Journal of Computational Science 81 (2024) 102388

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

IP-GCN: A deep learning model for prediction of insulin using graph convolutional network for diabetes drug design

Farman Ali [a,*], Majdi Khalid [b], Abdullah Almuhaimeed [c,*], Atef Masmoudi [d], Wajdi Alghamdi [e], Ayman Yafoz [f]

[a] *Department of Computer Science, Bahria University Islamabad Campus, Pakistan*
[b] *Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah 21955, Saudi Arabia*
[c] *Digital Health Institute, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*
[d] *Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia*
[e] *Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia*
[f] *Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

ARTICLE INFO

Keywords:
Insulin protein
Deep learning
Machine learning
Graph convolutional network

ABSTRACT

Insulin is a kind of protein that regulates the blood sugar levels is significant to prevent complications associated with diabetes, such as cancer, neurodegenerative disorders, cardiovascular disease, and kidney damage. Insulin protein (IP) plays an active role in drug discovery, medicine, and therapeutic methods. Unlike experimental protocols, computational predictors are fast and can predict IP accurately. This work introduces a model, called IP-GCN for IP prediction. The patterns from IP are extracted by K-spaced position specific scoring matrix (KS-PSSM) and the model training is accomplished using powerful deep learning tool, called Graph Convolutional Network (GCN). Additionally, we implemented Pseudo Amino Acid Composition (PseAAC) and Dipeptide Composition (DPC) for feature encoding to assess the predictive performance of GCN. To evaluate the efficacy of our novel approach, we compare its performance with well-known deep/machine learning algorithms such as Convolutional Neural Network (CNN), Extremely Randomized Tree (ERT), and Support Vector Machine (SVM). Predictive results demonstrate that the proposed predictor (IP-GCN) secured the best performance on both training and testing datasets. The novel computational would be fruitful in diabetes drug discovery and contributes to research for therapeutic interventions in various Insulin protein associated diseases.

1. Introduction

Insulin protein (IP) composed of amino acids, holds immense significance in human physiology and healthcare. It regulates glucose metabolism and maintains optimal blood sugar levels. The importance of insulin extends beyond its role in diabetes management, as it influences various biological processes and has broader implications in both health and disease [1]. IP serves as a key regulator of blood glucose levels. After a meal, when blood sugar rises, pancreas releases insulin to signal body tissues, particularly muscle, adipose tissue, and liver for glucose absorption from the bloodstream. This process stores excess glucose as glycogen in the muscles and liver, preventing blood sugar from reaching dangerously high levels. It also decomposes the stored glycogen for energy production [2].

Insulin Protein is indispensable in the treatment of diabetes. Insulin, a hormone synthesized by the pancreas, is vital for regulating blood sugar levels. Insulin therapy entails the administration of insulin through injections or an insulin pump to compensate for the inadequate or absent natural insulin. This therapy aids in reducing blood sugar levels and enables glucose to enter the cells, where it can be utilized for energy [3].

Elevated insulin levels and insulin resistance have been associated with several kinds of cancers including pancreatic, breast, and colorectal cancer. Insulin and IGFs can stimulate cell division and inhibit cell death, potentially contributing to tumor growth [4]. Disruptions in insulin resistance and the signaling pathways associated with insulin in the brain have the capacity to influence glucose metabolism, resulting in dysfunction of brain cells and triggering neuroinflammation [5].

Insulin has anticoagulant properties that help to prevent excessive blood clotting. It inhibits the production of clotting factors and promotes

* Corresponding authors.
E-mail addresses: farman335@yahoo.com, farman.buic@bahria.edu.pk (F. Ali), muhaimed@kacst.edu.sa (A. Almuhaimeed).

https://doi.org/10.1016/j.jocs.2024.102388
Received 23 October 2023; Received in revised form 21 June 2024; Accepted 9 July 2024
Available online 14 July 2024

the release of fibrinolytic agents that dissolve blood clots. Impaired insulin signaling can disrupt this balance and increase the risk of abnormal clotting [6]. During growth and development, particularly in childhood and adolescence, insulin plays a vital role. Its primary function is to facilitate the uptake of amino acids into cells, which is essential for protein synthesis and the growth and repair of tissues [7]. Proper insulin signaling is essential for normal growth, development, and maturation of various organ systems [8]. A series of computational models have been proposed for many biological problems [9–13]. Considering the above significance, the accurate identification of insulin protein is indispensable. The identification via experimental methods is costly, laborious, and time-expensive [14]. This work proposes a promising identification model ofor insulin protein.

