Grenze International Journal of Engineering and Technology, June Issue

Diabetes Detection using Machine Learning

Pranali Dhawas[1], Dhananjay Bhagat[2], Laxmikant Yenchalwar[3], Jayesh Nehare[4], Arya Lanjewar[5] and Sakshi Pawar[6]
Email: pranalidhawas10@gmail.com, dhananjaybhagat84@gmail.com
lyenchalwar@gmail.com, neharejayesh@gmail.com, aryalanjewar1@gmail.com, sakshi.pawar.aiml@ghrce.raisoni.net

***Abstract*— Diabetes is the most common disease worldwide. It typically arises when there is an excess of blood glucose or blood sugar levels. Glucose serves as the primary energy source for your body and is generated within the body from the food you consume, primarily through processes like glycogenolysis and gluconeogenesis. In the United States, around 37.3 million people, constituting approximately 11% of the population, grapple with diabetes, primarily Type 2 diabetes, which accounts for 90% to 95% of all diabetes cases. Globally, about 537 million adults live with diabetes presently, a figure projected to rise to 643 million by 2023 and potentially escalate to 783 million by 2045. Our model utilizes machine learning algorithms to calculate the occurrence of diabetes, based on a dataset encompassing demographic and medical information, alongside the diabetes status of patients categorized as either positive or negative. Leveraging various machine learning techniques proves advantageous in prognosticating outcomes by crafting models from patient-derived datasets. Within the scope of this endeavor, we will employ machine learning classification and ensemble techniques on a dataset to anticipate the onset of diabetes. These methodologies encompass K-Nearest Neighbor, Logistic Regression, Naïve Bayes, and Random Forest. Notably, each model exhibits distinct accuracy rates when juxtaposed with others. The maximum accuracy we attained is 98.6%. The culmination of our project endeavors to present a model that attains superior accuracy, effectively demonstrating its capacity to prognosticate diabetes.**

***Index Terms*— Diabetes detection, Machine learning, Ensemble, Random forest.**

I. INTRODUCTION

Diabetes, medically mentioned to as diabetes mellitus, is a persistent metabolic illness that touches how your body functions manages glucose, the primary fuel source for your cells. This condition arises when there are issues with the hormone insulin or how your body responds to it. The pancreas secretes insulin, which is essential for controlling blood sugar levels. [1] According to the latest National Diabetes Statistics report, In the US, 28.5 million adults—or 28.7 million people—have a diabetes diagnosis.

Types of diabetes

1. Type 1 Diabetes: The immune system of the body unintentionally targets and kills the beta cells of the pancreas, which are in charge of producing insulin, in this autoimmune disease. As a result, there is a minimal or, at times, complete absence of insulin production. Individuals afflicted with Type 1 diabetes want to use an insulin pump or injections to control their blood sugar levels. This type of diabetes is incurable and typically manifests in childhood or adolescence.

2. Type 2 Diabetes: Type 2 diabetes stands as the prevailing variant of diabetes, typically manifesting later in life,

Grenze ID: 01.GIJET.10.2.52

although its onset is possible at any age. When a person has Type 2 diabetes, their body becomes less sensitive to insulin and their pancreas may have difficulty producing enough insulin to maintain normal blood sugar levels. This type of diabetes exhibits a robust association with lifestyle elements like obesity, a sedentary way of life, suboptimal dietary habits, and genetic predisposition. Unlike Type 1 diabetes, Type 2 diabetes can frequently be effectively measured through a combination of lifestyle adjustments, oral medications, and, on occasion, insulin injections.

3. Gestational Diabetes: Some women who did not have diabetes before to getting pregnant may develop gestational diabetes during pregnancy. Pregnancy-related hormonal changes might result in insulin resistance, which raises blood sugar levels if the body fails to produce sufficient insulin. Most women with gestational diabetes can manage it through dietary changes, physical activity, and sometimes medication. Typically, the condition resolves after childbirth, however, there is a greater chance that women who have experienced gestational diabetes will go on to get Type 2 diabetes in the future. Diabetes often manifests as increased thirst, thirst, fatigue, hunger, and blurred eyesight in addition to unexplained weight loss.

How can we Prevent Diabetes?

