2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) | 979-8-3503-1590-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ACCAI58221.2023.10200323

Web application based Diabetes prediction using Machine Learning

G Ravi Kumar[1], Reddyvari Venkateswara Reddy[1], Jayarathna M[2], N Pughazendi[2], S Vidyullatha[3], and *Pundru Chandra Shaker Reddy[4]

[1]Computer Science and Engineering, CMR College of Engineering & Technology, Hyderabad, India
[2]Computer Science and Engineering, Panimalar Engineering College, Poonamalle, Chennai, India
[3]Computer Science and Engineering, BVRIT Hyderabad College of Engineering for Women, Hyderabad
[4]Shcool of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.

E-mail: ravicmrcse@gmail.com, pughazendi@gmail.com, venkatreddyvari@cmrcet.ac.in, vidyullatha.1988@gmail.com, rathnajaya98@gmail.com, chandu.pundru@gmail.com

Abstract- Diabetes is a worldwide epidemic that affects millions of people. Long-term consequences, such as cardiovascular disease and renal failure, are more likely to occur in people with diabetes. If this condition could be diagnosed at an early stage, people may live longer and better lives. Diabetic primary care can benefit from many supervised machine learning models educated on relevant datasets. Finding reliable classifier models for diabetes detection using clinical data is the focus of this investigation. This article introduces a number of machine learning techniques, such as the decision-tree (DT), naïve-bayes (NB), k-nearest neighbour (KNN), random-forest (RF), gradient-boosting (GB), logistic-regression (LR), and support vector machine (SVM) that may be taught using a variety of datasets. We have used effective pre-processing methods, such as label-encoding and normalisation, to raise the quality of the models' predictions. Additionally, we have isolated and prioritised a variety of risk variables utilising different feature selection methods. Extensive tests have been run on two separate datasets to evaluate the model's efficacy. When compared to other previous studies, our model shows improved accuracy, ranging from 3.71 percent to 15.13 percent, based on the dataset and the ML technique used. At last, a machine learning algorithm with the best accuracy is chosen for research and development. We use the python flask web development framework to incorporate this model into a web application. This study's findings provide preliminary evidence that using a suitable preprocessing pipeline on clinical data and using ML-based classification might improve the accuracy and efficiency of diabetes prediction.

Keywords— Android Application, Machine Learning, Diabetics, Accuracy, Prediction

I.INTRODUCTION

Hyperglycemia due to defects in insulin production, insulin action, or both characterises the metabolic disease class known as diabetes mellitus (DM) or simply diabetes [1]. Insulin deficit, in which the - cells in the pancreas fail to make enough insulin, is a contributing factor to this medical disease; this kind of diabetes is known as type I diabetes. Type II diabetic mellitus (T2DM) is the other major cause, and it is produced by a combination of insulin resistance and an inadequate compensatory secretary response to insulin [2]. Almost 422 million individuals had DM in 2014, and its effects contributed to a 5 percent increase in the premature death rate between 2000 and 2016, per the World Health Organization (WHO). As a result, DM has been ranked by the WHO as the seventh leading cause of mortality worldwide in 2016. More than a million individuals worldwide have lost a limb to it, and many more have been left visually impaired or blind [3].

Despite the staggering USD 760 billion dollars spent on diabetes, every other diabetic goes undetected. When untreated, diabetes mellitus (DM) can wreak havoc on the kidneys, blood vessels, and eyes. With over half of persons affected by diabetes being unaware of their status until complications emerge; early diagnosis of the disease is crucial [4]. Machine learning (ML) can play a significant role in preventing type2 diabetes, with previous studies estimating that around 80% of T2DM can be prevented or at least postponed if persons at risk are recognised and addressed early. Several real-world issues have been addressed with the use of ML &DL methods. By controlling data from people who already have T2DM, ML can analyse a person's health and forecast whether or not they will get the disease [5]. Since this is something that millions of people are interested in, there are a lot of studies being conducted and that have already been published. The authors of this study report the creation of a deep learning model for T2DM prediction and its incorporation into cloud-based software available to users all over the world. Due to the life-saving potential of early diagnosis, this study has the potential to make a momentous involvement to the state of the art medical care system [6].

