scientific reports

Check for updates

OPEN

Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model

Furqan Rustam[1,11], Ahmad Sami Al-Shamayleh[2,11], Rahman Shafique[3,11], Silvia Aparicio Obregon[4,5,6], Ruben Calderon Iglesias[4,7,8], J. Pablo Miramontes Gonzalez[9,10] & Imran Ashraf[3]✉

Diabetes is a persistent health condition led by insufficient use or inappropriate use of insulin in the body. If left undetected, it can lead to further complications involving organ damage such as heart, lungs, and eyes. Timely detection of diabetes helps obtain the right medication, diet, and exercise plan to lead a healthy life. ML approach has been utilized to obtain rapid and reliable diabetes detection, however, existing approaches suffer from the use of limited datasets, lack of generalizability, and lower accuracy. This study proposes a novel feature extraction approach to overcome these limitations by using an ensemble of convolutional neural network (CNN) and long short-term memory (LSTM) models. Multiple datasets are combined to make a larger dataset for experiments and multiple features are utilized for investigating the efficacy of the proposed approach. Features from the extra tree classifier, CNN, and LSTM are also considered for comparison. Experimental results reveal the superb performance of CNN-LSTM-based features with random forest model obtaining a 0.99 accuracy score. This performance is further validated by comparison with existing approaches and k-fold cross-validation which shows the proposed approach provides robust results.

Keywords ZeroShot learning, Transfer learning, Spider mites detection, Plants health, Zeroshot CNN

Diabetes is a persistent medical condition attributed to insufficient production of insulin by the pancreas or impaired/improper utilization of the insulin by body cells[1]. Insulin plays a crucial role in regulating blood glucose levels[2]. When diabetes is not managed properly, it often results in hyperglycemia, also known as elevated blood glucose or high blood sugar levels[3]. Prolonged high sugar levels can lead to significant damage to various biological systems, including nerves and blood vessels. Diabetes can be caused by either inadequate insulin production or the body's ineffective utilization of insulin, which is necessary for using glucose as fuel[4]. The most common kind is type 2 diabetes, which is commonly found in adults. Symptoms include frequent urination, increased thirst, and hunger. If not treated, diabetes can lead to a variety of complications in the human body[5] like cardiovascular disease (CVD), foot ulcers, nerve damage, etc. Diabetes-led complications may pose a real threat to human life and lead to even death.

In 2014, a study revealed that 8.5% of adults aged 18 and above were diagnosed with diabetes, with nearly half of diabetic fatalities occurring before the age of 70. Diabetes was accountable for 460,000 deaths related to renal disease and contributed to 20% of CVD mortality due to elevated blood glucose levels[6]. By the year 2020, the global diabetic population reached approximately 2.85 billion individuals, representing 6.4% of the world's populace. Without significant advancements in preventive measures and therapeutic interventions,

[1]School of Systems and Technology, Department of Software Engineering, University of Management and Technology, Lahore 54770, Pakistan. [2]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al Ahliyya Amman University, Amman 19328, Jordan. [3]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea. [4]Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain. [5]Universidad Internacional Iberoamericana, 24560 Campeche, Mexico. [6]Universidad Internacional Iberoamericana, Arecibo, Puerto Rico 00613, USA. [7]Universidade Internacional do Cuanza, Cuito, Bie, Angola. [8]Universidad de La Romana, La Romana, República Dominicana. [9]Departamento de Medicina, Facultad de Medicina, Universidad de Valladolid, Valladolid, Spain. [10]Servicio de Medicina Interna, Hospital Universitario Río Hortega, Valladolid, Spain. [11]These authors contributed equally: Furqan Rustam, Ahmad Sami Al-Shamayleh, Rahman Shafique. ✉email: imranashraf@ynu.ac.kr

experts project this figure to escalate to 430 million. This concerning trend is attributed to factors such as the adoption of Western dietary habits and urban lifestyles in emerging economies, alongside inadequate public awareness regarding diabetes[7,8]. In India, the prevalence of diabetes has surged, with the current number of affected individuals increasing from 70 million in 2019 to 101 million. This increase in diabetic patients poses a real threat to public health and governments need additional budget to handle it[9]. Besides type 1 and type 2, diabetes encompasses gestational, monogenic, and cystic fibrosis-related diabetes[10]. The accumulation of thick mucus can cause pancreatic complications thereby leading to cystic fibrosis diabetes which adversely affects insulin production[11]. Monogenic is another form that arises from a single gene mutation that disrupts insulin synthesis or action[12].

Advanced computational techniques, including data mining, machine learning (ML), statistics, and database systems, are instrumental in identifying individuals at heightened risk of developing Diabetes Mellitus. These methodologies are tailored to achieve crucial objectives such as pattern detection and clustering. Particularly, data mining endeavors to uncover concealed insights from vast databases through automated procedures. The efficacy of analysis hinges upon the availability of high-quality raw data and processing approaches. Integration of data mining within healthcare elevates the quality of care, diminishes costs, and facilitates the prediction of diabetes-related issues[13]. A plethora of computer algorithms have been devised for classifying diabetes, leading to improved diagnostic accuracy, cost efficiency, and more efficacious treatment options[14].

A large database is essential to automate the diagnosis and assess the severity of diabetes. This database encompasses data on the impact of diabetes on various organs within the human body. Early detection of diabetes enables timely treatment and the development of a suitable dietary plan to prevent later complications such as cardiovascular and kidney issues. Previous diabetes detection systems have utilized ML and DL techniques, but they often lack high detection accuracy and generalization. In addition, due to the versatility of the datasets, the adopted models should be generalizable offering similar detection accuracy on the unseen data. Existing approaches to diabetes detection lack robustness and generalizability thereby requiring further efforts to improve diabetes detection.

Automating the diagnosis of diabetes detection and assessing the severity of diabetes using ML and DL require large datasets. This automated system can enable early detection of diabetes with lower costs and higher accuracy. Early diagnosis facilitates timely treatment and the development of suitable dietary plans to prevent complications such as cardiovascular and kidney issues. Previous diabetes detection systems that utilized ML and DL techniques often lacked high detection accuracy and generalization. Motivated by these challenges, this study analyzes how significant results in diabetes prediction can be achieved by using feature engineering and an ensemble of different datasets with state-of-the-art (SOTA) ML algorithms. To overcome these limitations and find suitable models for accurate diabetes detection, this work adopts a DL approach. This study contributes as follows

- A novel DL approach is introduced to extract appropriate feature sets from the dataset. The feature extraction approach is based on an ensemble of convolutional neural networks and long short-term memory (LSTM). The models are joined in a stacked manner to obtain the final feature set.
- ML models like decision tree (DT), logistic regression (LR), support vector classifier (SVC), and random forest (RF) are used for experiments. In comparison to the proposed feature extraction approach, models are evaluated using features obtained from an extra tree classifier (ETC), long short-term memory (LSTM), convolutional neural network (CNN), and an ensemble of CNN and LSTM (CNN-LSTM).
- Throughout the experiments, we utilized three distinct datasets to assess the effectiveness of the suggested feature extraction method. These include Aravindpcoders' diabetes dataset, Mathchi's diabetes dataset, and Ishandutta's early-stage diabetes risk prediction dataset. Additionally, we conducted a performance comparison with existing models.Section "Literature review" delves into the literature review, while section "Materials and methods" presents an outline of the proposed strategy and pertinent details. Subsequently, section "Results and discussion" delineates the results, and section "Conclusions and future work" furnishes the conclusion.

Literature review

In the literature conducted by[15], a DL architecture was developed to differentiate between diabetic and normal heart rate variability (HRV) signals. The study provides an extensive overview of diabetes, HRV, and related research in the domain of automated non-invasive diabetes detection. Employing a combination of LSTM, CNN, and their integrated architectures, the objective is to extract intricate temporal and dynamic features from the input Heart Rate Variability (HRV) data. The proposed CNN-LSTM with SVM classification system demonstrates a remarkable accuracy of 95.7% in diagnosing diabetes using electrocardiography (ECG) signals. In another study[16], meticulous scrutiny was directed toward the utilization of ML and deep learning (DL) methodologies for predicting diabetes. The author utilized conventional ML techniques, notably SVM and RF, against DL approaches utilizing a CNN model. The evaluation was grounded on the publicly accessible Pima Indians Diabetes database, comprising 768 samples. The experimental findings revealed that RF surpassed DL and ML models with an accuracy of 83.67%, compared to 76.81% and 65.38% accuracy of CNN and SVM, respectively.

