boosting (XGB)  Support Vector Machine (SVM) | Proposes a DL model using the ESDR dataset and Grid Search algorithm for the prediction of diabetes. | The dataset needs to be more diverse. |
| 9 | **[9]** | Support Vector Machine (SVM) | Proposes a review of the workings of the various ML models concerning various factors concerning DM. | Clinical data sets were mainly used; however, they can be more generalised. |
| 10 | **[10]** | Convolutional Neural Networks (CNNs) | Proposes a model that detects DM by thermal images of diabetic foot ulcer (DFU) based on a fusion schema | The dataset can be further enriched to generalise strategies for making fusions of smaller CNNs and allowing embedding in mobile devices. |
| 11 | **[11]** | Convolutional Neural Networks (CNNs) | Proposes a model that detects diabetic foot using pictures captured from infrared cameras in a smartphone. | Both feet thermogram works better compared to a single one. |

**Table 1: The characteristics of some of the pre-existing works.**

**III.METHODOLOGY**

Diabetes is one of the most harmful diseases known to man, which, while it doesn't cause an immediate death, can cause a chronic death. The only possible way to deal with such a disease is to discover it early and receive prompt, effective treatment. It is not always possible with traditional testing methods, as they are time-consuming and expensive. While diabetes itself does not show any noticeable symptoms in its early stages, this makes it difficult for patients to get tested. Therefore, technology such as the machine learning models for the early detection of diabetes has become a vital part of healthcare and has been researched by many before us. Building on previous studies, this study aims to create an ensemble model of two high-performing models—ANN and Catboost—using renowned data balancing techniques, SMOTE-ENN and SMOTETOMK-ENN, to determine the best variant of the ensemble model for achieving the highest prediction accuracy of disease. The proposed model aims at using multiple machine learning techniques like KNNImputer to clear up the missing values, one hot Encoding method to encode the data, LOF and Isolation Forest to remove outliers, Z score normalisation for the ANN data form, Train Test Split model to work the model and 3 cross fold validation method to simulate the model and Bayesian Optimization to fine tune the parameters. Once these processes are over, both data forms of ANN and Catboost undergo both SMOTE-ENN AND SMOTETOMK-ENN data balancing methods to figure out the better working model with the better balancing technique among them and combined using the weighted voting system to deliver the best possible ensemble model to achieve the highest results of correct diabetes prognosis.

**3.1 Data Collection**

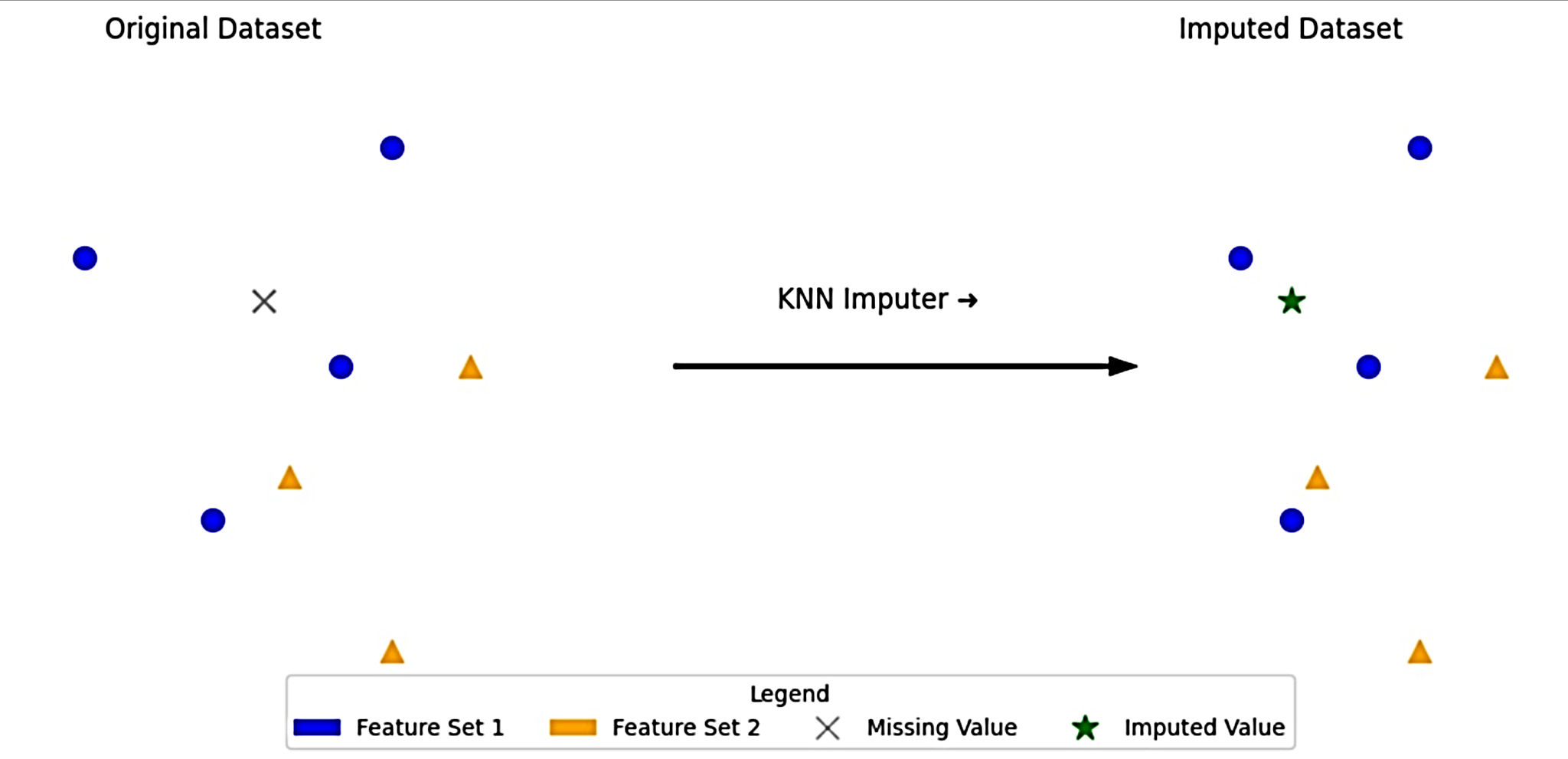
Obtaining information and organising it systematically to test a hypothesis or verify a statement is called Data Collection. In this study, the data has been collected from a publicly available data source, Kaggle, comprising a total of 100,000 records [26]. Adding the class feature, it has 8 features. The associated class variable is a binary variable (diabetes) that directly indicates whether the patient has diabetes. Wherein the variable for "no diabetes" is 0 and for "having diabetes" is 1.

**3.2 Data preprocessing**

The process of changing raw data into a format that the machine learning model for training can understand. Its primary aim is to address data issues such as missing values, [focusing on improving data quality](https://lakefs.io/data-quality/improve-data-quality/), and preventing any trouble that may arise when used by machine learning algorithms. It is because the model's outcome can only be as good as the data that is entered into it. Algorithms are not built to handle incomplete or noisy data. Therefore, only by preprocessing the data can the quality of the inputs improve, ultimately leading to a more accurate model.

In this research, the missing values have been handled using the KNNImputer function from the impute model of scikit-learn. Using the Euclidean distance matrix, it identifies the nearest neighbours for the required missing values. The Euclidean matrix calculates the straight-line distance between two points in a Cartesian coordinate system.

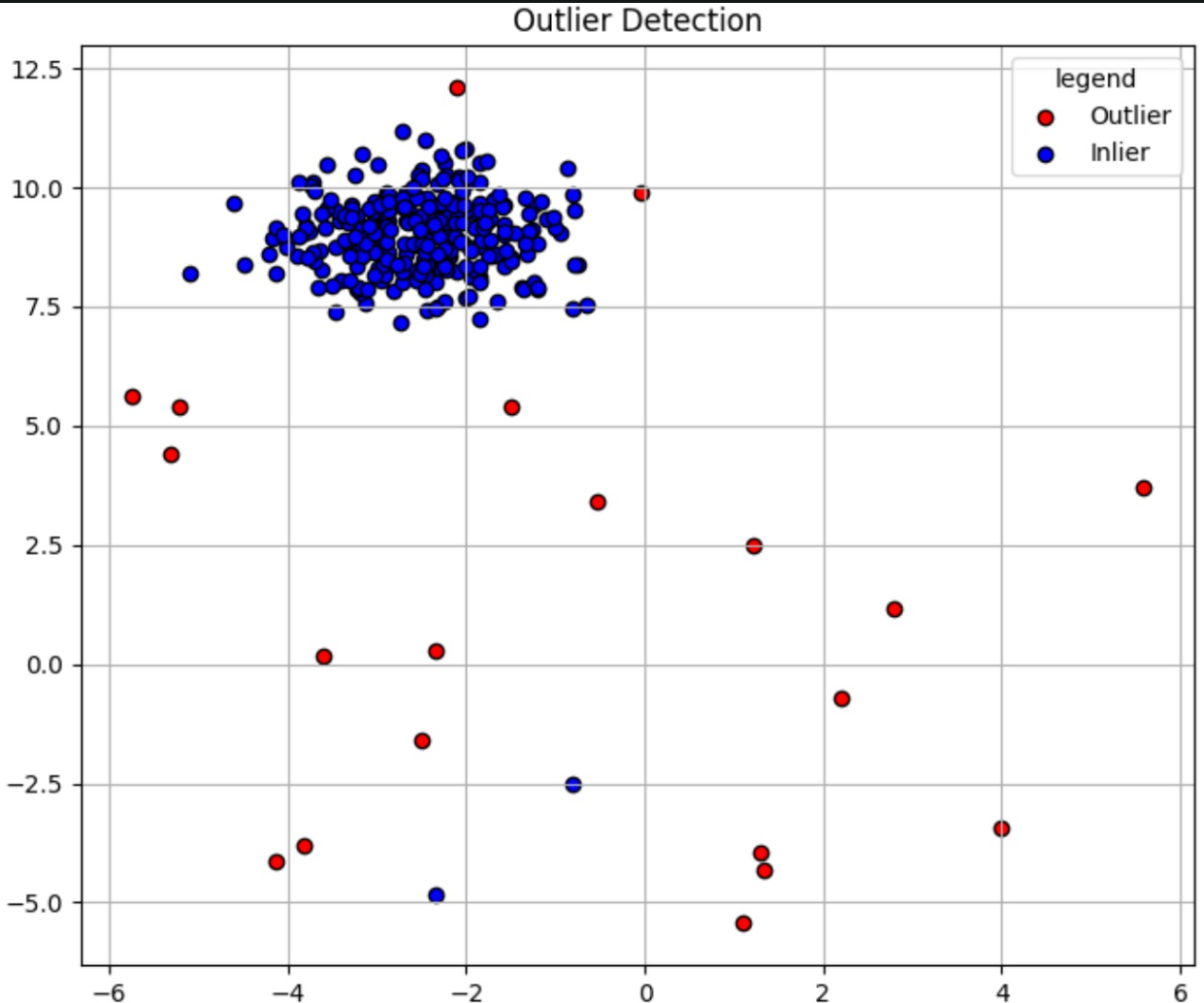




**Figure 1: Illustration of KNNImputer**

The next step was Data encoding, where any data in text form is converted to its numerical format - a format where algorithms have an easier time processing, as most machine learning algorithms cannot deal with text or categorical values as input. In this study, a specific method called one-hot encoding has been used, where a binary column is created for each unique category found in a variable. And when the criteria for the category are met, the adjacent binary column is set to 1. If not, it will be set to 0.

