RES MILITARIS Social Science Journal

CNN: An Approach for Prediction of Diabetes and its Consequences at an Early Stage Using Advanced Deep Learning Techniques with Image Data

Mr. Somendra Tripathi
Department of Computer Science and Engineering
Rama University
Kanpur, Uttar Pradesh
somendra.tripathi@gmail.com

Prof. (Dr.) Hari Om Sharan
Department of Computer Science and Engineering
Rama University
Kanpur, Uttar Pradesh
drsharan.hariom@gmail.com

Prof. (Dr.) C. S. Raghuvanshi
Department of Computer Science and Engineering
Rama University
Kanpur, Uttar Pradesh
drcsraghuvanshi@gmail.com

Abstract: Diabetes is a prolonged health ailment that affects how your body turns food into energy, and the majority of the world's population suffers from it and its consequences. Medical experts have classified diabetes as a non-reversible disease that can be better controlled if diagnosed as early as possible. Technology advancement is helping in every aspect of human life by solving problems in an efficient and cost effective manner. Researchers are working to make it possible to diagnose or predict the disease at an early stage by proposing many prediction models. The majority of the models in the state of the art use machine learning techniques like KNN, Naive Bayes Classifier, and so on. All these models take input as a numerical feature like BMI index, blood pressure, fasting and PP sugar levels, lifestyle, etc. Advancement in technology proves to be more accurate prediction and classification by using input as an image in deep learning techniques. The proposed work uses advanced deep learning classification techniques, CNN, and proves to be more accurate in predicting the probability of being diabetic in the future and also predicting the future consequences that will occur if the patient is diagnosed as diabetic. It also recommends some treatment and precautions to be followed for a better cure of the disease. Our proposed model works in three steps, first, select the best feature to be considered in prediction and convert it into image form; second, input the best deep learning classification model, CNN; and last, based on the classification result, predict diabetes and its consequences with treatment and precaution. Experimental analysis with existing machine learning prediction models in the literature, using KNN, random forest, SVM, and decision tree, is carried out on the India Diabetic Dataset PIMA and diabetes type data sets and shown to be more accurate in prediction.

Keyword: Deep learning, Deep Neural Networks, Convolution Neural Network, Diabetes, Machine learning.

1. Introduction:

Diabetes mellitus, commonly known as diabetes, is a chronic metabolic disorder that affects millions of people worldwide. It is characterized by elevated blood glucose levels due to the body's inability to produce enough insulin or properly utilize it. The prevalence of diabetes has been steadily increasing over the past decades, posing a significant global health challenge. Early detection and accurate prediction of diabetes can play a crucial role in preventing complications and improving the quality of life for affected individuals.

In recent years, advancements in data science and machine learning have opened up new opportunities to harness the power of data for medical applications, including disease prediction and diagnosis. Leveraging these technologies, we present a novel approach for predicting diabetes risk in individuals based on their demographic, clinical, and lifestyle factors.

Our proposed approach involves the integration of various machine learning algorithms and data analysis techniques to identify patterns and risk factors associated with diabetes. The development of this predictive model is based on a comprehensive dataset containing information from a diverse group of individuals, including medical records, genetic markers, and lifestyle data.

Key components of our approach include data preprocessing, feature selection, and model training. To ensure the accuracy and generalizability of the predictive model, we employ cross-validation techniques and perform rigorous evaluations on both training and test datasets.

The ultimate goal of our research is to create a user-friendly, reliable, and accessible tool that can assist healthcare professionals in identifying individuals at high risk of developing diabetes. This early prediction can lead to timely interventions, lifestyle modifications, and personalized healthcare plans, which, in turn, may help prevent or delay the onset of diabetes and its associated complications.

In this paper, we will present the details of our approach, the methods used, the dataset utilized, and the performance metrics of our predictive model. We will also discuss the implications and potential challenges of implementing this approach in real-world healthcare settings.

We believe that our work can contribute significantly to the field of diabetes prediction and highlight the importance of leveraging data-driven approaches in combating the global burden of diabetes. By empowering individuals and healthcare providers with accurate predictions, we can take proactive steps towards a healthier future for those at risk of diabetes.

Diabetes is a prevalent global public health issue that impacts 74% of the world's population. The aetiology of diabetes remains uncertain, however, researchers hypothesise that it is associated with genetic predisposition and environmental influences. According to the International Diabetes Federation (IDA), there are already 460 million individuals with diabetes globally, and an additional 815 million individuals will acquire the condition during the next 25 years. Early detection of the illness is crucial to halt its progression. Early identification is crucial in halting the progression of diabetes, as there is currently no solution for this chronic condition. Early diagnosis allows for effective management of the condition by appropriate therapy, consistent diet, and medication. Nevertheless, a diagnosis that is postponed might lead to cardiovascular ailments and significant damage to other organs. Clinical and physical data, such as plasma glucose concentration, serum insulin levels, body mass index (BMI), and age, are commonly utilised for the early diagnosis of diabetes. Based on these data, a physician does the diagnosis of the illness. Nevertheless, the process of making a medical diagnosis is an arduous undertaking for physicians and can be time-consuming. Furthermore, the doctor's choices may be fallacious and influenced by personal prejudice. Hence, the disciplines known as data mining and machine learning are commonly employed as a means of decision support to swiftly and precisely identify illnesses based on data.

Machine learning methods are commonly employed to analyse large datasets using artificial intelligence, with the goal of interpreting the data using regression, classification, or clustering techniques. These algorithms enable the learning of the link between them by using samples and observations of the data. Commonly employed machine learning techniques for this purpose include artificial neural networks (ANN), support vector machines (SVM), k-nearest neighbours (k-NN), decision trees (DT), and naïve Bayes (NB). These approaches immediately establish the relationship between the input and target data. Nevertheless, due to advancements in artificial intelligence and computer processors over the past decade, artificial neural networks (ANN) have been enhanced and deep learning, which combines feature extraction and classification, has gained prominence. Deep learning has provided a significant edge over standard machine learning approaches, particularly in big data applications. The convolutional neural network (CNN) is the most often employed model in deep-learning-based medical diagnostic and detection applications. CNN models are highly sought after because of their complex structure and ability to capture detailed features.

As the architecture of CNN is meant to be end-to-end, it takes raw data as input and produces classes as output. Hence, the meticulously crafted architecture plays a crucial role in determining the performance of the CNN model. Recently, researchers have started using transfer learning applications and implementing well-known CNN architectures as ResNet, GoogleNet, Inception, Xception, VGGNet, and others. Utilising pre-trained or pre-designed convolutional neural network (CNN) architectures has proven to be advantageous in several data-driven research, offering improved performance and ease.

1.1. Existing Research Utilising Artificial Intelligence for Diabetes Prediction

This work utilises deep learning techniques to predict diabetes by utilising the PIMA dataset. Typically, research conducted on predicting diabetes relies on either machine learning or deep learning techniques.