Similarly[17], the researchers embarked upon a comprehensive exploration of diverse machine-learning methodologies aimed at the diagnosis of diabetes. Their primary objective was to leverage sophisticated ML algorithms to proficiently diagnose patients afflicted with diabetes, utilizing an extensive array of medical prognostic data. The study harnessed a dataset replete with intricate medical predictor variables, including gravidity, Body Mass Index (BMI), insulin concentrations, and chronological age, among other factors. A repertoire of six classifiers, namely LR, DT, SVM, extreme gradient boosting (XGBoost), RF, and Adaboost, were meticulously implemented. Their performances were rigorously scrutinized employing a multitude of

evaluation metrics such as accuracy, F1-score, recall, precision, and area under the curve (AUC). The empirical findings unveiled that Adaboost emerged as the preeminent classifier, boasting the highest accuracy rate of 83% amongst its counterparts. The study conducted by[18] significantly contributes to the realm of health informatics by demonstrating the efficacy of ML algorithms in accurately diagnosing diabetes. The authors adeptly optimize the performance of classification models through the implementation of feature selection techniques, notably principle components analysis (PCA). They underscore the pivotal importance of feature selection in the development of interpretable models and the augmentation of data mining efficacy. The comparative analysis encompasses a diverse array of ML models, including Bayes Net, DT, Gradient Boosting Machine (GBM), J48, KNN, JRIP, LDA, and LR, alongside DL models such as ANN, CNN, MLP, and Deep Neural Network (DNN), among others. Remarkably DNN achieves an exceptional accuracy rate of 98.1%, underscoring its efficacy in diabetes diagnosis.

The study[19] aimed to develop an automated method for diagnosing brain tumors from MRI scans using a deep-learning CNN model. Tests were performed on the Br35H dataset, which contains a large number of MRI images. Enhanced CNN models were created by implementing different activation functions, hyperparameters, and data augmentation techniques such as rotation, flipping, and rescaling. The Adam optimizer was used to improve learning speed, and the dropout technique was applied to prevent overfitting. The proposed model (CNN) outperformed existing models, achieving high scores: 99.18% recall, 99.45% precision, 99.31% F1 score, and 99.28% accuracy.

The authors[20] delve into the prospective applications of intelligent systems in advancing medical technologies for the detection of diabetes mellitus. The discourse elucidates the employment of oversampling methodologies alongside data dimensionality reduction via feature selection strategies to augment the precision and dependability of diabetes identification processes. Furthermore, it underscores the significance of formulating models capable of scrutinizing attributes gleaned from datasets, advocating for the fusion of ML and DL algorithms as a promising avenue. Leveraging an ANN, the researchers attained a commendable accuracy rate of 99.13%. The study[21] presents a research endeavor concerning the utilization of ML algorithms to prognosticate the probability of diabetes onset. The inquiry heralds ML as a potent instrument for scrutinizing data replete with latent patterns and concentrates on three distinct algorithms: SVM, NB, and RF. Empirical findings delineate that the RF algorithm attains a predictive accuracy rate of 88.14% in discerning diabetes occurrences. Additionally, the author furnishes a succinct overview encompassing various diabetes classifications, comprising Type 1, Type 2, and Type 3 (gestational diabetes). Furthermore, statistical insights regarding the global prevalence of diabetes are presented, underscoring the imperative for precise prognostication and diagnosis methodologies.

The study[22] demonstrates a notable strength in their pragmatic approach towards forecasting diabetes onset. Through the utilization of ML methodologies, the authors proffer a promising avenue to aid healthcare practitioners in early diagnostic endeavors. The lucid exposition of dataset attributes along with their corresponding statistical metrics lends credence to the research endeavor. Delving into the application of ML techniques, particularly the KNN algorithm, for diabetes prognosis, the authors expound upon the rationale behind its selection, underscoring its congruity with the dataset characteristics and its efficacy in furnishing precise prognostications. The reported accuracy rate stands at a commendable 89.5%. The study by[23] delves into the utilization of ML algorithms for prognosticating diabetes occurrences, drawing upon datasets sourced from Bangladesh, India, and Germany. It tackles the challenges intrinsic to precise diabetes prediction, including the scarcity of labeled data and the presence of outliers within the datasets. Various ML models are scrutinized, encompassing boosting techniques such as AdaBoost, CB, Gradient Boosting, and XGBoost, alongside foundational models like RF and DT. Empirical observations reveal that the Bangladesh dataset demonstrates enhanced performance with boosting algorithms, notably with CB achieving a noteworthy accuracy score of 0.99 in forecasting the prevalence of diabetes.

Similarly,[24] contribute significant insights into the utilization of ML algorithms for diabetes prognosis. Commencing with a discourse on the two primary types of diabetes and their etiological factors, the study delineates the deleterious impacts of diabetes on diverse bodily organs, underscoring the imperative nature of early prediction and intervention. The authors proceed to implement and juxtapose two widely employed ML models, namely RF and LR, for diabetes prognostication. They elucidate the experimental outcomes, revealing an impressive accuracy rate of 99% attained by the RF algorithm. This underscores the inherent potential of ML algorithms in precisely forecasting diabetes occurrences and facilitating timely intervention. measures. In a similar vein the author[25] underscores the criticality of early detection in diabetes, recognizing its often asymptomatic nature, which predisposes it to underdiagnosis. Leveraging a dataset inclusive of both newly diagnosed diabetic individuals and those deemed at risk, the study employs five distinct ML methodologies to forecast the propensity for diabetes development. Among these methodologies, the RF approach emerges as the most efficacious, yielding an overall accuracy rate of approximately 99%. Furthermore, an interpretable machine-learning technique is deployed to elucidate the correlation between the response variable and the predictors.

The study[26] introduces a transfer learning-based method for detecting Monkeypox in images, with InceptionV3 achieving up to 98% accuracy, demonstrating its potential as a standard in medical imaging. The transfer learning method not only saves resources but also allows for easy updates as new data becomes available. Similarly, the authors[27] fine-tuned the hyperparameters of VGG-19, Inception-v3, and XceptionNet architectures on the CK+, JAFFE, and FER2013 datasets, to enhance the performance of image sentiment analysis systems. Transfer learning proved especially effective with smaller datasets, suggesting the potential for future advancements in handling larger datasets and automating hyperparameter tuning in sentiment analysis. The study encapsulates the findings and constraints of preceding research endeavors in Table 1.

Ref.	Year	Classifier	Accuracy	Limitation
[15]	2018	CNN-LSTM with SVM	95.7%	Small dataset, lack of external validation, limited interpretability of DL algorithms, and practical challenges in clinical implementation.
[16]	2019	SVM,RF,CNN	83.67% by RF	Small dataset
[17]	2021	SVM, Xgboost, RF, LR, ADA, DT	83% by ADA	Dataset biases, limited feature selection, constrained model selection, evaluation metric reliance.
[18]	2023	GBM, J48, KNN, JRIP, LDA, LM, LR, GBM, MDR, RF	98.1% by DNN	Small dataset limitations, no validation
[20]	2023	ANN	99.13%	Small dataset
[21]	2022	SVM, NB, RF	88.14% by RF	Limited evaluation of other illnesses, no discussion on model execution time.
[22]	2023	KNN	89.5%	Limited scope of models, lack of information on dataset and generalizability, limited evaluation of other illnesses
[23]	2023	XGBoost, CatBoost	99% by CatBoost	Limited population representation, limited labeled data availability, potential for improvement with larger datasets, scope for integrating additional features, exploration of hybrid models
[24]	2022	RF, LR	99% by RF	Potential overfitting, lack of comprehensive model comparison, reliance on accuracy as the sole evaluation metric, limited scope to diabetes prediction, and lack of interpretability.
[25]	2023	RF	99%	Limited dataset diversity, potential overfitting, limited model comparison

Table 1. Overview of recent studies on diabetes prediction. Adopted approaches in the studies along with their results are stated in addition to the limitations.