[1] While it's not possible to prevent autoimmune and genetic variations of diabetes, there are definite measures you can take to reduce your likelihood of growing prediabetes, Type 2 diabetes, and even gestational diabetes.

- Take healthy diet and nutrition rich diet.
- Engage in physical activity. Spend at least five days a week, 30 to 45 minutes a day of physical activity.
- Keep working to maintain your weight that is healthy according to age and height.
- Quite Smoking, and alcoholic drinks.
- Get adequate and deep sleep (typically 7 to 9 hours) and ensure treatment of sleep disorders, if any.
- Follow your healthcare provider's instructions regarding the use of medications to control existing risk factors associated with heart disease.

II. Literature Review

Heart disease and diabetes are the most common diseases worldwide. It became crucial and foremost step to predict such diseases at the early stage. Many studies have already proved the true potential of machine learning algorithms in prediction of various diseases.

KM Jyoti Rani in [2] presented a study entitled ‘Diabetes prediction using machine learning’ aimed at comparing the different approaches for predicting diabetes disease. The proposed methodology encompasses five machine learning algorithms named KNN, Logistic Regression, Decision tree, Random Forest. After employing these five distinct methods on John diabetes database, testing accuracy achieved was 78%, 78%, 99%, 97%, 77% respectively. Decision tree yielded maximum remarkable accuracy of 99%.

[3] Chilupuri Anusha’s project of 2023 is aimed to design and implement diabetes mellitus prediction using machine learning algorithms and ensemble learning methods. In this proposed approach, logistic regression achieved maximum accuracy of 82.7% compared to Random Forest and KNN. This study highlighted the crucial role algorithm selection in enhancing accuracy and effectiveness within the systems.

Among the three types of Diabetes, Study [4] mainly focused on predicting gestational diabetes using SVM algorithm. Their main concern was on increasing the number of features to achieve outstanding accuracy; hence they trained the model with 16 attributes, improvised the performance of SVM algorithm and achieved 93.26% accuracy. In this study, the authors introduced an advanced diabetes prediction model that incorporates supplementary criteria alongside the conventional ones, such as glucose levels, body mass index (BMI), age, and insulin levels. Instead they used attributes like genital thrush, muscle stiffness, itching and others. The utilization of these additional criteria resulted in an enhanced ability to categorize diabetes accurately when compared to previous approaches using a different dataset.

[5] Recent study of April, 2023 published in IEEE entitled ‘Risk Prediction of Diabetes Based on Spark and Random Forest Algorithm’ focused on improving diabetes detection and prediction by splitting dataset in ratio of 7:3 for training and testing[5]. On the spark platform of big data computing, Random Forest model was implemented yielding an accuracy of 97.12%. The outcome of the analysis indicates that Random Forest achieving an F1 score of 0.9714, exhibited superior performance compared to the other classifiers.

[6] Md. Tanvir Islam and colleagues introduced The Random Forest approach aimed at forecasting diabetes, proved to be the most effective with training data that had 10 features after several dataset issues were resolved and data analysis and cleaning were carried out utilizing machine learning techniques. In a tabular comparison of existing systems using the suggested methodology, the random forest algorithm was shown to contribute most to the greatest accuracy of 98.24%. They constructed a system that can accurately anticipate diabetes onset for

patients through the application of the Random Forest algorithm within the framework of machine learning. The results illuminated that the suggested approach greatly improves the precision of predicting diabetes onset.

Limitations of existing systems

- The John diabetes database available at [7] used in [2] encompasses only 2000 cases and 8 features. In this paper, our proposed methodology model is trained with 100000 cases and 7 features.
- The accuracy of model in [3] and [5] could have been improved by applying further methods of data pre-processing, data mining and pipelining.
- The accuracy of diabetes prediction was successfully increased in [4] by increasing the amount of characteristics, although the dataset only contains 520 patient information, of which 320 were positive for diabetes and 200 were negative. By utilising other algorithms, such as Logistic Regression and ensemble techniques, increasing the number of data samples, and determining the relevance of the features, the accuracy of this project can still be increased.

III. PROPOSED METHODOLOGY

The objective of the study is to predict weather a person is diabetic or not. The overall workflow of the analysis has been shown in Figure 1. The whole process can be divided into 3 main phases.