With the use of machine learning algorithms, this study intends to offer analytical findings on the physical factors and environmental elements that escort to the onset of diabetes in people. Pre-processing the datasets advances the model's exactness more than what is possible with just current research [7]. This approach of selecting features from a dataset is based on correlation, which helps uncover missing data points. The web app uses the classifier that performs best across all datasets. In light of this knowledge gap, we propose the following study objectives. For comparing the PIMA Indian dataset to other datasets, how does the exactness of ML techniques change when anticipating diabetes? Is the most accurate method for identifying diabetes dependent on any particular factors? How can we locate a powerful machine learning method to do this in a web app?

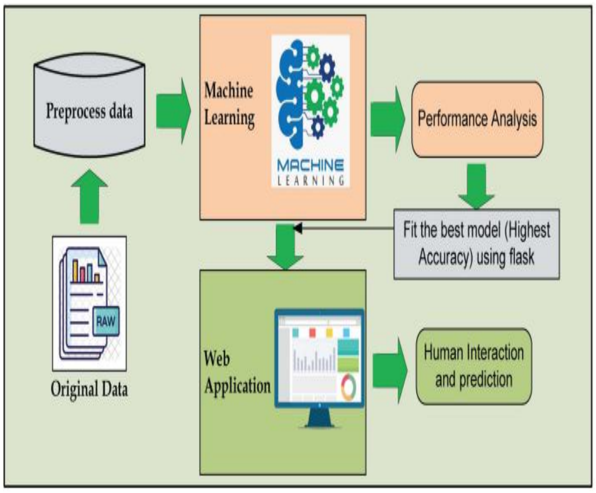


Fig.1. Concept diagram of Machine Learning

We have implemented and compared the efficacy of multiple machine learning algorithms for using individual-level diabetes risk factors to make predictions. To construct a predictive model, we compared the effectiveness of seven distinct models: NB, DT, RF, SVM, LR, GB, &KNN techniques. The ML strategies are trained on two datasets containing information such as glucose-level, insulin-level, blood-pressure, body mass-index, &age [8]. Several performance measures are used to assess the algorithms' outputs. Finally, an internet-based programme is created to make diabetes predictions for everyone depending on the accuracy level. This programme may be used by anybody, on any device, to acquire a diabetes prognosis. Each step of the implemented ML dependent diabetes anticipation design is depicted in Fig. 1.

Here are some of the findings that this study contributes to:

- To use four distinct clinical datasets to train a variety of ML approaches for diabetes detection. A variety of pre-processing methods are used on each of the datasets.
- Precision, Recall, f1-score, and accuracy are only few of the metrics used to evaluate the ML algorithms' results on the four datasets. Down addition, we have used a variety of feature selection techniques, including correlation, chi-square, etc., to zero in on a number of critically significant characteristics or traits. The most strongly connected aspects to diabetes illness are identified by the feature selection techniques.
- The effectiveness of each ML method was evaluated in light of the smaller dataset of features. Depended on the outcomes, a web-application is created to diagnose diabetes in individuals.

For the sake of clarity, this paper will follow these sections: In Part II, we briefly review the relevant literature and compare and contrast the various approaches used by the various studies we review. In Part III, we provide the presented approach, system design, and a high-level outline of the entire system. Section IV details the results and analysis of the suggested technique, while section V discusses the conclusion and future efforts.

II.RELATED WORKS

Brief discussions of a few relevant works are presented below. Many studies have utilised the Pima Indians Diabetes Dataset (PIDD) for diabetes forecasting. Predictions of diabetes were made using a number of supervised machine learning techniques [9]. Some examples are the SVM using a radial basis function (RBF) kernel, the ANN, the MDR, the linear SVM, and the k-NN. LR has been utilized to classify diabetes risk variables based on p value &odds ratio. Patients with diabetes may now be anticipated using one of four classifiers: NB, DT, Adaboost, or RF. K2, K5, and K10 partitioning techniques were also implemented, with each iteration used in 20 separate trials. Classifier performance was evaluated using accuracies and area-under-curve (AUC). Authors [10] compared popular regression models for predicting T2DM, including Glmnet, RF, XGBoost, and LightGBM. The purpose of this research was to compare traditional regression methods to cutting-edge machine learning approaches to see if the latter could more accurately predict impaired fasting glucose and fasting plasma glucose (FPGL) levels. Select four categories—NB, DT, Adaboost, and random forest—for the prediction of diabetes patients. Three distinct partition schemes also used these techniques (K2, K5, and K10). Both accuracy (measured by ACC) and curve surface area are used to evaluate these classifiers (AUC).