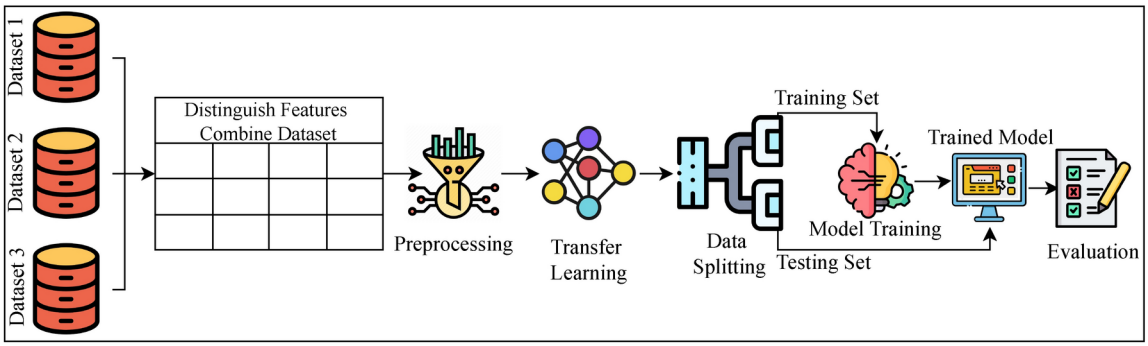

Fig. 1. The workflow of the proposed approach indicates the process of feature engineering, preprocessing, and adoption of transfer learning for diabetes detection.

Limitations and gaps in the existing research

The recent studies on diabetes prediction using ML and DL have limitations such as small datasets, lack of interpretability in DL models, and practical challenges in clinical implementation. Additional issues include dataset biases, limited model selection, and reliance on specific metrics. These studies also face challenges like narrow scope in model execution time discussion, limited model types, and potential for overfitting. Furthermore, there are limitations in population representation, labeled data availability, and model interpretability. Our study emphasizes the importance of creating larger datasets, enhancing model interpretability, and conducting broader model evaluations.

Challenges in diabetes detection

A rich variety of technologies are used for diabetes detection ranging from blood tests, and continuous glucose monitoring (CGM), to smart insulin pens. The challenges of diabetes detection across various modalities include issues such as limited accuracy and reliability in traditional clinical tests, and expertise in imaging methods. There are also concerns over sensor accuracy and user adoption in wearable technologies like CGM. Additionally, data-driven approaches, while promising, face obstacles like limited datasets, poor generalizability, and the difficulty of effective feature extraction and multimodal data integration. The current study focuses on these challenges by employing a novel ensemble of CNN and LSTM models, leveraging a comprehensive dataset, and integrating multiple feature extraction techniques to enhance accuracy and generalizability, demonstrating superior performance validated through cross-validation and comparison with existing methods.

Materials and methods

This study performs experiments for diabetes prediction, where data is extracted, optimal features are selected, and an ML approach is employed for diabetes detection. The architecture of the proposed approach is illustrated in Fig. 1.

Initially, we downloaded three publicly available datasets and merged them to form a combined dataset. Upon this combined dataset, we applied state-of-the-art ML algorithms. The classification task involved two classes: 1 for Diabetic and 0 for Non-Diabetic. Among the ML algorithms utilized are DT, RF, LR, and SVC. Notably,

Dataset	Description	Classes	Attributes
[29]	Medical data related to diabetes, including glucose levels and more.	2 classes (diabetic, non-diabetic)	9
[30]	Medical data with features such as glucose, blood pressure, and more.	2 classes (diabetic, non-diabetic)	9
[31]	Dataset for predicting early-stage diabetes risk.	3 classes (positive, negative, uncertain)	17

Table 2. Summary of the datasets used for experiments. It includes a description of the datasets, number of classes, and attributes in each dataset.

Urea	HbA1c	Chol	HDL	LDL	VLDL	target
59	5.2	10.9	2.1	1.1	2.5	1
55	6.1	8.5	2.1	0.9	3.8	1
57	4.6	6.8	6.0	2.5	3.5	1
45	3.248749	5.073910	1.021650	1.147828	2.618088	0
39	3.2	5.0	1.3	1.0	3.0	0
31	3.6	3.7	2.1	1.0	2.4	0

Table 3. Some of the medical metrics and their corresponding diabetes status for non-diabetic individuals.

tree-based algorithms demonstrated superior performance. In the medical domain, even slight improvements in accuracy can yield significant outcomes[28]. Hence, to enhance accuracy, we employed feature engineering. This involved the selection of the most critical features to train models, aiming to optimize accuracy. For feature selection, we utilized the ETC with hyperparameter settings: n_estimators = 300 and max_depth = 20. Through this process, we identified the top 12 features that had the most impact on the model's performance. Subsequently, we re-applied ML models using these refined features, resulting in noticeable improvements in accuracy across all classifiers. We partitioned the dataset into training and testing sets using an 80:20 ratio, allocating 80% of the data for training purposes and reserving the remaining 20% for testing.

Datasets

Three types of datasets are employed in this research. All datasets are publicly accessible through Kaggle.

- Aravindpcoder's Diabetes Dataset.
- Mathchi's Diabetes Data Set.
- Ishandutta's Early Stage Diabetes Risk Prediction Dataset.

Aravindpcoder's diabetes dataset

This dataset contains medical data related to diabetes, including glucose levels, blood pressure, and other health metrics. This dataset has 2 classes (diabetic, and non-diabetic) and includes 9 attributes[29]. Columns like 'AGE', 'Urea', 'HbA1c', 'Chol', 'TG', 'HDL', 'LDL', 'VLDL', 'BMI', and 'CLASS' has been utilized from this dataset.

Mathchi's diabetes dataset

Mathchi's Diabetes dataset is a collection of medical data focusing on features such as glucose, blood pressure, and skin thickness, among others. Like the previous dataset, this one also has 2 classes (diabetic, and non-diabetic) and consists of 9 attributes[30]. Attributes like 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', and 'Outcome' has been utilized from this dataset.

Ishandutta's early stage diabetes risk prediction dataset

Ishandutta's Early Stage Diabetes Risk Prediction dataset is designed to predict the risk of developing early-stage diabetes. It includes 3 classes (positive, negative, and uncertain) and contains 17 attributes[31]. We selected identical attributes and extracted records from this dataset that contain only two classes, ensuring symmetry to combine all datasets. Columns 'visual blurring', 'Itching', 'Irritability', 'delayed healing', 'partial paresis', 'muscle stiffness', 'Alopecia', 'Obesity', and 'class' were used from this dataset.

Table 2 provides a concise summary of the dataset. We merged these datasets by selecting 200 features from each. The resulting dataset has a shape of 400 rows and 28 columns, indicating 400 rows with 28 identical columns, encompassing 27 features and one class with two values, diabetic or non-diabetic.

Table 3 shows sample values of the dataset. Increased levels of Urea in the blood may indicate kidney dysfunction, a common complication of diabetes. Higher HbA1c levels, reflecting poorer blood sugar control over the past few months, are a hallmark of diabetes. In terms of cholesterol levels, high total cholesterol coupled with low levels of HDL (known as 'good' cholesterol) is associated with insulin resistance and metabolic syndrome, precursors to type 2 diabetes. Elevated levels of LDL (commonly referred to as 'bad' cholesterol) have been correlated with a heightened risk of cardiovascular disease, a condition frequently observed in individuals with diabetes. Additionally, high VLDL levels can indicate insulin resistance, contributing to the development of diabetes. These markers collectively suggest an increased risk or presence of diabetes, potentially due to insulin resistance, poor blood sugar control, and associated metabolic changes. However, a comprehensive evaluation

considering these markers alongside other clinical information is essential. In the 'Target' column, '1' signifies a positive diagnosis of diabetes, while '0' indicates a non-diabetic status.

Challenges with dataset

The primary challenge with the dataset is its limited size and the insufficiency of diverse attributes. These limitations restrict the model's ability to capture complex patterns and may impact its generalization capabilities. To address this, expanding the dataset by increasing the number of samples and incorporating more comprehensive and diverse attributes would significantly enhance the model's performance. A more robust dataset would allow the model to learn from a broader range of features, ultimately improving its accuracy and ability to generalize to unseen data.

