



Application of Deep Learning Autoencoders as Features Extractor of Diabetic Foot Ulcer Images

Abbas Saad Alatrany^{1,2(✉)}, Abir Hussain^{1,3}, Saad S. J. Alatrany⁴,
and Dhiya Al-Jumaily¹

¹ School of Computer Science and Mathematics,
Liverpool John Moores University, Liverpool, UK
a.s.alatrany@2020.ljmu.ac.uk

² University of Information Technology and Communications, Baghdad, Iraq

³ Department of Electrical Engineering, University of Sharjah, Sharjah, UAE

⁴ Imam Ja'afar Al-Sadiq University, Baghdad, Iraq

Abstract. Diabetic Foot Ulcer is one of the most common diabetic complications that can lead to amputation if not treated appropriately and timely. When diagnosed by professionals, diabetic foot ulcers can be extremely successful, but the diagnosis comes at a great cost. Therefore, early automated detection tools are required to help diabetic people. In the current work, we test the ability of deep learning autoencoders to extract appropriate features that can be fed to machine learning algorithms to classify normal or abnormal skin areas. The proposed model was trained and tested on 754-foot photos of healthy and diabetic ulcer-affected skin from several individuals. By benchmarking various machine learning algorithms, our extensive research and experiments showed that the features extracted from autoencoder models can generate high accuracy when passed to a support vector machine with a polynomial kernel for early diagnosis, with 0.933 and 0.939 for accuracy and F1 score, respectively.

Keywords: Autoencoder · Deep learning · Machine learning · DUF · Diabetic foot ulcer

1 Introduction

Diabetes mellitus, or diabetes, is a metabolic condition characterised by elevated blood sugar levels [1]. Insulin is a hormone that transports sugar from the bloodstream into cells, where it can be stored or used for energy. When you have diabetes, your body either generates inadequate insulin or is unable to adequately use the insulin it does produce.

Diabetes comes in a variety of forms: Diabetes type 1 is an autoimmune illness. In the pancreas, where insulin is produced, the immune system fights and destroys cells. It's still unclear what's causing this onslaught. This type of diabetes affects approximately 10% of diabetics [2]. The second kind occurs when sugar builds up in the bloodstream and the body becomes insulin resistant, known as type 2 diabetes (T2D) [3]. Finally, prediabetes

is defined as having blood sugar levels that are higher than normal but not high enough to be diagnosed with T2D. Each form of diabetes has its own set of symptoms, causes, and treatments [4].

Diabetic peripheral neuropathy (DPN) is a serious consequence of diabetes that impairs the sensory nerve supply to the feet, causing infections, structural alterations, and the development of diabetic foot ulcers (DFUs). It affects 30–50 percent of persons with diabetes [5]. Recurrent stress over a region prone to high vertical stress is a major cause of diabetic foot ulcers in patients with DPN [6].

According to research analysing 785 million outpatient visits by diabetic patients in the United States over a six-year period, diabetic foot ulcers and associated infections are a substantial risk factor for emergency department visits and hospital admissions [7].

Machine learning (ML) is the study of computer algorithms that aid in the formulation of correct predictions and reactions in specific situations, as well as the intelligent behaviour of humans. Machine learning, in general, is about learning to generate better future conditions based on what has been learnt in the past. Machine learning is the development of programmes that allow us to analyse data from various sources by selecting relevant data and using that data to predict the behaviour of the system in similar or different scenarios [8].

Although previous studies achieved good results, machine learning performance for the early diagnosis of diabetic foot ulcers still needs to be improved. This paper is divided into several sections. Section 2 outlines recent literature. Section 3 defines the materials and methods conducted in the current work, while Sect. 4 demonstrates the results obtained from the experiment. Finally, Sect. 5 concludes the paper.

2 Related Work

Goyal et al. [9] propose a novel CNN architecture named (DFUNet) that combines significant aspects of CNN design. They used convolution layers in both depth and parallel to improve the extraction of important features for DFU classification. Their idea is to decrease the number of layers in the network, and use a larger size filter to learn the feature map from the input images. Parts of the network employ layers in a parallel manner to extract concatenated features using tenfold CV as internal validation, in which the deep learning model reached an AUC of 0.96, outperforming other deep learning architectures such as Alexnet and LeNet.

