Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(7): 16446–16462.
DOI: 10.3934/math.2023840
Received: 15 December 2022
Revised: 10 April 2023
Accepted: 20 April 2023
Published: 09 May 2023

Research article

Deep intelligent predictive model for the identification of diabetes

Salman khan[1], Muhammad Naeem[2,*] and Muhammad Qiyas[3]

[1] Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan
[2] Department of Mathematics Deanship of Applied Sciences Umm Al-Qura University, Makkah, Saudi Arabia
[3] Department of Mathematics, Riphah International University Faisalabad Campus, Pakistan

* **Correspondence:** Email: mfaridoon@uqu.edu.sa.

Abstract: Diabetes mellitus is a severe, chronic disease that occurs when blood glucose levels rise above certain limits. Many complications arise if diabetes remains untreated and unidentified. Early prediction of diabetes is the most high-quality way to forestall and manipulate diabetes and its complications. With the rising incidence of diabetes, machine learning and deep learning algorithms have been increasingly used to predict diabetes and its complications due to their capacity to care for massive and complicated facts sets. This research aims to develop an intelligent computational model that can accurately predict the probability of diabetes in patients at an early stage. The proposed predictor employs hybrid pseudo-K-tuple nucleotide composition (PseKNC) for sequence formulation, an unsupervised principal component analysis (PCA) algorithm for discriminant feature selection, and a deep neural network (DNN) as a classifier. The experimental results show that the proposed technique can perform better on benchmark datasets. Furthermore, overall assessment performance compared to existing predictors indicated that our predictor outperformed the cutting-edge predictors using 10-fold cross validation. It is anticipated that the proposed model could be a beneficial tool for diabetes diagnosis and precision medicine.

Keywords: deep learning; PseKNC; PCA; diabetes mellitus; diabetes complications
Mathematics Subject Classification: 68T05, 68T45, 97Rxx

1. Introduction

Diabetes is a metabolic illness characterized by inadequate insulin production and secretion abnormalities [1], with hyperglycemia as the primary symptom. The physiological system will be harmed by long-term exposure of organs to hyperglycemia. Damage to tissues and organs, such as the eyes, kidneys, nerves, heart, and blood vessels, can result in organ and tissue failure. Diabetes mellitus is currently split into two types: diabetes mellitus type 1 and diabetes mellitus type 2. Type 2 diabetes is the most prevalent type of diabetes, accounting for approximately 95% of diabetic people [2,3]. The major causes of diabetes are environmental factors and poor lifestyle choices. Furthermore, diabetes is triggered by aging, poor diet, and inadequate activity [4].

The normal transition from health to diabetes mellitus occurs in three stages: health, pre-diabetes, and type 2 diabetes [5]. After being diagnosed with diabetes, patients' blood glucose levels will continue to rise, and pharmaceutical therapy is difficult to discontinue [6,7]. On the other hand, can maintain blood glucose control and even restore health with artificial intervention. Many studies have demonstrated that the most effective technique for avoiding and controlling diabetes is to detect and treat it as soon as possible. As a result, early detection and lifestyle adjustments are crucial in diabetes care and the avoidance of complications, including diabetic retinopathy [8–11].

Early diagnosis reduces the relative and absolute risk of heart attacks and mortality, as per the ADDITION-Europe analytical model analysis [12]. Diabetes mellitus is a global public health concern, with the International Diabetes Federation estimating that 463 million people worldwide have diabetes, with a 51% increase projected by 2045. Furthermore, it is predicted that for every diagnosed person with diabetes, there is one undiagnosed individual [13,14]. Understanding the evolving pattern of diabetes at all stages and extracting useful diabetes information from physical examination data is crucial to diabetes prevention and treatment. As a result, several scholars have tried to develop diabetes prediction models. Traditional statistical learning procedures, such as linear regression, were used to create the first models. Traditional statistical learning systems, on the other hand, provide predictions, but predictive accuracy is not one of their strong suits.

The toolbox was recently enhanced to cover a variety of machine-learning approaches. These approaches anticipate future occurrences based on patterns discovered in training data from previous examples. Furthermore, different techniques are sometimes combined to construct ensemble models that outperform single models in terms of prediction. For example, Kalsch et al. [14] established a relationship between liver damage markers and diabetes and used random forests to predict diabetes using serum data. Zou et al. [15] utilized principal component analysis (PCA) and minimal redundant maximum (mRMR) correlation to screen risk factors, and decision tree (DT) and Random forest (RF) to predict diabetes. However, they are based on traditional learning models, which need much human expertise and the ability to extract crucial parts automatically.

Furthermore, in the presence of nonlinearity in the dataset, the predictive power of the models mentioned above is reduced. As a result, a more sophisticated and robust computer model in bioinformatics is necessary to detect diabetic mellitus. Using an improved deep learning algorithm and fast feature extraction methods, this research provides a solid and trustworthy computational prediction for reliably recognizing diabetic mellitus. In this paper, we proposed a computational model using a multilayer DNN technique. The proposed model performance is thoroughly examined using the 10-fold cross-validation approach, which is based on several performance assessment indicators. The suggested predictor uses several modes of PseKNC as a feature formulation approach to construct the provided samples into a features set, including Single/Di/tri nucleotide composition. Second, an unsupervised PCA technique is used to eliminate unnecessary and noisy features, resulting in the selection of efficient features. The suggested predictor attained an average prediction accuracy of 88.65%

based on the testing findings. Furthermore, the performance of the proposed predictor is compared to that of recently published predictors. The suggested model beat the current predictors in accuracy and other performance measures, according to the comparative findings. The major contributions of the paper are summarized as follow:

- Propose a deep generative model for classification of the probability of diabetes in patients at an early stage.
- The proposed model considered non-linearity in dataset using multi-stack processing layers and non-linear activation function.
- The performance of the proposed model is extensively evaluated using different performance measurement metrics such as Accuracy (ACC), Area under the Curve (AUC), Sensitivity (SN), Specificity (SP) and Mathew's Correlation Coefficient (MCC).

2. The proposed model's design

The section outlines the suggested model's prototype and architecture. Figure 1 depicts the suggested model's design, which includes various elements that are addressed in-depth below.