ML models

In this study, we employed several models to propose a diabetes prediction system. We utilize DT, RF, LR, and SVC for classification tasks, and ETC, CNN, and LSTM for feature extraction. These models are configured with their optimal hyperparameters, as detailed in Table 4. Each algorithm is explored to address specific challenges in classification tasks, from decision tree depth to neural network architecture design.

Decision trees

The DT classifier is a well-known non-linear supervised classification method recognized for its tree-like structure[32]. In this approach, the branch nodes within the tree represent distinct scenarios, while the leaf nodes correspond to classified records. DT is characterized by its simplicity in comprehension and visualization, rendering it suitable for a diverse array of applications. However, it is important to note that DT is sensitive to variations in data, as even minor changes in the dataset can result in different tree structures. In the case of the presented approach we used this algorithm configured with a single hypermeter which is max_depth of 6, indicating that each tree created by the algorithm will have a maximum depth of 6. To obtain the optimal performance, during the tuning process, the algorithm explores a range of depths from 5 to 200 to find the optimal tree depth. The max_depth parameter controls how deep the tree can grow, affecting the complexity of the model. A higher max_depth allows the algorithm to capture more intricate patterns in the data, but it also increases the risk of overfitting if not properly tuned. The set max_depth parameter is suitable for the dataset used in this study.

Random forest

RF is a supervised learning algorithm utilized for both classification and regression tasks. Among the various algorithms available, RF stands out as one of the simplest and most user-friendly options. An RF model consists of an ensemble of decision trees, where the term "forest" refers to this collection of trees. This algorithm has demonstrated impressive performance, with the quality of predictions improving as the number of trees in the forest increases. The RF process involves constructing decision trees based on randomly selected subsets of the diabetes data, making predictions with each tree into diabetes and non-diabetes, and then aggregating the results to determine the final prediction of whether the sample belongs to the diabetic or non-diabetic person. This aggregation method often involves a voting approach, where the most commonly predicted outcome is selected as the final prediction[33]. RF employs the bagging algorithm, which involves training multiple decision trees using different bootstrap samples from the training dataset. This process ensures that each tree is trained on a diverse subset of the data, enhancing the model's robustness and generalization. The bagging algorithm is mathematically represented as shown in Eqs. (1) and (2).

$$W = \text{mode } J_1(Y), J_2(Y), \ldots, J_t(Y) \tag{1}$$

Algorithm	Hyperparameters	Tuning range
DT	max_depth = 6	max_depth = {5 to 200}
RF	n_estimators = 300, random_state = 42, max_depth = 1	n_estimators = {100, 200, 300, 400, 500}, max_depth = {5, 10, 20, 30}
LR	random_state = 1000, solver = 'liblinear', C = 2.0	No tuning needed for these parameters
SVC	kernel = 'rbf', C = 3.0, random_state = 500	No tuning needed for these parameters
ETC	n_estimators = 300, max_depth = 20, criterion = 'entropy'	n_estimators = {100, 200, 300, 400, 500}, max_depth = {5, 10, 20, 30}
LSTM	LSTM (filters = 128), optimizer = 'adam', loss = 'binary_crossentropy', Dropout = {0.3, 0.4, 0.5}, activation = 'sigmoid', epochs = 100	
CNN	Conv1D (filters = 64, 128), kernel = 3×3, maxpooling1D = 2×2, optimizer = 'adam', loss = 'binary_crossentropy', Dropout = 0.5, activation = 'sigmoid', epochs = 100	
LSTM+CNN	Conv1D(filters = 64, 128, 128), kernel = 3×3, maxpooling1D = 2×2, optimizer = 'adam', loss = 'binary_crossentropy', Dropout = 0.5, activation = 'sigmoid', epochs = 100	

Table 4. Settings for hyperparameters in ML and DL models using which the models showed the best performance.

$$W = \text{mode} \sum_{k=1}^{l} J_t(Y) \tag{2}$$

where W is employed as the final prediction, relying on the maximum decision from an ensemble of DT denoted as $J_1(Y), J_2(Y)$, and so forth, all of which contribute to the prediction process.

In this experiment, RF is set with n_estimators at 300, indicating it will create 300 trees in the forest. During tuning, the algorithm explores the number of trees from 100 to 500 to find the optimal balance between model performance and computational efficiency. The random_state parameter is set to 42 for reproducibility, ensuring that the same results are obtained each time the algorithm is run with the same inputs. Additionally, max_depth is set to 1 for each tree in the forest, controlling the maximum depth of individual trees to prevent overfitting.

Logistic regression

LR is employed to estimate the probability of an event occurring based on a specified set of independent variables. The outcome variable ranges between 1 and 0, representing probabilities. LR utilizes a logistic function, which is characterized by an S-shaped curve and maximizes the likelihood of the predicted outcomes[34]. The mathematical expression of the logistic function is provided in Eq. (3).

$$X = \frac{w}{1 + K^{-e(r-r_0)}} \tag{3}$$

where K integrates Euler's constant, $r_0$signifies the central value of the sigmoid, w denotes the peak of the curve, and e delineates the curvature's steepness.

For current experiments, the random_state is set to 1000, ensuring the reproducibility of results for diabetes detection. The 'solver' parameter is set to 'liblinear', which is suitable for small datasets, as is the case with the current study which has a binary classification problem. For the multi_class parameter, the 'ovr' (One-vs-Rest) strategy is employed, allowing the algorithm to handle multiclass classification tasks. The C parameter is set to 2.0, representing the regularization strength in the model. Higher C values indicate less regularization, allowing the model to fit the training data more closely.

Support vector classifier

SVC is a flexible ML technique utilized for both regression and classification tasks[35]. The primary aim of this model is to establish a hyperplane within an N-dimensional space. In this experiment, the SVM algorithm is configured with a radial basis function (RBF) kernel, specified by the 'kernel' parameter. The C parameter is set to 3.0, controlling the trade-off between achieving a low error on the training data and maximizing the decision boundary's margin between the samples of diabetic and non-diabetic class samples. A higher C value allows the algorithm to fit the training data more closely, potentially leading to overfitting. The random_state parameter is set to 500 for result reproducibility.

Extra trees classifier

ETC aids in feature extraction by providing feature importance rankings, handling noisy and collinear features effectively, capturing non-linear relationships, leveraging its ensemble approach, and allowing for robust evaluation through cross-validation techniques. These properties make ETC a valuable tool for identifying the most informative features in the diabetes dataset[36]. In this configuration, n_estimators is set to 300, indicating the number of trees in the forest. During tuning, the algorithm explores the number of trees from 100 to 500 and produces the best results with 300 estimators. The max_depth parameter is set to 20, controlling the maximum depth of individual trees to balance model complexity and overfitting. The criterion parameter is set to entropy which measures the information gained to make decisions.

Long short-term memory

LSTM is a type of recurrent neural network (RNN) commonly used for sequence modeling tasks like text and time series analysis[37]. In this LSTM configuration, the model architecture includes an LSTM layer with 128 filters. The optimizer is set to 'adam', a popular optimization algorithm for DL. For training, the model uses binary_crossentropy as the loss function, suitable for binary classification tasks. Dropout regularization is applied with rates of 0.3, 0.4, and 0.5 to prevent overfitting. The activation function sigmoid is used for the LSTM layer, typical for binary classification. The model is trained for 100 epochs.

Convolutional neural network

CNNs are well-suited for image and sequence processing tasks[38]. This CNN configuration includes Conv1D layers with 64 and 128 filters, applying 3×3 kernels. Maxpooling1D with a 2×2 window is used to reduce spatial dimensions. The optimizer is set to 'adam', a commonly used optimization algorithm. The loss function binary_crossentropy is employed for binary classification tasks. Dropout with a rate of 0.5 is applied for regularization, reducing overfitting. The activation function sigmoid is used for the final layer. The model is trained for 100 epochs.

Ensemble LSTM+CNN model
This architecture combines LSTM and CNN layers for improved performance in sequence analysis tasks[39]. The configuration includes Conv1D layers with 64, 128, and 128 filters, applying 3×3 kernels. Maxpooling1D with a 2×2 window is utilized for dimension reduction. The optimizer is set to 'adam', a popular choice for DL models. The loss function binary_crossentropy is used for binary classification. Dropout regularization with a rate of 0.5 is applied to prevent overfitting. The activation function sigmoid is used for the final layer. The model is trained for 100 epochs.