The next issue is the outliers. Outliers are defined as data points that deviate significantly from the other data points in a dataset. The reasons for its existence are varied, from measurement errors to natural deviation in the data. However, they cannot be allowed to go unchecked, as this will compromise the data analysis and ultimately impact the final results produced by the machine learning model. The method used for removing outliers is called Isolation Forest, which utilises trees to select features randomly and split the data according to threshold values to isolate outliers. It is beneficial when dealing with large datasets where abnormalities are rare and of unique types. Another method used is the Local Outlier Factor (LOF), which, according to the relativeness of a point with its neighbours, calculates the local density. It is because these outliers have a significantly lower density compared to their neighbours. Figure 2 illustrates the presence of outliers, and Figure 3 explains the isolation forest and LOF algorithm via a flowchart.



**Figure 2: Illustration of outliers**



**Figure 3: Flow chart of isolation forest and LOF algorithm**

Then, two copies of the data frames are created, one for the ANN model and another for the Catboost model. While doing so, the data is required to be standardized, a process where the data's features are brought onto the same scale. This approach ensures that all features are given equal importance, preventing any single feature from dominating the others and resulting in improved model performance.

For the ANN model, a method called the Z-score normalization has been used. This method converts its data such that its mean becomes 0 and the standard deviation becomes 1. The data is thus adjusted for these two factors, considering how far each value deviates from the mean.

In which,

* (Z)- Z-score.
* (X)- Value of the data point.
* (μ)- Mean of the dataset.
* (σ)- Standard deviation of the dataset.

For the Catboost model, the data frames can be created as is, owing to the fact that it can handle the values directly.

**3.3 Train-test split with 3-fold cross validation**

The train-test split, as the name suggests, is a validation procedure that both trains and tests the machine learning model using a simulated dataset split (80 per cent for training and 20 per cent for testing, in this case). It tests the model's ability and its resulting accuracy. Sometimes, it also requires a validation set, from which the data can be used to further improve the ML model's performance.

Going through the process step by step:

***1. Arrange the Data***

Let the data be arranged in a way that the train-test split model can work with- here in scikit-learn, it means that you separate the dataset into 'features' and 'targets.

***2. Split the Data***

The dataset will then be divided into two parts: a training set and a testing set. Eighty per cent of the dataset is subjected to rounds of random sampling in the training set without replacement. The rest are included in the test set.

***3. Train the Model***

Data in the training set undergoes training in the model.

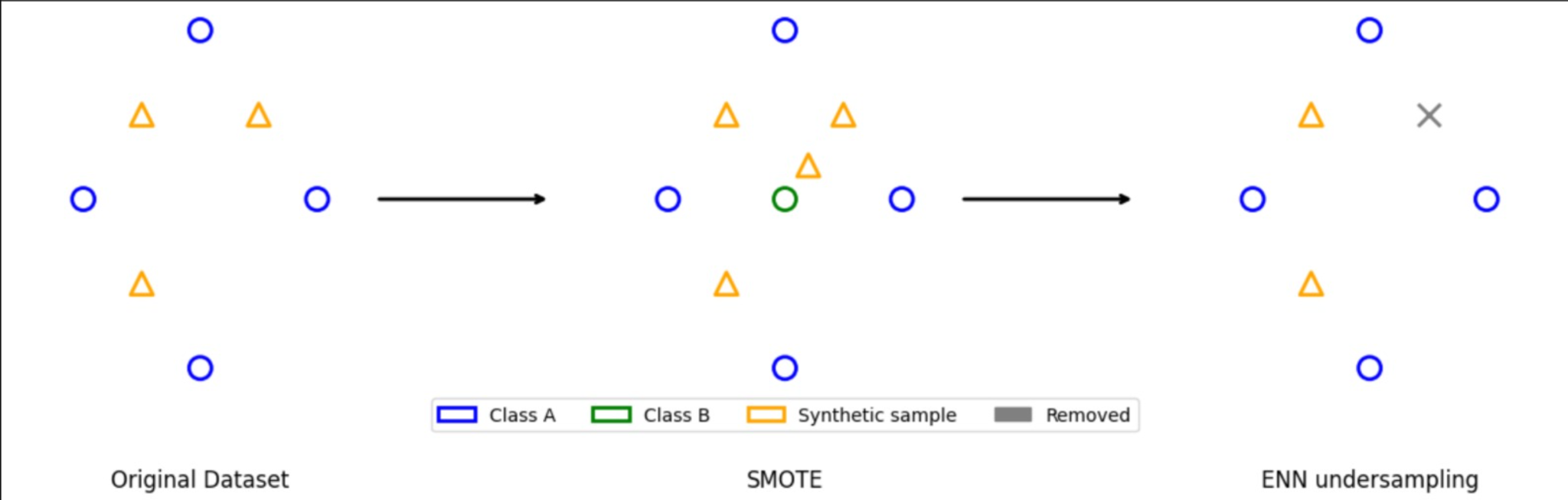
***4. Test the Model***

The data in the testing set is now tested in the model. Subsequently, this has been paired with another method called three-fold cross-validation. This method is particularly used for evaluating the model's performance on unseen data. It is done by splitting the dataset into several random parts and then undergoing training, validation and testing randomly and repeatedly. Then the results of each round of testing's mean is given as an estimate of the model's ability. It prevents overfitting, meaning the model works not only with the preassigned data but also handles real-world data.

**3.4 Data balancing using SMOTE+ENN AND SMOTETOMK+ ENN**

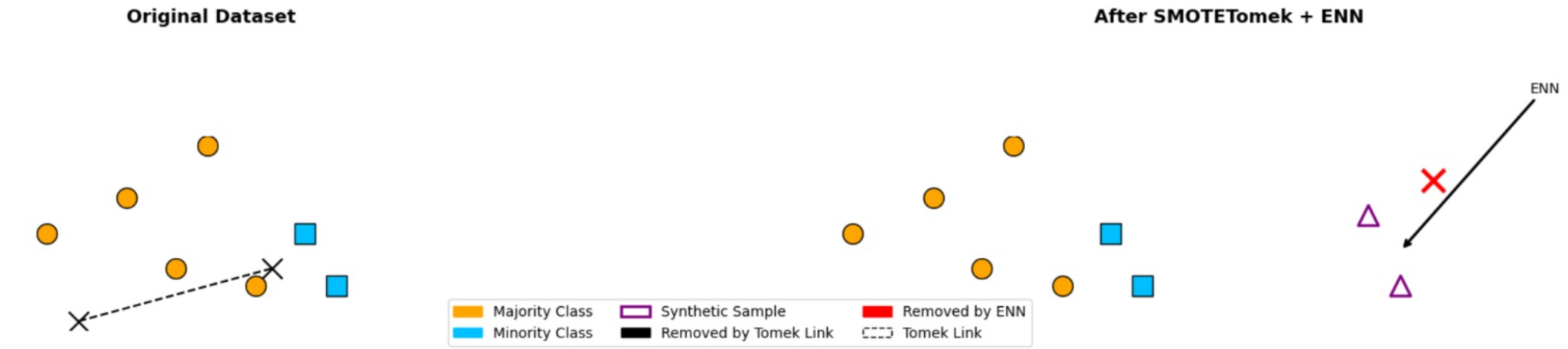
Data balancing is considered a necessity when dealing with machine learning models due to the presence of imbalanced data, which refers to data where the distribution of observations across the target class is uneven. One class label has a significantly larger number of observations compared to another, which can cause the ML models to become biased in their predictions.

In this study, methods like the SMOTE-ENN method and the SMOTETOMK-ENN method have been used extensively for the same. The former merges the abilities of both SMOTE and ENN, where SMOTE oversamples the minority class and ENN undersamples the majority class. ENN also discards all samples that have differing class types from their neighbours, allowing all inaccurate classified values to be removed. SMOTE-ENN yields better results than SMOTE alone. Figure 5 explains the working of SMOTE-ENN.



**Figure 5: Illustration of SMOTE-ENN**

While in the SMOTETOMEK+ENN method, the SMOTE part focuses on generating new values based on its neighbours, and Tomek focuses on detecting the boundary between classes of nearest neighbours and increasing the separation between classes. As mentioned above, ENN clears out inaccurate values. Figure 6 explains the working of SMOTETOMK-ENN.



**Figure 6: Illustration of SMOTETOMK + ENN**

By applying these techniques separately to both models — ANN and Catboost — a solution will be determined as to which method is more effective and which model is best suited for creating an ensemble.

**3.5 Hyperparameter fine-tuning**

Selecting the ideal values as the hyperparameters of a machine learning model is known as hyperparameter fine-tuning. These are often set before the start of the training session and govern various areas of the learning process. At its core, it entails rigorously identifying the most suitable set of hyperparameters that goes hand in hand with the model to better its performance. Unlike model parameters, these hyperparameters are not automatically learned throughout the training process. They are predetermined instead. Their exact arrangement has the power to influence the model's performance significantly, separating mediocre models from exceptional ones. The model's results can be affected by these variables in terms of both speed and quality. The ideal solution may be missed if the model converges too rapidly due to a high learning rate. In addition to requiring more time and computer resources, a poor learning rate may cause slower convergence.