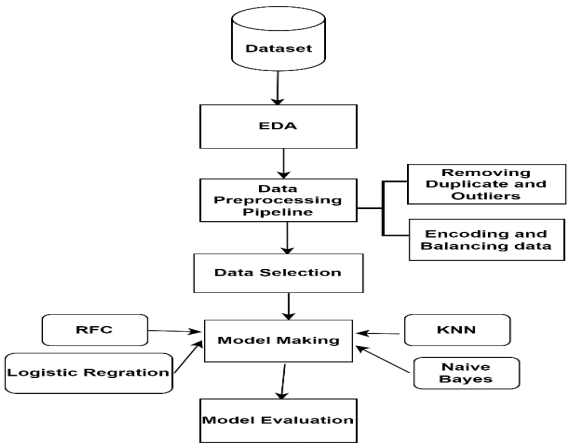

Figure 1: System Architecture

1. Data Collection and analysis
2. Data pre-processing
3. Model training and testing

Dataset Description: The Diabetes dataset includes patient demographic and medical data as well as a binary classification of their diabetes status (positive or negative). This dataset presents a wide range of characteristics linked to 100,000 patients, including age, gender, body mass index (BMI), blood glucose level, smoking history, hypertension, heart disease, and HbA1c level. (Reference Table 1).

Brief information of features:

- *Gender* - An individual's gender, or biological sex, may have an impact on how susceptible they are to diabetes.
- *Age* - Age is a significant factor, with diabetes is detected in older persons more frequently. Our dataset includes ages ranging from 3 to 80 years.
- *Hypertension* - Hypertension, characterized by consistently high blood pressure in the arteries, is indicated by values of 0 (no hypertension) and 1 (hypertension present).
- *Heart Disease* - Diabetes is another medical condition linked to an increased risk of heart disease, specifically myocardial infarction.
- *Smoking History* - It is also known that a person's smoking history increases the likelihood of developing diabetes and its associated problems.
- *BMI* - Body Mass Index, or BMI, calculates body fat from height and weight. A higher BMI is linked to a higher chance of developing diabetes.

- *HbA1c Level* - Haemoglobin A1c level determines the average blood sugar level of a person for the preceding two to three months.
- *Blood Glucose Level* - It indicates the amount of glucose in the bloodstream at a precise moment. Elevated levels are associated with an increased risk.
- *Diabetes* – It's a target variable for prediction, with a value of 1 indicating its presence and 0 indicating its absence.

TABLE I: DATA DESCRIPTION

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

Data Analysis: Explore the data visually and statistically to understand its characteristics. This includes creating histograms, scatter plots, summary statistics, and identifying patterns.
Data Analysis on Single Feature:

- *Gender*: The ratio of Male & Female to have diabetes is same.
- *Age*: The lowest age to have diabetes is 3 years.12% of Positive diabetes patients are of age greater than 79.
- *Hypertension*: A person has a 30% probability of developing diabetes if they have hypertension. A person has a 6% probability of having diabetes if they do not have hypertension.
- *Heart disease*: A person has a 32% chance of having diabetes if they have any heart disease, meanwhile, if they do not have any, probability reduces to 6%.
- *BMI*: Most people have BMI of 27.23 which falls in category of overweight.
- *Diabetes*: In our dataset the ratio of positive and negative samples is 1:10.

Data Analysis between Two Features:

1. bmi vs HbAlc_level

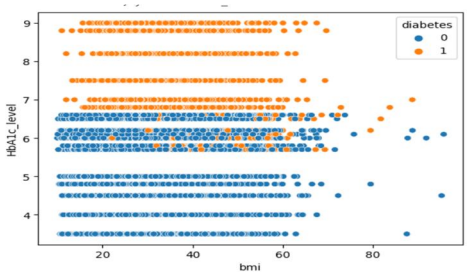

Figure 2: bmi vs hbaic_level

From Figure 2 we can conclude that those have HbAlc_level greater than 6.5 have the 100% chance of developing diabetes.