Das et al. [10] suggested DFU_SPNet, a deep learning model to classify healthy and unhealthy samples. Their network utilised stacked parallel layers with different kernel sizes. This variety in sizes aided in the successful learning of both global and local feature abstractions. There were 1679 image patches in the dataset, the majority of which were annotated as abnormal, and 641 were annotated as normal. The dataset was divided into 80% for training and 20% for testing. The authors trained the network with various optimizers and hyperparameters, and the best combinations reached an AUC of 0.97.

For DFU detection, Yap and his team [11] employed Faster R-CNN [12]. The ML models trained using the dataset from the DFU grand challenge [13] consist of 4000 images divided equally for training and testing. All images were reduced to 640×480 pixels to boost the efficiency of the deep learning algorithms and to minimise processing

expenses. A modified R-CNN model called deformable convolution achieved the best results with an f1-score of 0.74. The paper also concludes that using ensemble technique based on several deep learning algorithms can improve the F1-Score but does not have much effect on the mean average precision.

From a single foot thermograms, the detection results of six deep CNN models for categorizing the thermograms into control and diabetes groups is reported by Khandakar et al. [14]. The dataset contains 167-foot pair thermogram images from 122 diabetic individuals and 45 healthy controls. This study employed five-fold CV, with each fold separated into an 80% training and 20% testing set. The validation set was made up of 20% of the training data. DenseNet201 exceeds the other five CNN models studied, with an overall sensitivity of 94.01% for the detection of Diabetes foot ulceration.

Scebba and colleagues [15] Present a deep learning model named “detect and segment” to segment clinical wound areas in images. Deep neural networks were used to determine the wound’s location and separate it from the background. The model was tested on various datasets of clinical wounds. The Matthews’ correlation coefficient of 0.85 was reported for the diabetic foot ulcer data set.

The following are the aims of the present work:

- To propose and implement a deep learning model by employing autoencoder, based on the idea of compress features volume for the purpose of facilitating the training process of state-of-the-art machine learning algorithms.
- To enhance the classification performance of normal vs. abnormal skin areas.

3 Materials and Methods

3.1 Dataset

The diabetic foot ulcer dataset used in this paper is requested from [16–18]. The data was gathered in the form of a standardised collection of colour images of diabetic foot ulcers from various individuals. The dataset includes 754 photos of diabetic patients’ feet with DFU and healthy skin from the diabetic centre at Nasiriyah Hospital in the south region of Iraq. These photographs were taken using a Samsung Galaxy Note 8 and an iPad at various brightness levels and angles.

Regions of Interest (ROI) was cropped into small patches. This is a considerable area around the ulcer that comprises essential tissues from both normal and abnormal skin classes, and the patches were then annotated by a specialist. A medical professional marked the ground-truth labels in two types of normal and abnormal skin patches. A total of 1609 skin patches were collected, 542 of which were normal and 1067 of which were aberrant DFU. For computational simplicity, the patches have been rescaled to 128×128 pixels in the current work.

3.2 Background of Machine Learning Algorithms

Machine learning is a branch of artificial intelligence in which computer algorithms are used to learn from data independently. This section provides an overview of the ML algorithms used in the current work to classify healthy and unhealthy skin areas.

Support Vector Machine

Machine-learning algorithms based on support vector machines (SVM) are among the most accurate and robust [19]. When undertaking the two-class learning task, SVMs are used to identify the most appropriate function of classification for separating the classes in the training data. By utilizing different kernel functions, different degrees of nonlinearity and flexibility can be incorporated into the model. Support vector machines have attracted a lot of academic attention in recent years since they can be developed from advanced statistical theories and limitations on the generalisation error can be computed for them. In the medical literature, performance comparable to or better than that of other machine learning algorithms has been documented [19].

Random Forest

Random Forest Regression is an ensemble approach that uses a group of decision trees to predict output for each tree, and then takes the average of all the forecasts to get the random forest model's output prediction [20]. It employs the bagging principle, i.e., Bootstrap and Aggregation. The term "bootstrap" refers to the process of selecting random samples from a dataset and replacing them with new ones. Aggregation is the process of combining all of the predictions to obtain the final result. Bagging aids in the reduction of model overfitting [20].