In this case, the Bayesian optimization approach has been employed, an automated optimization methodology that treats the search process as an optimisation problem to discover the best hyperparameters. It seeks to maximise an objective function f(x), which is especially useful for functions that are considered "black boxes," meaning that their underlying structure is unknown, and therefore computationally costly to assess. The capacity of Bayesian Optimisation to incorporate prior evaluations into its decision-making process, thereby selecting the subsequent set of hyperparameter combinations, is one of its primary characteristics. A probabilistic model, which calculates the likelihood of an objective function's outcome given a collection of hyperparameters, is used to accomplish this. This model, represented by P(y | x), is referred to as a "surrogate" for the objective function. Figure 7 illustrates the operation of the Bayesian algorithm through an example. There are multiple steps in the Bayesian Optimisation algorithm:

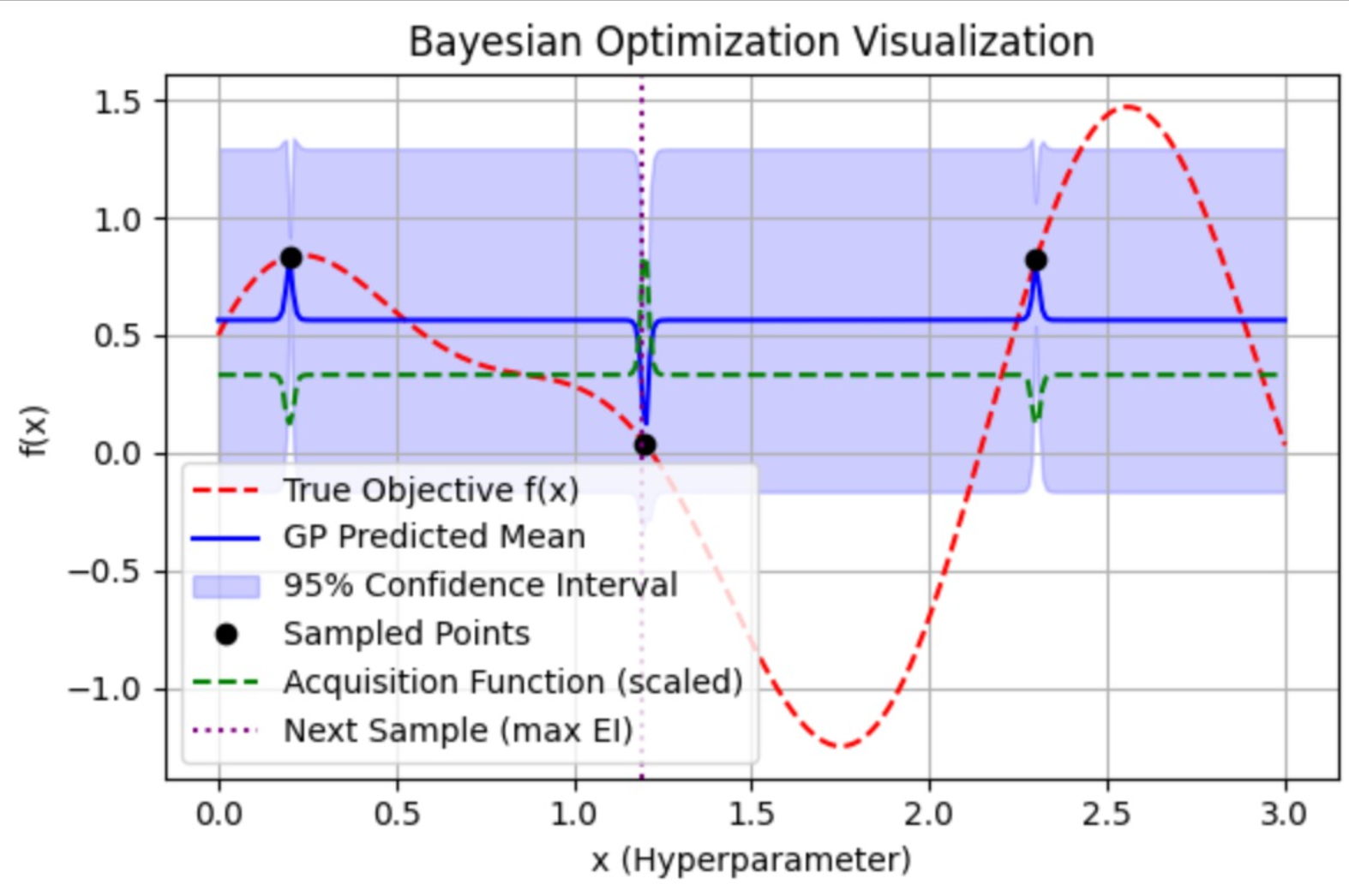
1. *Create a Probability Model*: Using data from previous analyses, create a probability model of the objective function.

2. *Determine Optimal Hyperparameters*: Based on the probability model, determine which hyperparameters work best.

3. *Apply Hyperparameters*: Put the chosen hyperparameters into practice and assess how well they work with the real goal function.

4. *Update Probability Model*: Apply the most recent findings to the probability model.

5*. Repeat*: Repeat steps 2-4 until you reach the time restriction or maximum number of repeats.



**Figure 7: Bayesian optimisation visualisation**

**3.5 Model deployment**

**3.5.1 ANN- Artificial Neural Network**:

A system based on the intricate connections of nerves in a human brain. Activation function is a known mathematical function that calculates the weighted sum of inputs with an added bias. It introduces non-linearities to the output and helps in making complex predictions. Here, in particular, the sigmoid function has been used for the output layer. It is mathematically defined as

It provides an advantage in handling complex patterns that linear equations cannot. It is also typically used for binary classification, as its output ranges from 0 to 1. Then, for the hidden layers, the Prelu function has been used, which is the Parametric Rectified Linear Unit, known to generalise the usual rectified unit with negative slope values.





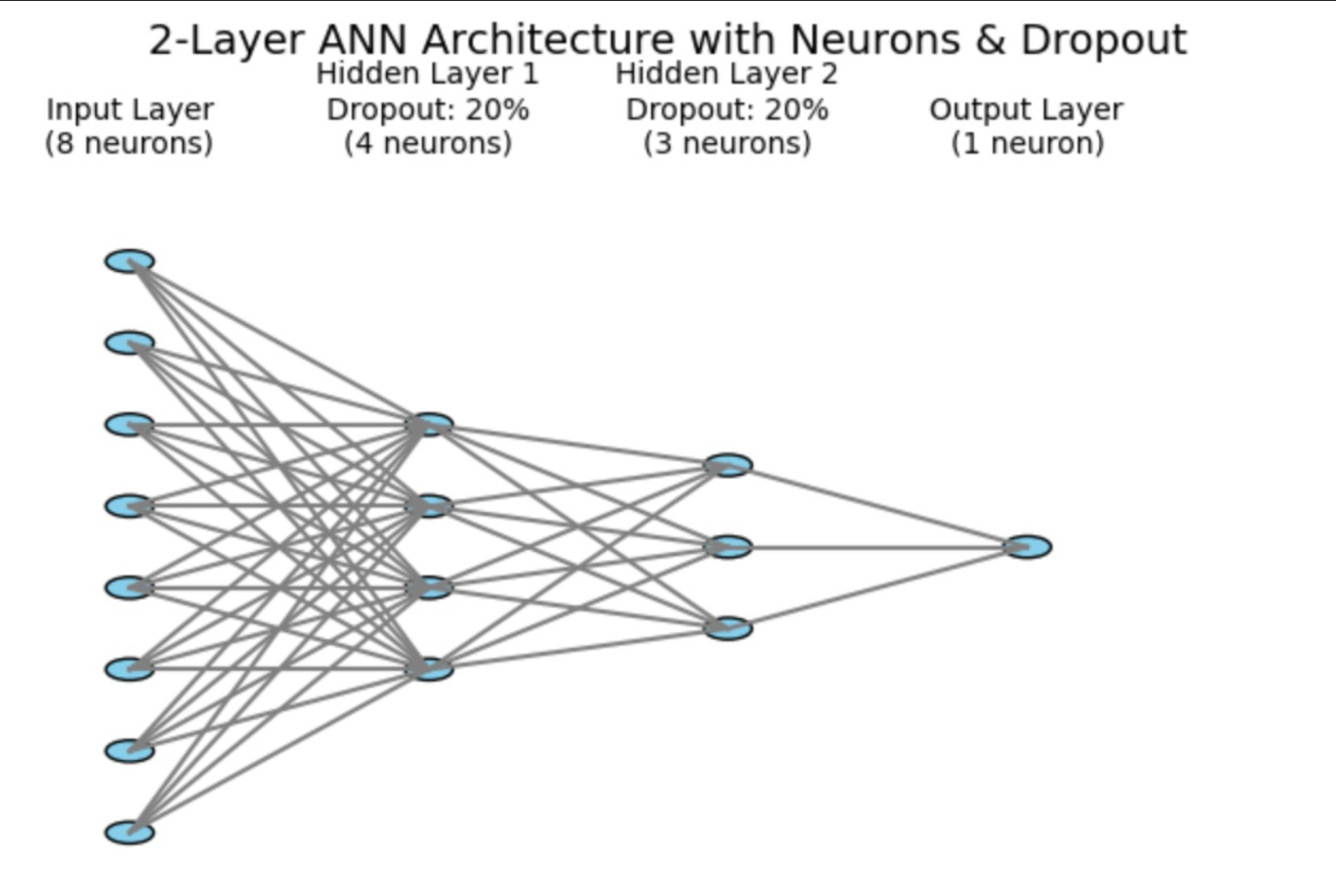
A loss function is also a mathematical function that calculates the difference between the final values on paper and what the model has achieved. Its primary goal lies in minimising the loss due to the model's inaccuracy. Here, the Binary Cross-Entropy (BCE) has been used as the loss function, as it is most suited for binary classifications and thus yields only two outcomes: 0 or 1. Mathematically, Binary Cross-Entropy (BCE) is defined as:



Where:

* N = number of observations
* = Actual binary label (0 or 1) of the observation.
* ​ = Predicted probability of the observation being in class 1.