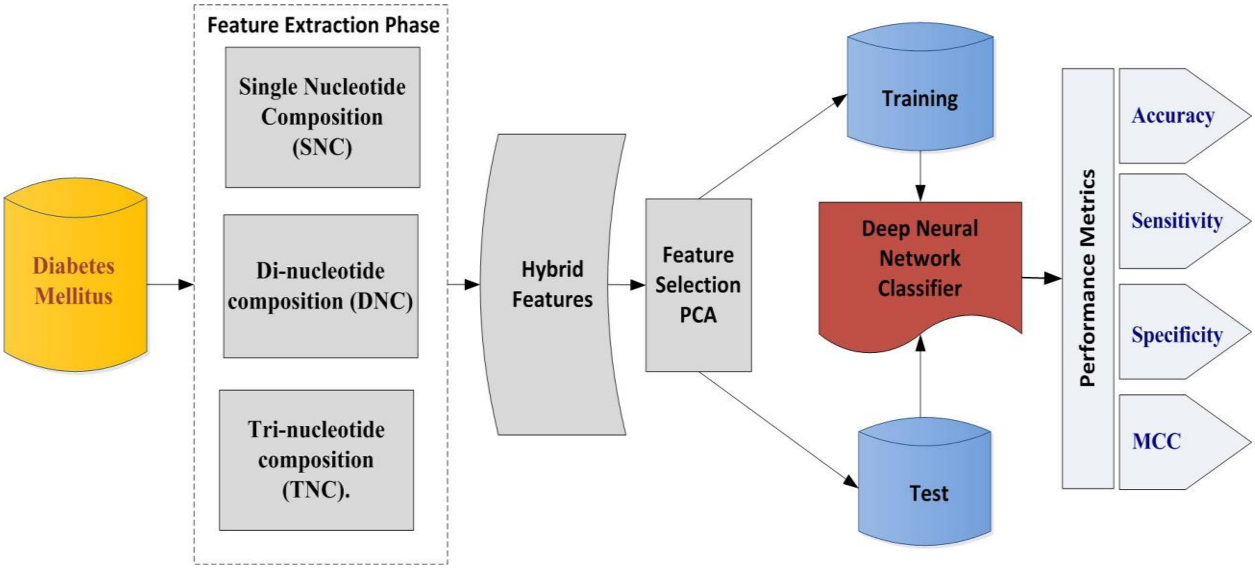

Figure 1. Proposed model architecture.

2.1. Benchmark dataset

A valid and trustworthy benchmark dataset is required to design a robust and resilient computational model. Therefore, we select a useful benchmark dataset available in [16–19] for training and validating the outcomes of the proposed model. Following Eq (1) represented the mathematical form of the chosen benchmark dataset.

$$D = D^{Positive} \bigcup D^{Negative} \tag{1}$$

The benchmark data were collected from medical and physical examinations and separated into healthy persons and diabetics. We used one of the physical examination data that contains 64,000 instances as the validation set. In the other dataset, 10,000 samples were randomly selected as an independent test set. We consider 14 different variables, i.e., age, pulse rate, breathe, left systolic pressure (LSP), proper systolic pressure (RSP), left diastolic pressure (LDP), right diastolic pressure (RDP), height, weight, physique index, fasting glucose, waistline, low-density lipoprotein (LDL), and high-density lipoprotein (HDL).

It is to be noted that the benchmark dataset is imbalanced (i.e., the positive samples are in the minority class, and the negative samples are in the majority class). The classification is biased toward the majority class in the imbalanced dataset, referred to as a classification problem, as shown in Figure 2. Various approaches, i.e., data-level approach and algorithm-level approach, have been considered in the literature for the class imbalance issue; however, the data-level system, including over-sampling and under-sampling methods, is widely employed the issue of the imbalanced dataset. The over-sample method could increase the risk of model over-fitting due to data sample replication. The under-sampling method can offer a modest solution in most cases. Hence, in this study, we applied the under-sampling method with the Near Miss procedure implemented in Python to balance the benchmark dataset.

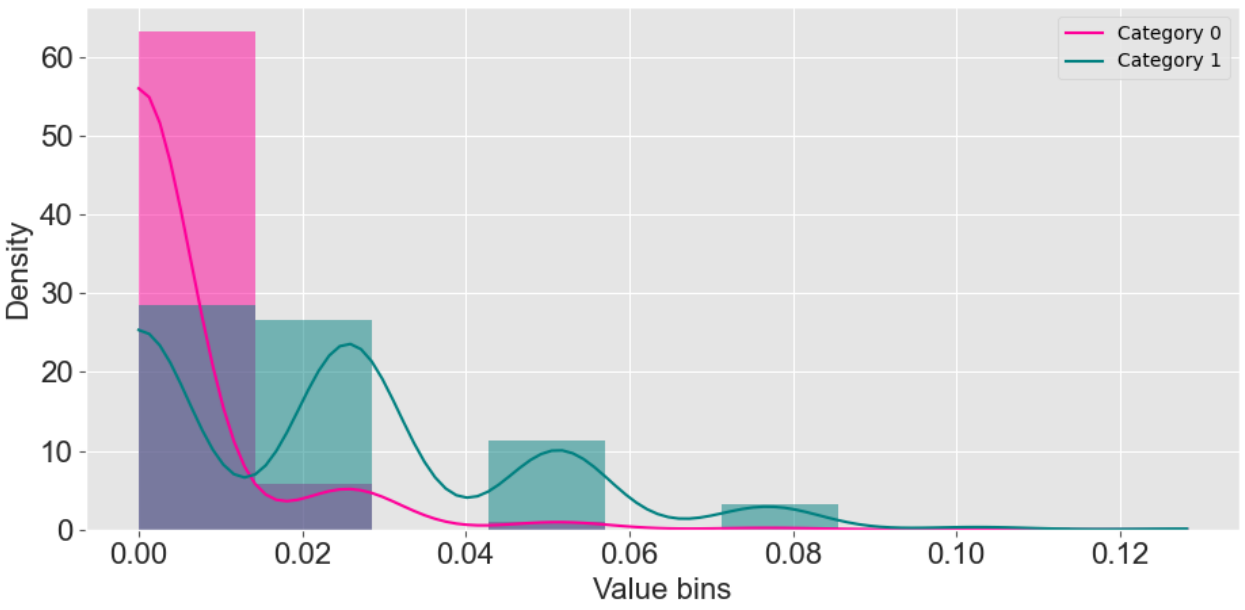

Figure 2. Benchmark dataset density versus value bins.