Here are some examples of research that used machine learning approaches to predict diabetes using the PIMA dataset [33] conducted diabetes detection using a combination of Support Vector Machines (SVM) and feedforward neural network, known as an ensemble. To do this, the outcomes derived from the separate classifiers were merged using a predominant voting approach. The ensemble technique yielded superior results compared to the individual classifiers, with a success rate of 78.58%. Sneha and Gangil utilised several machine learning techniques, including naïve Bayes (NB), SVM, and logistic regression, to predict diabetes. The highest level of accuracy was achieved using Support Vector Machine (SVM) with a precision of 69.56%. Furthermore, the authors implemented feature selection techniques on the PIMA dataset. The characteristics exhibiting a poor correlation were eliminated. Edeh et al. conducted a comparative analysis of four machine learning algorithms, namely Bayes, decision tree (DT), SVM, and random forest (RF), for the purpose of predicting diabetes. The analysis was performed on two distinct datasets. The experimental findings using the PIMA dataset showed that the Support Vector Machine (SVM) achieved the maximum accuracy, reaching 76.25%. Chen et al. applied preprocessing techniques to reorganise the PIMA data and utilised the k-means method for data reduction by removing misclassified data. Subsequently, the reduced data was categorised using a decision tree (DT). The study accurately predicted diabetes with a precision of 94.57%. Dadgar and Kaardaan suggested a hybrid methodology for predicting diabetes. Initially, the UTA algorithm was employed to do feature selection. Next, the chosen characteristics were inputted into a two-layer neural network (NN) and its weights were adjusted using a genetic algorithm (GA). The estimation of diabetes was supplied with a precision of 91.22%. Zou et al. employed decision tree (DT), random forest (RF), and neural network (NN) models to forecast diabetes. In addition, they employed principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) techniques to decrease the number of dimensions. RF outperformed the other methods, with a higher accuracy rate of 79.57%. To explore additional suggested research endeavours centred around machine learning, one might analyse the studies conducted by Choudhury and Gupta, as well as Rajeswari and Prabhu. Below are few research that employ deep learning models with the PIMA dataset: Ashiquzzaman et al. developed a network design for predicting diabetes, which includes an input layer, fully connected layers, dropouts, and an output layer. The PIMA dataset characteristics were directly inputted into the proposed MLP, resulting in an accuracy of 89.57% at completion of the application. Massaro et al. generated synthetic recordings and applied long short-term memory (LSTM) to classify this data, resulting in LSTM-AR. The classification result of LSTM-AR, reported as 86.54%, outperformed both LSTM and the multi-layer perceptron (MLP) with previously conducted cross validation. Kannadasan et al. developed a sophisticated deep neural network that utilises stacked autoencoders to extract information and use softmax for diabetes classification. The implemented deep architecture achieved an accuracy of 89.21%.

Rahman et al. introduced a model that relies on convolutional LSTM (Conv-LSTM). In addition, they conducted experiments using regular LSTM and CNN models in order to compare the outcomes. The grid search technique was utilised for hyperparameter optimisation in deep models. All models utilised a one-dimensional (1D) input layer. Following the separation of training and test data, the Conv-LSTM model demonstrated superior performance compared to other models, with an accuracy of 88.68%. Alex and colleagues developed a one-dimensional convolutional neural network (CNN) structure for the purpose of predicting diabetes. Nevertheless, the missing numbers were rectified through the process of outlier identification. Next, the data was preprocessed using the synthetic minority oversampling method (SMOTE) to address the imbalance in the data. Subsequently, the processed data was inputted into the 1D CNN architecture, resulting in an accuracy of 86.29%. To explore further applications of deep learning for diabetes prediction, it is recommended to review the works conducted by Zhu et al. and Fregoso-Aparicio et al.

Previous research indicates that the PIMA dataset is often employed in the fields of machine learning, 1D-CNN, and LSTM architectures. The PIMA dataset's numerical structure imposes limitations on the range of classification and extraction of features techniques available to researchers. In this study, the researchers address this constraint by transforming numerical data into graphic representations. Therefore, the PIMA numerical dataset may be used with well-known CNN models as ResNet, VGGNet, and GoogleNet.

1.2. The Structure, Purpose, Differences and Contribution of the Study

Upon reviewing the prior research stated in Section 1.1., it becomes evident that a range of machine learning and deep-learning-based applications have been able to accurately predict diabetes using the PIMA dataset, which consists of clinical data records. Like the PIMA dataset, much clinical data in the medical domain consist of numerical values. Conventional machine learning techniques commonly utilise numerical values directly. In studies involving machine learning models like SVM, NB, RF, DT, etc., raw data or minimally processed data is fed directly to the model. The output is then assigned target values ranging from 0 (negative) to 1 (positive). Studies employ deep architecture to construct models that utilise the same data to feed the PIMA features, either through the 1D convolution layer or the fully linked layers. The researchers Massaro, Maritati, Giannone, Convertini, and Galiano utilised a recurrent-neural-network (RNN)-based LSTM to analyse the PIMA dataset, which consists of 1D data. However, LSTM was specifically created for analysing sequential data, but the PIMA dataset consists of independent data points.

Deep learning has excelled traditional machine learning approaches in several aspects, leading to its increased popularity in recent years. Due to their advanced features, especially deep CNN models, they have demonstrated superior performance, particularly in computer vision applications. Nevertheless, because to the presence of numerical values in the PIMA dataset, researchers have been compelled to develop one-dimensional convolutional neural network (1D CNN) models.

CNN models are primarily designed for computer vision tasks, specifically for processing 2D data. As a result, the input layer of these models only takes 2D data. These models are used in transfer learning applications. Consequently, the use of popular CNN models for feature extraction and diabetes prediction based on this PIMA dataset, which comprises independent numerical data, has not been demonstrated. Hence, to enhance the accuracy of diagnoses, the raw data might be subjected to modification using well-known CNN models.

In order to circumvent this constraint, this work transforms each sample in the PIMA dataset into pictures, specifically diabetic images. Every picture related to diabetes has cells that reflect certain characteristics in the PIMA dataset. The ReliefF feature selection technique was employed to enhance the prominence of features with strong correlation in the picture. Once each feature is positioned on the picture based on its significance, data augmentation is implemented on these images. One key addition of this work is the straightforward implementation of data augmentation for diabetic data. This is particularly significant since, in comparison to numerical data, data augmentation for photos is a more often used and simpler approach. The enhanced visual data is subsequently inputted into the ResNet18 and ResNet50 convolutional neural network models for the purpose of diabetes prediction. To enhance the existing outcomes, the characteristics of both models are combined and categorised using Support Vector Machine (CNN-SVM). Ultimately, the ReliefF method is employed for feature selection from a multitude of fusion features, and these chosen features are then categorised using SVM. Upon completion of the research, all of these outcomes are compared. The results indicate that the CNN-SVM structure, using chosen fusion characteristics, achieves superior diabetes prediction compared to other methods. Furthermore, the efficacy of the suggested approach is demonstrated by comparing its results with those of earlier research. The suggested method's contributions can be summarized as follows:

- A diabetes prediction application with a comprehensive framework is proposed.
- The PIMA dataset, which contains numeric values, is transformed into pictures.
- The use of numerical diabetic data in conjunction with well-known CNN models is made available.
- The significance of the characteristics is considered during the conversion to the picture.
- The proposed technique surpasses the majority of prior investigations.