Features are discovered using the methods namely PseAAC, DPC, and KS-PSSM. We also adopted three classifiers including CNN, ERT, and SVM to compare the performance of the IP-GCN. Each method's results are validated with 5-fold cross-validation (CV). The results analysis illustrated that our predictor surpassed all the classifiers in terms of performance. The graphical view of the proposed predictor is shown in Fig. 1.

2. Material and methods

2.1. Dataset construction

The efficacy of a novel protocol heavily relies on the construction of reliable datasets. Considering this, we have created a new dataset using sequences of insulin proteins (IP) and non-insulin proteins (non-IP). The dataset consists of two classes: the positive class, which includes IP sequences, and the negative class, which includes non-IP sequences. We obtained these samples from UniProt database. Similarly, the same process was repeated for the negative samples. First, we downloaded 745 insulin protein and 734 non-insulin protein. Second, we applied CD-HIT tool [15] for removing 25 % sequence similarity index. Third, sequences containing less than 50 amino acids were eliminated. Finally, the remaining set contains 849 IP and 832 non-IP sequences which were considered as final dataset.

The training set comprised 718 IP sequences and 703 non-IP sequences, while the testing set included 131 IP sequences and 129 non-IP sequences.

2.2. Feature encoding method

Direct prediction of the class of protein sequences is not feasible for learning algorithms [16]. Protein sequences are converted to numerical form to extract feature using feature descriptors [17]. In this connection, we used KS-PSSM to explore meaningful information for effective model development.

PSSM is a matrix that represents the statistical occurrence of each amino acid at a particular position in a protein sequence [18,19]. It is derived from multiple sequence alignments, which capture evolutionary information by comparing related protein sequences [20]. PSSM provides a quantitative [21] measure of the conservation or variation of each amino acid residue in a protein family, enabling researchers to identify functionally important regions and make predictions about protein function [22]. In this connection, PSSM was used in several