2.2. Feature formulation technique

The data is mainly analyzed in discrete form by quantitative machine learning techniques [20]. Formulating biological sequences into discrete forms is difficult in computational biology [21]. Various approaches [22,23] for transforming biological sequences into discrete feature vectors have been proposed in the literature. Chou's suggested approach has been widely adopted in current application development and is now used in all sectors of computational biology [24–26]. Chou's Pseudo K-tuple Nucleotide Composition (PseKNC) approach for representing RNA/DNA sequences into a discrete feature vector is widely used in computational biology [27,28]. We used the PseKNC approach to encode diabetes samples into a discrete feature set in this research.

To encode the following sequence D into a discrete feature vector, consider the following:

$$D = d_1 d_2 d_3 .... d_i ... d_l \tag{2}$$

Here d1 represents the very first nucleotide in D, d2 represents the next nucleotide in D, and dL represents the last nucleotide in D. Equation (3) may be stated in the PseKNC's basic form utilizing Eq (4).

$$D = [\phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_z \quad]^T \tag{3}$$

Where, T is the transpose, z is the integer number and ϕ_u is feature vector. The selected vector can be calculated by Eq (5).

$$\varphi_u = \begin{cases} \dfrac{f_u^{KTuple}}{\sum_{i=1}^{4^k} f_u^{K-tuple} + w\sum_{j=1}^{\lambda}\theta_j} (1 \le u \le 4^k, u = 1,2,3,...) \\ , \dfrac{w\theta_{u-4^k}}{\sum_{i=1}^{4^k} f_u^{K-tuple} + w\sum_{j=1}^{\lambda}\theta_j} (4^k + 1 \le u \le 4^k + \lambda) \end{cases} \tag{4}$$

$$\theta_j = \begin{cases} \dfrac{1}{L-K-(\lambda-1)} \displaystyle\sum_{i=1}^{L-K-(\lambda-1)} C_{i,i+j} (j \quad \to \quad 1,2,..\lambda; \quad \lambda < L-K \end{cases} \tag{5}$$

Where, Eq (6) computed the coupling factor i.e., $C_{i,i+j}$

$$C_{i,i+j} = \frac{1}{u}\sum_{\xi=1}^{\lambda} \times \left[H_\xi (N_i N_{i+1} .. N_{i+K-1}) - H_\xi (N_{i+j} N_{i+j+1} .. N_{i+j+K-1}) \right]^2 \tag{6}$$

Finally, PseKNC technique to calculate the samples by using K=1, 2, 3 in Eq (6).

$$D_{PseSNC} = \big[\mathbf{f}_j^{1Tuple} {}_{j=1,..4D} \xrightarrow{f} (U, C, G, A) \tag{7}$$

$$D_{PseDNC} = \big[\mathbf{f}_j^{2Tuple} {}_{j=1,..16D} \xrightarrow{f} (UU, CC, GA, UA, \ldots, GG) \tag{8}$$

$$D_{PseTNC} = \big[\mathbf{f}_j^{3Tuple} {}_{j=1,..64D} \xrightarrow{f} (ACG, CAA, GUU, AUU, \ldots, GAU) \tag{9}$$

2.3. Feature selection

The feature vector may contain noisy, redundant, and irrelevant features significantly impacting a classifier's performance. Hence, in this paper, we perform feature selection using Principal Component Analysis (PCA) algorithm. The PCA algorithm [29–34] is a multivariate data processing method that computes covariance matrices and eigenvectors to reduce feature vector dimensionality to low-dimensional vectors. The PCA algorithm's primary goal is to decrease a feature vector's dimension while retaining as many discriminative features as possible. For instance consider i × j dimension as feature vector, "i" represent extracted features and "j" represent total samples. Using Eq (10) consider the input feature vector D:

$$\left[D_1, D_2, D_3, \ldots, D_n\right] \tag{10}$$

Using the PCA algorithm, we took the following procedures to reduce the feature vector dimensionality.

Step 1: Using Eq (11), we first calculate the feature vector mean value

$$\bar{D} = \frac{1}{i}\sum_{n=1}^{i} D_n \tag{11}$$

Step 2: Using Eq (12), from D_n we subtract the mean value i.e., $\bar{D}$

$$D_n^{'} = D_n - \bar{D} \qquad \text{n} = (1, 2, 3 \ldots) \tag{12}$$

Step 3: Using Eq (13), covariance matrix C_n is computed

$$C_n = D_n^{'}.(D_n^{'})^T BB^T \tag{13}$$

Where, $B = \left\{D_1^{'},\ D_2^{'},\ \ldots,\ D_p^{'}\right\}(i\ *\ j)$ and T= transpose of a matrix.

Step 4: Eigenvalues are computed in this step. In other words, the first Eigen must have a value larger than that of the second Eigen, and so on.

$$\left\{\gamma_1 > \gamma_2 > \gamma_3 \cdots \gamma_n\right\}$$

Step 5: Eigenvector is compute in this step as:

$$c_n : \left\{\Phi_1, \Phi_2, \Phi_3, \ldots, \Phi_n\right\} \tag{14}$$

Last step: We choose k values with the highest Eigenvalues. Eigen vector with the highest Eigen values was chosen to reconstruct the new set of feature vector to accomplish a better prediction performance.

2.4. Fully connected deep neural network

DNN is a branch of AI algorithms provoked by the functioning mechanism and activities of the human brain [35–37], as shown in Figure 3. Hidden layers are significant components of the DNN model and are actively involved in the learning process [38–40]. Using more hidden layers in the training phase can improve the performance of a model; however, it may cause significant issues such as model complexity, computation cost, and over-fitting [41–43]. The DNN model can extract relevant features automatically from the given unlabeled or unstructured dataset without the involvement of human engineering and expertise using standard learning procedures [37,44–46]. It has been reported by several investigators that the DNN model performed better than the conventional learning methods applied for various complex classification problems [47–49]. Moreover, deep learning algorithms are successfully applied in several domains, such as speech recognition [50,51], image recognition [52,53], bio-engineering [54,55], and natural language processing [56,57]. Inspired by the tremendous performance of deep learning in various domains for a complex classification problem, this paper has applied the DNN algorithm as a prediction engine.