Hyperparameter tuning
We fine-tuned the models' hyperparameter settings using the grid search method. The models were tuned over specific ranges. After tuning the models within these ranges, we identified the best settings. For example, for the decision tree, the optimal max_depth was 6, which was fine-tuned between a range of 5 to 200. Table 4 provides all the hyperparameters used for all models used in this study.

Feature extraction

Figure 2 shows the architecture of the proposed feature extraction approach that utilizes CNN and LSTM to make an ensemble. The models are stacked to obtain the most impactful features. Initially, we introduced DL models such as LSTM and CNN separately. The objective was to leverage these models to automatically extract relevant features for more efficient training. While the accuracy improved for some ML models with the introduction of these DL models, it was not a universal enhancement across all models. In the final stage, we adopted a fusion approach, combining the strengths of both LSTM and CNN. This fusion technique allowed us to select the most suitable features from both models, leading to a remarkable increase in accuracy. Specifically, the results showed a significant jump to 99% accuracy, marking a substantial improvement from previous stages of the methodology.

Feature embedding

We sourced three datasets from a public repository and meticulously extracted unique features from each. The PIMA Indians Diabetes Dataset[30] served as our base dataset. To increase dimensionality, we incorporated additional attributes from the Diabetes Dataset[31] and the Early Stage Diabetes Risk Prediction Dataset[29]. While PIMA originally provided 9 attributes, we identified 15 unique attributes from the other two datasets that were not present in PIMA. We integrated these features for both diabetic and non-diabetic categories, resulting in a total of 24 attributes plus one target variable as diabetic and non-diabetic, as shown in Fig. 3.

The key takeaway from this process is the pivotal role of feature engineering in enhancing model accuracy. By carefully selecting and refining features, we were able to significantly boost the performance of both traditional ML and DL models. Additionally, the fusion of LSTM and CNN proved to be a potent strategy, resulting in a highly accurate classification model for diabetic and non-diabetic cases.

Results and discussion

The experiments utilized Python 3.10.4 and Jupyter Notebook, employing libraries including Seaborn, Scikit-learn, Pandas, and NumPy. The experiments were conducted on an HP machine running Windows, equipped with 64 GB RAM and a 2 TB SSD.

Various evaluation metrics were utilized to assess the performance of ML classifiers, including precision, accuracy, recall, and F1 score[40].

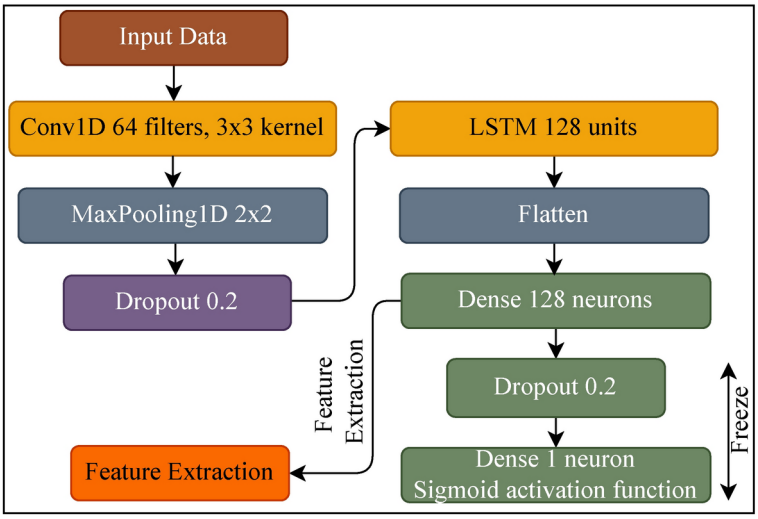

Fig. 2. Proposed feature extraction approach where CNN and LSTM models are combined.

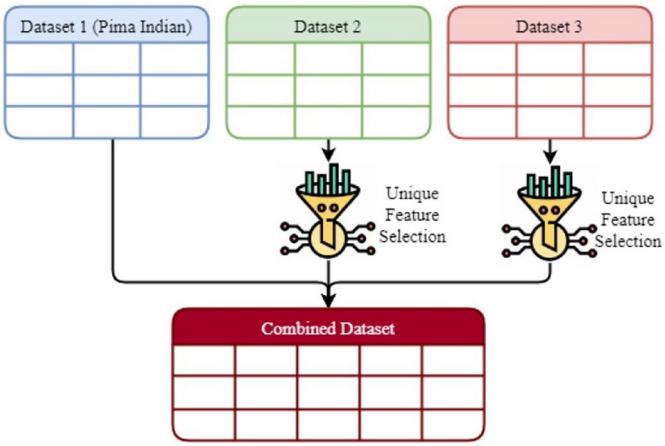

Fig. 3. Process followed in this study for embedding features from three datasets.

Model	Accuracy	Class	Precision	Recall	F1 score
DT	0.95	0	0.94	0.94	0.94
		1	0.95	0.95	0.95
		Micro. avg	0.95	0.95	0.95
RF	0.96	0	0.97	0.94	0.96
		1	0.96	0.98	0.97
		Micro. avg	0.96	0.96	0.96
LR	0.96	0	0.97	0.94	0.96
		1	0.96	0.98	0.97
		Micro. avg	0.96	0.96	0.96
SVC	0.91	0	0.87	0.94	0.91
		1	0.95	0.89	0.92
		Micro. avg	0.91	0.92	0.91

Table 5. Experimental results using ML models when using the original features from the dataset.

$$Accuracy = \frac{T_rP_o + T_rN_e}{T_rP_o + T_rN_e + F_aP_o + F_aN_e} \tag{4}$$

$$P = \frac{T_rP_o}{T_rP_o + F_aP_o} \tag{5}$$

$$R = \frac{T_rP_o}{T_rP_o + F_aN_e} \tag{6}$$

$$F1\text{-}score = 2 * \frac{P * R}{P + R} \tag{7}$$

where,

- T_rP_o: is true positive in which the sample belongs to Dibatic and the model also predicts it as Dibatic.
- F_aP_o: is a false positive in which the sample belongs to NON-Dibatic but the model predicts it as Dibatic.
- T_rN_e: is true negative in which the sample belongs to Dibatic and the model also predicts it as NON-Dibatic.
- F_aN_e: is a false negative in which the sample belongs to Dibatic but the model predicts it as NON-Dibatic.

Performance comparison of ML models with original features

Table 5 presents the results of ML models utilizing the original feature set. Both LR and RF perform admirably with an accuracy score of 0.96. RF, being a tree-based algorithm, excels due to its capacity to manage features without one-hot encoding, its ensemble approach to capturing intricate interactions, and its resistance to overfitting. LR shines in binary classification tasks, offering interpretable probabilities and aligning with linear assumptions for certain categorical data. DT also achieves good performance with an accuracy of 0.95, slightly lower than RF and LR, attributable to its inclination to overfit without the benefits of ensemble learning and

Model	Accuracy	Class	Precision	Recall	F1 score
DT	0.97	0	1.00	0.94	0.97
		1	0.96	1.00	0.98
		Micro. avg	0.98	0.97	0.97
RF	0.97	0	1.00	0.94	0.97
		1	0.96	1.00	0.98
		Micro. avg	0.98	0.97	0.97
LR	0.97	0	0.97	0.97	0.97
		1	0.98	0.98	0.98
		Micro. avg	0.97	0.97	0.97
SVC	0.90	0	0.87	0.92	0.89
		1	0.93	0.89	0.91
		Micro. avg	0.90	0.90	0.90

Table 6. Experimental results using ML models when trained and tested using ETC feature selection approach.

Model	Accuracy	Class	Precision	Recall	F1 score
DT	0.96	0	0.95	0.97	0.96
		1	0.98	0.95	0.97
		Micro. avg	0.96	0.96	0.96
RF	0.95	0	0.94	0.94	0.94
		1	0.95	0.95	0.95
		Micro. avg	0.95	0.95	0.95
LR	0.94	0	0.92	0.94	0.93
		1	0.95	0.93	0.94
		Micro. avg	0.94	0.94	0.94
SVC	0.94	0	0.94	0.92	0.93
		1	0.93	0.95	0.94
		Micro. avg	0.94	0.94	0.94

Table 7. Experimental results using ML models when features selected using LSTM model are used.

regularization. Conversely, the SVC lags with an accuracy of 0.91, attributed to its lower noise tolerance and inefficiency in handling large datasets.