2. PIMA Indians Diabetes Dataset

This study utilises the PIMA Indians Diabetes dataset, sourced from the Kaggle data repository, which is commonly employed for predicting diabetes. The date of access was 10 September 2022. The dataset was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to determine whether a patient has diabetes or not, using specific diagnostic metrics included in the collection. All patients in this facility are specifically PIMA Indian women who are 19 years of age or older.

The dataset comprises a collection of clinical and physical features, together with their corresponding measures and ranges. The variables measured in this study include the number of pregnancies (ranging from 0 to 19), glucose levels (ranging from 0 to 210), blood pressure (measured in mm Hg and ranging from 30 to 145), skin thickness (measured in mm and ranging from 0 to 90), insulin levels (measured in mu U/mL and ranging from 0 to 972), BMI (measured in kg/m2 and ranging from 0 to 74.32), diabetes pedigree function (PDF) values (ranging from 0.065 to 3.87), age (measured in years and ranging from 19 to 78), and outcome (represented as a Boolean value, either 0 or 1). The dataset consists of 768 samples, each containing 7 numerical features. Table 1 displays a limited number of examples extracted from the dataset.

Table 1. Some examples of Pima Indians Diabetes dataset

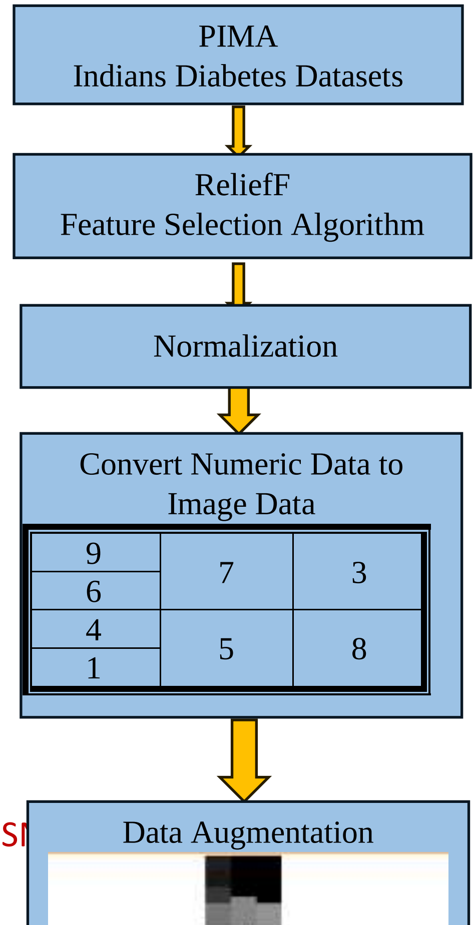
Pregnancies [0-19]	Glucose [0-210]	Blood pressure [30-145]	Skin thickness [0-90]	Insulin [0-972]	Body mass index [0-74]	Diabetes pedigree function [0.065-3.87]	Age [19-78]	Outcome [0-1]
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1

10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1

3. Methodology

This section will provide a detailed explanation of the procedures employed to ascertain the diabetes status of patients. The procedure of the suggested approach is illustrated in Figure 1. The feature selection process begins by selecting the most valuable features from the numerical data, as demonstrated.

As seen in Figure 1. After normalising the numerical data, the bounds of all features are adjusted for the conversion from numeric to picture. The process of converting numerical data to images is implemented in a manner that prioritises the most impactful features identified by the feature selection algorithm. Data augmentation strategies are employed to enhance the classification accuracy of deep ResNet models. The three ResNet-based methodologies proposed in this work are employed for the purpose of categorising data at the ultimate phase. Here, we will provide a more detailed explanation of each of these procedures.



0 1 0 1 0 1

Figure 1. Application steps of proposed methods.

3.1. ReliefF Feature Selection Algorithm

In order to enhance the ability to classify, researchers have investigated various techniques for reducing the number of features in the existing body of knowledge. ReliefF is a distance-based feature selector commonly used in literature. ReliefF, created by Kira and Rendell in 1992, is a highly effective technique for filtering features.

Dimension reduction techniques assist in eliminating unnecessary attributes from a dataset. These technologies facilitate data compression, resulting in storage space conservation.

Additionally, it decreases the time needed for computational complexity and minimises the duration required to achieve the same objective.

In 1994, Kononenko enhanced the algorithm for addressing multi-class problems. This algorithm facilitates successful feature selection. The ReliefF algorithm is extremely efficient and does not impose any limitations on the characteristics of the data. The ReliefF approach aids in resolving many types of problems by selecting the closest neighbouring samples from each sample in every category. ReliefF aims to uncover the relationships and coherence present in the attributes of the dataset. Moreover, it is possible to identify the important characteristics in the dataset by creating a model that considers the closeness to samples of the same class and the distance to samples of different classes. The ReliefF model selects neighbouring attributes that are closest to each other from samples of unique quality. In this model, the dataset is partitioned into two components: training data and test data. Ri random samples are taken from the training set, and the difference function di f f is employed to determine the closest neighbours of both the same and different classes, in order to identify the nearest neighbours to the chosen Ri sample, as shown in Equation (1). The di f f function is used to calculate the distance between instances while determining nearest neighbours. The total distance is calculated by adding up all attribute discrepancies, which is also known as the Manhattan distance.

Equation (1) is utilised to calculate the disparity between two distinct samples, I1 and I2, for the attribute A, and to ascertain the minimum distance between the samples. The closest neighbour H from the same class and the closest neighbour M from a different class are selected. The proximity between adjacent sample Af within the same class and across different classes is assessed by considering the values of Ri, M, H, and the weighting vector of the dataset. The WAf weight is determined using a comparison process that assigns lower importance to traits that are further away. The aforementioned processes are executed m times for every attribute, resulting in the computation of weight values for each attribute. The weights are modified using Equation (2).

$$dif\ f\left(A, I_{1,}I_2\right) = \frac{|value(A, I_1) - value(A, I_2)|}{\max(A) - \min(A)} \quad - - - - - - - - - - - -(1)$$

$$W_{new}(A_f) = W_{old}(A_f) + \frac{diff\ (A_f, R_i, M)}{m} - \frac{diff\ (A_f, R_i, H)}{m} - - - - - - (2)$$

After using the ReliefF feature selection method on the PIMA dataset features, the significance weight of each feature is displayed in Figure 2. The number of nearest neighbours was determined to be 10. According to Figure 2, the most

The ReliefF technique was used to identify the effective features from the PIMA numerical data.

