International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 15s,2025
https://theaspd.com/index.php

Optimized Diabetes Prediction Using Soft Computing: CURE-ADASYN for Imbalance and Advanced Deep Learning Classification Methods

Mrs.V.Abinaya[1], **Dr.K. Chitra**[2]
[1]Research Scholar, Department of Computer Science, RVS College of Arts and Science (Autonomous), Sulur, Coimbatore, veeruabi93@gmail.com
[2]Assistant Professor, School of Computer Studies (PG), RVS College of Arts and Science, (Autonomous), Sulur, Coimbatore, chitra.k@rvsgroup.com

Abstract
Diabetes has become a major global health concern, leading to a number of catastrophic effects such as cardiovascular problems, kidney disease, and vision loss. Deep learning algorithms have shown potential in medical services for precise disease detection and treatment, which will ease the burden on medical personnel. Rapid advancements in diabetes forecasting have opened up new avenues for patient empowerment and early intervention. In order to do this, this research suggests a novel diabetes prediction model that uses an improved LSTM classifier, feature selection using Grey Wolf Optimization, and Particle Swarm Optimization. Using performance metrics including accuracy, precision, recall, and F1 score, our method is thoroughly assessed.
Keywords: Diabetes, Deep Learning, Prediction Model, Classifier

I.INTRODUCTION

Diabetes is a metabolic disease in which the secretion or action of insulin is compromised, resulting in high blood sugar levels. Diabetes damages the brain, kidneys, heart, and nerves, among other tissues. One of the main reasons of eyesight loss in diabetics is diabetes illness. For therapy and illness management to be effective, early diagnosis is essential. Diabetes is a long-term illness that directly impacts the pancreas, preventing the body from producing insulin. Insulin is the main component that controls blood glucose levels. Numerous factors, such as being overweight, not exercising, having high blood pressure, and having abnormal cholesterol levels, might contribute to diabetes. Urine production is one of the most common problems it may cause. Diabetes may cause damage to the eyes, nerves, and skin. It can also lead to diabetic retinopathy, an eye disorder, and kidney failure if treatment is not received. According to figures from the International Diabetes Federation (IDF), 537 million people globally had diabetes in 2021.The recommended study methodology provides a fresh approach to diabetes prediction by fusing robust classification techniques with enhanced feature selection. Using Lightweight Self-Adaptive Multi-Trajectory Particle Swarm Optimization (LMT-PSO) for feature selection and hybrid classification models is the main objective in order to enhance prediction performance. In order to achieve consistent feature scaling, the method begins with data pre-processing and imbalance handling, which includes advanced imputation techniques for missing values and normalization. Synthetic data augmentation techniques like SMOTE are used to reduce class imbalance and increase the reliability of subsequent classification stages. These phases offer a strong foundation for forecasting that is both fair and accurate.The following will be the arrangement of the remaining sections. To illustrate the nominal research conducted in this field, we shall provide an overview of pertinent studies in section II. In part III, we will go into depth about the methodology we used, including the data source, pre-processing procedures, and how the deep learning model was implemented. The confusion matrix for the test and validation data sets, together with the results of our experiment employing a number of performance measures, is shown in Section V. We then examine these results. The research presented here is summarized in Section VI.

II. Literature Survey

Ritika Bateja [1] proposed a diabetes prediction system that uses SVM in conjunction with collaborative filtering and particle swarm optimization to make medicine recommendations. They employed data cleansing, null value deletion, and feature selection as preprocessing techniques. Motivated by their work, we have improved our diabetes prediction model using the SVM method. J.

Abdullahi [2] employed feature selection with PSO to predict diabetes. Consequently, he also used a variety of machine learning approaches to study three different medical fields. Their results showed that Decision Tree, Random Forest, and Naïve Bayes were the best algorithms in terms of accuracy and error rate. In order to highlight the scientific merit of their work as well as the practicality of our findings in clinical practice, Talukder, M.A., et al. [3] carried out an extensive investigation on diabetes diagnosis utilizing ML approaches. Their contributions include resolving dataset imbalance, avoiding overfitting, showcasing higher performance through rigorous testing, and the result of an enhanced pre-processing pipeline. Pima Indian, Austin Public, Tigga, and Mendeley are the four datasets on which they have thoroughly tested a variety of machine learning models.The scientific community has been exploring the use of deep learning and machine learning techniques for diabetes prediction. Researchers have developed methods to enhance effectiveness using various data sources and algorithms. The Pima Indian diabetes dataset is used in a study by V. Chang, J. Bailey, and Q [4] to evaluate machine learning techniques for diabetes prediction, focusing on improving prediction accuracy using similar methodologies.Gangani Dharmarathne [5] proposed an easy-to-use interface for diabetes diagnosis utilizing machine learning models, including Decision Tree, KNN, SVC, and XGB. They explained the interpretability of the model using the Shapley Additive technique, and the XGB model produced some promising findings.