Performance comparison of ML models with ETC feature selection

Table 6 displays the results of ML models using the best-selected features. A set of 12 top features was extracted, resulting in an accuracy of 0.97, showcasing strong performance from DT, RF, and LR. Feature extraction is pivotal for enhancing the value of results, involving the selection and transformation of the most pertinent information from raw data. This process reduces dimensionality, highlights crucial patterns, and improves model efficiency and interpretability. ETC, a variant of RF, played a key role in this feature extraction, as it identifies and ranks important features. By focusing on the most informative features, ETC enhances model performance, reduces noise, improves accuracy, and accelerates computation, resulting in more effective decision-making and insights from the data. SVM's suboptimal performance on features extracted from ETC could be due to the complexity and non-linearity of the features, lack of flexibility in adapting to ETC's patterns, sensitivity to noise, and differences in algorithms.

Performance comparison of ML models with LSTM feature selection

Table 7 illustrates the performance of ML models on features extracted from LSTM. The use of LSTM for feature extraction can lead to decreased model performance compared to ETC, attributed to the mismatch between LSTM's sequential data design and the tabular data typical in traditional ML tasks. LSTM, tailored for sequences like time series or text, may create complex representations unsuitable for traditional ML models, resulting in diminished performance.

On the other hand, ETC excels in tabular feature extraction, offering clear feature importance rankings that enhance model interpretability and prevent overfitting. DT achieved the highest accuracy of 0.96, followed by a slight decrease for RF to 0.95. LR and SVC showed drops in performance as well, with an accuracy of 0.94. Figure 4 illustrates the explanation of training and testing validation loss.

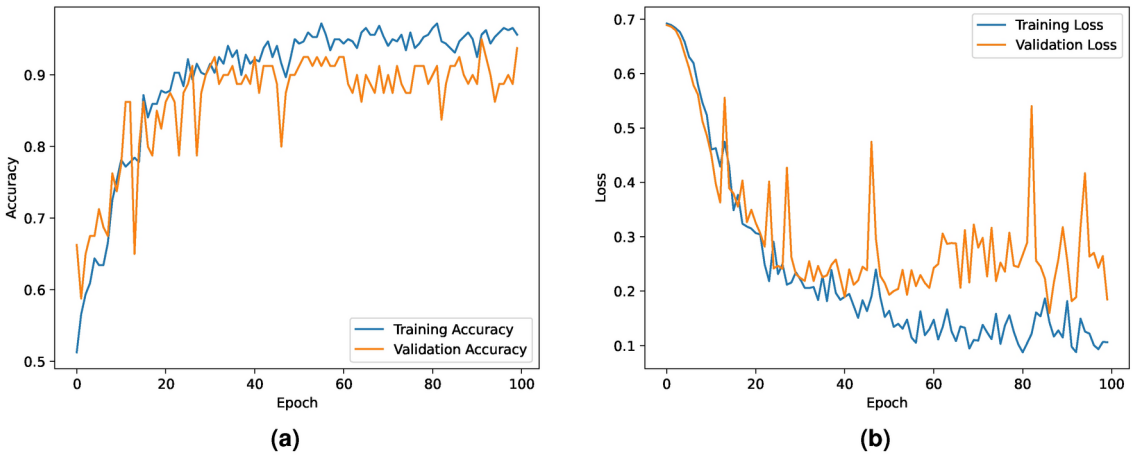

Fig. 4. Training and validation accuracy, and loss per epoch, (**a**) Training and validation accuracy per epoch with LSTM, and (**b**) Training and validation loss per epoch with LSTM.

Model	Accuracy	Class	Precision	Recall	F1 score
DT	0.95	0	0.94	0.94	0.94
		1	0.95	0.95	0.95
		Micro. avg	0.95	0.95	0.95
RF	0.97	0	0.97	0.97	0.97
		1	0.98	0.98	0.98
		Micro. avg	0.97	0.97	0.97
LR	0.96	0	0.97	0.94	0.96
		1	0.96	0.98	0.97
		Micro. avg	0.96	0.96	0.96
SVC	0.97	0	0.97	0.97	0.97
		1	0.98	0.98	0.98
		Micro. avg	0.97	0.97	0.97

Table 8. Experimental results using ML models when CNN-extracted features are used for model training.

Performance comparison of ML models with CNN feature selection

Table 8 provides a comparative overview of ML models using CNN features. Utilizing CNN for feature extraction improved RF and SVC performance compared to LSTM due to CNN's ability to extract hierarchical, non-linear features, which align better with tabular data characteristics, thus enhancing model accuracy and generalization. Both RF and SVC achieved a performance of 0.97. Conversely, DT and LR performed less with 0.95 and 0.96, respectively.

DT model shows poor performance when used with features extracted using the CNN model. DT's lower performance on CNN features is attributed to their complex and hierarchical nature, which may not align well with DT's simple splitting rules. However, DT excels with LSTM features, benefitting from its sequential data handling capabilities. Figure 5 illustrates the explanation of training and testing validation loss.

Performance evaluation of ML models with CNN-LSTM feature selection

Table 9 presents ML models utilizing CNN-LSTM-based features. The notable increase in accuracy can be attributed to the synergistic strengths of CNNs and LSTMs.

CNNs excel at capturing spatial patterns, while LSTMs specialize in learning sequential patterns. The combination allows CNNs to extract hierarchical spatial features, subsequently processed by LSTMs to capture temporal dependencies. This integration of spatial and sequential information enables a deeper understanding of complex patterns, resulting in a substantial accuracy improvement. The models DT, RF, and LR achieved an accuracy of 0.99, while SVC yielded 0.96. Figure 6 illustrates the explanation of training and testing validation loss.

Figure 7 displays the confusion matrix of the top-performing model in each scenario. In the case of the original feature, RF exhibited the best performance with 77 correct predictions and 3 incorrect predictions. Similarly, with ML-based features, RF continued its strong performance with 78 correct and 2 incorrect predictions. When features were extracted from LSTM, DT showed significant improvement with 77 correct predictions and 3 incorrect predictions. Extracting features from CNN, RF again performed well with 78 correct predictions and 2

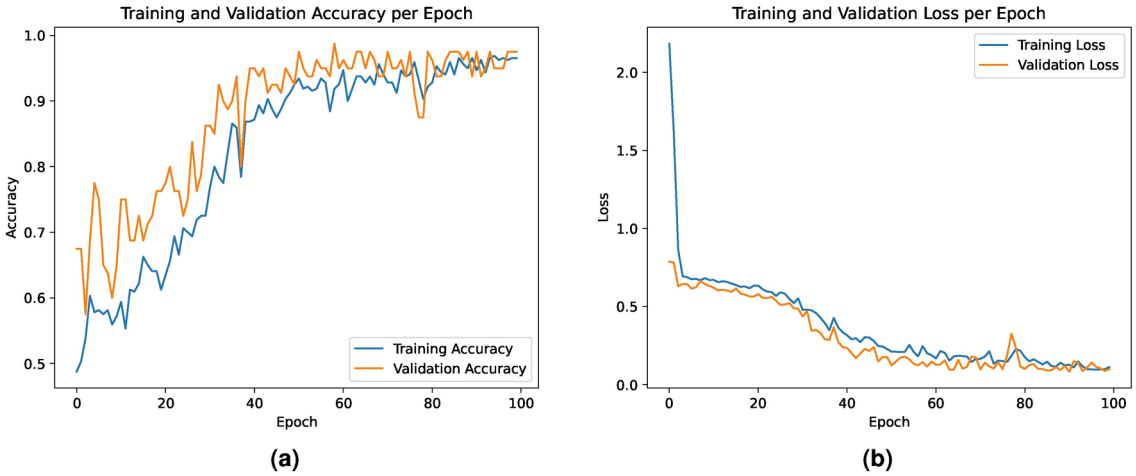

Fig. 5. Training and validation accuracy, and loss per epoch, (**a**) Training and validation accuracy per epoch using CNN, (**b**) Training and validation loss per epoch using the CNN model.