Using a soft voting classifier, the authors of [11] were able to ensemble the ML approaches RF, LR, &NB. On the Pima Indians Diabetes Database, the soft voting classifier outperformed the individual methods in terms of accuracy, recall, precision, &F1-score. The greatest results for detecting diabetes on PIDD were achieved with DL, DT, Artificial Neural Network (ANN), and NB. The study [12] claims to be the first to use convolutional Long Short Term Memory (Conv-LSTM) for diabetes prediction; it employs the Boruta algorithm for feature selection from PIDD, compares its results to those obtained using Convolutional-Neural-Network (CNN), LSTM, and CNN-LSTM methods, and concludes that the former yields superior results. NB, SVM, and DT are all methods recommended for use with PIDD in [13], although NB offers superior performance. On order to diagnose T2DM in a small dataset of 149 patients from a Taiwanese hospital, employed Deep-Neural-Network (DNN), SVM, RF, and other ensemble methods. However, the DNN model underperformed the others due to the limited size of the dataset. Using a preprocessing approach called Synthetic Minority Oversampling Technique, the authors of study [14] used models such as DT, multi-layer perceptron (MLP), bagging, and SVM (SMOTE). In particular, DT and MLP benefitted from this preprocessing strategy, while all models improved overall. Improved DM prediction accuracy for PIDD was achieved using an ensemble model based on DT. The authors used the median as a preprocessing step to make up for PIDD's missing values, making the dataset more stable while increasing accuracy. Improved accuracy over more conventional LSTM and MLP networks was one of the main goals of the work presented in [15], which combined LSTM with AR to stabilize training datasets. Several machine learning classifiers, such as KNN, DT, RF, logistic regression, &SVM, were utilized to predict T2DM in the PIDD dataset in the research. When comparing the accuracy of the several used classifiers for predicting Type 2 Diabetes, the RF classifier emerged as the clear victor. Overall, the accuracy of T2DM prediction from PIDD using MLP with a genetic algorithm was rather high. In order to predict T2DM, the scientists employed a genetic-neural network and a UTA optimization technique for feature selection [16].

III. Proposed Methodology

Prediction of diabetes using the suggested framework is shown in Figure 2. In the first step, we preprocess two independent data sets. To uncover characteristics that will be helpful in diagnosing diabetes, the pre-processing phase involves analysing the association between properties of the datasets. Then, we split the data into two groups: training and testing. Various ML approaches are utilized to the training data to generate predictive ML models. After that, we evaluate the proposal using a variety of standards. The final step is integrating the top ML model into a flask-based web app. We then provide a quick overview of the operation of each component:

Data Collection: To ensure the model's stability, we gathered data from two independent sources, each with a unique collection of attributes or parameters. Statistics on diabetes and other health characteristics were collected from people all around the world and from a wide range of health research institutions to create the databases [17].

Data Analysis and Preprocessing: In order to increase the performance of the ML technique, several pre-processing strategies are done to the datasets before being fed into the design. Common examples of pre-processing activities include filtering for anomalies, handling missing data, standardising and encoding the data, etc [18].

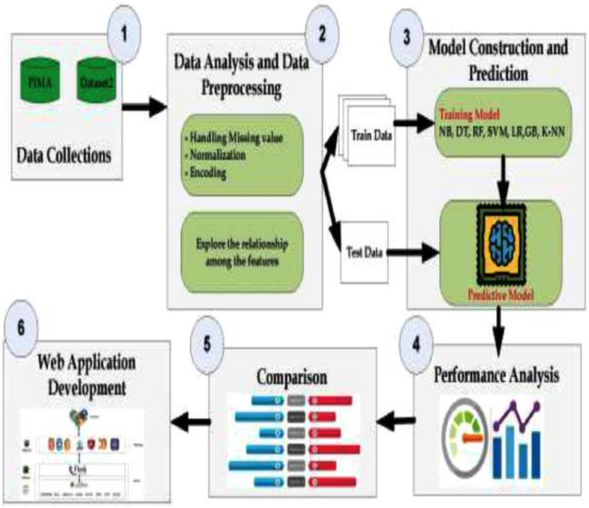

1
PIMA
Dataset2
Data Collections
2
Data Analysis and Data Preprocessing
- Handling Missing value
- Normalization
- Encoding
Explore the relationship among the features
Train Data
Test Data
3
Model Construction and Prediction
Training Model
NB, DT, RF, SVM, LR,GB, K-NN
Predictive Model
4
Performance Analysis
5
Comparison
6
Web Application Development

Fig. 2. Work flow of the presented model

Model Building and Prediction: Eighty percent of the preprocessed data was utilised for training the predictive model, while the remaining twenty percent was used for testing.