III. Methodology

Deep learning models for diabetes prediction use cutting-edge machine learning and deep learning techniques to improve the accuracy and timeliness of diabetes diagnosis. The following objectives must be fulfilled for the system to function at its best: it should process patient data to predict diabetes risk efficiently. The workflow diagram of the proposed methodology is presented in the Figure 1 as follows:

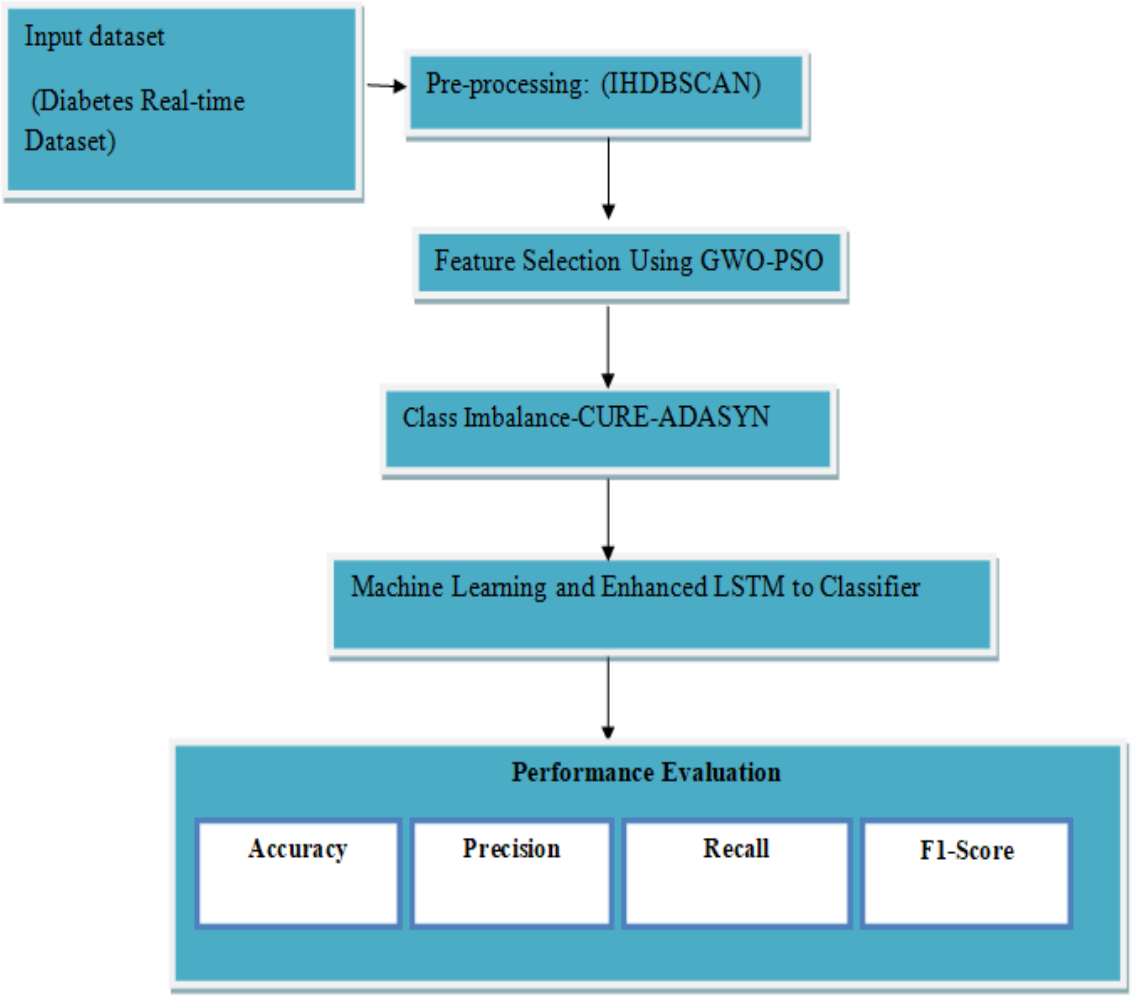

Figure 3.1 Workflow of Proposed Methodology

3.1 Data Collection

This study primarily uses two datasets: the Realtime Diabetes dataset, collected from a hospital, and the PIMA Diabetes dataset, sourced from the UCI Machine Learning Repository. These datasets were selected to ensure the robustness and generalizability of the projected model, providing a diverse range of demographic and health-related information.