2. HbA1c_level vs blood_glucose_level

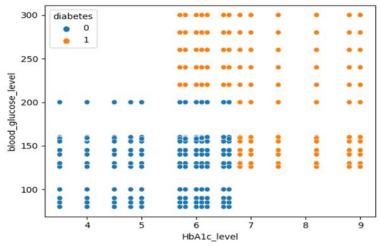

Figure 3: Blood_Glucose_Level Vs Hb1Ac_Level

Furthermore, if a person has blood_gluocose_level greater than 250, are in 100% chance of having diabetes.

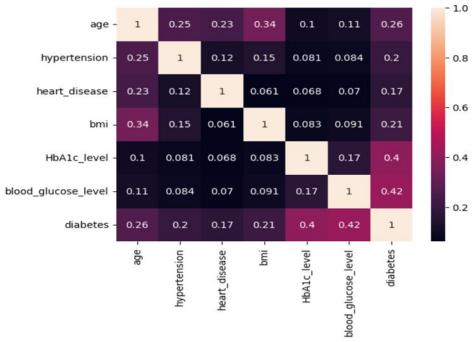

Figure 4: Correlation matrix for all features:

Data pre-processing: Data pre-processing holds paramount significance, particularly in healthcare-related contexts. This phase addresses issues like missing values and data impurities that could compromise data efficiency. Enhancing quality and efficiency post-mining is the primary objective of data pre-processing. This step is essential to ensure accurate outcomes and successful predictions when employing machine learning techniques on a dataset.

Data cleaning: Data cleaning involves extracting raw information and employing various techniques to refine the data, including removing duplicates and irrelevant entries.

Feature Transformation: Raw data often needs to be transformed, cleaned, and pre-processed to make it suitable for analysis. In 'bmi' feature 27.23 value count is 25495 and it is present near to center of distribution of data, because of this most of reasonable values detect as an outliers[8]. So, we define frequency and convert It in categorical. Frequencies are underweight (below18.5), normal (18.5, 25), overweight (25, 30), obesity class1 (30, 34), obesity class2 (35, 40), obesity class3 (above 40).

Data balancing: Data balancing is crucial since imbalanced classifications can pose challenges for predictive modelling. In our dataset ratio is 1:10 of having diabetes and not having diabetes. So, we create samples of minority class same as majority class.

Column Transformer Pipeline: it is a sequential steps transformer that applied to the specific column without changing others parameter. These pipelines help automate and standardize the data transformation process, making it easier to manage and reproduce data pre-processing steps. Column transformation pipelines are a key component of the data engineering and data pre-processing workflow.

Feature Encoding: It involves converting categorical or nominal data into a numerical presentation that machine learning algorithms can comprehend and process with ease. In this context, we encode the gender column using Ordinal Encoder and the BMI category and smoking features using One Hot Encoding.

Employing Machine Learning : Once the data is prepared, the application of machine learning techniques ensues. A diverse array of classification and ensemble methods is employed for diabetes prediction, all of which are applied to diabetes dataset. The overarching objective revolves regarding the use of machine learning methods to assess the usefulness of these methods and ascertain their correctness.

Random Forest (RF): In the realm of RF ensemble learning methodology, applicable to classification and various other tasks, decision trees are built through training alongside optimal tree approximations. During the classification phase, the majority voting approach is employed to yield significant outcomes in the identification of diabetic conditions. Outcomes for a non-afflicted patient can be categorized as indicative of good health or a medical ailment. In most of cases RF perform well than other algorithms, in our model we used Random Forest Classifier with max_depth equal to 30 and achieved the accuracy of 98.6%. This, in comparison to all other algorithms, has the highest accuracy.

Logistic Regression: Logistic regression, similarly, to supervised classification techniques, is employed to assess the likelihood of a binary outcome, relying on single or multiple predictor variables, which can take on continuous or discrete values. This approach finds utility when the objective is to categorize data points into distinct groups. The binary classification nature involves assigning either a 0 or 1, often applied in scenarios like determining diabetes positivity in patients. Within the Logistic Regression framework, the sigmoid function assumes a pivotal role by assigning each data point a place on an S-shaped curve. The sigmoid function's representation is outlined in the following equation:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

This model is best suitable for classification problems but in this case we attained accuracy of 87%.