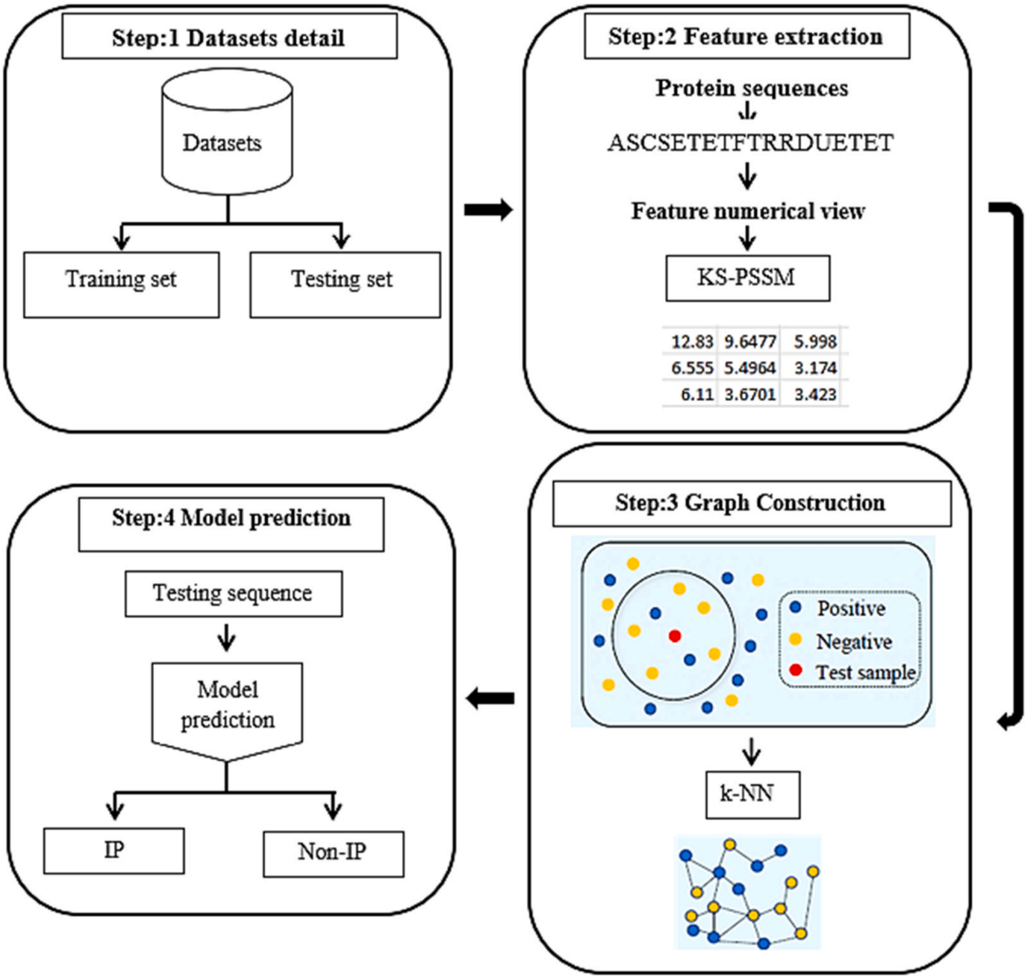

Fig. 1. Flow work of the IP-GCN.

biological prediction tasks [23–27].

The K-spaced PSSM feature descriptor takes the PSSM one step further by considering not only the direct neighbors of each amino acid but also the amino acids that are K positions away. By incorporating this additional context, the K-spaced PSSM captures long-range interactions and dependencies within the protein sequence, which improves model performance [28]. We set K values ranging from 1 to 7 and find the best results at K=2.

2.3. Feature representation via graph

Graph Convolutional Networks (GCN) have experienced a surge in popularity when it comes to addressing a wide range of research problems involving graph and network data [29,30]. Recent research has demonstrated that by leveraging correlated information extracted from topological structures using k-nearest neighbors (k-NN) [31,32], the effectiveness of GCN can be enhanced, particularly in certain classification tasks [33]. Motivated by the aforementioned study, we incorporated the k nearest neighbors (k-NN) to create a feature graph, enabling us to capture structural information within the feature space [34]. Our approach involved several steps. Firstly, we computed distance between all pairs of nodes using Euclidean formula [35]. Afterward, a k-NN classifier was utilized to create the graph, determining the k nearest neighbors for each node [36]. This allowed us to establish edges connecting each node with its respective k nearest neighbors [37]. It is important to note that the selection of k can significantly influence the outcome [38]. To investigate how the number of neighbors affects insulin protein prediction, we conducted experiments by systematically varying the value of k [39,40]. To maintain the integrity of the training phase [41,42], we modified the labels of the test data to 0, ensuring they were not included in the training set. This step was taken to prevent any potential contamination or bias during the training process [24].

2.4. Graph convolutional network

Graph Convolutional Network (GCN) is a deep learning tool which is widely applied in several biological problems like drug-protein interaction [43,44], identification of anticancer peptides [33], and DNA-binding protein prediction [45]. GCN leverages neighborhood aggregation to capture both local and global information from the graph. By aggregating information from adjacent nodes, GCN can capture the local structure and relationships within a neighborhood [46]. At the same time, they can incorporate information from distant nodes to capture global patterns and dependencies. We consider a homogeneous graph, represented as G = (V, E), where E and V indicate the edge set and node set, respectively, capturing the intricate associations among the nodes [47,48].

Graph convolution in the context of graphs can be understood as the product of an input signal $X \in R^{N\times C}$ where C denotes the dimensionality of the feature vector for each node, with a Fourier domain filter $g_{\varnothing}$. The GCN model relies on two essential inputs: the feature matrix X and the adjacency matrix A [33]. The normalized graph Laplacian is computed as

$$L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{1}$$

where, I_N represents the identity matrix and $D_{ii} = \sum_j A_{ij}$ represents the degree matrix. Assuming U represents the eigenvector matrix of L and M denotes the diagonal matrix of its eigenvalues, L can be transformed into $L = UMU^T$. As a result, the convolutional operation is defined as $g_{\varnothing} \times X = Ug_{\varnothing}U^T\mathrm{X}$. To perform the convolutional operation, we begin by obtaining the eigenvector matrix U and the diagonal matrix of eigenvalues M. With these matrices at hand, we execute the convolution by multiplying U with $g_{\varnothing}$ and U^T with X.