3.2 Data Preprocessing

The preprocessing process involves using IHDBSCAN (Hierarchical DBSCAN), a robust clustering technique, to identify and isolate noise in data. This method is particularly effective in complex datasets where standard methods may struggle to distinguish between noise and valid data points. IHDBSCAN enhances the quality of the data, crucial for building accurate predictive models. Additionally, data reconstruction techniques are applied to address missing values, using advanced imputation strategies like mean/mode imputation, k-nearest neighbor imputation, or deep learning-based approaches. This combination ensures a high-quality dataset, ensuring the reliability of input data for model training and evaluation. This comprehensive preprocessing pipeline enhances machine learning model performance by minimizing noise, ensuring uniformity, and addressing data gaps.

3.3 Feature selection using GWO-PSO

Feature selection is a vital phase in machine learning, aiming to improve model performance by selecting relevant and informative features while reducing dimensionality. Metaheuristic algorithms like Grey Wolf Optimization (GWO) and Particle Swarm Optimization (PSO) have gained prominence due to their ability to explore and exploit the search space efficiently. GWO, inspired by the social hierarchy and hunting strategies of grey wolves, provides a structured exploration of the search space, identifying clusters of relevant features by balancing global exploration and local exploitation. PSO, on the other hand, is a velocity-based optimization algorithm inspired by bird flocks or fish schools, excelling in fine-tuning solutions by exploiting the most promising areas of the search space identified during the exploration phase.A hybrid GWO-PSO feature selection approach is created by combining the strengths of GWO and PSO, ensuring that the final feature subset is both compact (to reduce dimensionality) and predictive (to improve accuracy). This two-stage process addresses key challenges like over-fitting and computational inefficiency.When applied to a real-time diabetes dataset, the hybrid GWO-PSO technique proves particularly effective in identifying the most influential features associated with diabetes prediction. The approach involves an exploration phase where GWO evaluates all possible combinations of features and identifies clusters of relevant features, followed by an exploitation phase where PSO fine-tunes these feature subsets by iteratively improving their relevance, ensuring that redundant or non-informative features are excluded from the final selection.

The hybrid GWO-PSO technique offers a powerful solution for feature selection in machine learning, combining the structured exploration capabilities of GWO with the refinement abilities of PSO, ensuring that selected features are highly relevant and non-redundant.

3.4 Class Imbalance with CURE-ADASYN

CURE-ADASYN is a hybrid approach that combines the strengths of Clustering Using Representatives (CURE) and Adaptive Synthetic Sampling (ADASYN) to address class imbalance in machine learning. CURE identifies clusters within the minority class by grouping similar instances and reducing the influence of outliers, ensuring that synthetic samples generated by ADASYN are meaningful and aligned with the true distribution of the minority class. ADASYN is a widely used oversampling technique that generates synthetic samples to balance class distributions, focusing on creating synthetic instances near hard-to-learn instances. CURE-ADASYN enhances class balance and mitigates the risks of overfitting and loss of information associated with traditional oversampling techniques. Its advantages include improved minority class representation, robustness to outliers, preservation of data patterns, and enhanced model performance. When applied to imbalanced datasets like healthcare, fraud detection, or anomaly detection, CURE-ADASYN significantly improves predictive performance.

3.5 Improved LSTM Classification

The Long Short-Term Memory (LSTM) network is a novel approach to sequential data analysis that addresses the challenges of such tasks. It incorporates advanced techniques like attention mechanisms, feature selection, hyperparameter tuning, and regularization methods to improve the model's robustness and capability to handle multifaceted datasets. The improved LSTM architecture overcomes limitations

like vanishing gradients during training, making it suitable for classification tasks, particularly when dealing with sequences with noise, missing values, or irregular time steps. Key benefits of the improved LSTM include better handling of long-term dependencies, increased accuracy and efficiency, robustness to noise and variability, and generalization to unseen data. These improvements have been applied in various real-world applications, including healthcare, financial market prediction, cybersecurity, and video and image classification. In healthcare, improved LSTMs can classify patient records, speech recognition, sentiment analysis, and language translation. In financial market prediction, models must understand long-term patterns and temporal relationships. In video and image classification, they can process temporal sequences of frames to classify activities or events over time.