Performance Analysis: We have assessed the suggested model's output in terms of a variety of performance measures. The optimal algorithm for the construction of a web application is chosen as the one that delivers the highest prediction accuracy.

Performance comparison: At this point, we have compared the proposal's accuracy to that of several previous researches dealing with diabetes prediction. When compared to the most up-to-date studies in the field, the performance results show that the suggestion is an improvement.

Creation of Web-Based Programs: We've utilized the Flask micro-framework and the top approach to create a sophisticated web app. If you want to know if you'll get diabetes, you'll need to fill out a form with a certain minimum number of diabetes-related characteristics. By utilising the machine learning model, the server-hosted application may make accurate predictions. Here, we detail the specific machine learning methods that have been put into use.

Naive Bayes: Classification using Naive Bayes is based on Bayes' Theorem. This model takes into account the possibility that predictors are not truly independent of one another. Naive Bayes uses a conditional probability model to label data points in a problem instance [19].

Decision Tree: One of the most used and an effective strategy for classification &prediction is the DT. A decision tree's internal nodes stand for attribute testing. The leaf nodes with the relevant class names indicate the results of the testing [20].

Random Forest Classifier: Classification models like random forest are used to model forecasts and analyse behavioural traits. In the RF, numerous decision trees are used, and each tree is a representation of a different instance. The examples aid the random forest in classifying the data that is fed into it. Using a voting system, random forests do blind evaluations of data and then deliver the forecast with the most votes [21].

Support Vector Machine: Classification and regression issues may be handled with the help of Support Vector Machines.

Logistic Regression: In logistic regression, the probability is used to identify whether or not a specific data point is in the "1" category [22].

Gradient Boosting: Another machine learning method that can deal with regression and classification issues is gradient boosting. Several simple prediction models, often decision trees, are combined to form this one. The model is built step by step, just as previous boosting methods. Then, an optimization strategy based on an arbitrary differentiable loss function is used to further reduce the complexity of the model [23].

K-Nearest Neighbor: The algorithm is a popular example of a fundamental Machine Learning algorithm that relies on the Supervised Learning approach. It is commonly used for classification but also for regression. When determining how to proceed with a new case or set of data, the K-NN method takes into account how similar it is to the previous case or set of data. Afterwards, the new instance is filed under the heading that most resembles it [24].

IV. Results and Discussions

In this part, we put our suggested model through its paces utilizing a number of different ML strategies, such as the NB, DT, RF, SVM, LR, GB, &K-NN. We have evaluated the efficiency using four datasets, each with a unique combination of attribute types and numbers. There are four possible answers in the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The next step is to evaluate the proposed model using the following criteria.

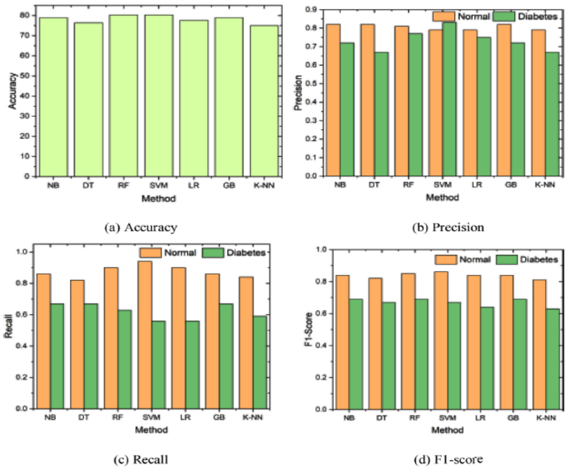
Experiment results in terms of accuracy, precision, recall, and f1-score are shown in Figure 2. Models range in accuracy from 77.63% to 80.25%, and from 78.95% to 75%. SVM performs better than NB, DT, RF, SVM, LR, GB, and KNN. In comparing the ROC (receiver operating characteristic) curves (graph 3(e)) for different models, we observed that SVM offers superior performance. Unexpected advances in model accuracy are ensured by the proficient preprocessing pipeline's attention to outlier's elimination, missing-value management, and label-encoding. All machine learning techniques used here exceeded the best previous efforts. The sum of the selected five variables is displayed in Table1 &Figure 3. In comparison to other algorithms, SVM clearly excels, with an accuracy of 83.13 percent. Table1 &Figure 4 displays a comparison of dataset accuracy performance. Accuracy is increased by 5.57 percentage points, 12.81 percentage points, 2.71 percentage points, 4.99 percentage points, and 15.13 percentage points for NB, DT, RF, SVM, &KNN, correspondingly, compared to LR.