Considering the computational complexity of computing the decomposition of L, Rao et al. [33] proposed a framework known as Chebyshev graph convolution. This framework utilizes a truncated expansion of Chebyshev polynomials to approximate the filter. The filter $g_{\varnothing}$ can be approximated as

$$g_{\varnothing} \approx \sum_{k=0}^{k} \varnothing_k T_k(\widetilde{L}) \tag{2}$$

where, $\varnothing_k T_k$ is the Chebyshev polynomials with Chebyshev coefficients, and $\widetilde{L}$ is defined as $\widetilde{L} = \frac{2}{\lambda\mathrm{max}}L - I_N$ with λmax being the largest eigenvalue of L. To further simplify Eq. (2), Kipf et al. [49] limited K to 1 and approximated $\lambda\mathrm{max} \approx 2$. As a result, formula Eq. (2) becomes

$$g_{\varnothing} = I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{3}$$

To address numerical instabilities that arise when applying this operator repeatedly, they proposed renormalization trick: $I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ was replaced by $\widetilde{D}^{-\frac{1}{2}} \widetilde{A}\widetilde{D}^{-\frac{1}{2}}$, where $\widetilde{A} = A + I_N$ and $\widetilde{D}_{ii} = \sum_j A_{ij}$. Consequently, the following equation is used to represent the final convolutional operation.

$$g_{\varnothing} \times X = \varnothing\widetilde{D}^{-\frac{1}{2}} \widetilde{A}\widetilde{D}^{-\frac{1}{2}} X \tag{4}$$

In our GCN model, we utilized the final form of graph convolution along with the layer-wise propagation rule which is described by Eq. (5).

$$H^{(l+1)} = \sigma(\widetilde{D}^{-\frac{1}{2}} \widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{5}$$

where, $H^{(l)}$ and $W^{(l)}$ indicate the feature matrix and trainable weight matrix of the 1th layer, respectively, and ReLU activation function is denoted by σ. For prediction of insulin protein, we constructed a two-layer GCN model. The forward propagation formula is expressed as

$$\widehat{Y} = softmax(\widehat{A}ReLU(\widehat{A}XW^{(0)})W^{(1)}), \quad \widehat{A} = \widetilde{D}^{-\frac{1}{2}} \widetilde{A}\widetilde{D}^{-\frac{1}{2}} \tag{6}$$

The probability distribution is computed by softmax function, which is defined as

$$softmax(x_i) = \frac{1}{\sum_i \exp(x_i)}\exp(x_i) \tag{7}$$

To train the network, we used cross-entropy ($\mathbb{C}$) loss function which is formulated as

$$\mathbb{C} = -\frac{1}{|Tr|}\sum[y_i.\log(\widehat{Y}_l) + (1 - Y_i) .\log(1 - \widehat{Y}_l)] \tag{8}$$

where, Tr represents the training set, Y_i denotes the label of node i, and $\widehat{Y}_l$ represents the predicted probability of node i being insulin protein. To minimize loss function, the Adam optimizer is employed. Additionally, dropout is used to prevent overfitting during the training process. Other hyperparameters like learning rate and epochs have described in the result and discussion section.

2.5. Model validation

After designing a novel predictor, its performance is examined by a reliable method [50]. 5-fold CV is extensively used for assessment of predictor performance in bioinformatics [51–55] and other research areas [56–59]. In this connection, we implemented 5-fold with evaluation metrics including accuracy (Acc), sensitivity (Sn), specificity (Sp), and Mathew correlation coefficient (MCC). [60–62]. CM generates true negative (TN), false positive (FP), true positive (TP), and false negative (FN) which can be used for calculation of the said parameters. TP indicates the true prediction of IP and TN represents the non-IP. FP denotes the IP which is detected by model as non-IP. Similarly, FN has been used to show the non-IP which is identified by model as IP. The evaluation parameters are defined below.

$$Accuracy = TP + TN/TP + FP + TN + FN \tag{8}$$

$$Sensitivity = TP/TP + FN \tag{9}$$

$$Specificity = TN/TN + FP \tag{10}$$

$$MCC = (TN \times TP) - (FN \times FP)\Big/\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)} \tag{11}$$

3. Results and discussion

In this section, we discussed results obtained by classification algorithms including GCN and machine learning (ML) classifiers and their comparative performance analysis. The details of predictive results are listed below.