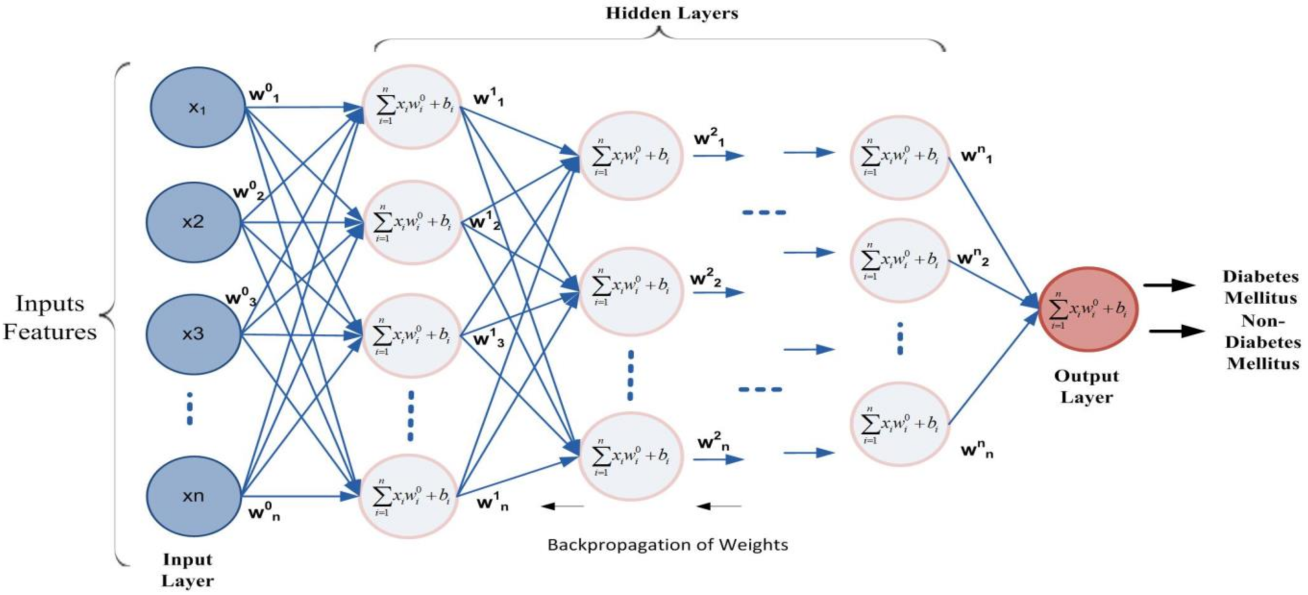

Figure 3. Deep neural network configuration.

Inspired by the effective employment of deep learning models in numerous domains for composite classification problems, this paper considered the DNN model to predict diabetes mellitus using a benchmark dataset. As shown in Figure 3, the proposed deep neural network method has six hidden layers, an input and output layer. Each layer in the DNN topology is configured with numerous nodes i.e., take input features set and produce output using the following equation.

$$y_a = f(B_a + \sum_{b=1}^{m} x_b w_b^a) \tag{15}$$

Where, y_a represent output at a layer a, B represent bias value, w_b^a mean weight used at a layer

a by a neuron b, x_b^a represent input feature and f represent a non-linear activation Tanh function and it can be calculated using Eq (16).

$$f(i)=\frac{e^{i}}{1+e^{i}} \tag{16}$$

3. Performance evaluation metrics

Any statistical machine learning-based model's performance can be evaluated using some measurement metrics before the model deploy in a natural production environment [58]. Several performance measurements have been proposed in the literature to assess a learning model. Accuracy is the primary criterion for evaluating the algorithm's effectiveness in every one of these measures. More than the accuracy metrics for a model performance evaluation is required. Hence, different performance metrics set is considered in the literature to evaluate a model [59–61]. These performance metrics, along with accuracy, are included: Area under the Curve (AUC), Sensitivity (SN), Specificity (SP), and Mathew's Correlation Coefficient (MCC). We considered all the performance as mentioned earlier measurement metrics to assess the proposed Deep-Sumo performance as these metrics were widely adopted in a series of publications [25,46,62–66]. The performance measurement metrics can be calculated using the following equations.

$$Accuracy=1-\frac{D_{-}^{+}+D_{+}^{-}}{D^{+}+D^{-}} \tag{17}$$

$$\text{Specificit y}=1-\frac{D_{-}^{+}}{D^{+}} \tag{18}$$

$$\text{Sensitivit y}=1-\frac{D_{+}^{-}}{D^{-}} \tag{19}$$

$$MCC=\frac{1-(\frac{D_{-}^{+}+D_{+}^{-}}{D^{+}+D^{-}})}{\sqrt{(1+\frac{D_{-}^{+}+D_{+}^{-}}{D^{+}})(1+\frac{D_{-}^{+}+D_{+}^{-}}{D^{-}})}} \tag{20}$$

Where, D^{+} represent the total number of diabetes mellitus samples, D^{-} represent total non- diabetes mellitus samples. Similarly, D_{-}^{+} represents total mellitus that is wrongly recognized as non- diabetes mellitus, D_{+}^{-} represent total non- diabetes mellitus that wrongly recognized as diabetes mellitus.

4. Experimental results

4.1. Feature analysis

We first performed an analysis of the feature on the selected benchmark dataset. The success rates of Pseudo-DNC, Pseudo-TNC, and Pseudo-Tetra NC and fusion feature space on classification methods are listed in Table 1. Among all hypothesis learners, deep neural network using fusion features in the case of k-folds cross-validation test obtained the highest accuracy of 84.11%, specificity of 82.72%, the sensitivity of 89.11%, and MCC of 0.645. Similarly, using individual feature space of Pseudo-DNC accomplished the performance ratio of 82.03%, with an MCC of 0.641, a sensitivity of 85.66%, and a specificity of 78.89%, respectively.

Table 1. Performance analysis by different extraction approaches.