Table. 1. Performance Metrics of ML approaches

Model	Accuracy	Precision	Recall	F1-score
NB	79	0.7	0.65	0.7
DT	78	0.75	0.71	038
RF	80	0.79	0.66	0.69
SVM	80	0.85	0.58	0.67
LR	78	0.74	0.58	0.62
KNN	76	0.68	0.61	0.67

Table 2. Accuracy comparison with existing Models

Model	Accuracy
NB	79
DT	78
RF	80
SVM	80
LR	78
KNN	76



(a) Accuracy (b) Precision

(c) Recall (d) F1-score

Fig. 3. Performance of proposed models

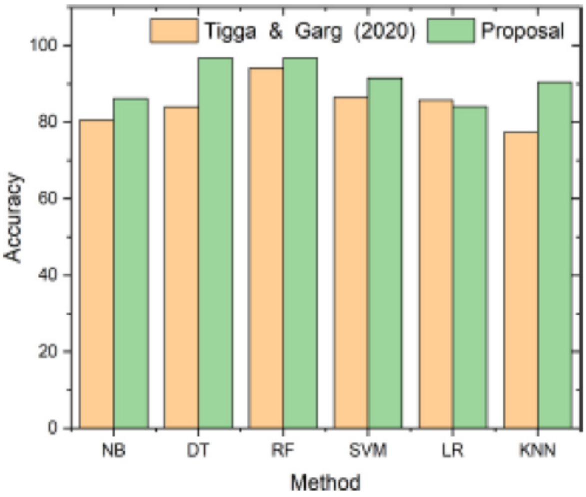
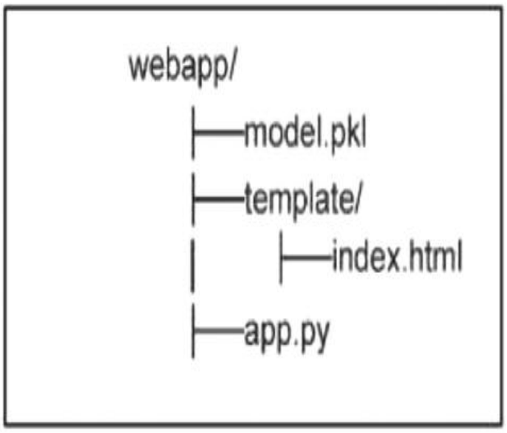


Fig. 4. Performance comparison with existing models

Implementation of web application

This section details the steps taken to create a web application that can accurately forecast a diabetic patient's future health complications. We begin with a quick summary of the web app's creation process. Then, we detail the practical applications of the programme with a few straightforward experiments. Flask is a microweb framework written in Python that makes it easy to add features to a project as if they were originally part of the framework. The application's fundamental file structure is depicted in Figure 5, and the development process includes the following four software modules.



```
webapp/
    ├──model.pkl
    ├──template/
    |       ├──index.html
    ├──app.py
```

Fig. 5. Web Application Scheme

The diabetes prediction model is stored in the model.pkl file. Since SVM achieved 78.125% precision across all characteristics, we will use it as the prediction model in the model.pkl file.

- **app.py** - This package contains Flask APIs that take in Diabetes data via a graphical user interface (GUI) or API requests, then use our model to forecast a result and return it.
- **Template** - This folder has an HTML form (index.html) that the user may use to enter their diabetic information and see the expected results.