Naive Bayes

In its simplest form, the naive Bayes model is a Bayesian probability model. Naive Bayes classifiers are based on the principle of high independence. This indicates that the likelihood of one characteristic has no impact on the likelihood of the other. The Naive Bayes classifier makes 2^n independent assumptions given a set of n features. The classifier determines whether a mathematical mapping between a set of attributes and a set of labels is valid by evaluating a particular problem domain with n characteristics and m classifications. The classifier calculates a prior probability of each datapoint for each class. Then it classifies the datapoint as belonging to the class with the highest probability [21].

Logistic Regression

A machine learning model usually used for problems that have two classes output. The classifier calculates the maximum likelihood depending on the datapoint given to the model [22].

Autoencoders

An autoencoder (AE) is a type of unsupervised learning artificial neural network. AEs learn features from unlabeled datapoints automatically, and their principal falls within the field of data dimensionality reduction. A typical autoencoder is made up of three or

more layers: an input layer, a set of hidden layers, and a reconstruction or output layer. A shallow or simple structured autoencoder is a single hidden layer neural network that uses an encoding process to convert original data (input values) to compressed data (lower dimensionality than the original data), which is then mapped to an output layer to approximate the original data via a decoding process [23].

The encoder half derives the feature vector from the input datapoints as follows:

$$y = f(w^e x + b^e) \quad (1)$$

whereas Eq. 2 describes the decoder half.

$$\hat{y} = f(w^d y + b^d) \quad (2)$$

where w^e and w^d represent the weights of the encoder half and decoder half respectively, whereas b^e and b^d represent the biases of encoder and decoder components. These parameters are learnt by minimizing the input–output error. The loss function can be mathematically described as:

$$\text{Loss Function} = \|y - \hat{y}\|^2 \quad (3)$$

3.3 Proposed Machine Learning Model

The proposed model architecture of the encoder and overall model is illustrated in Fig. 1 and Fig. 2. This architecture is proposed as a way to improve the extraction of key characteristics linked to DFU classification. The autoencoder's value comes from the fact that it reduces noise from the input images, leaving only a high-value representation. Because the algorithms can learn the patterns in the data from a smaller selection of high-value inputs, the performance of the machine learning algorithms can be improved. The comparison feature provided by autoencoders reduces the training time significantly, allowing the building of lighter ML models that can work efficiently on low-performance computers.

The suggested encoder contains 28 layers which are divided into four blocks. The various layers types are defined as:

Input layer: 128×128 patches of two classes: normal and abnormal.

Convolutional layer: this is the first layer after the input layer, which takes the inputs and extracts the various features. In this layer, a mathematical process is performed between the input and a filter. A filter which can be of any size of the form $N \times N$ is slid over the inputs. Dot product is produced between the filter and the parts of the images in terms of filter size. A feature map is created because of this procedure, which comprises information about the pictures such as their corners and edges. This feature map is then passed on to further layers, which use it to learn various features from the input image. In the current work, a filter size of 3×3 is used. A layer of batch normalization and activation follows each convolutional layer.