3.1. Performance of GCN with different hyperparameters

The selection of the best parameters could enhance a model performance as reported in several studies [63–66]. In this connection, we tuned the hyperparameters of the proposed approach. To construct a promising GCN model, we applied different hyperparameters including hidden layers, number of neighbor nodes (k-NN), dropout layer, epochs, learning rate, and optimizer. The number of hidden layers in a GCN is a crucial architectural parameter that can significantly impact the model's performance. In this connection, we implemented hidden layers ranging from 32, 64, 128, 256,512, and 1024 and recorded their outcomes in Table 1. Table 1 indicates that increase in the number of hidden layers leads to improvement in performance of the model and the best performance was achieved by model on 256. Further increase in the hidden layers causes decrease in model performance which means that redundant features are extracted.

Similarly, the number of k-nearest neighbors has a great impact on the model performance. We set k=2,3,4,5, and 6 and summarized their results in Table 2. The GCN generated 90.45 % and 91.91 % accuracies using 2 and 3 values of k, respectively.

The Dropout layer is a regularization technique commonly used in neural networks, including GCN. By reducing overfitting, Dropout improves the model's generalization capabilities. It encourages the network to learn more diverse and independent features. Dropout helps the model generalize better to unseen examples, leading to improved performance on testing data. We used dropout values ranging from 0.1 to 0.5. The result generated by model on these values is listed in Table 3. On 0.1, the model produced 90.34 % Acc, 90.65 % Sn, 85.04 % Sp, and 0.80 MCC. The promising performance was shown by model on 0.2 and secured the best 92.52 % accuracy and highest success rate on all evaluation parameters. Further enhancement in dropout value gradually decreases is reported by all parameters. Hence, 0.2 is considered the best value for dropout layer for generating the highest results of the model. Onward, we also analyzed the best values for learning rate (0.001), epochs (1000), and adam optimizer.

Table 1
Filter-wise results of the GCN.

Filter	Acc (%)	Sn (%)	Sp (%)	MCC
32	88.46	89.06	84.94	0.78
64	90.44	90.53	84.67	0.79
128	91.23	91.34	85.97	0.80
256	**92.52**	**92.06**	**86.15**	**0.81**
512	92.16	91.60	86.03	0.81
1024	91.89	91.96	85.88	0.81

On k=3, the highest prediction was observed. However, with k=5, 6 model gradually reduced the performance. Hence, the optimum value for k=3.

Table 2
Results of GCN with different numbers of KNN values.

Knn value	Acc (%)	Sn (%)	Sp (%)	MCC
2	90.45	90.67	84.23	0.79
3	91.91	91.89	85.87	0.80
4	**92.52**	**92.06**	**86.15**	**0.81**
5	92.04	91.99	85.84	0.81
6	91.43	91.76	85.45	0.80

Table 3
Dropout-wise results of the GCN.

Dropout	Acc (%)	Sn (%)	Sp (%)	MCC
0.1	90.34	90.65	85.04	0.80
0.2	**92.52**	**92.06**	**90.15**	**0.84**
0.3	91.56	91.65	89.85	0.83
0.4	91.11	91.21	88.22	0.82
0.5	90.97	91.32	84.65	0.80

3.2. Comparative analysis of GCN with machine learning algorithms

GCN is compared with deep learning framework i.e., CNN and traditional two ML classifiers including SVM) and ERT and shown their comparative performance in Table 4. Additionally, two feature encoders PseAAC and DPC are adopted to compare their performance with KS-PSSM feature descriptor using GCN and other classification algorithms. From Table 4, it can be seen that SVM generated 87.54 % accuracy using PseAAC. ERT slightly decrease its performance and produce 87.25 % Acc. CNN obtained 88.35 % Acc, 88.01 % Sn, 88.76 % Sp, and 0.78 MCC. GCN demonstrates the highest performance with 90.17 % Acc, 90.12 % Sn, 88.96 % Sp, and 0.79 MCC.

With DPC feature descriptor, SVM and ERT classifiers obtained 87.45 % and 88.14 % accuracies, respectively. Compared with PseAAC, both classifiers showed improved results. CNN also performs well with an accuracy of 89.79 %, sensitivity of 87.31 %, specificity of 88.76 %, and MCC of 0.80. GCN continues to exhibit remarkable performance, achieving 90.58 % Acc, 89.67 % Sn, 89.91 % Sp, and 0.82 MCC. Among all classifiers, the predictive results of GCN are promising and consistently showed the best performance on both PseAAC and DPC descriptors.