Techniques	**Acc (%)**	**Sn (%)**	**Sp (%)**	**MCC**
Before Feature Selection				
Pseudo-SNC	80.63	82.71	78.61	0.584
Pseudo-DNC	82.03	85.66	78.89	0.641
Pseudo-TNC	80.38	86.97	78.41	0.677
Fusion feature	84.11	89.11	82.72	0.645
After Feature Selection				
Pseudo-SNC	81.31	83.21	78.99	0.594
Pseudo-DNC	84.11	86.34	80.88	0.681
Pseudo-TNC	83.56	87.71	79.12	0.687
Fusion Features	88.65	81.50	89.41	0.766

The dimensions of the fusion features vector were lowered utilizing the feature selection approach to increase the performance of the suggested model even further (i.e., PCA). Consequently, the suggested deep neural network model's success rate increased dramatically, with an average accuracy of 88.65%, specificity of 89.41%, sensitivity of 81.50%, and MCC of 0.766. Integrating all selected traits we highlight, contributes to enhancing MCC value.

4.2. Performance analysis

In addition, we examined the performance of the deep neural network model in terms of AUC. The AUC is a frequently used metric for evaluating the effectiveness of a classification algorithm, and it is calculated in the range of 0 to 1. In comparison to a predictor that computes a lower value, a classifier that computes a more significant rate is deemed the best predictor [66]. Figure 4 depicts the DNN's effectiveness in AUC utilizing various sequence formation strategies. Figure 4 shows that, compared to alternative sequence formulation approaches, the DNN method produced the most outstanding results when applying fusion characteristics. The proposed method, for example, generates the maximum AUC values utilizing fusion features 0.946 compared to the second maximum value of 0.891, using Pseudo-DNC. These findings revealed that the deep neural network model performs the most remarkable prediction by employing fusion features.

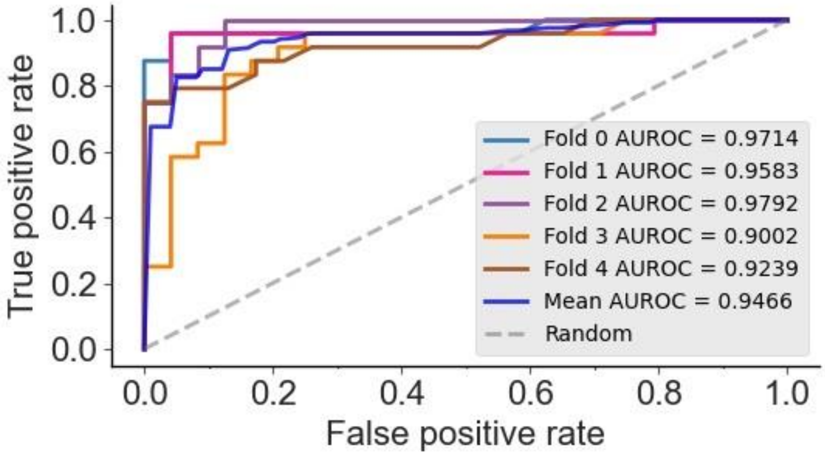

Figure 4. AUROC of proposed deep neural network model.

Furthermore, the suggested model highlights the issues of under fitting and overfitting, as well as how linear regression using polynomial features may approximate nonlinear functions. The graphic depicts the function we wish to approximate, a cosine function component, as shown in Figure 5.

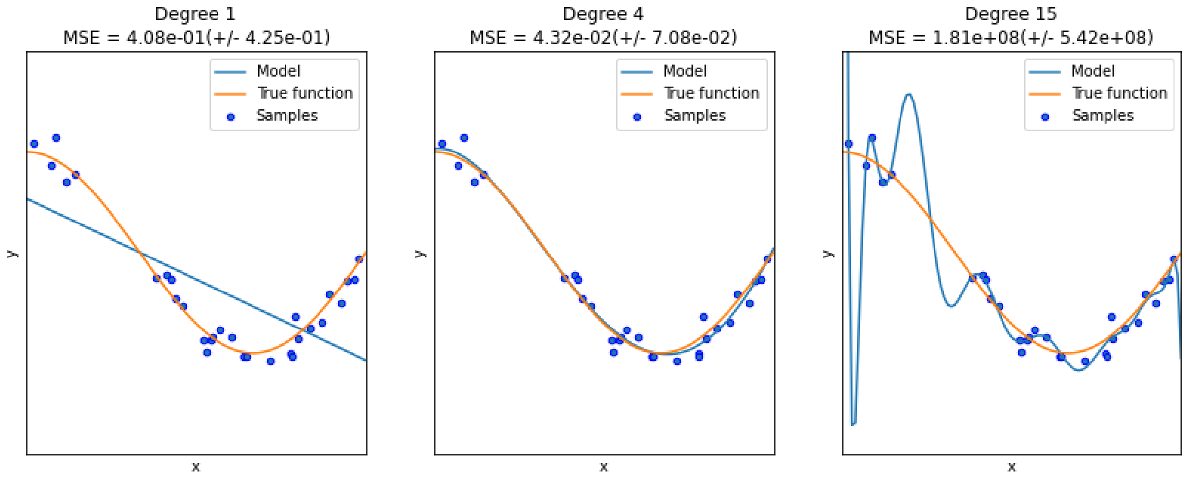

Figure 5. Proposed model under fitting and over fitting analysis.

As shown in Figure 5, a linear function (polynomial of degree 1) is insufficient to match the training data. This is known as under fitting. A polynomial of degree 4 nearly exactly approximates the underlying function. For higher degrees, however, the model will over fit the training data, i.e., it will learn the noise in the training data. The mean squared error (MSE) on the validation set is calculated; the more significant the MSE, the less probable the model generalizes successfully from the training data.

4.3. Performance comparison with machine learning techniques

To authenticate the efficiency of our suggested framework, we accompanied comparative experimentations with several well-known methods. Table 2 shows the performance measure value for each categorization algorithm. Table 2 shows that the overall performance accuracy of all categorization techniques is greater than 70%. The DNN model, for example, had an average accuracy of 88.65% compared with other ML methods. Similarly, MCC, which shows the model stability, the DNN model had the highest score of 0.815. Moreover, Gradient Boosting and Extra Trees Classifier both show better accuracy for both testing methods, while the SVM-Linear Kernel model came in second last with a score of 72.49%. Furthermore, the Naive Bayes model performed the worst regarding gauging performance.