Model	Accuracy	Class	Precision	Recall	F1 score
DT	0.99	0	1.00	0.97	0.99
		1	0.98	1.00	0.99
		Micro. avg	0.99	0.99	0.99
RF	0.99	0	1.00	0.97	0.99
		1	0.98	1.00	0.99
		Micro. avg	0.99	0.99	0.99
LR	0.99	0	1.00	0.97	0.99
		1	0.98	1.00	0.99
		Micro. avg	0.99	0.99	0.99
SVC	0.96	0	1.00	0.92	0.96
		1	0.94	1.00	0.97
		Micro. avg	0.97	0.96	0.96

Table 9. Experimental results using ML models when features from the proposed CNN-LSTM model are used.

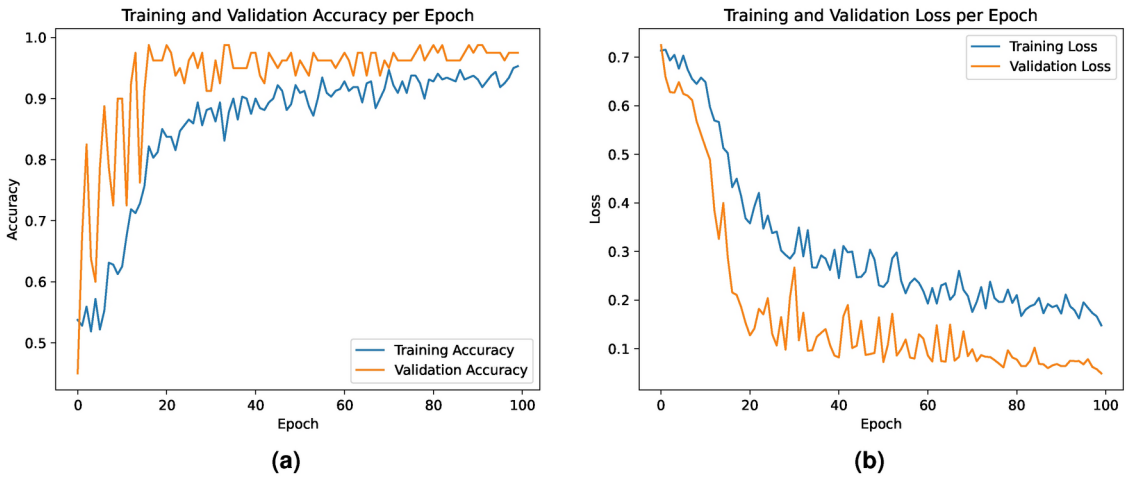

Fig. 6. Training and validation accuracy, and loss per epoch, (**a**) Training and validation accuracy per epoch with CNN-LSTM, and (**b**) Training and validation loss per epoch with CNN-LSTM models.

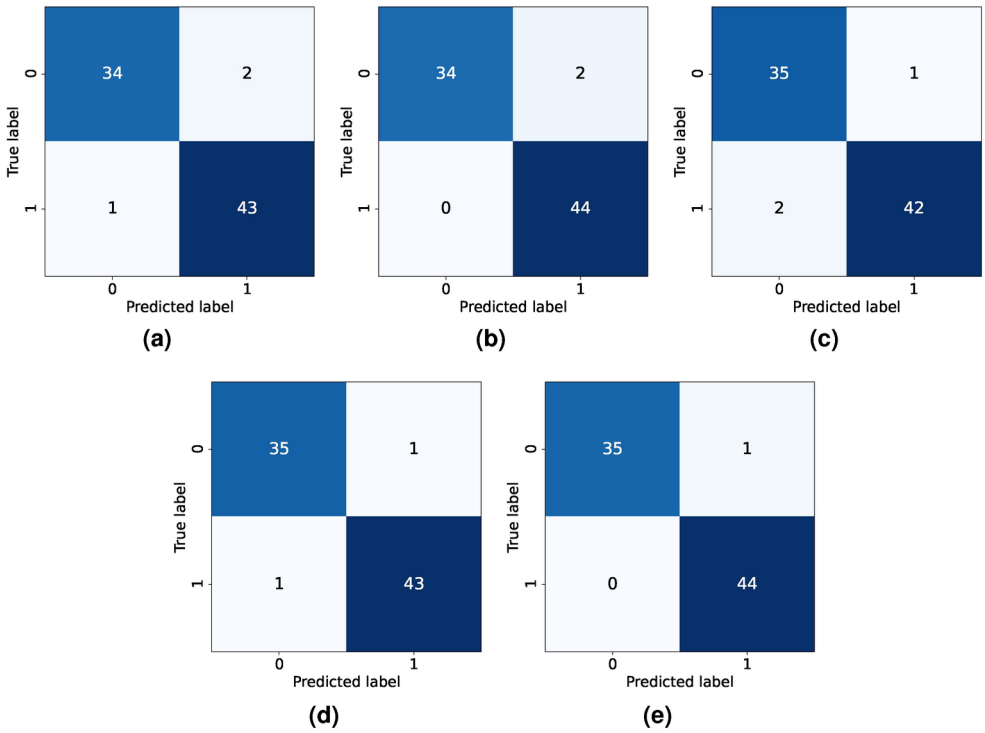

Fig. 7. The confusion matrices for best-performing models, (**a**) RF using the original dataset, (**b**) RF using ETC-based features, (**c**) DT using LSTM-extracted features, (**d**) RF in the case of CNN features, and (**e**) RF using the features from CNN-LSTM model.

incorrect predictions. Finally, using CNN-LSTM-based features, the model achieved 79 correct predictions with only 1 incorrect prediction.

Validation of the proposed approach

Table 10 summarizes k-fold cross-validation results for ML models using different feature sets: Original, ETC, LSTM, CNN, and CNN-LSTM. In the CNN-LSTM feature set, all models achieve 99% accuracy, with low standard deviations indicating consistent performance. The CNN feature set also shows strong results, with RF and LR reaching 97% accuracy. Models in the Original and ETC sets perform competitively, with accuracies ranging from 90-97%. SVC consistently exhibits lower accuracy and higher variability across feature sets. Overall, CNN-LSTM emerges as the most effective feature set for these models.

Performance comparison

Table 11 presents a comparison of accuracy levels from various studies in the literature on diabetes prediction. Each row represents a study, detailing the reference, year, features used, dataset type, classifier employed, and the reported accuracy. Notable findings include a 95.7% accuracy using LSTM-CNN in 2018, 98.1% with DNN in 2023, and 99.13% with ANN in 2023. The proposed model in 2024, utilizing CNN-LSTM features with an RF classifier, also achieves an accuracy of 99%, showcasing its performance alongside previous studies.

Conclusions and future work

Timely detection of diabetes can greatly help in determining the proper medication for the patients and avoid further complications related to the heart, lungs, eyes, etc. The use of ML and DL holds significant potential for accurate and rapid diabetes detection. In this direction, this study introduces a novel feature extraction approach to boost the performance of ML models by making an ensemble of LSTM and CNN models. The proposed model helps obtain highly contributing features from a combined dataset, comprising three different datasets. Consequently, model overfitting, low accuracy, and generalizability issues are resolved, in addition to obtaining higher accuracy. Experiments involve various scenarios where original features and features obtained from LSTM, CNN, and ETC are used for model training and testing. Experimental findings reveal the superb performance of the proposed CNN-LSTM feature engineering approach with a 0.99 accuracy score, thereby outperforming other approaches. Results from k-fold cross-validation, along with a comparison with existing methodologies, further illustrate the enhanced performance of the proposed method. In future work, we intend to experiment with transfer learning approaches for diabetes detection. In addition, diabetes detection in a real-time environment is also under consideration. Future research could focus on enhancing the ensemble level of classifiers, which may lead to further accuracy improvements. Additionally, expanding the dataset is crucial; a larger dataset would enable more effective model training, potentially resulting in even higher accuracy. Building

Features	Model	Accuracy	Stander deviation
Original	DT	0.95	±0.21
	RF	0.96	±0.19
	LR	0.96	±0.23
	SVC	0.91	±0.38
ETC	DT	0.97	±0.20
	RF	0.97	±0.17
	LR	0.97	±0.21
	SVC	0.90	±0.30
LSTM	DT	0.96	±0.15
	RF	0.95	±0.14
	LR	0.94	± 0.14
	SVC	0.94	±0.15
CNN	DT	0.95	±0.19
	RF	0.97	±0.17
	LR	0.96	±0.19
	SVC	0.97	±0.16
CNN-LSTM	DT	0.99	±0.19
	RF	0.99	±0.16
	LR	0.99	±0.16
	SVC	0.96	±0.17

Table 10. K-fold cross-validation results. Given values are averaged for 10 folds.