Steps in Improved LSTM-Based Classification

1. **Input Layer**: The layer receives pre-processed and normalized data as input, starting the neural network and ensuring sequential data is ready for subsequent layers' analysis.
2. **Optimized LSTM Layer**: The LSTM layer, a key component of the prototype, is optimized using Particle Swarm Optimization to handle sequential dependencies and identify critical temporal patterns for diabetes prediction.
3. **Dropout Layer**: This layer randomly sets a fraction of input units to zero during training to prevent overfitting and ensure model generalization to unseen data.
4. **Batch Normalization Layer**: The LSTM layer normalizes the output, stabilizing the training process, enhancing convergence speed, and improving model performance.
5. **Fully Connected Layer**: The dense layer aggregates features from LSTM and dropout layers, connecting every neuron in the previous layer to the next, facilitating comprehensive feature learning.
6. **SoftMax Layer**: The layer simplifies classification by converting logits into probabilities, generating a probability distribution over the output classes.
7. **Output Layer**: The final layer uses learned patterns and features to classify data and predict a patient's diabetes status based on the processed input data.

3.6 Performance Evaluation

The Improved Long Short-Term Memory (LSTM) network outperforms traditional LSTM networks and other machine learning algorithms in the diabetes dataset, which contains dynamic and time-series data. The improved LSTM is designed to handle long-term dependencies in sequential data, letting it to make more precise predictions. It integrates feature selection techniques to focus on relevant variables, reducing noise from irrelevant features and improving model interpretability. Techniques like dropout and early stopping prevent overfitting, making the Improved LSTM a more robust solution for diabetes prediction. The evaluation of various algorithms on the diabetes dataset demonstrates the advantages of the Improved LSTM:

1. Accuracy: The Improved LSTM consistently outperforms all other models in terms of accuracy.
2. Precision: The Improved LSTM excels in precision, measuring the proportion of true positive predictions out of all positive predictions.
3. F1-score: The Improved LSTM performs strongly in this area, offering an optimal balance between precision and recall, ensuring accurate and reliable predictions.

In conclusion, the Improved LSTM is a superior choice for diabetes prediction, offering enhanced capabilities over traditional LSTM networks and other machine learning algorithms.

IV. RESULTS AND DISCUSSION

Accuracy

It checks the ratio of properly classified occurrences among all the occurrences in the dataset. The formula for classification accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Prediction}}$$

In mathematical terms:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision

It shows how accurately the model is able to find positive cases. In general, it shows us the ratio of the model's correct predictions among all the predicted positive classes.

The formula for precision is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

The recall, also called sensitivity, is about the model's capability to correctly detect positive patterns. It determines the percentage of true positive instances correctly predicted by the model to all instances, which were in fact positive ones.

The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 Score

F1 score is being the harmonic mean of precision and recall. The purpose of F1 Score is to achieve a balance between recall and precision hence it serves as a worthy performance indicator, particularly for cases where the classes are not equally represented. The range for F1 Score is [0, 1].

The mathematical formula of F1 score is:

$$\text{F1 Score} = \frac{2 \text{ X Precision X Recall}}{\text{Precision} + \text{Recall}}$$

TABLE I Accuracy, Precision, Recall and F1 Score of Various Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Improved LSTM	99.7	99	99	99	99.8
LSTM	97	96	95	96	97.5
Neural Network	92	91	90	90	93
Support Vector Machine	88	87	86	86	89

The table compares various classification algorithms for diabetes prediction, including Improved Long Short-Term Memory (LSTM), LSTM, Neural Network (NN), and Support Vector Machine (SVM). Key assessment metrics include accuracy, precision, recall, F1-Score, and AUC. The Improved LSTM model outperforms traditional LSTM and other machine learning algorithms in capturing long-term dependencies and handling time-series data, demonstrating its effectiveness in predicting diabetes instances and handling dynamic health data.