Table 2. The suggested Deep Neural Network method comparison with machine learning techniques.

Classifier	**ACC (%)**	**MCC**	**SP (%)**	**SN (%)**
Deep Neural Network	88.65	0.815	89.41	76.58
Gradient Boosting Classifier	75.56	0.613	79.69	72.4
Extra Trees Classifier	74.79	0.697	78.75	81.32
SVM-Linear Kernel	72.49	0.653	75.22	69.17
Naive Bayes	71.12	0.634	73.98	75.84

Based on the data, the primary reason of DNN model outperformed traditional learning approaches is the multi-stack processing. Whereas traditional learning methods employ single-layer processing, which is inadequate for commerce with a complex nature dataset with high nonlinearity.

4.4. Performance comparison with the existing predictors

The performance of the proposed model was compared with the existing predictors, as mentioned in [16–19]. The mentioned latest methods build prediction models based on machine learning algorithms. The performance of our proposed model and the existing benchmark models are evaluated on benchmark datasets by using 10-fold cross-validation. For facilitating comparison, Table 3 shows the corresponding results obtained by the existing state of the art methods.

Table 3. The DNN model comparison with the existing learning techniques.

Classifier	**ACC (%)**
Deep Neural Network	88.65
LSTM [16]	87.26
Artificial neural network [19]	85.09
Naive Bayes [18]	76.3
RB-Bayes [17]	72.9

The results demonstrated that the proposed model outperformed the recently published predictors. For example, the proposed model achieved the highest accuracy, i.e., 88.65%, compared with the existing predictor achieving the second highest success rate, i.e., 87.26%. These results confirm that the proposed model performed superior to the existing models, with an average success rate improvement of 5.61%.

5. Conclusions

The medical issues present during diagnosis determine how an individual with metabolic disease is classified, and people may not always neatly fall into a specific category. Some people, for instance, cannot differentiate between type 1 and 2 diabetes. Diabetes has a broad spectrum of clinical symptoms and disease progression. Considering the importance of PseKNC, PCA, and deep learning algorithms, this paper proposed a robust and accurate predictor to improve the performance of diabetes prediction. A hybrid feature vector with significant undertaking dimensional feature space was created by combining the feature vectors created using the various PseKNC techniques. PCA algorithm was applied to select only prominent features by eliminating noisy and irrelevant parts. Several machine learning algorithms have been applied for piRNA prediction; however, the DNN model produced the best performance using a 10-fold cross-validation test. In addition, the proposed optimized DNN classifier algorithm outperformed existing state-of-the-art models with an accuracy improvement of 5.61%. Furthermore, the proposed model might be utilized for a wide range of commonly used medications that raise insulin resistance and impair beta cell activity, potentially leading to diabetes mellitus in vulnerable persons.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work under grant code 22UQU4310396DSR60.

Conflict of interest

The authors declare that they have no competing interests.

References

1. J. M. Lachin, D. M. Nathan, Understanding metabolic memory: The prolonged influence of glycemia during the diabetes control and complications trial (DCCT) on future risks of complications during the study of the epidemiology of diabetes interventions and complications (EDIC), *Diabetes Care*, **44** (2021), 2216–2224. http://doi.org/10.2337/dc20-3097
2. C. Greenhill, How does leptin decrease hyperglycaemia in T1DM and T2DM?, *Nat. Rev. Endocrinol.*, **10** (2014), 511. http://doi.org/10.1038/nrendo.2014.104
3. J. Schofield, J. Ho, H. Soran, Cardiovascular risk in type 1 diabetes mellitus, *Diabetes Ther.*, **10** (2019), 773–789. http://doi.org/10.1007/s13300-019-0612-8
4. H. Cho, C. H. Kim, E. Q. Knight, H. W. Oh, B. Park, D. G. Kim, et al., Changes in brain metabolic connectivity underlie autistic-like social deficits in a rat model of autism spectrum disorder, *Sci. Rep.*, **7** (2017), 13213. http://doi.org/10.1038/s41598-017-13642-3
5. M. Huber, L. Beyer, C. Prix, S. Schönecker, C. Palleis, B.-S. Rauchmann, et al., Metabolic correlates of dopaminergic loss in dementia with Lewy bodies, *Mov. Disord.*, **35** (2020), 595–605. http://doi.org/10.1002/mds.27945