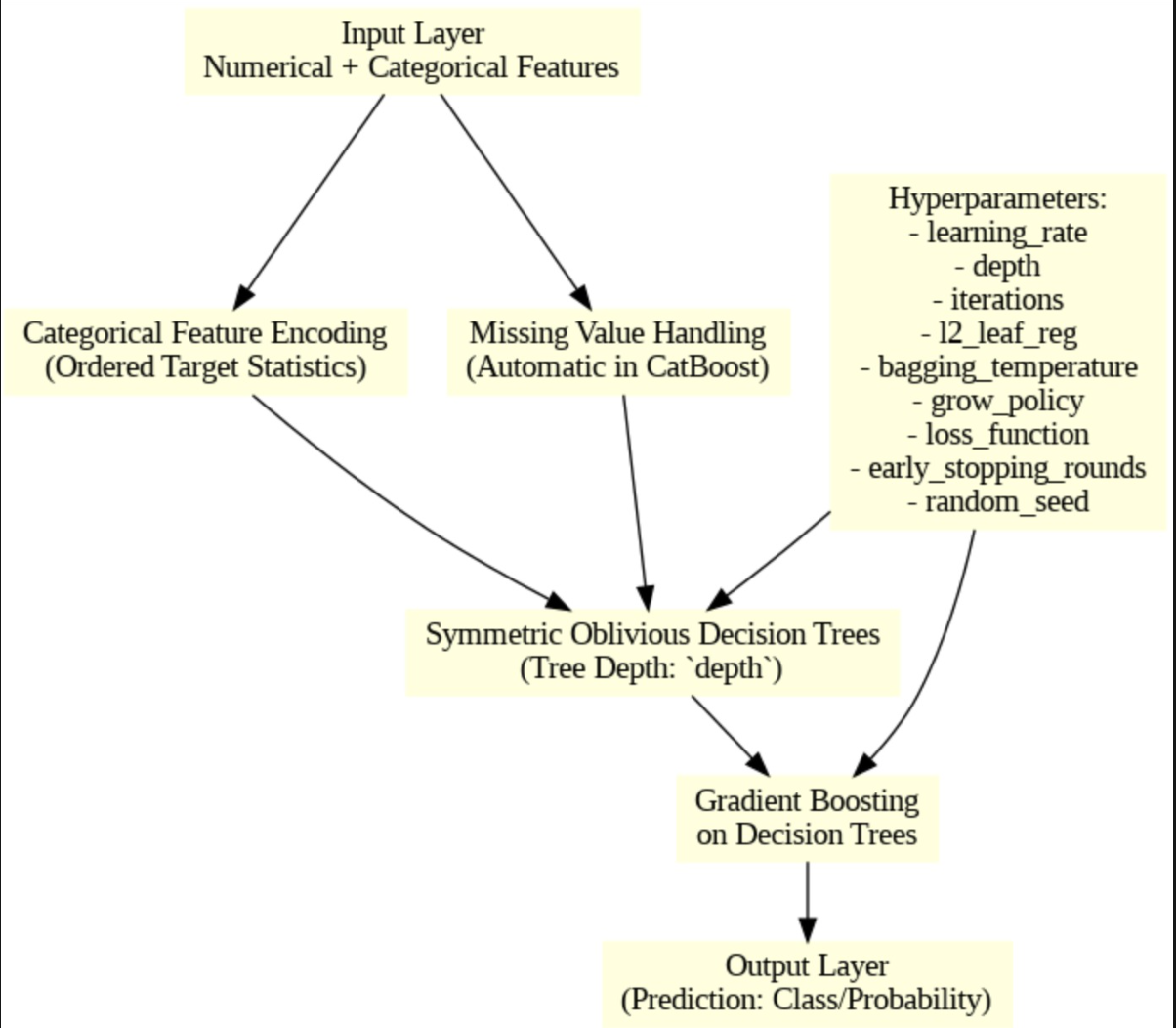
It is known to sometimes yield inaccurate values due to its logarithmic nature. The hyperparameters for the above include neurons1(no of neurons in layer 1),neurons2(no of neurons in layer 2), learning rate(rate that indicates the model on how much it is required to change following its errors), dropout rate(the random amount of neurons set to 0 during training sessions to avoid overfitting), batch size(the number of training samples it undertakes in a specified amount of time), epochs(The number of times the machine learning model goes through the complete dataset while training). All these were passed during the Bayesian optimisation in order to get the optimal values. Figure 8 is an example of an ANN architecture that has 8 input layers, 1 output layer, and 2 hidden layers.



**Figure 8: ANN architecture with 2 hidden layers**

**3.6.2 Catboost**:

An open-sourced model known for its ability to handle categorical data effectively and efficiently. The loss function in Catboost is optimised during the model's training. It is defined as "LOGLOSS" for binary classification and 'RMSE' for regression tasks. Figure 9 illustrates the working of the CatBoost model. The hyperparameters used for the above are iterations (number of repeated iterations in the form of trees used for training), Learning rate(the step size determiner), Depth ( maximum Depth of individual trees used in the Ensemble),l2\_leaf\_reg ( prevents overfitting by removing the larger parameter values).



**Figure 9: Working of the catboost model**

**3.6.3 Ensemble Model**

An ensemble model is a combination of multiple base models to create a highly accurate and robust model for predicting the corresponding problem statement and delivering results that solve the stated problem. Its main goal is thus to improve the accuracy of the model's predictions. In this study, an attempt is made to combine an ANN model and a Catboost model (both with SMOTE+ENN or SMOTETOM+ENN, depending on which method yields better results for the respective model) using a weighted voting system.

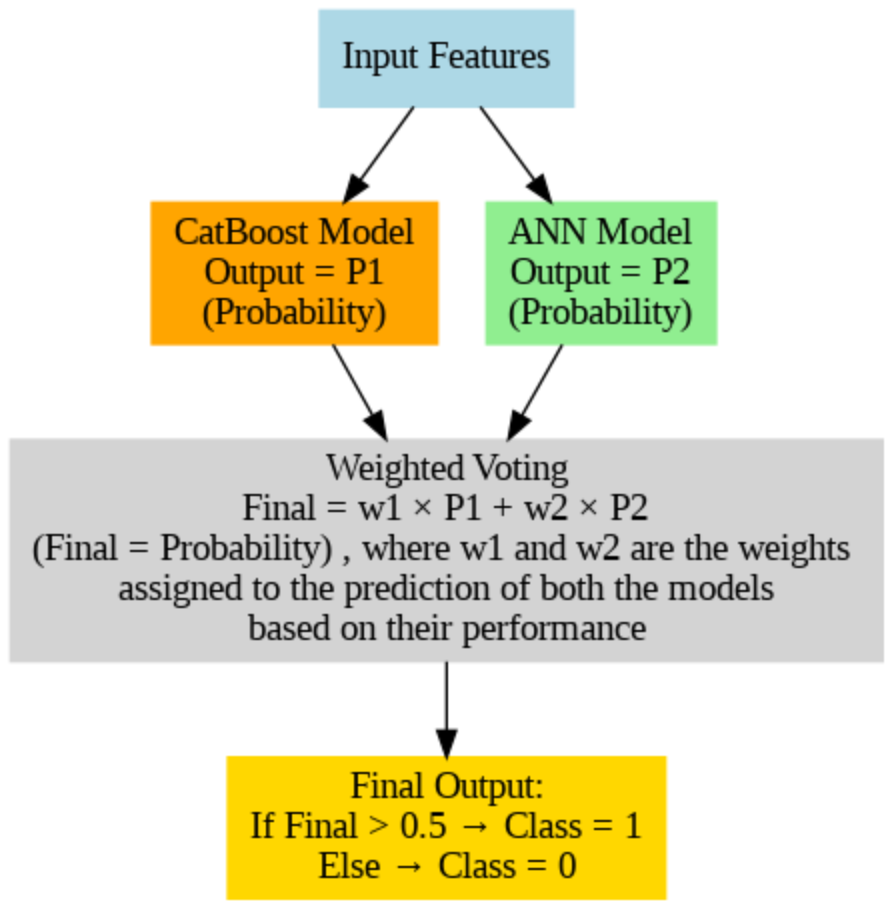
The weighted voting system works on the basis that one group of voters have a higher power as compared to other voters in the same situation. Thus, the weighted voting system aims to dispel the bias by balancing out these powers. Figure 10 illustrates the workflow of the weighted voting system.

A weighted voting system is usually written in the form:

[q:w1​,w2​,w3​,...,wn​]

Where:

* q = quota (minimum number of votes needed to pass a motion)
* w1 ​,w2​,...,wn​ = weights (number of votes assigned to each voter)
* n = number of voters



**Figure 10: Algorithm of the ensemble model creation using a weighted voting system**

**3.7 Evaluation metrics**

*Accuracy*: Refers to the proportion of correct predictions made by the model. It can be mathematically defined as:

Where TP= True Positive, TN-True Negative, FP=False Positive, FN=False Negative

*Precision:* It evaluates the accuracy of optimistic predictions made by a model



*Recall:* It measures a model's ability to identify all relevant instances (positive cases) within a dataset.



*F1 score:* It is the harmonic mean of precision and recall, providing a balance between these two measures, used to evaluate the performance of machine learning models.