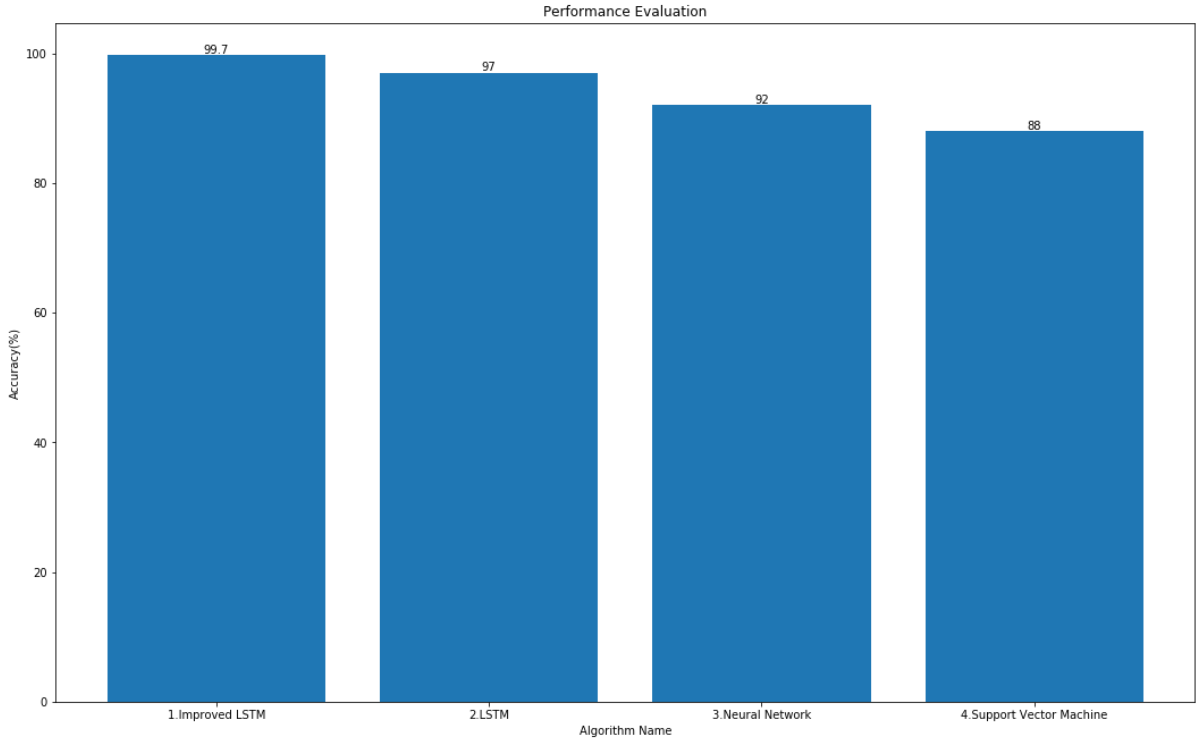

Figure 2. Accuracy of Classification Models

The figure 2 shows four classification models - Improved LSTM, LSTM, Neural Network, and Support Vector Machine (SVM) - evaluated for accuracy in a diabetes prediction task. The Improved LSTM model achieved the highest accuracy at 99.7%, outperforming the standard LSTM, Neural Network, and SVM, indicating its superior predictive capability.

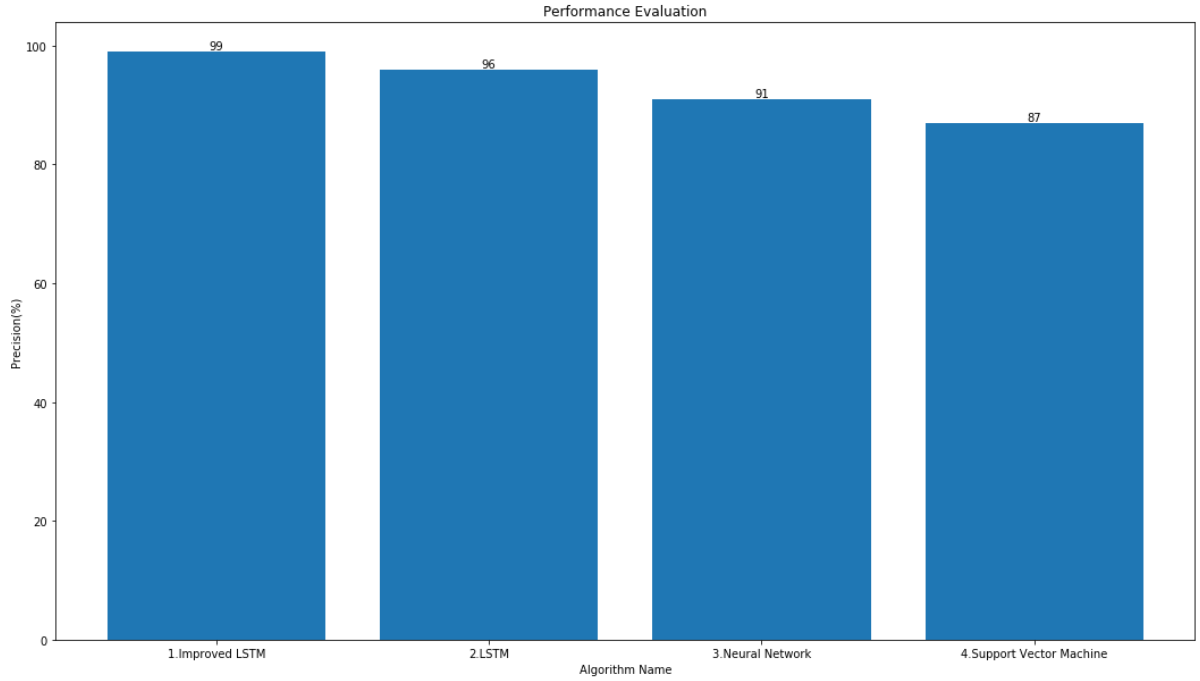

Figure 3. Precision Evaluation of Classification Models

The figure 3 showcases the precision evaluation of four classification models and the Improved LSTM model achieves the highest precision at 99%, followed by the LSTM at 96%, the Neural Network at 91%, and the SVM at 87%, making it the most effective approach.