Further analyzing the classifier's outcomes on KS-PSSM, it is noted that all classifiers raised the performance. For instance, SVM and ERT yielded an accuracy of 88.78 % and 89.24 %, respectively, which are better than PseAAC and DPC. The accuracy achieved by CNN is 90.31 % which is higher than SVM and ERT. The best efficacy was reflected by GCN producing 92.52 % Acc, 92.06 % Sn, 90.15 % Sp, and 0.84 MCC which are the highest success rate than other classifiers.

We depicted the ROC curves of all classifiers in Fig. 2. ROC curve of the GCN outperformed the ROC curves of other classifiers that reflect the robustness of the GCN model.

Table 4
Classifiers performance using training set.

Classifier	Feature descriptor	Acc (%)	Sn (%)	Sp (%)	MCC
SVM	PseAAC	87.54	84.40	88.68	0.77
ERT		87.25	85.30	89.29	0.78
CNN		88.35	88.01	88.76	0.78
GCN		90.17	90.12	88.96	0.79
SVM	DPC	87.75	86.63	89.33	0.76
ERT		88.14	88.58	87.41	0.78
CNN		89.79	87.31	88.76	0.80
GCN		90.58	89.67	89.91	0.82
SVM	KS-PSSM	88.78	86.97	90.54	0.78
ERT		89.24	89.54	87.56	0.80
CNN		90.31	90.03	88.76	0.82
GCN		**92.52**	**92.06**	**90.15**	**0.84**

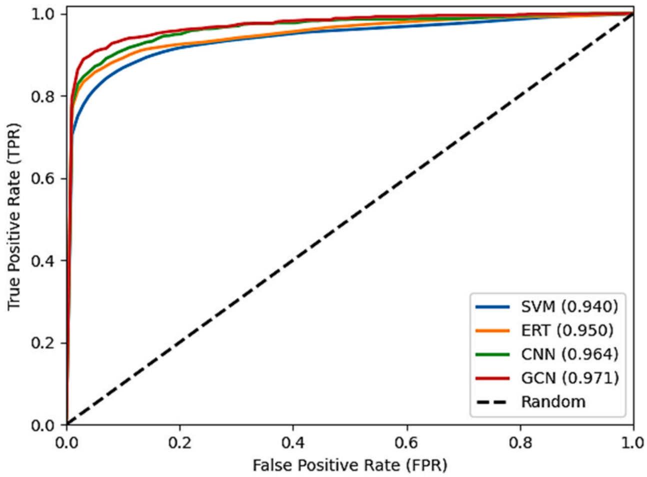

Fig. 2. ROC curves of applied classifiers.

Compare with PseAAC and DPC, KS-PSSM can capture meaningful information that results in higher performance on all classifiers. The superior performance of KS-PSSM over PseAAC and DPC is due to exploring evolutionary patterns and extracting the correlation information in consecutive and K-spaced amino acids. Similarly, GCN consistently shows the highest performance across the different feature descriptors, indicating its effectiveness in capturing complex patterns in the data.

3.3. *Comparative performance analysis of classifiers on the testing dataset*

The testing dataset provides an independent set of samples that were not used during the training phase. It allows for an unbiased assessment of how well the predictor generalizes to unseen data. By evaluating the predictor on the testing dataset, its performance can be accurately measured and compared against other predictors. In this connection, the performance of the IP-GCN predictor was further evaluated using the testing dataset and compared its outcomes with the applied other classification algorithms. From Table 5, we can observe that GCN achieved 87.68 % Acc, 86.79 % Sn, 87.06 % Sp, and 0.76 MCC which are 2.61 %, 0.68 %, 2.19 %, and 2 % higher than second-best predictor (CNN) in term of accuracy, sensitivity, specificity, and MCC. GCN improved the Acc, Sn, Sp, and MCC by 3.7 %, 2.39 %, 1.88 %, and 3 % to ERT model. Similarly, GCN surpassed SVM predictor on all evaluation parameters. These results reveal that proposed (IP-GCN) is promising for unseen data and higher generalization power than other predictors.

In summary, the remarkable performance of GCN with different feature descriptors using training dataset and testing dataset indicates that GCN model can accurately predict insulin protein.