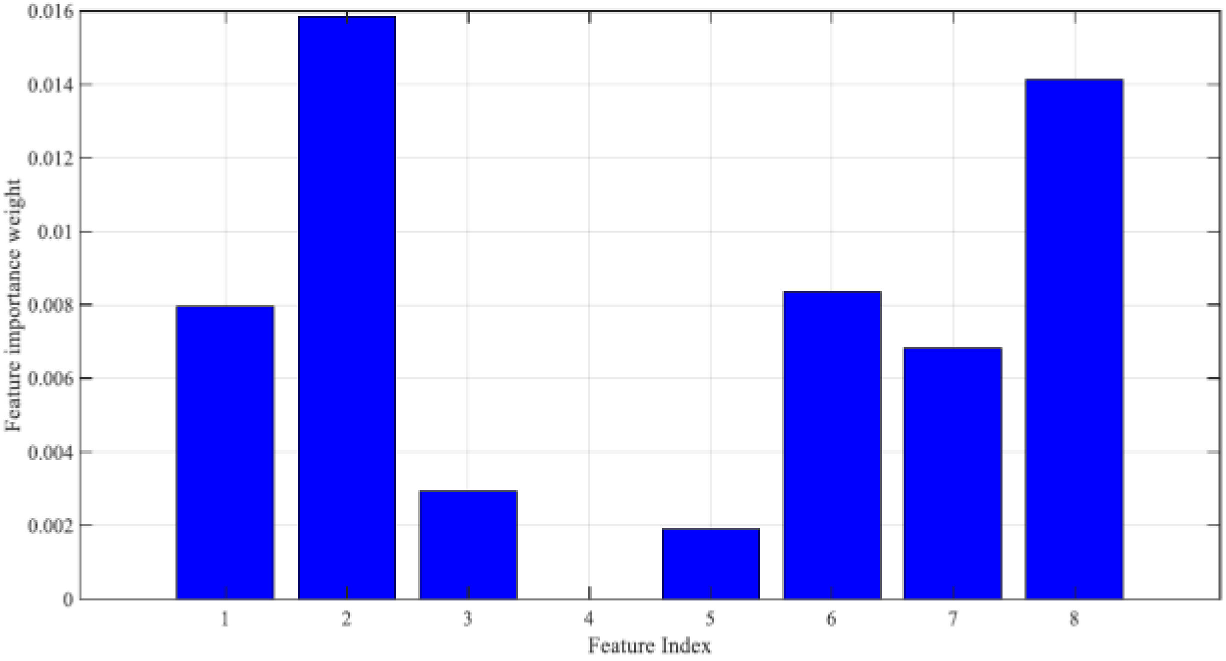


Figure 2. Importance weight of features in the PIMA dataset

3.2. Normalization of Data

Normalising data with a large number of features is a well-established procedure in the field of artificial intelligence. Various functionalities have distinct limitations. Normalising the characteristics by setting them to the same or similar range enhances the learning performance. The PIMA dataset as well

Exhibits distinct minimum and maximum values, as depicted in Table 1. Therefore, it is imperative to normalize these values. Furthermore, normalization is crucial for the conversion process of numeric-to-image in the suggested method, as it ensures that each feature's value is accurately placed on the corresponding image representing the sample. The cell in the associated image appears brighter in colour based on the amplitude of the feature. Thus, it is necessary for all features to have identical maximum and lowest values.

Feature scaling is the recommended approach for normalizations. This method involves rescaling the feature values to a specific range. The feature scaling method employed in this investigation is the min-max normalizations method. In this approach, the updated value of the sample (ˆ x) is

The determination is made based on the maximum (x_{max}) and minimum (x_{min}) values of the characteristics.

Normalisation ensures that all features are scaled to a range of 0 to 1. During the application step, the process of normalisation is implemented on eight specific features inside the PIMA dataset. Figure 3 illustrates that following the normalisation process, the glucose levels range from 0 to 210 and the blood pressure data range from 0 to 145, both scaled to a range of 0 to 1. Formula (3) presents the equation for the min-max normalisation technique.

$$\hat{x} = \frac{X - X_{min}}{x_{max} - x_{min}} ------------------ (3)$$

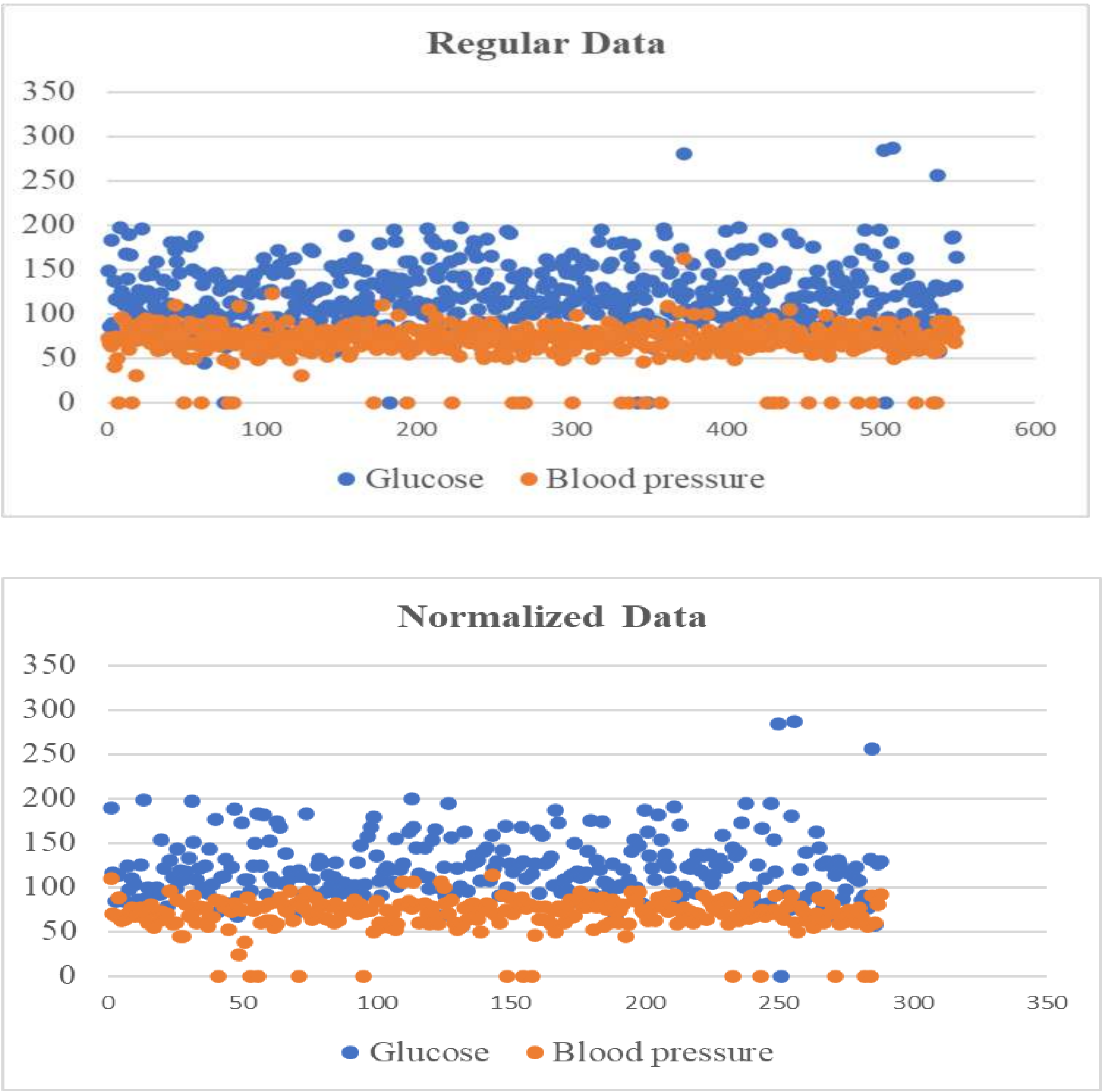


Figure 3. Min–max normalization of PIMA dataset

3.3. Conversion of Numeric Data to Image Data

Despite a significant recent growth in the quantity of picture data in the medical profession, there remains a substantial amount of numerical data available. While numerical numbers can be readily and inexpensively acquired, the analysis and understanding of this data is often conducted.

using machine learning techniques. Recently, there has been a preference for using 1D CNN structures in deep architecture investigations. These structures are utilised to process numerical values as input. This preference arises because popular CNN models, which have shown considerable advances in computer vision, cannot be directly applied to this type of data. For these models, the input layer requires 2D data as input.