K-Nearest Neighbour: - KNN also falls under the umbrella of supervised machine learning methodologies. It is adept at addressing both classification and regression tasks. Functioning as a lazy prediction technique, KNN operates under the assumption that entities sharing similarities tend to be situated in proximity. This proximity translates to data points with resemblances often being spatially adjacent. KNN facilitates the categorization of novel instances based on measures of likeness.

KNN gifted us the second highest accuracy (95%), achieved with features of n_neighbours = 10, weight = 'distance', algorithm =" ball tree".

Naive Bayes (NB): It is a probabilistic based classification technique grounded in Bayes' theorem, employed for binary and multiclass classification endeavors. It earns the moniker "naive" due to its assumption of feature independence, a presumption that may not consistently align with real-world conditions. Nonetheless, despite this simplification, Naive Bayes frequently exhibits commendable performance across diverse applications, particularly when working with relatively limited datasets. This algorithm helped us achieving accuracy of 81%.

Confusion Matrix: In the realm of machine learning, When evaluating a classification algorithm's efficacy, the Confusion Matrix is essential. The matrix takes on a tabular format, with the predicted class labels represented in columns and the actual class labels in rows. The following definitions pertain to specific terminology within the comprehensive Confusion Matrix Structure. These terms will subsequently serve as the foundation for assessing the efficacy of each classifier.

- *Actual Class*: The class label designating the starting class prior to the development of the classifier.
- *Predicted Class*: Following the construction of the classifier, the predicted class is represented by the class label.
- *True Positive*: The quantity of cases that were accurately counted as positive.
- *False Positive*: The quantity of cases that are mistakenly projected as positive when they are actually negative is known as the "False Positive" rate.
- *True Negative*: The quantity of cases that were accurately predicted to be negative.
- *False Negative*: The quantity of cases that are positively correlated but are mistakenly forecasted as negative.

Models	Accuracy score	Precision	Recall	F1 score
RFC	0.98	0.97	1.0	0.99
KNC	0.95	0.92	1.0	0.96
GaussianNB	0.81	0.79	0.85	0.81
Logistic Regression	0.87	0.87	0.87	0.87

Figure 5: Accuracy Of Algorithms

IV. Experimental Results

By trying all possible method and EDA (Exploratory data analysis) we noticed that the Random Forest Classifier algorithm is best suit with max_depth equal to 30 and it yields 98.6% of Accuracy that is helpful for medical experts to correctly identify patient with diabetes and based on given features predict as well.

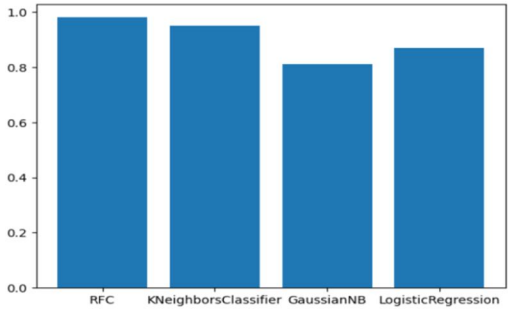

Figure 6: Accuracy Result Of Algorithms

This characteristic assumed a significant part in the estimate process associated with the random forest algorithm. The cumulative significance of each prominent attribute contributing to diabetes prediction was graphed, with the

importance of individual attributes depicted along the Y-axis and the corresponding attribute names along the X-axis.

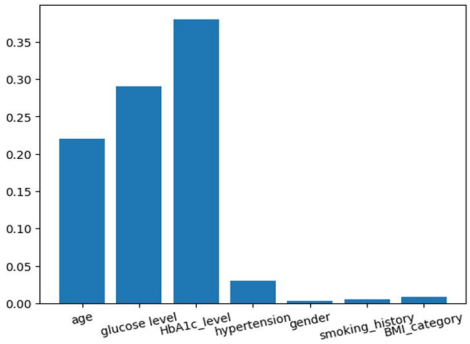

Figure 7: Feature Importance Plot for Random Forest

V. Conclusion

Developing and implementing a diabetes detection system based on machine learning techniques was the primary objective of this effort with a focus on assessing the effectiveness of the method. We have successfully met this objective. The devised approach encompasses diverse classification and ensemble learning methods, encompassing KNN, Random Forest, Naïve Bayes, and Logistic Regression classifiers. The dataset had 100000 rows which was quite imbalanced due to more number positive samples than negatives one's. Although, we faced these issues, we overcame and successfully performed analysis. Ultimately, a classification accuracy rate of 98.6% was attained. The findings of the study may open the way for early healthcare prognostics and well-informed diabetic treatment decisions, helping to preserve human lives.