Ref	Year	Features	Datasets	Classifier	Accuracy (%)
[15]	2018	LSTM-CNN	Single	SVM	95.7
[16]	2019	Original	Single	RF	83.67
[17]	2021	Original	Single	ADA	83
[18]	2023	Original	Single	DNN	98.1
[20]	2023	Original	Single	ANN	99.13
[21]	2022	Original	Single	RF	88.14
[22]	2023	Original	Single	KNN	89.5
[23]	2023	Original	Single	XGBoost, CatBoost	99
[24]	2022	Original	Single	RF	99
[25]	2023	Original	Single	RF	99
Proposed	2024	CNN-LSTM	Multiple	RF	99

Table 11. Comparing the accuracy levels reported in recent studies in the literature in comparison to the proposed approach in this study.

on our work with feature ensembling, these approaches could significantly boost the model's performance and reliability.

Data availability

The data can be requested from corresponding authors.

Received: 13 May 2024; Accepted: 25 September 2024
Published online: 07 October 2024

References

1. Davison, L. Diabetes mellitus and pancreatitis-cause or effect?. *J. Small Anim. Pract.* **56**, 50–59 (2015).
2. Sonksen, P. & Sonksen, J. Insulin: understanding its action in health and disease. *Br. J. Anaesth.* **85**, 69–79 (2000).
3. Inzucchi, S. E. Management of hyperglycemia in the hospital setting. *N. Engl. J. Med.* **355**, 1903–1911 (2006).
4. NYU Langone Health. Diagnosing type 2 diabetes. NYU Langone Health (n.d.).
5. Bajaj, A., Sethi, A., Rathor, P., Suppogu, N. & Sethi, A. Acute complications of myocardial infarction in the current era: diagnosis and management. *J. Investig. Med.* **63**, 844–855 (2015).
6. Mohebbi, A. et al. A deep learning approach to adherence detection for type 2 diabetics. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2896–2899 (IEEE, 2017).

7. Deberneh, H. M. & Kim, I. Prediction of type 2 diabetes based on machine learning algorithm. *Int. J. Environ. Res. Public Health* **18**, 3317 (2021).
8. Howlader, K. C. et al. Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. *Health Inf. Sci. Syst.* **10**, 2 (2022).
9. Talaei-Khoei, A. & Wilson, J. M. Identifying people at risk of developing type 2 diabetes: a comparison of predictive analytics techniques and predictor variables. *Int. J. Med. Inform.* **119**, 22–38 (2018).
10. Allalou, A. et al. A predictive metabolic signature for the transition from gestational diabetes mellitus to type 2 diabetes. *Diabetes* **65**, 2529–2539 (2016).
11. Tsao, H.-Y., Chan, P.-Y. & Su, E.C.-Y. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinform.* **19**, 111–121 (2018).
12. Kavakiotis, I. et al. Machine learning and data mining methods in diabetes research. *Comput. Struct. Biotechnol. J.* **15**, 104–116 (2017).
13. Fan, Y. & Long, E. Machine learning approaches to predict risks of diabetic complications and poor glycemic control in nonadherent type 2 diabetes. *Front. Pharmacol.* **12**, 665951 (2021).
14. Rai, V. et al. Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. *Mol. Metab.* **32**, 109–121 (2020).
15. Swapna, G., Vinayakumar, R. & Soman, K. Diabetes detection using deep learning algorithms. *ICT Express* **4**, 243–246 (2018).
16. Yahyaoui, A., Jamil, A., Rasheed, J. & Yesiltepe, M. A decision support system for diabetes prediction using machine learning and deep learning techniques. In *2019 1st International informatics and software engineering conference (UBMYK)*, 1–4 (IEEE, 2019).
17. Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F. & Sayadi, M. J. Diabetes diagnosis using machine learning. *Front. Health Inform.* **10**, 65 (2021).
18. Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. & Juwono, F. H. Diabetes detection based on machine learning and deep learning approaches. *Multimed. Tools Appl.*, 1–33 (2023).
19. Meena, G., Mohbey, K. K., Acharya, M. & Lokesh, K. An improved convolutional neural network-based model for detecting brain tumors from augmented MRI images. *J. Auton. Intell.* **6** (2023).
20. Singh, P., Silakari, S. & Agrawal, S. An efficient deep learning technique for diabetes classification and prediction based on Indian diabetes dataset. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 487–491 (IEEE, 2023).
21. Jain, V. Diabetes prediction using support vector machine, naive bayes and random forest machine learning models. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, 837–841 (IEEE, 2022).
22. Rathi, B. & Madeira, F. Early prediction of diabetes using machine learning techniques. In *2023 Global Conference on Wireless and Optical Technologies (GCWOT)*, 1–7 (IEEE, 2023).
23. Shampa, S. A., Islam, M. S. & Nesa, A. Machine learning-based diabetes prediction: A cross-country perspective. In *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 1–6 (2023).
24. Mangal, A. & Jain, V. Performance analysis of machine learning models for prediction of diabetes. In *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, 1–4 (IEEE, 2022).
25. Islam, M. S., Alam, M. M., Ahamed, A. & Meerza, S. I. A. Prediction of diabetes at early stage using interpretable machine learning. In *SoutheastCon 2023*, 261–265 (IEEE, 2023).
26. Meena, G., Mohbey, K. K. & Kumar, S. Monkeypox recognition and prediction from visuals using deep transfer learning-based neural networks. *Multimed. Tools Appl.*, 1–25 (2024).
27. Meena, G. & Mohbey, K. K. Sentiment analysis on images using different transfer learning models. *Procedia Comput. Sci.* **218**, 1640–1649 (2023).
28. Shafique, R. et al. Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning. *Cancers* **15**, 681 (2023).
29. Aravindpcoder. Diabetes dataset. https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset. Accessed: March 26, 2024.
30. Mathchi. Diabetes data set. https://www.kaggle.com/datasets/mathchi/diabetes-data-set. Accessed: March 26, 2024.
31. Ishandutta. Early stage diabetes risk prediction dataset. https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset. Accessed: March 26, 2024.
32. Goethals, S., Martens, D. & Evgeniou, T. The non-linear nature of the cost of comprehensibility. *J. Big Data* **9**, 1–23 (2022).
33. Sahin, E. K. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* **2**, 1308 (2020).
34. Maalouf, M. Logistic regression in data analysis: an overview. *Int. J. Data Anal. Tech. Strateg.* **3**, 281–299 (2011).
35. Hadem, P., Saikia, D. K. & Moulik, S. An SDN-based intrusion detection system using SVM with selective logging for IP traceback. *Comput. Netw.* **191**, 108015 (2021).
36. Alfian, G. et al. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers* **11**, 136 (2022).
37. Sagheer, A. & Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* **323**, 203–213 (2019).
38. Krichen, M. Convolutional neural networks: A survey. *Computers* **12**, 151 (2023).
39. Kim, T. & Kim, H. Y. Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *PLoS One* **14**, e0212320 (2019).
40. Shafique, R., Rustam, F., Murtala, S., Jurcut, A. D. & Choi, G. S. Advancing autonomous vehicle safety: Machine learning to predict sensor-related accident severity. *IEEE Access* **12**, 25933–25948 (2024).

Author contributions

FR: conceptualization, formal analysis, and writing—original draft. ASAS: conceptualization, data curation, and writing—original draft. RS: data curation, formal analysis, and methodology. SAO: funding acquisition, methodology and visualization. RCI: investigation, software, and visualization. JPMG: formal analysis, Investigation and project administration. IA: Supervision, validation, and writing—review and editing. All writers reviewed and approved the final manuscript.

Funding

This research was funded by the European University of Atlantic.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.