CNN models like ResNet, VGGNet, GoogleNet, and others are specifically created with an architecture that is well-suited for processing image data. Hence, the incapacity to examine datasets comprising 1D samples using these robust models is a significant drawback in terms of both the range of applications and the accuracy of predictions. This section explores the process of converting numerical data into visual representations.To address this constraint in the PIMA dataset, which consists of numerical data, a solution needs to be found.

When converting PIMA data to pictures, the brightness of a given region (cell) in the image is determined based on the amplitude of each feature. Indeed, every characteristic can be regarded as a component of the puzzle formed by the example image. Each sample in the PIMA dataset utilises a 120 x 120 picture structure, as depicted in Figure 4.

The index assigned to each cell corresponds to the feature index in the PIMA dataset. Figure 4 displays the positions of features in a sample image. Figure 4 depicts the precise location and size of features, which are defined based on their significance rather than being chosen arbitrarily. Figure 2 displays the order of relevance of features resulting from the ReliefF algorithm as 2-8-6-1-7-3-5-4. Consequently, a larger cell is designated for the trait that is deemed more significant. The colour of each cell is determined by the magnitude of the relevant feature value. Since the data has been normalised, every feature value falls within the range of 0 to 1. The brightness values of the cells in the photos are obtained by multiplying each feature value by 255, resulting in values ranging from 0 to 255. Consequently, the images that are produced are in shades of grey. Figure 4 displays several exemplar photos related to diabetes.

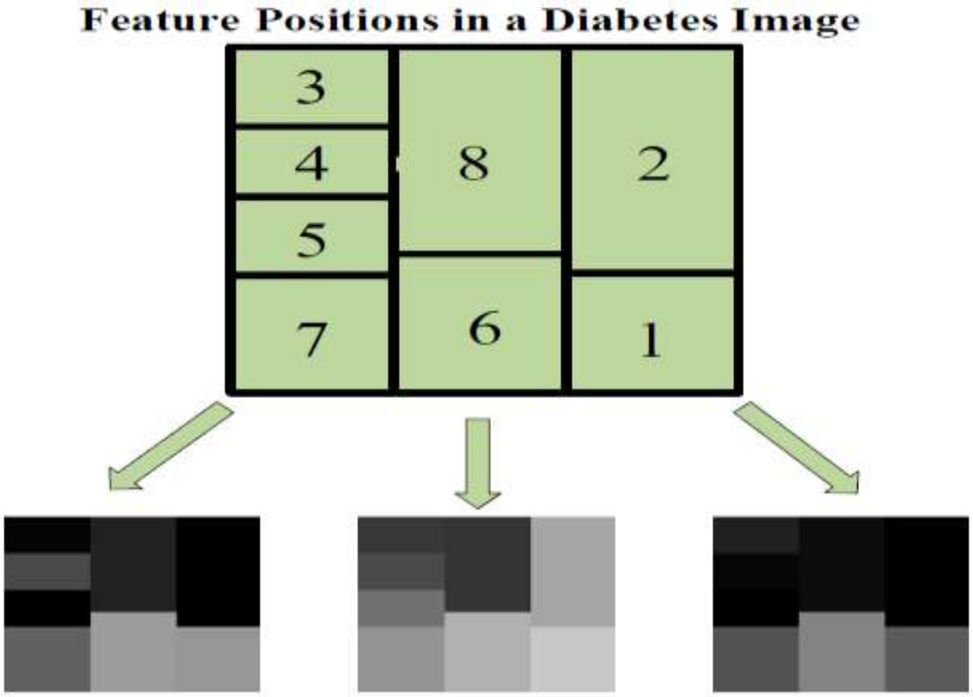


Figure 4. Conversion selected features to image (numeric to image).

3.4. Data Augmentaion

The quantity of samples has a direct impact on the efficacy of deep learning methods.Nevertheless, it is not always feasible to retrieve a substantial amount of data. Consequently, researchers enhance the size of a training dataset by generating altered versionsof the photos contained in the collection. Data augmentation approaches refer to the application of certain methods to raw photos in order to achieve this objective.

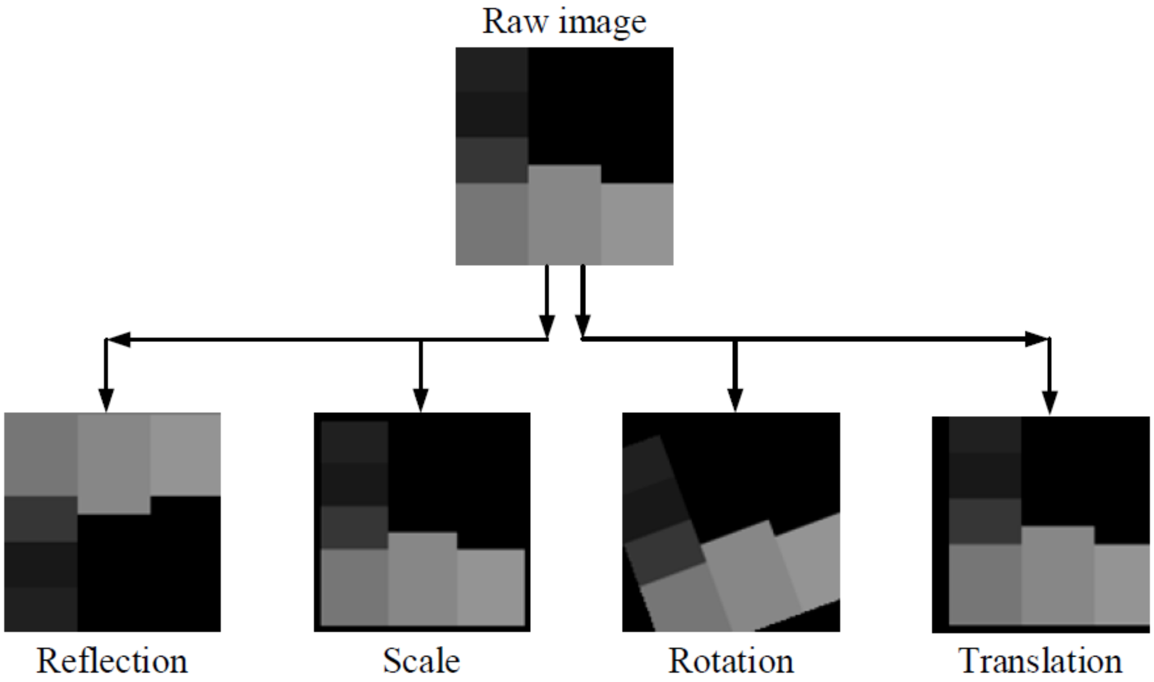


Figure 5. Data augmentation methodologies and sample augmented images.