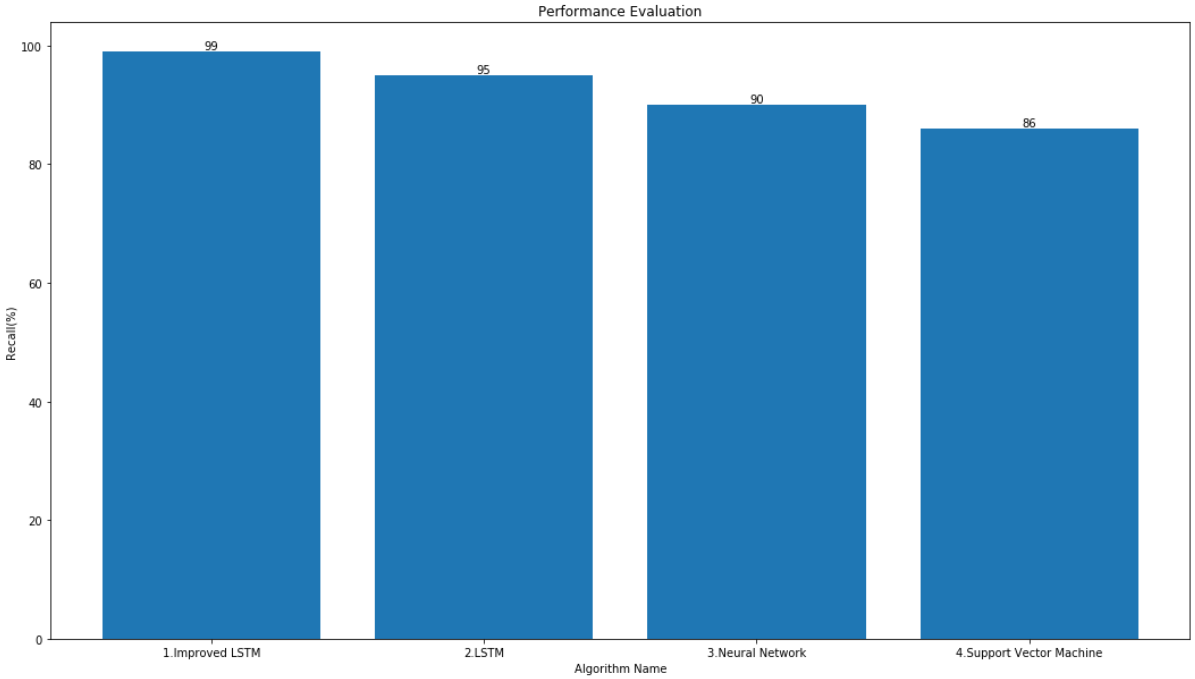

Figure 4. Recall Evaluation of Classification Models

The figure 4 shows the recall evaluation of four classification models in a graphical representation. The Improved LSTM model achieves the highest recall at 99%, followed by the LSTM at 95%, the Neural Network at 90%, and the SVM at 86%, showcasing its superior performance.