References

[1] National Diabetes Statistics Report by Centre for Disease Control and Prevention(CDC) at https://www.cdc.gov/diabetes/data/statistics-report/index.html, last reviewed on June 29, 2022

[2] Rani, KM. (2020). Diabetes Prediction Using Machine Learning. 'International Journal of Scientific Research in Computer Science, Engineering and Information Technology'. 294-305. 10.32628/CSEIT206463.

[3] Chilupuri, Anusha. (2023). A Machine Learning Approach for Prediction of Diabetes Mellitus. 10.30534/ijeter/2023/031162023.

[4] Sharma, Shivani & Rai, Bipin Kumar & Gupta, Mahak & Dinkar, Muskan. (2023). DDPIS: Diabetes Disease Prediction by Improvising SVM. International Journal of Reliable and Quality E-Healthcare. 12. 1-11. 10.4018/IJRQEH.318090.

[5] Z. Dai, Z. Hu, T. Shen and Y. Zhang, "Risk Prediction of Diabetes Based on Spark and Random Forest Algorithm," 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2023, pp. 535-539, doi: 10.1109/EEBDA56825.2023.10090801.

[6] M. T. Islam, M. Raihan, F. Farzana, N. Aktar, P. Ghosh and S. Kabiraj, 'Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm,' 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-6, doi: 10.1109/ICCCNT49239.2020.9225430.

[7] John diabetes dataset taken from frankfurt hospital, Germany at: https://www.kaggle.com/johndasilva/diabetes

[8] Barse, Saurabh and Bhagat, Dhananjay and Dhawale, Krupali and Solanke, Yash and Kurve, Dikesh, "Cyber-Trolling Detection System" (January 10, 2023). Available at SSRN: https://ssrn.com/abstract=4340372 or http://dx.doi.org/10.2139/ssrn.4340372.

[9] F. Dsouza, A. Bodade, H. Kolhe, P. Chaudhari and M. Madankar, "Optimizing MRC Tasks: Understanding and Resolving Ambiguities," 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/PCEMS58491.2023.10136031.

[10] Hande, T., Dhawas, P., Kakirwar, B., & Gupta, A. , "Yoga Postures Correction and Estimation using Open CV and VGG 19 Architecture" 2023, International Journal of Innovative Science and Research Technology, Volume 8, Issue 4, April – 2023, ISSN No:-2456-2165, IJISRT23APR2280.

[11] R. Chitte, R. Mandal, R. Mathur, A. Sharma, and D. Bhagat, "Using Natural Language Processing (NLP) Based Techniques for Handling Customer Relationship Management (CRM)," vol. 10, no. 2, pp. 18–22, 2023.

[12] D. Bhagat et al., "SMS SPAM DETECTION Web Application Using Naive Bayes Algorithm & Streamlit," vol. 13, no. 1, pp. 276–280, 2023.

[13] P. Dhawas, "Big Data Preprocessing , Techniques , Integration , Transformation , Normalisation , Cleaning ," pp. 159–182, doi: 10.4018/979-8-3693-0413-6.ch006.

[14] P. Nirale and M. Madankar, "Design of an IoT Based Ensemble Machine Learning Model for Fruit and Quality Detection," 2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22), Nagpur, India, 2022, pp. 1-6, doi: 10.1109/ICETET-SIP-2254415.2022.9791718.
[15] P. Nirale and M. Madankar, "Analytical Study on IoT and Machine Learning based Grading and Sorting System for Fruits," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCICA52458.2021.9697161.
[16] R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 2014, pp. 1-5, doi: 10.1109/ICCIC.2014.7238427.
[17] Sahu, M., Dhawale, K., Bhagat, D., Wankkhede, C., & Gajbhiye, D. (2023). Convex Hull Algorithm based Virtual Mouse. Grenze International Journal of Engineering & Technology (GIJET), 9(2).