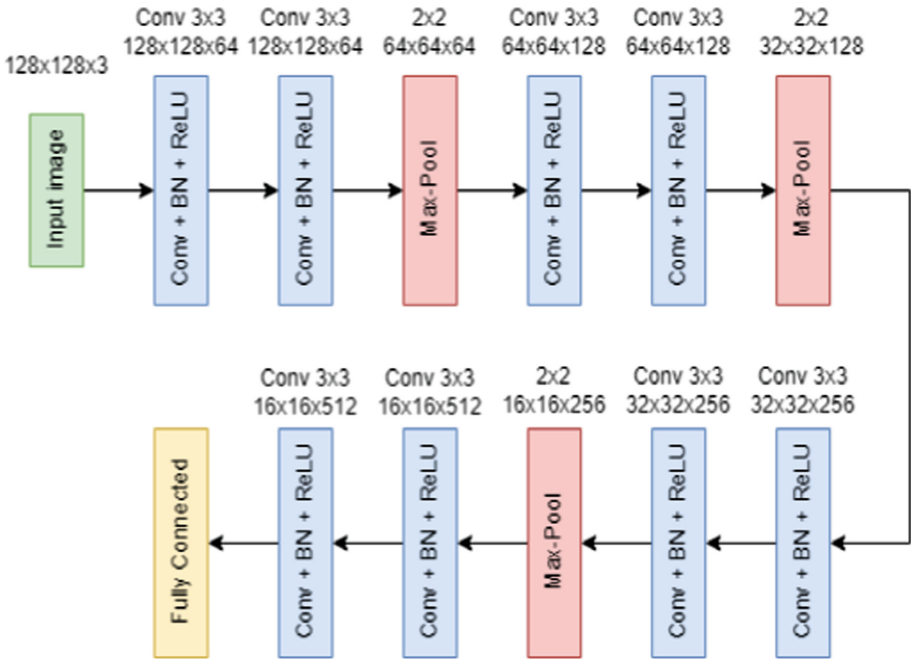


Fig. 1. The proposed encoder

Pooling layer: the primary purpose of the pooling layer is to minimise the size of the feature map produced by the convolutional layer in order to cut computational costs. There are different types of pooling procedures, and the max pooling operation is used in the current work to reduce the feature size.

Activation layer: They are used to learn and estimate any type of continuous and sophisticated network variable-to-variable linkage. There are various activation functions that are regularly utilised. In our case, the ReLU activation function was used.

Flatten layer: The preceding layers’ outputs are flattened and supplied to the machine learning models.

The features are taken from the bottleneck of the autoencoder and supplied to state of the art machine learning models along wide the labels of each datapoint, for binary classification of either normal or abnormal skin area.

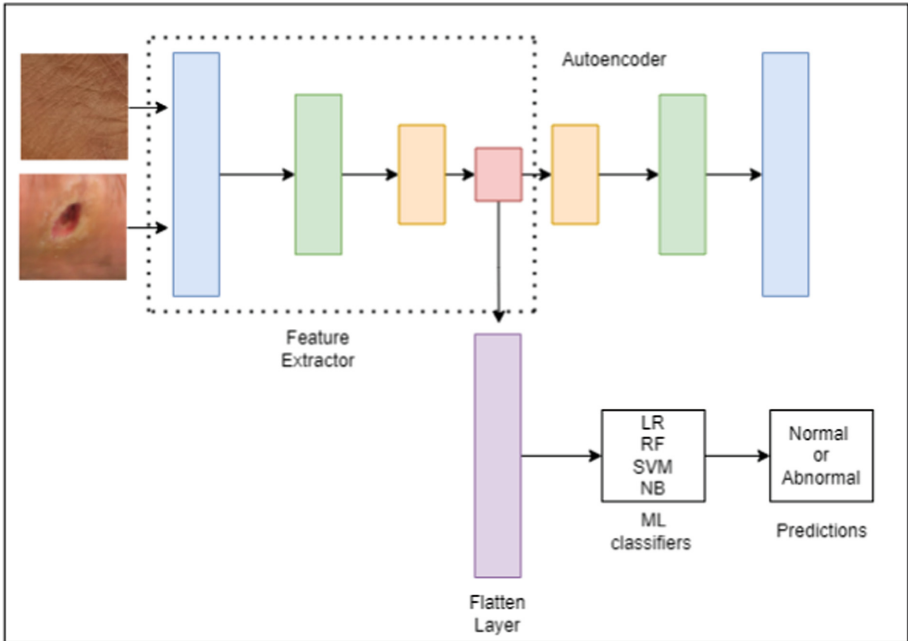


Fig. 2. The proposed model

4 Results

4.1 Evaluation Metrics

The experimental findings are assessed using five assessment indicators. Accuracy, Precision, Recall, F1 score, and AUC are the assessment measures that are used intensively in the literature [24, 25]. The Accuracy calculation formula is as follows:

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

The number of true positives, true negatives, false positives, and false negatives is represented as TP, TN, FP, and FN, respectively. Precision differs from Accuracy in that Precision is concerned solely with the number of positive samples projected to be positive. Precision is calculated using the following formula:

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

Data imbalance is a regular occurrence in the categorization of medical conditions, hence the Recall is required. The following is how the Recall is defined:

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

The F1 score is a weighted average of both Accuracy and Precision, using the following calculation formula:

$$\text{F1score} = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.2 Results

Figure 3 shows the area under the curve achieved by the proposed machine learning algorithms. The reported AUC ranged from 0.90 to 0.93. Both the logistic regression and Naïve Bayes had equivalent AUC measurements. Support vector machine was the second best classifier with an AUC of 0.9. Whereas SVM with Polynomial kernel achieved the best results with AUC of 0.934.

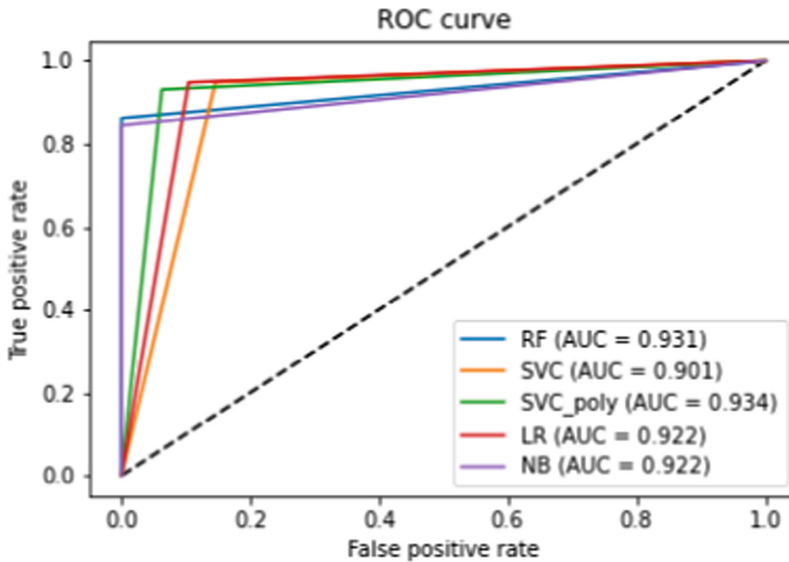


Fig. 3 AUC Plot of ML models performance

The performance results of the benchmarked ML algorithms based on the features retrieved by the autoencoder's encoder component are shown in Table 1. All the results are in the range of 0.86 to 1. Support vector machines with a polynomial kernel performed the best in terms of accuracy and F1 score, with 0.933 and 0.939, respectively. Both Random Forest and Naïve Bayes achieved a precision measure of 1, which means that the classifiers were able to correctly classify all cases. Logistic regression reached a value of 0.921, 0.916, 0.948, and 0.922 for accuracy, precision, recall, and F1 score, respectively. While SVM with a liner kernel performed the worst, with an accuracy of 0.905. A robust model can provide the assurance needed to put a ML model into production. Simply considering only model performance and disregarding model robustness might

have substantial consequences, particularly in critical ML applications such as disease risk prediction. Our model requires more testing to be considered a robust method. Such as testing it on a different dataset. This has been decided as a future work for the paper.

Table 1 Comparison of state-of-the-art algorithms

Classifier	Accuracy	Precision	Recall	F1-score
RF	0.924	1.00	0.862	0.925
SVM	0.905	0.887	0.948	0.916
SVM_Poly	0.933	0.947	0.931	0.939
LR	0.921	0.916	0.948	0.922
NB	0.915	1.00	0.844	0.915

Examples output of the best performing machine learning model is illustrate in Fig. 4.