  The CSS file with the formatting instructions for our HTML form can be found in the Static folder. Figure 6 depicts the processes involved in the proposal's application process.
- The user provides the required data for the programme via a Web page (Step-1). •The data is transmitted to a server in the backend (Step-2:).
- The outcomes may be predicted in advance using the flask server that has been implemented using a machine learning method (Step-3 and Step-4).
- In the end, the webpage displays the expected outcome (Step-5).

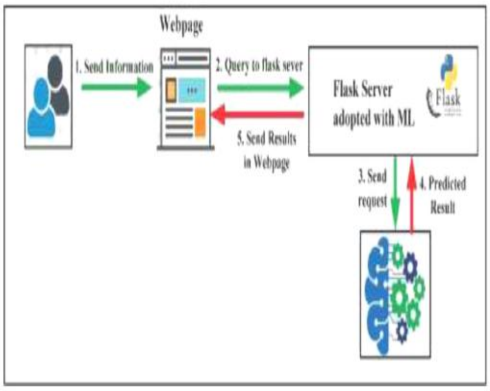


Fig. 6. Web Application work flow

V.CONCLUSIONS

As a preliminary step in our study, we implemented a number of machine learning algorithms and compared how well they could diagnose diabetes in individuals. Second, we have tested the plan extensively and analysed the results. After comparing SVM's performance to that of other algorithms, we determined that it was superior. Our findings are then used to inform the creation of a sophisticated web application for diabetes prediction. This web software allows anybody to upload their own health data and receive a prediction on whether or not they will develop diabetes. This programme might be helpful for those who are hesitant or who just wish to check in periodically. In this section, we compare our model to two current studies and show that, depending on the dataset and the ML approach employed, the proposed model can provide better accuracy ranging from 2.71 percentage points to 13.13 percentage points. Despite our extensive testing on two separate datasets, there is much more that can be learned from this work by applying new deep learning techniques. In the future, we want to deploy the web app on a cloud platform like as Heroku or AWS, making it openly available so that it can be assessed by actual users, and we will also study a bigger and deeper dataset for Bangladeshi patients with more features to increase accuracy.

REFERENCES

[1] Sujihelen, L., Boddu, R., Murugaveni, S., Arnika, M., Haldorai, A., Reddy, P.C.S., Feng, S. and Qin, J., 2022. Node Replication Attack Detection in Distributed Wireless Sensor Networks. Wireless Communications and Mobile Computing, 2022, pp.1-11.

[2] Liu, L., Shafiq, M., Sonawane, V.R., Murthy, M.Y.B., Reddy, P.C.S. and kumar Reddy, K.C., 2022. Spectrum trading and sharing in unmanned aerial vehicles based on distributed blockchain consortium system. Computers and Electrical Engineering, 103, p.108255.

[3] Singhal, A., Varshney, S., Mohanaprakash, T.A., Jayavadivel, R., Deepti, K., Reddy, P.C.S. and Mulat, M.B., 2022. Minimization of latency using multitask scheduling in industrial autonomous systems. Wireless Communications and Mobile Computing, 2022, pp.1-10.

[4] Sabitha, R., Shukla, A.P., Mehbodniya, A., Shakkeera, L. and REDDY, P.C.S., 2022. A Fuzzy Trust Evaluation of Cloud Collaboration Outlier Detection in Wireless Sensor Networks. Adhoc & Sensor Wireless Networks, 53.

[5] Shaker Reddy, P.C. and Sureshbabu, A., 2020. An Enhanced Multiple Linear Regression Model for Seasonal Rainfall Prediction. International Journal of Sensors Wireless Communications and Control, 10(4), pp.473-483.

[6] Kumar, K., Pande, S.V., Kumar, T., Saini, P., Chaturvedi, A., Reddy, P.C.S. and Shah, K.B., 2023. Intelligent controller design and fault prediction using machine learning model. International Transactions on Electrical Energy Systems, 2023.

[7] Madan, Parul, Vijay Singh, Vaibhav Chaudhari, Yasser Albagory, Ankur Dumka, Rajesh Singh, Anita Gehlot, Mamoon Rashid, Sultan S. Alshamrani, and Ahmed Saeed AlGhamdi. "An optimization-based diabetes prediction model using CNN and Bi-direction

[8] Ashok, K., Boddu, R., Syed, S.A., Sonawane, V.R., Dabhade, R.G. and Reddy, P.C.S., 2022. GAN Base feedback analysis system for industrial IOT networks. Automatika, pp.1-9.