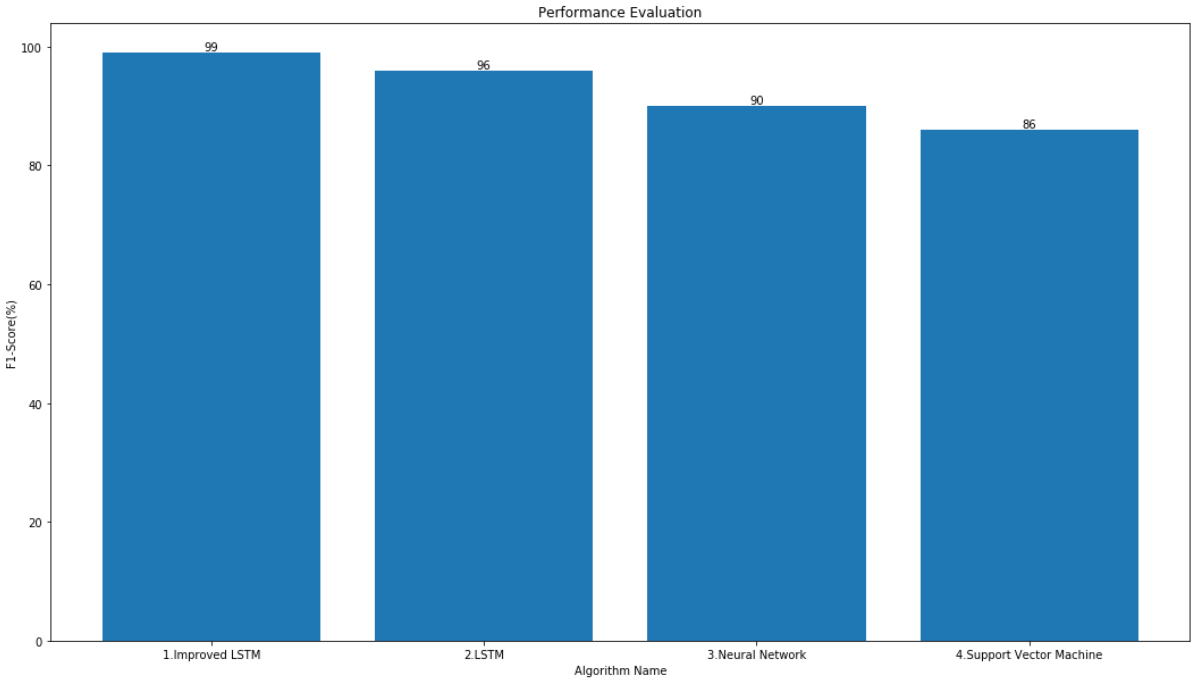

Figure 5. F1 Score Evaluation of Classification Models

The figure 5 shows the F1 score evaluation of the four classification models, with the Improved LSTM model achieving the highest F1-score at 99, demonstrating its ability to accurately and consistently classify instances while minimizing errors. The LSTM model had a slightly lower F1-score of 96, while the Neural Network had a moderate classification capability of 90. The Support Vector Machine (SVM) had the lowest F1-score of 86, indicating the lowest performance among the evaluated models. The Improved LSTM model demonstrated superiority in balancing precision and recall for high classification accuracy.