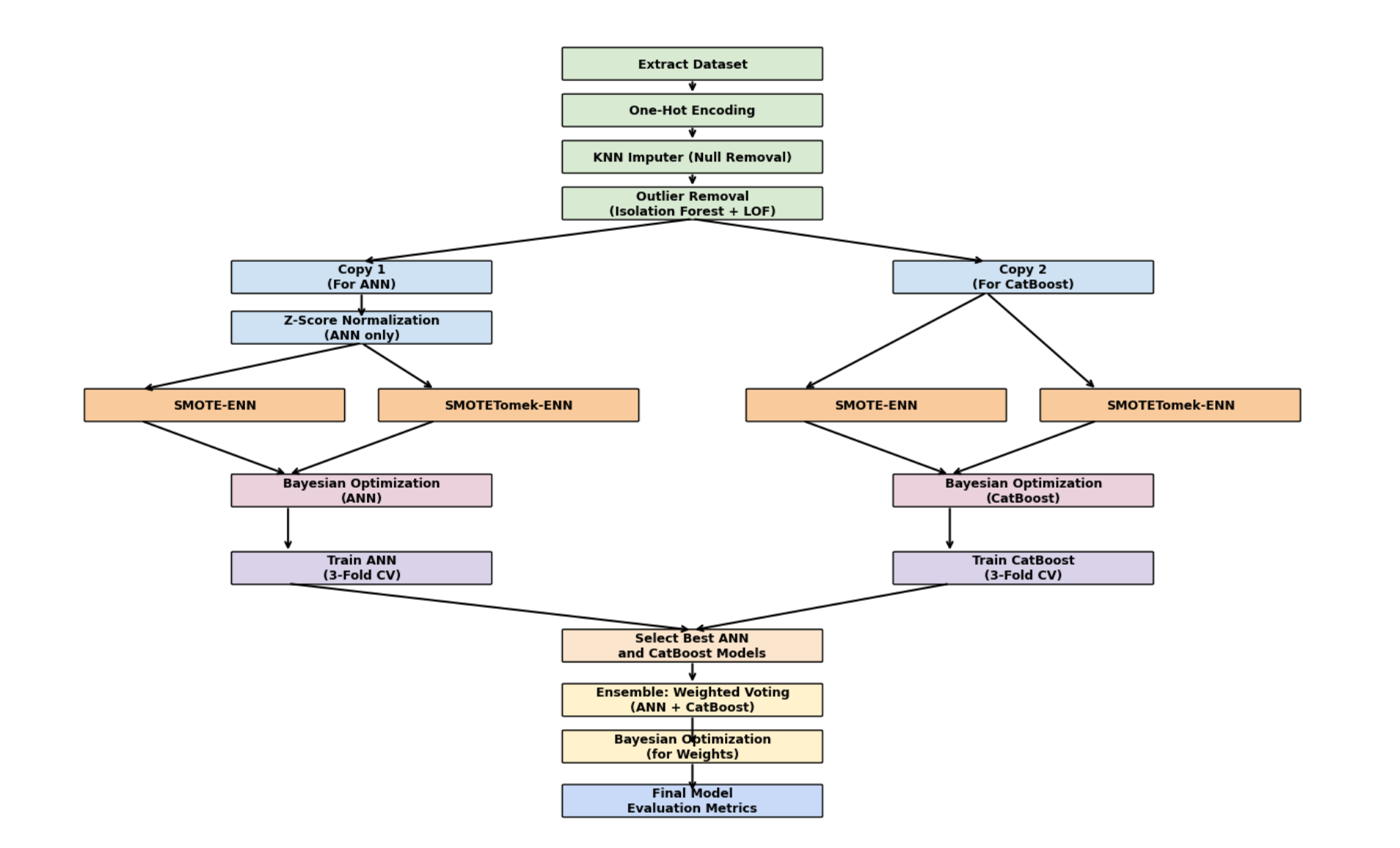
*Negative Predictive Value (NPV):* The proportion of accurate pessimistic predictions relative to the total negative predictions.



*Matthews Correlation Coefficient (MCC):* This measure assesses the correlation between actual and predicted values. It ranges from -1 to 1, with -1 being the lowest correlation and +1 being the highest correlation.



Figure 11 illustrates the entire research methodology using a flowchart.



**Figure 11: Overall research methodology flow chart**

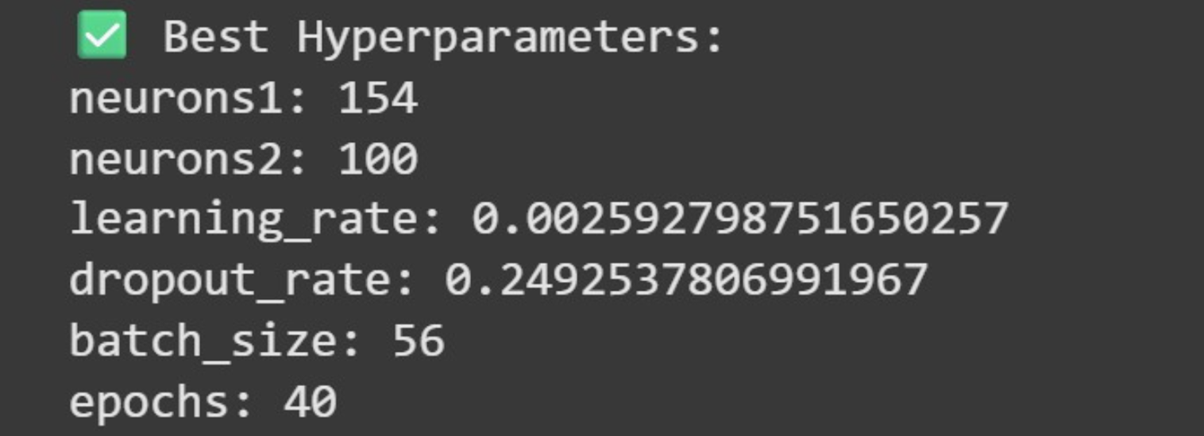
**IV RESULT AND DISCUSSION**

This study uses Google Colab with a T4 GPU (free) present in it.

**4.1 ANN model**

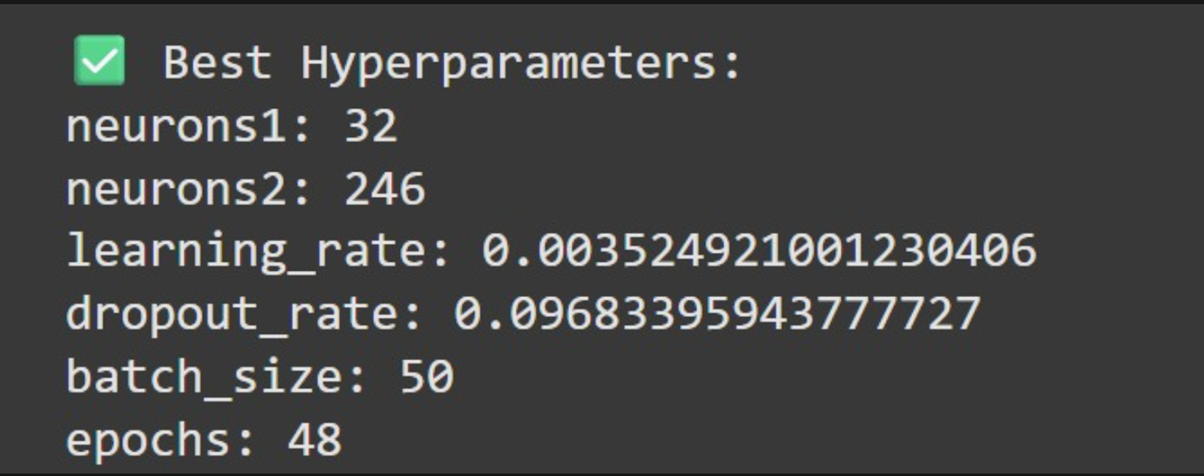
After fine-tuning the hyperparameters of the ANN model with SMOTE+ENN and SMOTETOMK +ENN separately using Bayesian optimization, the following results are obtained:

**4.1.1 ANN with SMOTE+ENN**



**Figure 12:** The best hyperparameters for ANN with SMOTE+ENN

**4.1.2. ANN with SMOTETOMK+ ENN:**

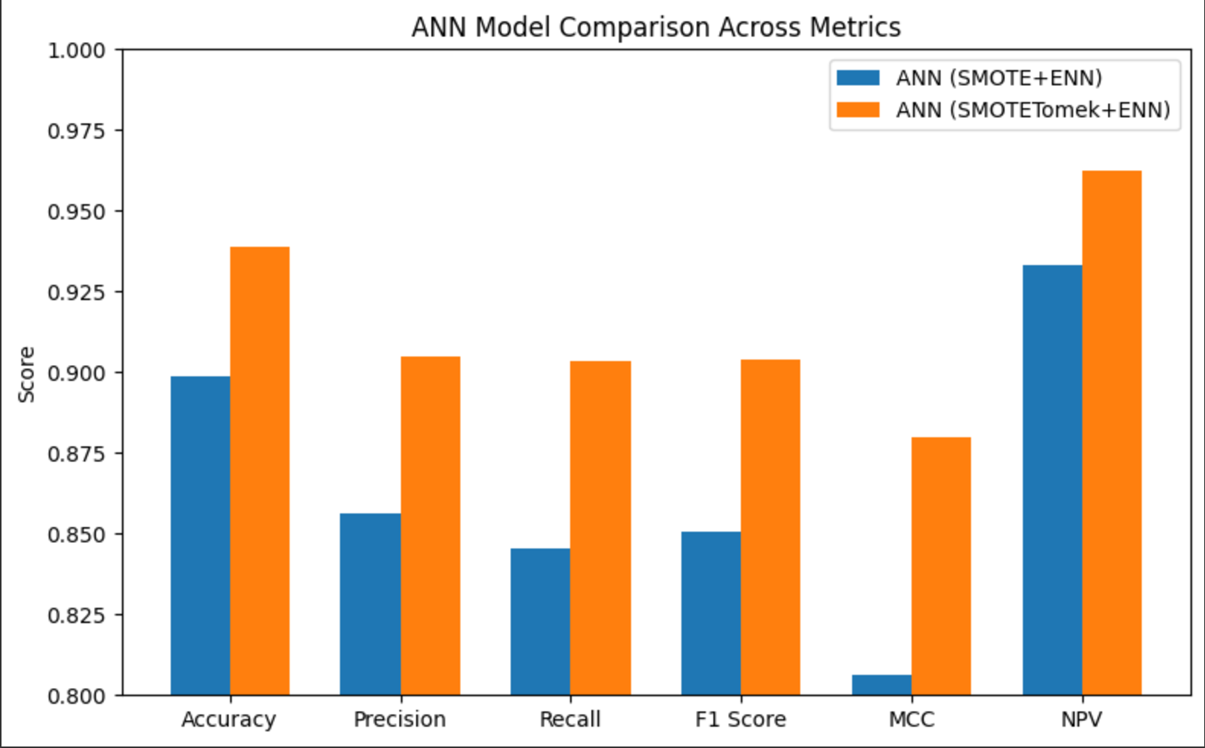


**Figure 13:** The best hyperparameters for ANN with SMOTETOMK+ENN

Figures 12 and 13 provide the best hyperparameters for ANN with SMOTE+ENN and ANN with SMOTETOMK+ENN, respectively. Now, these respective hyperparameters are fed into the corresponding ANN models and train them. The results obtained are recorded below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data sampling** | **Accuracy** | **Precision** | **Recall** | **F1 score** | **MCC** | **NPV** |
| ANN with SMOTE+ENN | 0.8989 | 0.8562 | 0.8452 | 0.8507 | 0.8061 | 0.9333 |
| ANN with SMOTETOMK+ ENN | 0.9389 | 0.9049 | 0.9034 | 0.9041 | 0.8798 | 0.9622 |

**Table 2:** ANN model results



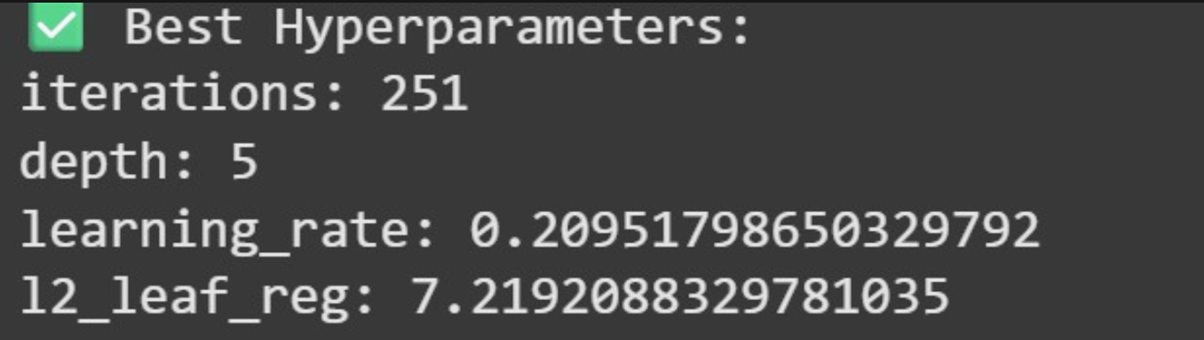
**Figure 14:** Comparison of evaluation metrics between ANN(SMOTE+ENN) and ANN(SMOTETOMK+ENN)

From the graph, it is very clear that for ANN, SMOTETOMK+ENN is a better data sampling technique than SMOTE+ENN.