6. F. S. Chiwanga, M. A. Njelekela, M. B. Diamond, F. Bajunirwe, D. Guwatudde, J. Nankya-Mutyoba, et al., Urban and rural prevalence of diabetes and pre-diabetes and risk factors associated with diabetes in Tanzania and Uganda, *Global Health Action*, **9** (2016), 31440. http://doi.org/10.3402/gha.v9.31440
7. A. Basit, A. Fawwad, H. Qureshi, A. S. Shera, Prevalence of diabetes, pre-diabetes and associated risk factors: second National Diabetes Survey of Pakistan (NDSP), 2016–2017, *BMJ Open*, **8** (2018), e020961. http://doi.org/10.1136/bmjopen-2017-020961
8. M. D. Campbell, T. Sathish, P. Z. Zimmet, K. R. Thankappan, B. Oldenburg, D. R. Owens, et al., Benefit of lifestyle-based T2DM prevention is influenced by prediabetes phenotype, *Nat. Rev. Endocrinol.*, **16** (2020), 395–400. http://doi.org/10.1038/s41574-019-0316-1
9. C. Ao, L. Yu, Q. Zou, Prediction of bio-sequence modifications and the associations with diseases, *Brief. Funct. Genomics*, **20** (2021), 1–18. http://doi.org/10.1093/bfgp/elaa023
10. M. Higazy, A. El-Mesady, A. M. S. Mahdy, S. Ullah, A. Al-Ghamdi, Numerical, approximate solutions, and optimal control on the deathly Lassa hemorrhagic fever disease in pregnant women, *J. Funct. Space.*, **2021** (2021), 2444920. http://doi.org/10.1155/2021/2444920
11. A. El-Mesady, A. Elsonbaty, W. Adel, On nonlinear dynamics of a fractional order monkeypox virus model, *Chaos Soliton. Fract.*, **164** (2022), 112716. http://doi.org/10.1016/j.chaos.2022.112716
12. I. Johansson, A. Norhammar, Diabetes and heart failure notions from epidemiology including patterns in low-, middle- and high-income countries, *Diabetes Res. Clin. Pract.*, **177** (2021), 108822. http://doi.org/10.1016/j.diabres.2021.108822
13. E. W. Gregg, N. Sattar, M. K. Ali, The changing face of diabetes complications, *Lancet Diabetes Endocrinol.*, **4** (2016), 537–547. http://doi.org/10.1016/S2213-8587(16)30010-9
14. J. Kälsch, L. P. Bechmann, D. Heider, J. Best, P. Manka, H. Kälsch, et al., Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort, *Sci. Rep.*, **5** (2015), 13058. http://doi.org/10.1038/srep13058
15. Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting diabetes mellitus with machine learning techniques, *Front. Genet.*, **9** (2018), 515. http://doi.org/10.3389/fgene.2018.00515
16. U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, H. H. R. Sherazi, Machine learning based diabetes classification and prediction for healthcare applications, *J. Healthc. Eng.*, **2021** (2021), 9930985. http://doi.org/10.1155/2021/9930985
17. R. Rajni, A. Amandeep, RB-Bayes algorithm for the prediction of diabetic in Pima Indian dataset, *International Journal of Electrical and Computer Engineering*, **9** (2019), 4866–4872. http://doi.org/10.11591/ijece.v9i6.pp4866-4872
18. D. Sisodia, D. S. Sisodia, Prediction of diabetes using classification algorithms, *Procedia Computer Science*, **132** (2018), 1578–1585. http://doi.org/10.1016/j.procs.2018.05.122
19. N. Pradhan, G. Rani, V. S. Dhaka, R. C. Poonia, 14-Diabetes prediction using artificial neural network, In: *Deep learning techniques for biomedical and health informatics*, Academic Press, 2020, 327–339. https://doi.org/10.1016/B978-0-12-819061-6.00014-8
20. K.-C. Chou, Impacts of bioinformatics to medicinal chemistry, *Med. Chem.*, **11** (2015), 218–234. http://doi.org/10.2174/1573406411666141229162834
21. P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets, *Int. J. Mol. Sci.*, **15** (2014), 3495–3506. http://doi.org/10.3390/ijms15033495

22. K.-C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, *Curr. Proteomics*, **6** (2009), 262–274. http://doi.org/10.2174/157016409789973707
23. K. Chou, An insightful recollection since the birth of Gordon Life Science Institute about 17 years ago, *Adv. Sci. Eng. Res.*, **4** (2019), 31–36. http://doi.org/10.33495/aser_v4i2.19.105
24. B. Liu, F. Liu, L. Fang, X. Wang, K.-C. Chou, repRNA: a web server for generating various feature vectors of RNA sequences, *Mol. Genet. Genomics*, **291** (2016), 473–481. http://doi.org/10.1007/s00438-015-1078-7
25. B. Liu, F. Liu, L. Fang, X. Wang, K.-C. Chou, RepDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects, *Bioinformatics*, **31** (2015), 1307–1309. http://doi.org/10.1093/bioinformatics/btu820
26. W. Chen, T. Y. Lei, D. C. Jin, H. Lin, K.-C. Chou, PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition, *Anal. Biochem.*, **456** (2014), 53–60. http://doi.org/10.1016/j.ab.2014.04.001
27. H. Lin, E. Z. Deng, H. Ding, W. Chen, K.-C. Chou, IPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, *Nucleic Acids Res.*, **42** (2014), 12961–12972. http://doi.org/10.1093/nar/gku1019
28. W. Chen, P. M. Feng, H. Lin, K.-C. Chou, ISS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, *Biomed Res. Int.*, **2014** (2014), 623149. http://doi.org/10.1155/2014/623149
29. J. Lu, R. T. Kerns, S. D. Peddada, P. R. Bushelet, Principal component analysis-based filtering improves detection for Affymetrix gene expression arrays, *Nucleic Acids Res.*, **39** (2011), e86. http://doi.org/10.1093/nar/gkr241
30. T. Postelnicu, Probit analysis, In: *International encyclopedia of statistical science*, Berlin, Heidelberg: Springer, 2011, 1128–1131. http://doi.org/10.1007/978-3-642-04898-2_461
31. R. Bro, A. K. Smilde, Principal component analysis, *Anal. Methods*, **6** (2014), 2812–2831. http://doi.org/10.1039/c3ay41907j
32. G. P. Zhou, D. Chen, S. Liao, R.-B. Huang, Recent progresses in studying helix-helix interactions in proteins by incorporating the Wenxiang diagram into the NMR spectroscopy, *Curr. Top. Med. Chem.*, **16** (2015), 581–590. http://doi.org/10.2174/1568026615666150819104617
33. P. Geladi, H. Isaksson, L. Lindqvist, S. Wold, K. Esbensen, Principal component analysis of multivariate images, *Chemom. Intell. Lab. Syst.*, **5** (1989), 209–220. http://doi.org/10.1016/0169-7439(89)80049-8
34. C. Goodall, Principal component analysis, *Technometrics*, **30** (1988), 351–352. http://doi.org/10.2307/1270093
35. X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2019, 4487–4496. http://doi.org/10.18653/v1/p19-1441
36. N. Tishby, N. Zaslavsky, Deep learning and the information bottleneck principle, In: *2015 IEEE Information Theory Workshop (ITW)*, Jerusalem, Israel, 2015, 1–5. http://doi.org/10.1109/ITW.2015.7133169