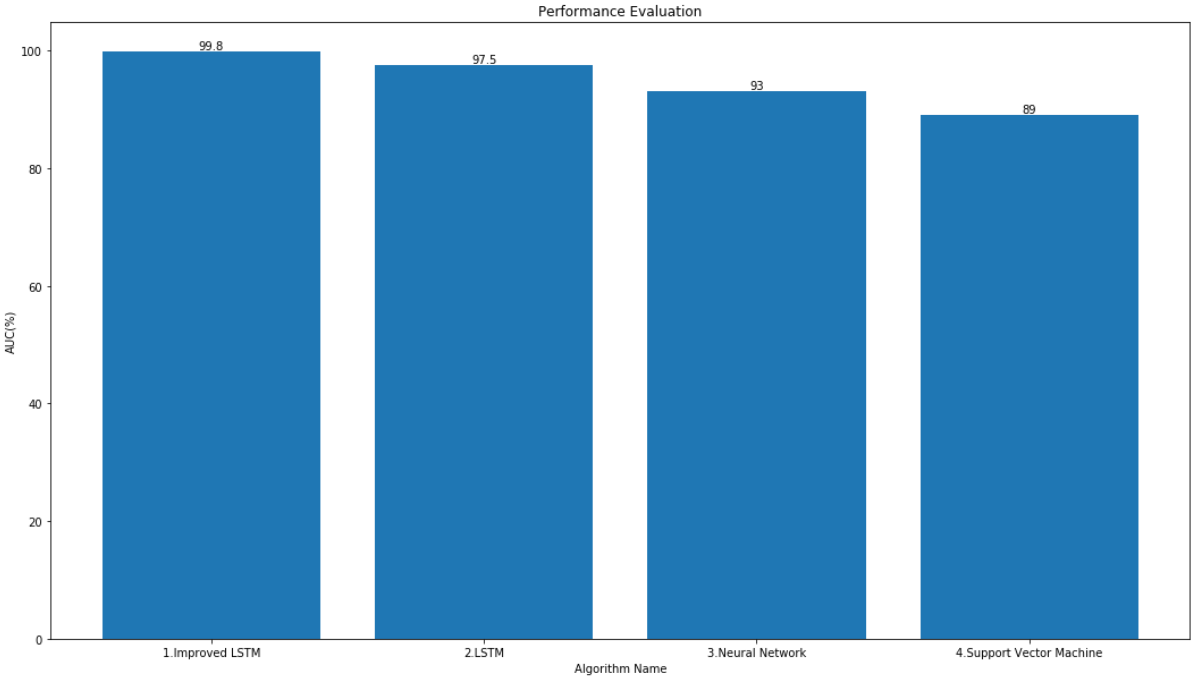

Figure 6. AUC Evaluation of Classification Models

A higher AUC indicates better discriminatory power, allowing for the separation of positive and negative classes. The figure 6 shows several AUC scores for different classification models: Improved LSTM (99.8), LSTM (97.5), Neural Network (93), and Support Vector Machine (89). The improved LSTM has the highest AUC, indicating near-perfect classification. The standard LSTM has a slightly lower AUC but still reflects excellent discrimination ability. The neural network shows good classification performance but a noticeable drop in class separation compared to the LSTM models. The Support Vector Machine has the lowest AUC, suggesting less effectiveness in class separation. AUC is crucial when dataset imbalances occur, as it evaluates the trade-off between true positive rates and false positive rates.

V. CONCLUSION

The operate objective was to investigate a variety of deep learning and machine learning methods that may be used to the diagnosis of diabetes by using the two datasets independently. With an accuracy of 90 to 99.8% across two datasets, it was seen that the suggested model outperformed the individual machine learning model in terms of accuracy. Our results suggest that using an ensemble approach to combine deep learning models may improve prediction accuracy in this situation. The next step for us is to further optimize the feature extraction process using autonomous deep learning approaches in order to improve model fitting and prediction accuracy. Beyond diabetes, these techniques effectively manage vast amounts of medical data and enhance healthcare outcomes.

REFERENCES

[1]. Ritika Bateja, Sanjay Kumar Dubey, and Ashutosh Kumar Bhatt, "Diabetes Prediction and Recommendation Model Using Machine Learning Techniques and MapReduce," Indian Journal of Science and Technology, vol. 17, no. 26, pp. 2747–2753, Jul. 2024, doi: https://doi.org/10.17485/ijst/v17i26.530.

[2]. J. Abdollahi and S. Aref, "Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms," SN Computer Science, vol. 5, no. 2, Jan. 2024, doi: https://doi.org/10.1007/s42979-023- 02545-y.

[3]. Talukder, M.A., Islam, M.M., Uddin, M.A., Kazi, M., Khalid, M., Akhter, A. and Ali Moni, M., 2024. Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. Digital Health, 10, p.20552076241271867.

[4]. V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," Neural Computing and Applications, Mar. 2022, doi: https://doi.org/10.1007/s00521-022-07049-z.

[5]. Gangani Dharmarathne, T. N. Jayasinghe, Madhusha Bogahawaththa, D.P.P. Meddage, and Upaka Rathnayake, "A novel machine learning approach for diagnosing diabetes with a self-explainable interface," Healthcare analytics, vol. 5, pp. 100301–100301, Jun. 2024, doi: https://doi.org/10.1016/j.health.2024.100301.

[6]. Chowdhury, P., Barua, P. and Uddin, M.N., 2024, September. Diabetes Prediction Using Machine Learning and Hybrid Deep Learning Ensemble Technique. In 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS) (pp. 1-7). IEEE.

[7]. Alam, M.A., Sohel, A., Hasan, K.M. and Islam, M.A., 2024. Machine Learning And Artificial Intelligence in Diabetes Prediction And Management: A Comprehensive Review of Models. Journal of Next-Gen Engineering Systems.

[8]. Eletter, S.F., Elrefae, A. and Aliter, H., 2024, September. Predicting Diabetes Status Using Deep Learning. In 2024 Global Digital Health Knowledge Exchange & Empowerment Conference (gDigiHealth. KEE) (pp. 1-4). IEEE.

[9]. Gaso, M.S., Mekuria, R.R., Khan, A., Gulbarga, M.I., Tologonov, I. and Sadriddin, Z., 2024, June. Utilizing Machine and Deep Learning Techniques for Predicting Re-admission Cases in Diabetes Patients. In Proceedings of the International Conference on Computer Systems and Technologies 2024 (pp. 76-81).

[10]. Naz, U., Khalil, A., Khattak, A., Raza, M.A., Asghar, J. and Asghar, M.Z., 2024, August. Deep Learning for Enhancing Diabetes Prediction. In 2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA) (pp. 1-7). IEEE.

[11]. Jeevaraja, J.D., Kavitha, P. and Kamalakkannan, S., 2024. Diabetes Prediction using Machine Learning Algorithms. Diabetes, 4(6).