3.5. Diabetes Prediction via ResNet Models

The CNN model is fed with images that have undergone data augmentation and are divided into 80% for training and 20% for testing. This study utilises the ResNet18 and ResNet50 models, commonly employed for comparative analysis, to estimate diabetes. ResNet models are widely utilised in numerous studies due to their inherent advantages. The advantage of ResNet lies in its ability to mitigate the issue of disappearing gradients by transmitting residual values to subsequent layers through the usage of residual blocks. ResNet models exist at various depths. The depths of the ResNet18 and ResNet50 models utilised in this investigation are 18 and 50, correspondingly.

This study utilises pre-existing models for diabetes detection instead of developing a novel convolutional neural network architecture. We apply current ResNet models to our work with just slight alterations, specifically fine-tuning. Both models undergo a process where the final two layers of the present models are eliminated and substituted with two fully linked output layers and a classification (softmax) layer. Furthermore, although the diabetic images generated have dimensions of 120 x 120, the required input size for ResNet models is 224 x 224. Consequently, all images related to diabetes are adjusted in size before and during the training process (refer to Figure 6). The findings section will provide an analysis of the outcomes achieved during the training and testing phases.

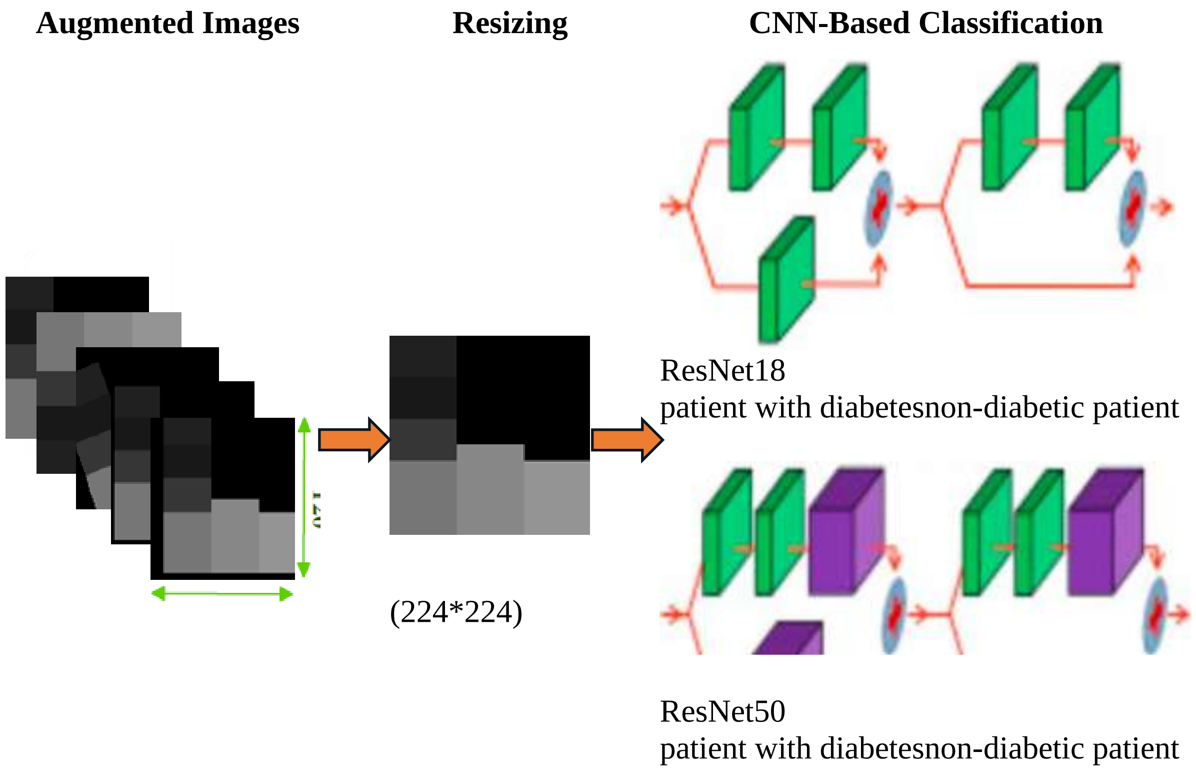

Augmented Images
Resizing
CNN-Based Classification
120
(224*224)
ResNet18
patient with diabetesnon-diabetic patient
ResNet50
patient with diabetesnon-diabetic patient

Figure 6. Classification of diabetes images as diabetic (1) and nondiabetic (0) with ResNet models.

4. Results and Discussion

The examined aspects of the suggested approach are outlined. The diabetes prediction deep learning programmes were executed on a laptop equipped with an Intel Core i7-7700HG processor, NVIDIA GeForce GTX 1050 4 GB graphics card, and 16 GB of RAM. The apps were constructed within the Matlab environment. Figure 8 can serve as a point of reference for the software design or code implementation of the proposed strategy. The algorithm for the method was devised according to Figure 8. The utilisation of toolbox and libraries directly during the developing process helped to mitigate programme complexity. The toolboxes utilised in this particular scenario are focused on Machine Learning.

The three toolboxes are Toolbox, Deep Learning Toolbox, and Image Processing Toolbox.

To showcase the superiority of the proposed strategy, results are generated using three distinct ways. Detailed methodological information regarding these three approaches has been provided in the preceding section. The initial method involves utilising fine-tuned ResNet models for classifying and predicting diabetes based on diabetic data.

Images following the process of data augmentation. To do this, the ResNet18 and ResNet50 models undergo finetuning, and the output layer is modified to align with the two classes. Next, the dataset of 5400 diabetic photos is split into two groups: 80% of the images are used as training data, while the remaining 20% are used as test data. The models are trained using the training data, and the performance of the network is evaluated using the test data. The second strategy involves extracting deep characteristics from two sources.The fusion features (2660) are classified by combining fine-tuned ResNet modelsby the SVM machine learning method.

5. Conclusions, Discussion and Future Works

Diabetes is a persistent medical condition that restricts individuals' everyday functioning, diminishes their overall well-being, and heightens the likelihood of mortality. Historically, machine learning and deep neural network (DNN) solutions have been created using clinical data, specifically for the purpose of predicting diabetes through various research.

has been conducted. Although these research have yielded promising results, the quantitative structure of clinical registry data has restricted the use of widely used CNN models. This study employed widely recognised CNN models to ascertain the diagnosis of diabetes. In this study, the numerical clinical patient data (PIMA dataset) were translated into pictures since CNN models require two-dimensional data input. Thus, every characteristic was incorporated into the example image. This technique was executed with deliberate intention, and the most impactful aspect was enhanced to be more prominent in the photograph. The ReliefF feature selection approach was employed to identify the most impactful features in this procedure. After augmenting the data and resizing the photos to fit the ResNet model, diabetes prediction was performed using three distinct methods.