37. S. Khan, M. Khan, N. Iqbal, M. Li, D. M. Khan, Spark-based parallel deep neural network model for classification of large scale RNAs into piRNAs and non-piRNAs, *IEEE Access*, **8** (2020), 136978–136991. http://doi.org/10.1109/ACCESS.2020.3011508
38. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence, *Sci. Rep.*, **6** (2016), 27755. http://doi.org/10.1038/srep27755
39. A. Majid, M. M. Khan, N. Iqbal, M. A. Jan, M. Khan, Salman, Application of parallel vector space model for large-scale DNA sequence analysis, *J. Grid Comput.*, **17** (2019), 313–324. http://doi.org/10.1007/s10723-018-9451-5
40. T. Hussain, H. F. Maqbool, N. Iqbal, M. Khan, Salman, A. A. Dehghani-Sanij, Computational model for the recognition of lower limb movement using wearable gyroscope sensor, *Int. J. Sens. Networks*, **30** (2019), 35–45. http://doi.org/10.1504/IJSNET.2019.099230
41. J. H. Miao, K. H. Miao, Cardiotocographic diagnosis of fetal health based on multiclass morphologic pattern predictions using deep learning classification, *Int. J. Adv. Comput. Sci. Appl.*, **9** (2018), 1–11. http://doi.org/10.14569/IJACSA.2018.090501
42. N. Inayat, M. Khan, N. Iqbal, S. Khan, M. Raza, D. M. Khan, et al., iEnhancer-DHF: identification of enhancers and their strengths using optimize deep neural network with multiple features extraction methods, *IEEE Access*, **9** (2021), 40783–40796. http://doi.org/10.1109/ACCESS.2021.3062291
43. F. Khan, M. Khan, N. Iqbal, S. Khan, D. M. Khan, A. Khan, et al., Prediction of recombination spots using novel hybrid feature extraction method via deep learning approach, *Front. Genet.*, **11** (2020), 1052. http://doi.org/10.3389/fgene.2020.539227
44. S. Khan, M. Khan, N. Iqbal, M. A. A. Rahman, M. K. A. Karim, Deep-piRNA: bi-layered prediction model for PIWI-interacting RNA using discriminative features, *Comput. Mater. Con.*, **72** (2022), 2243–2258. http://doi.org/10.32604/cmc.2022.022901
45. S. Khan, M. Khan, N. Iqbal, S. A. Khan, K.-C. Chou, Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC, *Chemom. Intell. Lab. Syst.*, **203** (2020), 104056. http://doi.org/10.1016/j.chemolab.2020.104056
46. S. Khan, M. Khan, N. Iqbal, T. Hussain, S. A. Khan, K.-C. Chou, A two-level computation model based on deep learning algorithm for identification of piRNA and their functions via Chou's 5-steps rule, *Int. J. Pept. Res. Ther.*, **26** (2020), 795–809. http://doi.org/10.1007/s10989-019-09887-3
47. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure-activity relationships, *J. Chem. Inf. Model.*, **55** (2015), 263–274. http://doi.org/10.1021/ci500747n
48. M. K. K. Leung, H. Y. Xiong, L. J. Lee, B. J. Frey, Deep learning of the tissue-regulated splicing code, *Bioinformatics*, **30** (2014), i121–i129. http://doi.org/10.1093/bioinformatics/btu277
49. M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, W. Denk, Connectomic reconstruction of the inner plexiform layer in the mouse retina, *Nature*, **500** (2013), 168–174. http://doi.org/10.1038/nature12346
50. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, et al., Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, *IEEE Signal Proc. Mag.*, **29** (2012), 82–97. http://doi.org/10.1109/MSP.2012.2205597

51. T. N. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, Deep convolutional neural networks for LVCSR, In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, 2013, 8614–8618. http://doi.org/10.1109/ICASSP.2013.6639347
52. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. http://doi.org/10.1145/3065386
53. C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2013), 1915–1929. http://doi.org/10.1109/TPAMI.2012.231
54. U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals, *Comput. Biol. Med.*, **100** (2018), 270–278. http://doi.org/10.1016/j.compbiomed.2017.09.017
55. Z. Zhu, E. Albadawy, A. Saha, J. Zhang, M. R. Harowicz, M. A. Mazurowski, Deep learning for identifying radiogenomic associations in breast cancer, *Comput. Biol. Med.*, **109** (2019), 85–90. http://doi.org/10.1016/j.compbiomed.2019.04.018
56. T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur, Extensions of recurrent neural network language model, In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, 5528–5531. http://doi.org/10.1109/ICASSP.2011.5947611
57. A. Bordes, S. Chopra, J. Weston, Question answering with subgraph embeddings, In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA, Association for Computational Linguistics, 2014, 615–620. http://doi.org/10.3115/v1/D14-1067
58. A. Baratloo, M. Hosseini, A. Negida, G. E. Ashal, Part 1: simple definition and calculation of accuracy, sensitivity and specificity, *Emergency*, **3** (2015), 48–49. http://doi.org/10.22037/emergency.v3i2.8154
59. J. Chen, H. Liu, J. Yang, K.-C. Chou, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, *Amino Acids*, **33** (2007), 423–428. http://doi.org/10.1007/s00726-006-0485-9
60. Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences, *Nucleic Acids Res.*, **36** (2008): 3025–3030. http://doi.org/10.1093/nar/gkn159
61. M. F. Sabooh, N. Iqbal, M. Khan, M. Khan, H. F. Maqbool, Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC, *J. Theor. Biol.*, **452** (2018), 1–9. http://doi.org/10.1016/j.jtbi.2018.04.037
62. P. M. Feng, W. Chen, H. Lin, K.-C. Chou, IHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition, *Anal. Biochem.*, **442** (2013), 118–125. http://doi.org/10.1016/j.ab.2013.05.024
63. Y. Xu, J. Ding, L.-Y. Wu, K.-C. Chou, iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, *PLoS One*, **8** (2013), e55844. http://doi.org/10.1371/journal.pone.0055844
64. W. Chen, P. M. Feng, H. Lin, K.-C. Chou, IRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, *Nucleic Acids Res.*, **41** (2013), e68. http://doi.org/10.1093/nar/gks1450

65. B. Liu, F. Yang, K.-C. Chou, 2L-piRNA: a two-layer ensemble classifier for identifying Piwi-interacting RNAs and their function, *Mol. Ther.-Nucl. Acids*, **7** (2017), 267–277. http://doi.org/10.1016/j.omtn.2017.04.008

66. A. R. Hedar, M. Almaraashi, A. E. Abdel-Hakim, M. Abdulrahim, Hybrid machine learning for solar radiation prediction in reduced feature spaces, *Energies*, **14** (2021), 7970. https://doi.org/10.3390/en14237970