**4.2 Catboost model**

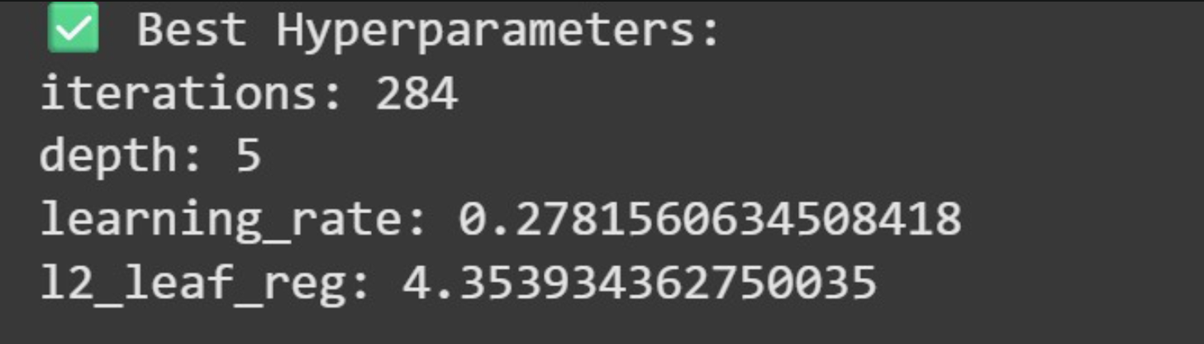
After fine-tuning the hyperparameters of the ANN model with SMOTE+ENN and SMOTETOMK +ENN separately using Bayesian optimization, the following results are obtained :

**4.2.1 Catboost with SMOTE+ENN :**



**Figure 15:** The best hyperparameters for Catboost with SMOTE+ENN

**4.2.2 Catboost with SMOTETOMK +ENN:**

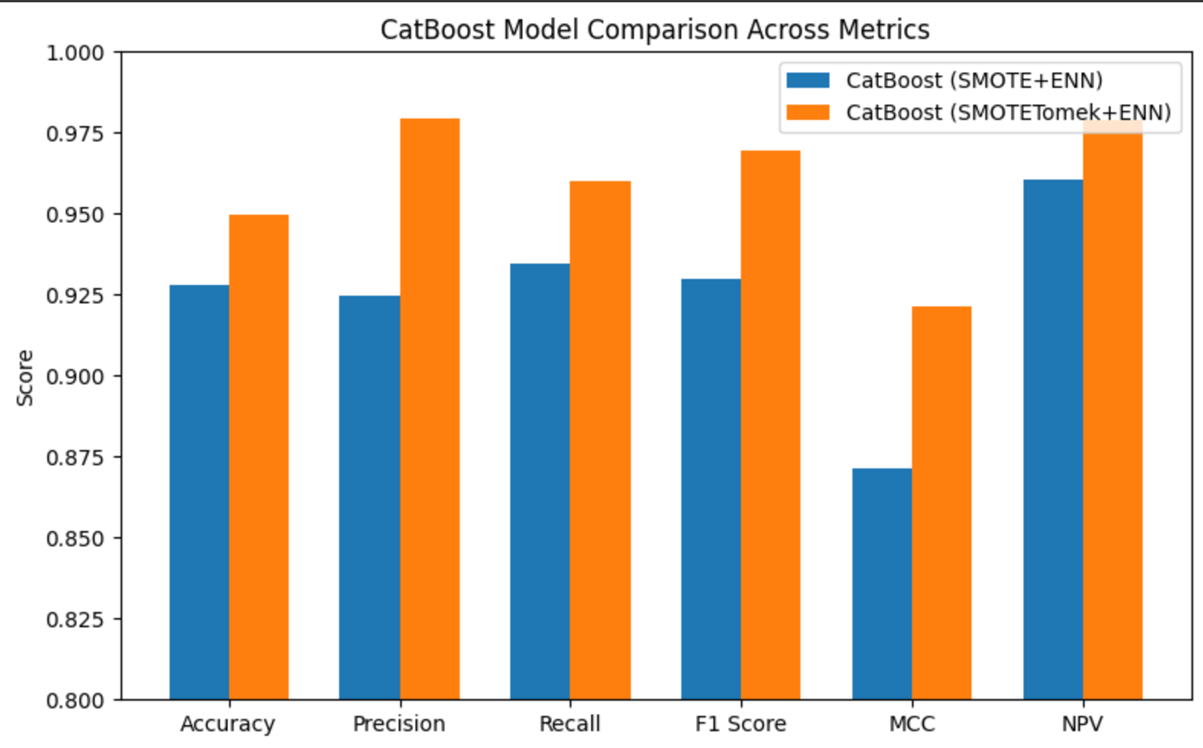


**Figure 16:** The best hyperparameters for Catboost with SMOTETOMK+ENN

Figures 15 and 16 provide the best hyperparameters for Catboost with SMOTE+ENN and Catboost with SMOTETOMK+ENN, respectively. Now, these hyperparameters are fed to the respective Catboost models and train them. The results obtained are recorded below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data sampling** | **Accuracy** | **Precision** | **Recall** | **F1 score** | **MCC** | **NPV** |
| Catboost with SMOTE+ENN | 0.9282 | 0.9248 | 0.9346 | 0.9297 | 0.8715 | 0.9607 |
| Catboost with SMOTETOMK+ ENN | 0.9495 | 0.9792 | 0.9599 | 0.9695 | 0.9212 | 0.9791 |

**Table 3:** Catboost model results

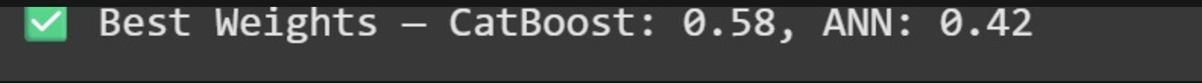


**Figure 17:** Comparison of evaluation metrics between Catboost (SMOTE+ENN) and ANN(SMOTETOMK+ENN)

From Figure 17, it is clear that for Catboost, SMOTETOMK+ENN is a better data sampling technique than SMOTE+ENN. From 4.1 and 4.2, it is very clear that SMOTETOMK+ENN is a better data sampling technique than SMOTE+ENN. So, both ANN and Catboost model use SMOTETOMK+ENN in the ensemble model creation .

**4.3 Ensemble Model**

The Ensemble was created by combining the ANN (with SMOTETOMK+ENN) and Catboost (with SMOTETOMK+ENN) using a weighted voting system. The optimal weights assigned to the ANN and Catboost models were determined using a Bayesian optimisation algorithm, whose results are shown below in Figure 18.

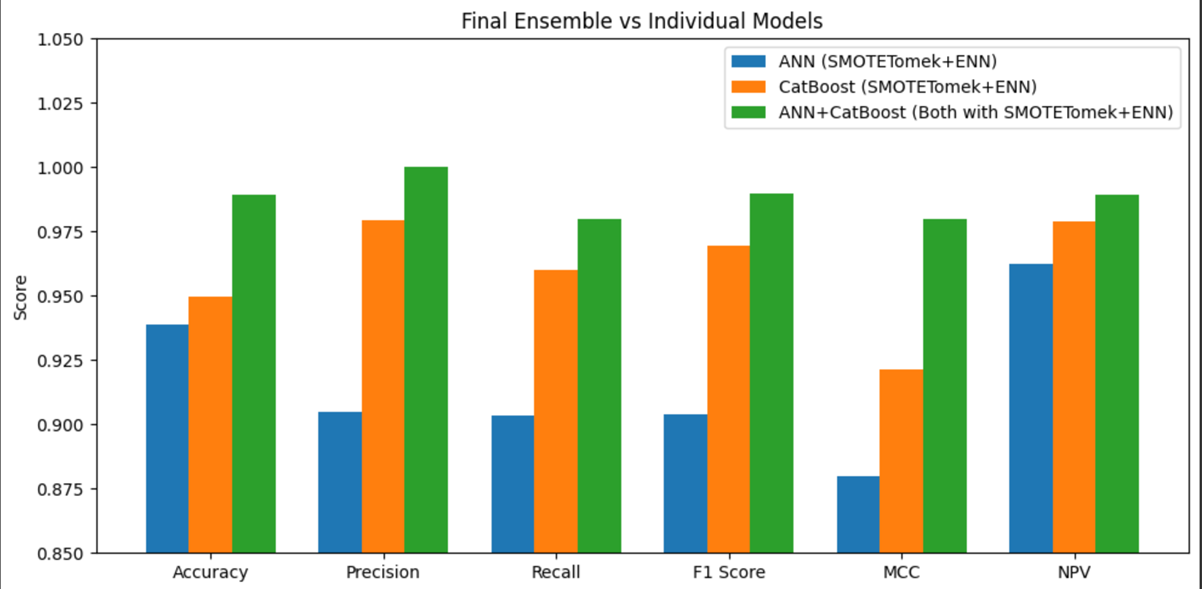


**Figure 18:** The best hyperparameters for weights of the ANN and Catboost model

And these were passed to the voting system to do the final testing, and the results obtained were as follows :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 score** | **MCC** | **NPV** |
| ANN(with SMOTETOMK+ENN) | 0.9389 | 0.9049 | 0.9034 | 0.9041 | 0.8798 | 0.9622 |
| Catboost with SMOTETOMK + ENN | 0.9495 | 0.9792 | 0.9599 | 0.9695 | 0.9212 | 0.9791 |
| Proposed Ensemble model (ANN +Catboost) | 0.9894 | 1.0000 | 0.9800 | 0.9899 | 0.9797 | 0.9895 |