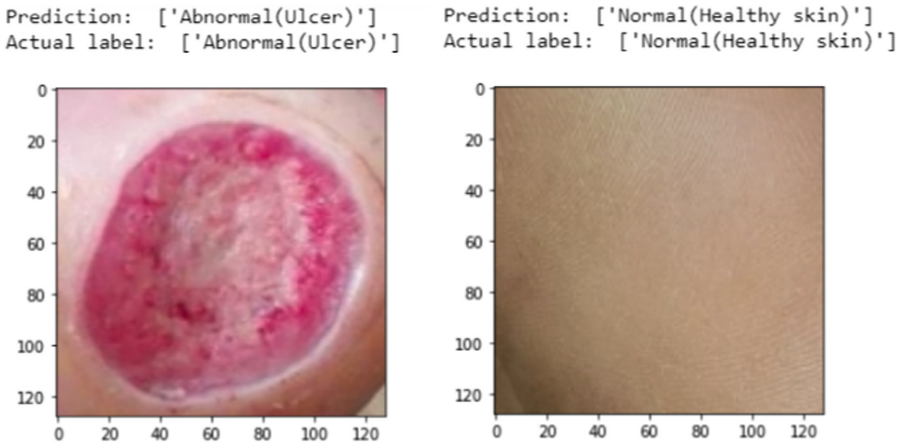


Fig. 4 Examples output of our proposed model

The current work was benchmarked with other models from the literature, as shown in Table 2. According to the F1 score, our proposed model comes in second place. Interestingly, the proposed model achieved the highest score of precision in comparison to its competitors. This is extremely important in the medical field, where accurate prediction of cases is critical. Despite the convergence of the results of our study with the results of other studies, our proposed model was less complex than the other proposed deep learning models. Goyal et al. [9] constructed a CNN model with 15 convolutional layers. While Das et al. [10] included 13 convolutional layers. Alzubaidi et al. [16] proposed a deep learning model with 17 convolutional layers that are arranged in both a parallel and sequential fashion. On the other hand, in the current work, the proposed

model only contains 8 convolutional layers, which results in an efficient light deep learning model.

Table 2 Comparison with previous work

Model	Accuracy	Precision	Recall	F1-score
DFUNet [9]	0.925	0.945		0.939
DFU_SPNet [10]	0.964	0.926	0.984	0.954
CA-DetNet [11]		0.719	0.768	0.743
DFU_QUTNet [16]		0.942	0.926	0.934
Proposed model	0.933	0.947	0.931	0.939

This study demonstrates that autoencoders can be used as an effective tool for feature extraction of diabetic foot ulcer images. By comparing multiple Machine Learning models applied to predict a foot ulcer, this has given us insight into the degree to which machine learning models are able in prediction an abnormal skin areas.

5 Conclusion

In this study, we used a deep learning technique as a feature extractor and subsequently trained multiple machine learning models to accurately classify healthy and DUF skin regions. The encoder component of the autoencoder architecture assisted in the extraction of key characteristics from the input images. To the best of our knowledge, this is the first time an autoencoder model has been applied to a diabetic foot ulcer classification. All of the machine learning models yielded high performance. The highest classification accuracy was attained using a support vector machine with a polynomial kernel.

References

1. Kaul, K., Tarr, J.M., Ahmad, S.I., Kohner, E.M., Chibber, R.: Introduction to diabetes mellitus. In: Ahmad, S.I. (ed.) *Diabetes: An Old Disease, a New Insight*, pp. 1–11. Springer New York, New York, NY (2013). https://doi.org/10.1007/978-1-4614-5441-0_1
2. Saberzadeh-Ardestani, B., et al.: Type 1 diabetes mellitus: cellular and molecular pathophysiology at a glance. *Cell J. (Yakhteh)* **20**(3), 294 (2018)
3. Chatterjee, S., Khunti, K., Davies, M.J.: Type 2 diabetes. *The Lancet* **389**(10085), 2239–2251 (2017)
4. Khan, R.M.M., Chua, Z.J.Y., Tan, J.C., Yang, Y., Liao, Z., Zhao, Y.: From pre-diabetes to diabetes: diagnosis, treatments and translational research. *Medicina* **55**(9), 546 (2019)
5. Tesfaye, S.: Neuropathy in diabetes. *Medicine* **43**(1), 26–32 (2015)