[9] Shaker Reddy, P.C. and Sucharitha, Y., 2022. IoT-Enabled Energy-efficient Multipath Power Control for Underwater Sensor Networks. International Journal of Sensors Wireless Communications and Control, 12(6), pp.478-494.

[10] Dhanalakshmi, R., Bhavani, N.P.G., Raju, S.S., Shaker Reddy, P.C., Marvaluru, D., Singh, D.P. and Batu, A., 2022. Onboard Pointing Error Detection and Estimation of Observation Satellite Data Using Extended Kalman Filter. Computational Intelligence and Neuroscience, 2022.

[11] Reddy, P.C.S., Suryanarayana, G. and Yadala, S., 2022, November. Data Analytics in Farming: Rice price prediction in Andhra Pradesh. In 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT) (pp. 1-5). IEEE.

[12] Balamurugan, D., Aravinth, S.S., Reddy, P.C.S., Rupani, A. and Manikandan, A., 2022. Multiview objects recognition using deep learning-based wrap-CNN with voting scheme. Neural Processing Letters, 54(3), pp.1495-1521.

[13] Sucharitha, Y. and Shaker Reddy, P.C., 2022. An Autonomous Adaptive Enhancement Method Based on Learning to Optimize Heterogeneous Network Selection. International Journal of Sensors Wireless Communications and Control, 12(7), pp.495-509.

[14] Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A., Ghazal, T.M. and Ahmad, M., 2022. Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, pp.8529-8538.

[15] Reddy, P.C.S., Yadala, S. and Goddumarri, S.N., 2022. Development of rainfall forecasting model using machine learning with singular spectrum analysis. IIUM Engineering Journal, 23(1), pp.172-186.

[16] Suresh, S., Prabhu, V., Parthasarathy, V., Boddu, R., Sucharitha, Y. and Teshite, G., 2022. A novel routing protocol for low-energy wireless sensor networks. Journal of Sensors, 2022, pp.1-8.

[17] Dhanalakshmi, R., Bhavani, N.P.G., Raju, S.S., Shaker Reddy, P.C., Marvaluru, D., Singh, D.P. and Batu, A., 2022. Onboard Pointing Error Detection and Estimation of Observation Satellite Data Using Extended Kalman Filter. Computational Intelligence and Neuroscience, 2022.

[18] Muthappa, K.A., Nisha, A.S.A., Shastri, R., Avasthi, V. and Reddy, P.C.S., 2023. Design of high-speed, low-power non-volatile master slave flip flop (NVMSFF) for memory registers designs. Applied Nanoscience, pp.1-10.

[19] Chillakuru, P., Madiajagan, M., Prashanth, K.V., Ambala, S., Shaker Reddy, P.C. and Pavan, J., 2023. Enhancing wind power monitoring through motion deblurring with modified GoogleNet algorithm. Soft Computing, pp.1-11.

[20] Sucharitha, Y., Reddy, P.C.S. and Suryanarayana, G., 2023. Network Intrusion Detection of Drones Using Recurrent Neural Networks. Drone Technology: Future Trends and Practical Applications, pp.375-392.

[21] Shanmugaraja, P., Bhardwaj, M., Mehbodniya, A., VALI, S. and Reddy, P.C.S., 2023. An Efficient Clustered M-path Sinkhole Attack Detection (MSAD) Algorithm for Wireless Sensor Networks. Adhoc & Sensor Wireless Networks, 55.

[22] Shaker Reddy, P.C. and Sucharitha, Y., 2023. A Design and Challenges in Energy Optimizing CR-Wireless Sensor Networks. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 16(5), pp.82-92.

[23] Reddy, P.C., Nachiyappan, S., Ramakrishna, V., Senthil, R. and Sajid Anwer, M.D., 2021. Hybrid Model Using Scrum Methodology for Softwar Development System. J Nucl Ene Sci Power Generat Techno, 10(9), p.2.

[24] Ashreetha, B., Devi, M.R., Kumar, U.P., Mani, M.K., Sahu, D.N. and Reddy, P.C.S., 2022. Soft optimization techniques for automatic liver cancer detection in abdominal liver images. International journal of health sciences, 6.