**Table 4:** Final ensemble model results



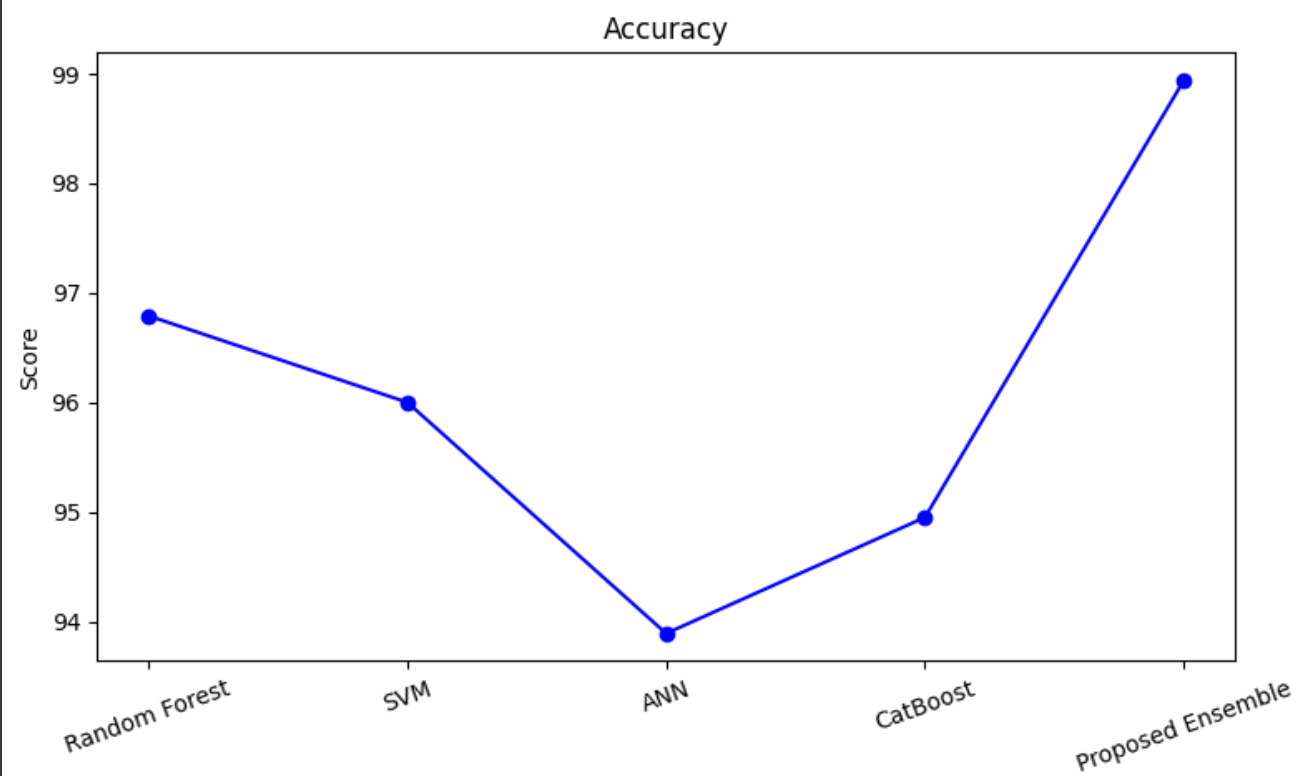
**Figure 19:** Comparison between the final ensemble model and the individual models

From Figure 19, it is clear that the study’s proposed ensemble model yields significantly better results than the individual models. Hence, the ensembling of the 2 given models was a huge success.

#### Table 5 shows how well different ML/DL models (which have been implemented on the same dataset), including the proposed models, perform in terms of Accuracy, Precision, recall, and F1 Score. Among the models that were looked at were Random Forest, Decision trees, and Logistic Regression, where Random Forest achieved the highest performance, so only metric values of Random Forest are included in the table [27], whereas, in [28], the Support Vector Machine(SVM) algorithm was used for the early diabetes prediction. The study’s suggested models were then compared to theirs. These were Catboost with SMOTE-TOMEK-ENN, ANN (with SMOTE-TOMEK-ENN), and the ensemble model of ANN and Catboost (each with SMOTE-TOMEK-ENN).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Accuracy** | **Precision** | **Recall** | **F1 score** |
| Random Forest [27] | 96.79 | 97.00 | 97.00 | 97.00 |
| SVM [28] | 96.00 | 96.00 | 96.00 | 95.00 |
| ANN (With SMOTETOMK-ENN) | 93.89 | 90.49 | 90.34 | 90.41 |
| Catboost (With SMOTETOMK-ENN) | 94.95 | 97.92 | 95.99 | 96.95 |
| Proposed ensemble model | 98.94 | 100.00 | 98.00 | 98.99 |

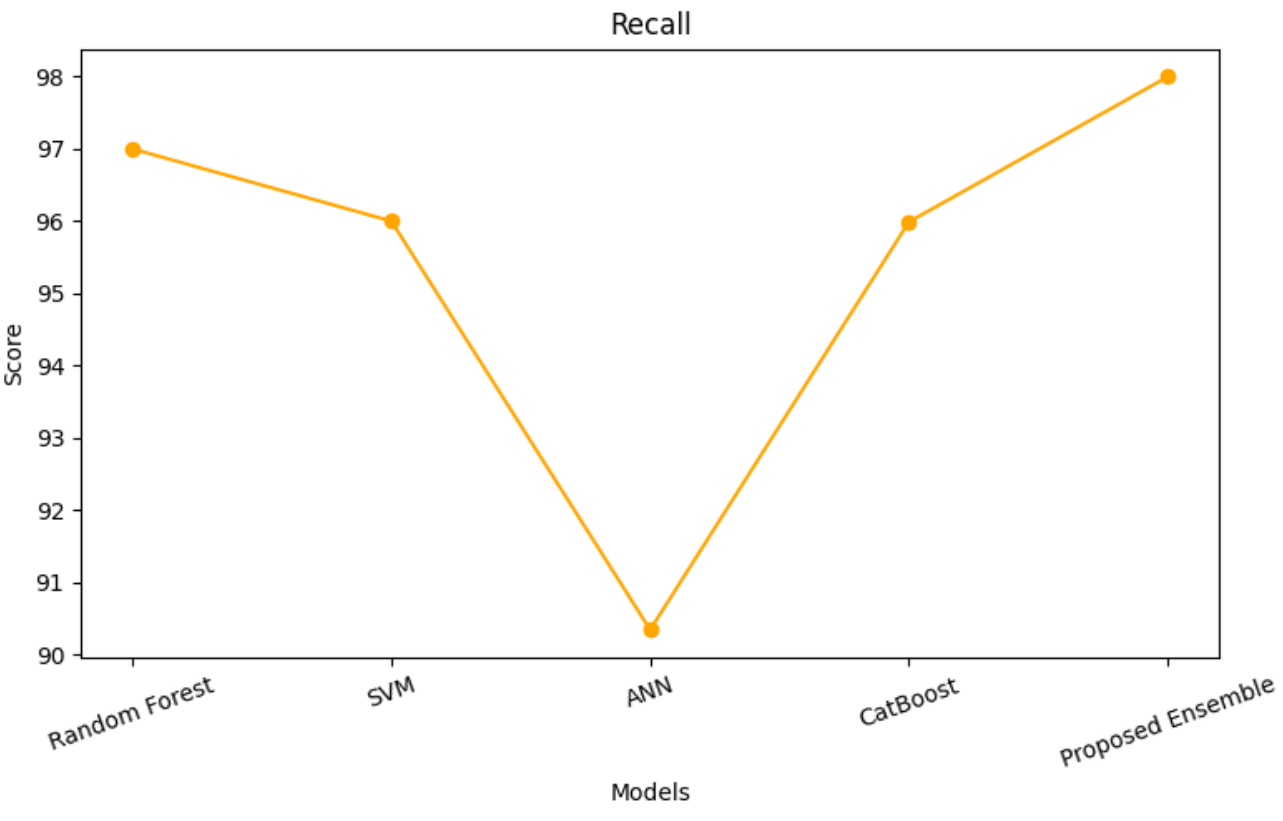
**Table 5:** Comparison between the proposed model and various other models