6. Bus, S.A., Ret al.: Footwear and offloading interventions to prevent and heal foot ulcers and reduce plantar pressure in patients with diabetes: a systematic review. *Diabetes/Metabol. Res. Rev.* **32**, 99–118 (2016)
7. Skrepnek, G.H., Mills, J.L., Sr., Lavery, L.A., Armstrong, D.G.: Health care service and outcomes among an estimated 6.7 million ambulatory care diabetic foot cases in the US. *Diabetes Care* **40**(7), 936–942 (2017)
8. Subasi, A.: Chapter 3 - Machine learning techniques. In: Subasi, A. (ed.) *Practical Machine Learning for Data Analysis Using Python*, pp. 91–202. Academic Press (2020)
9. Goyal, M., Reeves, N.D., Davison, A.K., Rajbhandari, S., Spragg, J., Yap, M.H.: DFUNet: convolutional neural networks for diabetic foot ulcer classification. *IEEE Trans. Emerg. Top. Comput. Intell.* **4**(5), 728–739 (2020)
10. Das, S.K., Roy, P., Mishra, A.K.: DFU_SPNet: a stacked parallel convolution layers based CNN to improve Diabetic Foot Ulcer classification. *ICT Express* (2021)
11. Yap, M.H., et al.: Deep learning in diabetic foot ulcers detection: a comprehensive evaluation. *Comput. Biol. Med.* **135**, 104596 (2021)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Presented at the Proceedings of the 28th International Conference on Neural Information Processing Systems, vol. 1. Montreal, Canada (2015)
13. Cassidy, B., et al.: The DFUC 2020 dataset: analysis towards diabetic foot ulcer detection. *TouchREVIEWS in endocrinology* **17**(1), 5–11 (2021)
14. Khandakar, A., et al.: A machine learning model for early detection of diabetic foot using thermogram images. *Comput. Biol. Med.* **137**, 104838 (2021)
15. Scebba, G., et al.: Detect-and-segment: A deep learning approach to automate wound image segmentation. *Inform. Med. Unlocked* **29**, 100884 (2022)
16. Alzubaidi, L., Fadhel, M.A., Olewi, S.R., Al-Shamma, O., Zhang, J.: DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network. *Multimedia Tools Appl.* **79**(21), 15655–15677 (2020)
17. Alzubaidi, L., Fadhel, M.A., Al-Shamma, O., Zhang, J., Santamaría, J., Duan, Y.: Robust application of new deep learning tools: an experimental study in medical imaging. *Multimedia Tools Appl.* **81**(10), 13289–13317 (2022). <https://doi.org/10.1007/s11042-021-10942-9>
18. Alzubaidi, L., et al.: Towards a better understanding of transfer learning for medical imaging: a case study. *Appl. Sci.* **10**(13), 4523 (2020)
19. Orru, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A.: Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. Biobehav. Rev.* **36**(4), 1140–1152 (2012)
20. Sarica, A., Cerasa, A., Quattrone, A.: Random forest algorithm for the classification of neuroimaging data in Alzheimer’s disease: a systematic review. *Front. Aging neurosci.* **9**, 329 (2017)
21. Salmi, N., Rustam, Z.: Naïve Bayes classifier models for predicting the colon cancer. In: *IOP Conference Series: Materials Science and Engineering*, vol. 546, no. 5, p. 052068. IOP Publishing (2019)
22. Christodoulou, E., Ma, J., Collins, G.S., Steyerberg, E.W., Verbakel, J.Y., Van Calster, B.: A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019)
23. Lopez Pinaya, W.H., Vieira, S., Garcia-Dias, R., Mechelli, A.: Chapter 11 – Autoencoders. In: Mechelli, A., Vieira, S. (eds.) *Machine Learning*, pp. 193–208. Academic Press (2020)

24. Alatrany, A., Hussain, A., Mustafina, J., Al-Jumeily, D.: A novel hybrid machine learning approach using deep learning for the prediction of alzheimer disease using genome data. In: Huang, D.-S., Jo, K.-H., Li, J., Gribova, V., Premaratne, P. (eds.) ICIC 2021. LNCS (LNAI), vol. 12838, pp. 253–266. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-84532-2_23
25. Alatrany, A.S., Hussain, A., Jamila, M., Al-Jumeiy, D.: Stacked machine learning model for predicting alzheimer’s disease based on genetic data. In: 2021 14th International Conference on Developments in eSystems Engineering (DeSE), 7–10 Dec 2021, pp. 594–598 (2021). <https://doi.org/10.1109/DeSE54285.2021.9719449>.