The first technique effectively categorised diabetes photos using the finetuned ResNet18 and ResNet50 models. The second strategy utilised Support Vector Machines (SVM) to classify a total of 2560 deep features that were retrieved from the fully linked layers of both ResNet models.

The previous method involved selecting the top 500 deep features using the ReliefF feature selection algorithm, and then classifying these selected features using SVM. The third strategy yielded the most successful prediction. The SVM/cubic model achieved a classification accuracy of 92.19% when employing 500 chosen features. The image data underwent all of these classifications. The conversion of the data to image format eliminated the restriction on the algorithm that may be used to the PIMA dataset. By utilising various CNN models, such as those capable of extracting intricate and advanced features, the PIMA dataset or similar numerical data can be effectively analysed. An application that includes image data can be analysed in a more varied and complete manner compared to an application that includes numerical data. This is due to the extensive range of artificial intelligence combinations that can be applied to the image data. The ResNet18 and ResNet50 models used in this study have shown superior outcomes compared to earlier studies. For instance, the quantity and diversity of features can be augmented by employing several CNN models. According to the available evidence, the experimental findings have demonstrated that the conversion of clinical data into visuals is a highly effective technique.

The methodology suggested in this research can also be utilised for various numerical datasets.Studies utilising deep learning techniques have decreased reliance on specific features, allowing the designed architecture to become more prominent. Nevertheless, the approach suggested in this research is valuable since it allows for the utilisation of intricate and all-encompassing structures for numerical data.

This application may require additional processing processes compared to studies that directly use raw data.Generating picture data allows for enhanced diabetes prediction performance by leveraging CNN models that can now be applied to numerical data across various architectures. Furthermore, diabetic photos can now be readily subjected to data augmentation. Furthermore, the application findings indicate that the integration of fusion features in the CNN-SVM architecture significantly enhances the level of success. In addition, by utilising specific characteristics, CNN-SVM offers higher accuracy in predictions while being more cost-effective. These situations can be explained by the significant tendency of experimental simulation studies. The system's performance is enhanced by the chosen fusion characteristics, despite their limited quantity. Furthermore, the CNN-SVM architecture is highly efficient. Utilising various applications with reduced complexity, enhanced efficiency, and greater versatility enhances the classification accuracy of the system.

REFERENCES

[1]. Data base in pima-indians-diabetes-database https://data.world/data-society/pima-indians-diabetes-database

[2]. Bhoi, S.K. Prediction of diabetes in females of pima Indian heritage: A complete supervised learning approach. Turk. J. Comput. Math. Educ. (TURCOMAT) 2021, 12, 3074–3084.

[3]. Mellitus, D. Diagnosis and classification of diabetes mellitus. Diabetes Care 2005, 28, S5–S10.

[4]. Madan, P.; Singh, V.; Chaudhari, V.; Albagory, Y.; Dumka, A.; Singh, R.; Gehlot, A.; Rashid, M.; Alshamrani, S.S.; AlGhamdi, A.S. An Optimization-Based Diabetes Prediction Model Using CNN and Bi-Directional LSTM in Real-Time Environment. Appl. Sci. 2022, 12, 3989. [CrossRef]

[5]. Pippitt, K.; Li, M.; Gurgle, H.E. Diabetes mellitus: Screening and diagnosis. Am. Fam. Physician 2016, 93, 103–109. [PubMed]

[6]. Xu, G.; Liu, B.; Sun, Y.; Du, Y.; Snetselaar, L.G.; Hu, F.B.; Bao, W. Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: Population based study. BMJ 2018, 362, k1497. [CrossRef] [PubMed]

[7]. Chiefari, E.; Arcidiacono, B.; Foti, D.; Brunetti, A. Gestational diabetes mellitus: An updated overview. J. Endocrinol. Investig. 2017, 40, 899–909. [CrossRef] [PubMed]

[8]. Mirghani Dirar, A.; Doupis, J. Gestational diabetes from A to Z. World J. Diabetes 2017, 8, 489–511. [CrossRef] [PubMed]

[9]. Sarwar, A.; Ali, M.; Manhas, J.; Sharma, V. Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int. J. Inf. Technol. 2020, 12, 419–428. [CrossRef]

[10]. Zhou, H.; Myrzashova, R.; Zheng, R. Diabetes prediction model based on an enhanced deep neural network. EURASIP J. Wirel. Commun. Netw. 2020, 2020, 148. [CrossRef]

[11]. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pract. 2018, 138, 271–281. [CrossRef]

[12]. Azrar, A.; Ali, Y.; Awais, M.; Zaheer, K. Data mining models comparison for diabetes prediction. Int. J. Adv. Comput. Sci. Appl. 2018, 9, 320–323. [CrossRef]

[13]. Larabi-Marie-Sainte, S.; Aburahmah, L.; Almohaini, R.; Saba, T. Current Techniques for Diabetes Prediction: Review and Case Study. Appl. Sci. 2019, 9, 4604. [CrossRef]

[14]. Jahani, M.; Mahdavi, M. Comparison of predictive models for the early diagnosis of diabetes. Healthc. Inform. Res. 2016, 22, 95–100. [CrossRef] [PubMed]

[15]. Ayon, S.I.; Islam, M. Diabetes Prediction: A Deep Learning Approach. Int. J. Inf. Eng. Electron. Bus. 2019, 11, 21.

[16]. Naz, H.; Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J. Diabetes Metab. Disord. 2020, 19, 391–403. [CrossRef]

[17]. Aslan, M.F.; Celik, Y.; Sabanci, K.; Durdu, A. Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. Int. J. Intell. Syst. Appl. Eng. 2018, 6, 289–293. [CrossRef]

[18]. Asiri, N.; Hussain, M.; Al Adel, F.; Alzaidi, N. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. Artif. Intell. Med. 2019, 99, 101701. [CrossRef] [PubMed]

[19]. Aslan, M.F.; Sabanci, K.; Durdu, A.; Unlersen, M.F. COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization. Comput. Biol. Med. 2022, 142, 105244. [CrossRef] [PubMed]

[20]. Baashar, Y.; Alkawsi, G.; Alhussian, H.; Capretz, L.F.; Alwadain, A.; Alkahtani, A.A.; Almomani, M. Effectiveness of artificial intelligence models for cardiovascular disease

prediction: Network meta-analysis. Comput. Intell. Neurosci. 2022, 2022, 5849995. [CrossRef]

[21]. Saba, T.; Mohamed, A.S.; El-Affendi, M.; Amin, J.; Sharif, M. Brain tumor detection using fusion of hand crafted and deep learning features. Cogn. Syst. Res. 2020, 59, 221–230. [CrossRef]

[22]. Abuhmed, T.; El-Sappagh, S.; Alonso, J.M. Robust hybrid deep learning models for Alzheimer's progression detection. Knowl. -Based Syst. 2021, 213, 106688. [CrossRef]

[23]. Kaur, S.; Singla, J.; Nkenyereye, L.; Jha, S.; Prashar, D.; Joshi, G.P.; El-Sappagh, S.; Islam, M.S.; Islam, S.M.R. Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. IEEE Access 2020, 8, 228049–228069. [CrossRef]

[24]. Mirbabaie, M.; Stieglitz, S.; Frick, N.R.J. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. Health Technol. 2021, 11, 693–731.