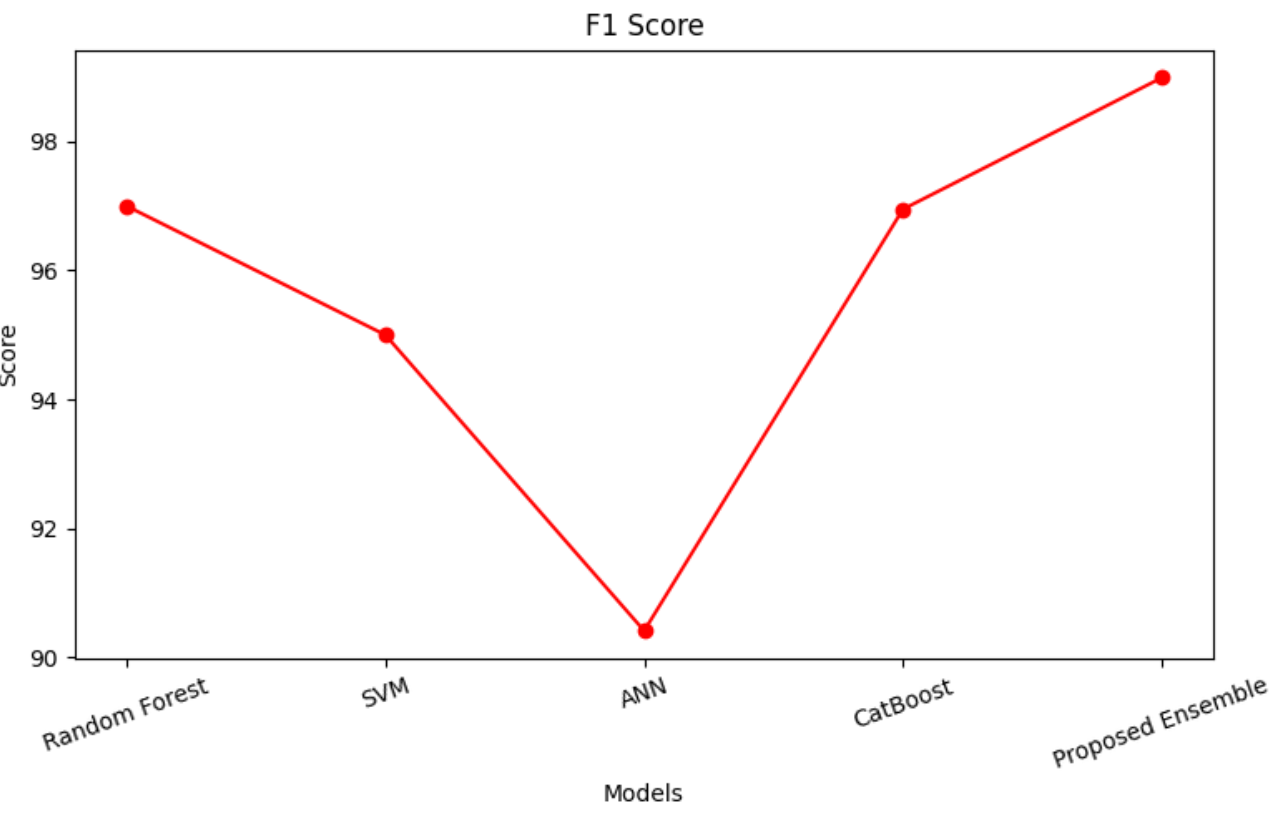
**Figure 20:** Comparison of various models based on accuracy



**Figure 21:** Comparison of various models based on precision



**Figure 22: C**omparison of various models based on recall



**Figure 23:** Comparison of various models based on F1 score

It is very clear from all the line graphs (Figures 20 to 23) that, based on various evaluation metrics, the proposed model achieves the highest values. Hence, it can be said that the proposed model has been an enormous success.

**V. CONCLUSION**

Diabetes is a chronic disease characterised by high levels of glucose produced by the body, and is taking the world in its clutches with an alarming rate of diagnosis. Since this disease has no cure, it relies on early diagnosis and effective treatment for its management and control. There have been many previous studies on the same topic, using different ML models and techniques. Inferring from them, this study thus proposes an ensemble model combining two powerful models, ANN and Catboost, which leverage their respective strengths: the former to model complex nonlinear interactions and the latter to handle categorical data, thereby providing interpretable results. The results align with the proposal with an impressive accuracy of 98.94%, a very high score compared to other models from other researchers. An additional twist added to the study's proposal was to try two different data balancing techniques, SMOTE-ENN and SMOTETOMK-ENN, to find out which one fits better with the models. Once a better fit was found, the two models were combined to form an ensemble model under a weighted voting system. The other evaluation metrics, such as Precision, Recall, F1 score, MCC, and NPV, all come out to staggering results of 100%, 98%, 98.99%, 97.97%, and 98.95%, respectively. All far higher than the metrics of the individual models using either data balancing technique.. By developing such innovative ensemble methods, this study makes a significant contribution to the field of early diabetes detection. The proposed model also performed better compared to other models implemented on the same dataset. These findings can serve as a foundation for additional research that can be conducted to enhance further and utilise these models in practical and real-time datasets.

**VI. REFERENCES**

1. Yunqi Pan, Meiyu Shao, Pan Li, Chen Xu, Jingjun Nie, Kai Zhang, Sen Wu, Dandan Sui, Fu-Jian Xu, Poly Aminoglycoside-mediated cell reprogramming system for the treatment of diabetes mellitus, Journal of Controlled Release, Volume 343, 2022, Pages 420-433, ISSN 0168-3659, <https://doi.org/10.1016/j.jconrel.2022.01.041>.

2. Leila Ismail, Huned Materwala, IDMPF: Intelligent diabetes mellitus prediction framework using machine learning, Journal of Applied Computing and Informatics, Volume 21, 2021, Pages 78-89, ISSN 2634-1964

3. Kumari, Madhu & Chaudhary, Prachi. (2021). DCPM: an effective and robust approach for diabetes classification and prediction. International Journal of Information Technology. 13. 10.1007/s41870-021-00656-4.

4. Kopitar, L., Kocbek, P., Cilar, L. *et al.* Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep* 10, 11981 (2020). <https://doi.org/10.1038/s41598-020-68771-z>

5. R. Alzoubi and S. Harous, "Machine Learning Algorithms for Early Prediction of Diabetes: A Mini-Review," 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2022, pp. 401-405, doi: 10.1109/ICECTA57148.2022.9990240

6. P. Singh, S. Silakari and S. Agrawal, "An Efficient Deep Learning Technique for Diabetes Classification and Prediction Based on Indian Diabetes Dataset," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 487-491, doi: 10.1109/ICTACS59847.2023.10390518.

7. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus with Machine Learning Techniques. Front Genet. 2018 Nov 6;9:515. doi: 10.3389/fgene.2018.00515. PMID: 30459809; PMCID: PMC6232260.

8. R. Özdemir, C. Öztürk, M. Taşyürek and V. Aslantaş, "A New Deep Learning Model for Early Stage of Diabetes Disease Prediction," 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), Sivas, Turkiye, 2023, pp. 1-4, doi: 10.1109/ASYU58738.2023.10296547.

9. Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal, Volume 15, 2017, Pages 104-116, ISSN 2001-0370, <https://doi.org/10.1016/j.csbj.2016.12.005>.

10. Munadi, K., Saddami, K., Oktiana, M., Roslidar, R., Muchtar, K., Melinda, M., Muharar, R., Syukri, M., Abidin, T. F., & Arnia, F. (2022). A Deep Learning Method for Early Detection of Diabetic Foot Using Decision Fusion and Thermal Images. Applied Sciences, 12(15), 7524. https://doi.org/10.3390/app12157524

11.Amith Khandakar, Muhammad E.H. Chowdhury, Mamun Bin Ibne Reaz, Sawal Hamid Md Ali, Md Anwarul Hasan, Serkan Kiranyaz, Tawsifur Rahman, Rashad Alfkey, Ahmad Ashrif A. Bakar, Rayaz A. Malik, A machine learning model for early detection of diabetic foot using thermogram images, Computers in Biology and Medicine, Volume 137, 2021, 104838, ISSN 0010-4825, <https://doi.org/10.1016/j.compbiomed.2021.104838>

12. M Ruthika, Mooli Sai Manogna, Nachukuri Pranathi, Kanala Mamatha, K Pavani, Mr A Venkatesan, Dr R Karuna Krishnapriya, Deep learning models on early detection of Diabetes Mellitus, International Journal for Research Trends and Innovation, Volume 10, Issue 4 April 2025, ISSN: 2456-3315

13.U. Ahmed et al., "Prediction of Diabetes Empowered with Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097

14 Alzboon, M.S., Al-Batah, M.S., Alqaraleh, M., Abuashour, A., Hamadah Bader, A.F. (2023). Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods. International Journal of Online and Biomedical Engineering (iJOE), 19(15), pp. 144–165 https://doi.org/10.3991/ijoe.v19i15.42417

15.Ginting JB, Suci T, Ginting CN, Girsang E. Early detection system of risk factors for diabetes mellitus type 2 utilisation of machine learning-random forest. J Family Community Med. 2023 Jul-Sep;30(3):171-179. doi: 10.4103/jfcm.jfcm\_33\_23. Epub 2023 Jul 24. PMID: 37675209; PMCID: PMC10479022.

16.P. Nuankaew, S. Chaising and P. Temdee, "Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction," in IEEE Access, vol. 9, pp. 137015-137028, 2021, doi: 10.1109/ACCESS.2021.3117269

17.Gowthami S, R Venkata Siva Reddy, Mohammed Riyaz Ahmed, Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus, Measurement: Sensors, Volume 31, 2024, 100983, ISSN 2665-9174, https://doi.org/10.1016/j.measen.2023.100983

18.U. Ahmed et al., "Prediction of Diabetes Empowered with Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097

19. I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim and S. Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," in IEEE Access, vol. 12, pp. 66516-66538, 2024, doi: 10.1109/ACCESS.2024.3398198

20.Ullah, Zahid, Saleem, Farrukh, Jamjoom, Mona, Fakieh, Bahjat, Kateb, Faris, Ali, Abdullah Marish, Shah, Babar, Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods, *Computational Intelligence and Neuroscience*, 2022, 2557795, 10 pages, 2022. <https://doi.org/10.1155/2022/2557795>

21.Olabanjo, O., Wusu, A., Olabanjo, O. *et al.* A novel deep learning model for early diabetes risk prediction using attention-enhanced deep belief networks with highly imbalanced data. *Int. j. inf. tecnol.* 17, 1933–1955 (2025). <https://doi.org/10.1007/s41870-025-02459-3>

22.H. Naz, R. Nijhawan and N. J. Ahuja, "DT-DL Based Hybrid Approach for Early Detection of Diabetes Using PIMA Dataset," 2022 10th International Conference on Reliability, Infocom Technologies and Optimisation (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9964904

23.K. Alnowaiser, "Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model," in IEEE Access, vol. 12, pp. 16783-16793, 2024, doi: 10.1109/ACCESS.2024.3359760

24.Swapna G., Vinayakumar R., Soman K.P.,Diabetes detection using deep learning algorithms,ICT Express,Volume 4, Issue 4, 2018, Pages 243-246,ISSN 2405-9595,[https://doi.org/10.1016/j.icte.2018.10.005](https://doi.org/10.1016/j.icte.2018.10.005.(https://www.sciencedirect.com/science/article/pii/S2405959518304624))

25.S. Chen and L. Cao, "Deep Learning-Based Genetic Detection and Pathogenesis of Diabetes," 2024 Fourth International Conference on Digital Data Processing (DDP), New York, NY, USA, 2024, pp. 1-4, doi: 10.1109/DDP64453.2024.00011

26.Diabetes Prediction Dataset, <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data>

27.Diabetes Prediction (EDA/ML), <https://www.kaggle.com/code/engahmedmohamed/diabetes-prediction-eda-ml-96>

28.Project Classification SVM: Diabetes, <https://www.kaggle.com/code/lilianakozlowska/projekt-classification-svm-diabetes>