[25]. Sun, H.; Liu, Z.; Wang, G.; Lian, W.; Ma, J. Intelligent Analysis of Medical Big Data Based on Deep Learning. IEEE Access 2019, 7, 142022–142037. [CrossRef

[26]. Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. Electron. Mark. 2021, 31, 685–695. [CrossRef]

[27]. Aslan, M.F.; Sabanci, K.; Durdu, A. A CNN-based novel solution for determining the survival status of heart failure patients with clinical record data: Numeric to image. Biomed. Signal Process. Control 2021, 68, 102716. [CrossRef] 27.

[28]. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

[29]. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

[30]. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

[31]. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.

[32]. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, preprint. arXiv:1409.1556.

[33]. Kırba¸s, ˙I.; Çifci, A. An effective and fast solution for classification of wood species: A deep transfer learning approach. Ecol. Inform. 2022, 69, 101633. [CrossRef]

[34]. Zolfaghari, R. Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm. Int. J. Comput. Eng. Manag/ 2012, 15, 2230–7893.

[35]. Sneha, N.; Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. J. Big Data 2019, 6, 13. [CrossRef]

[36]. Edeh, M.O.; Khalaf, O.I.; Tavera, C.A.; Tayeb, S.; Ghouali, S.; Abdulsahib, G.M.; Richard-Nnabu, N.E.; Louni, A. A Classification Algorithm-Based Hybrid Diabetes Prediction Model. Front. Public Health 2022, 10, 829519. [CrossRef]

[37]. Chen, W.; Chen, S.; Zhang, H.; Wu, T. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In Proceedings of the 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 24–26 November 2017; pp. 386–390.

[38]. Dadgar, S.M.H.; Kaardaan, M. A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes. Int. J. Mechatron. Electr. Comput. Technol. (IJMEC) 2017, 7, 3397–3404.

[39]. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 2018, 9, 515. [CrossRef] [PubMed]

[40]. Choudhury, A.; Gupta, D. A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques; Springer: Singapore, 2019; pp. 67–78.

[41]. Rajeswari, M.; Prabhu, P. A review of diabetic prediction using machine learning techniques. Int. J. Eng. Tech. 2019, 5, 2395-1303.

[42]. Ashiquzzaman, A.; Tushar, A.K.; Islam, M.; Shon, D.; Im, K.; Park, J.-H.; Lim, D.-S.; Kim, J. Reduction of overfitting in diabetes prediction using deep learning neural network. In IT Convergence and Security 2017; Springer: Singapore, 2018; pp. 35–43.

[43]. Tuncer, T.; Dogan, S.; Ozyurt, F. An automated residual exemplar local binary pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image. Chemom. Intell. Lab. Syst. 2020, 203, 104054. [CrossRef]

[44]. Kilicarslan, S.; Adem, K.; Celik, M. Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network. Med. Hypotheses 2020, 137, 109577. [CrossRef] [PubMed]

[45]. Kononenko, I. Estimating attributes: Analysis and extensions of RELIEF. In Proceedings of the European conference on machine learning, Catania, Italy, 6–8 April 1994; pp. 171–182.

[46]. Özyurt, F. Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures. J. Supercomput. 2020, 76, 8413–8431. [CrossRef]

[47]. Venkataramana, L.; Jacob, S.G.; Ramadoss, R.; Saisuma, D.; Haritha, D.; Manoja, K. Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data. Genes Genom. 2019, 41, 1301–1313. [CrossRef]

[48]. Kononenko, I.; Robnik-Sikonja, M.; Pompe, U. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. Artif. Intell. Methodol. Syst. Appl. 1996, 31–40.

[49]. Ardakani, A.A.; Kanafi, A.R.; Acharya, U.R.; Khadem, N.; Mohammadi, A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput. Biol. Med. 2020, 121, 103795. [CrossRef]

[50]. Koklu, M.; Unlersen, M.F.; Ozkan, I.A.; Aslan, M.F.; Sabanci, K. A CNN-SVM study based on selected deep features for grapevine leaves classification. Measurement 2022, 188, 110425. [CrossRef]

[51]. Srivastava, S.; Sharma, L.; Sharma, V.; Kumar, A.; Darbari, H. Prediction of Diabetes Using Artificial Neural Network Approach; Springer: Singapore, 2019; pp. 679–687.

[52]. Kalagotla, S.K.; Gangashetty, S.V.; Giridhar, K. A novel stacking technique for prediction of diabetes. Comput. Biol. Med. 2021, 135, 104554. [CrossRef]

[53]. Jakka, A.; Vakula Rani, J. Performance evaluation of machine learning models for diabetes prediction. Int. J. Innov. Technol. Explor. Eng. (IJITEE) 2019, 8, 1976–1980.

[54]. Massaro, A.; Maritati, V.; Giannone, D.; Convertini, D.; Galiano, A. LSTM DSS Automatism and Dataset Optimization for Diabetes Prediction. Appl. Sci. 2019, 9, 3532. [CrossRef]

[55]. Kannadasan, K.; Edla, D.R.; Kuppili, V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clin. Epidemiol. Glob. Health 2019, 7, 530–535. [CrossRef]

[56]. Rahman, M.; Islam, D.; Mukti, R.J.; Saha, I. A deep learning approach based on convolutional LSTM for detecting diabetes. Comput. Biol. Chem. 2020, 88, 107329. [CrossRef]

[57]. Alex, S.A.; Nayahi, J.; Shine, H.; Gopirekha, V. Deep convolutional neural network for diabetes mellitus prediction. Neural Comput. Appl. 2022, 34, 1319–1327. [CrossRef] 46. Zhu, T.; Li, K.; Herrero, P.; Georgiou, P. Deep Learning for Diabetes: A Systematic Review. IEEE J. Biomed. Health Inform. 2021, 25, 2744–2757. [CrossRef]

[58]. Fregoso-Aparicio, L.; Noguez, J.; Montesinos, L.; García-García, J.A. Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. Diabetol. Metab. Syndr. 2021, 13, 1–22. [CrossRef]

[59]. Saleem, T.J.; Chishti, M.A. Deep learning for the internet of things: Potential benefits and use-cases. Digit. Commun. Netw. 2021, 7, 526–542. [CrossRef]

[60]. O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.; Walsh, J. Deep learning vs. traditional computer vision. In Proceedings of the Science and Information Conference, Las Vegas, NV, USA, 2–3 May 2019; pp. 128–144.

[61]. Kira, K.; Rendell, L.A. A practical approach to feature selection. In Machine learning Proceedings 1992; Elsevier: Amsterdam, The Netherlands, 1992; pp. 249–256.

[62]. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. J. Biomed. Inform. 2018, 85, 189–203. [CrossRef] [PubMed]

[63]. Kira, K.; Rendell, L.A. The feature selection problem: Traditional methods and a new algorithm. In Proceedings of the Aaai, San Jose, CA, USA, 12–16 July 1992; pp. 129–134.