4. Conclusion

Insulin protein is crucial for glucose metabolism and the treatment of diabetes. To predict insulin protein accurately, we proposed a novel computational predictor using deep learning approach. The feature engineering is carried out by KS-PSSM followed by Graph construction strategy. The training and prediction are performed using GCN. We also implemented PseAAC and DPC for feature encoding to evaluate the predictive results of GCN. The performance of our novel predictor is compared with famous deep/machine learning including CNN, ERT, and SVM algorithms. On the training dataset, GCN generated superlative performance over all learning models using all feature-encoded methods. The proposed study generalization efficiency is assessed by testing datasets and showed remarkable performance.

Table 5
Classifiers performance on the testing set.

Classifier	Acc (%)	Sn (%)	Sp (%)	MCC
SVM	83.35	84.23	83.57	0.73
ERT	83.98	84.40	85.18	0.73
CNN	85.07	86.11	84.87	0.74
GCN	**87.68**	**86.79**	**87.06**	**0.76**

The best performance of the model is due to several factors including construction of reliable datasets, feature extraction by efficient technique, and classification by effective deep learning tool. In addition to being an accurate predictor for insulin protein, it will play an active role in drug discovery, medicine, and therapeutic methods.

5. Future work and limitations of the proposed study

While the proposed IP-GCN predictor demonstrates superior performance in identifying insulin protein, several limitations and areas for future research need to be addressed. The current computational methods, including the IP-GCN, offer a fast and accurate alternative to time-consuming and expensive experimental protocols. However, the study's reliance on KS-PSSM for pattern extraction and the use of PseAAC and DPC for feature encoding could be expanded. Future work should explore additional feature extraction methods to enhance predictive accuracy and robustness. Furthermore, improving the model's generalizability across diverse datasets is essential to ensure its broader applicability. Reducing the computational cost of the algorithm is another critical area to make the IP-GCN more accessible for widespread use in drug discovery and therapeutic research. Addressing these limitations will contribute to more effective identification and analysis of insulin proteins, ultimately aiding in the development of therapeutic interventions for various insulin protein-associated disease.

CRediT authorship contribution statement

Abdullah Almuhaimeed: Software, Supervision. **Atef Masmoudi:** Funding acquisition, Visualization. **Majdi Khalid:** Formal analysis, Resources. **Farman Ali:** Conceptualization, Data curation. **Ayman Yafoz:** Writing – review & editing. **Wajdi Alghamdi:** Methodology, Validation.

Declaration of Competing Interest

The authors declare no conflict of interests.

Data availability

Data will be made available on request.

Acknowledgment

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/424/45.

References

[1] A. Virkamäki, K. Ueki, and C.R.J.T.J. o c i. Kahn, "Protein–protein interaction in insulin signaling and the molecular mechanisms of insulin resistance," vol. 103, no. 7, pp. 931-943, 1999.
[2] X.J. Sun *et al.*, "Structure of the insulin receptor substrate IRS-1 defines a unique signal transduction protein," vol. 352, no. 6330, pp. 73-77, 1991.
[3] J.J.I.A. Avruch, "Insulin signal transduction through protein kinase cascades," pp. 31-48, 1998.
[4] M.G. Myers Jr and M.F.J.D. White, "The new elements of insulin signaling: insulin receptor substrate-1 and proteins with SH2 domains," vol. 42, no. 5, pp. 643-650, 1993.
[5] P.G. Drake, B.I.J.M. Posner, and C. biochemistry, "Insulin receptor-associated protein tyrosine phosphatase (s): role in insulin action," vol. 182, pp. 79-89, 1998.
[6] R.G. Fred, L. Tillmar, and N.J.C. d r. Welsh, "The role of PTB in insulin mRNA stability control," vol. 2, no. 3, pp. 363-366, 2006.

[7] M. Krüger, I. Kratchmarova, B. Blagoev, Y.-H. Tseng, C.R. Kahn, and M. J. P. o. t. N. A. o. S. Mann, "Dissection of the insulin signaling pathway via quantitative phosphoproteomics," vol. 105, no. 7, pp. 2451-2456, 2008.

[8] M.F. White and C.R.J.J. o B.C. Kahn, "The insulin signaling system," vol. 269, no. 1, pp. 1-4, 1994.

[9] P. Charoenkwan *et al.*, "PSRTTCA: A new approach for improving the prediction and characterization of tumor T cell antigens using propensity score representation learning," vol. 152, p. 106368, 2023.

[10] P. Charoenkwan, N. Schaduangrat, M.A. Moni, W. Shoombuatong, and B.J.I. Manavalan, "Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework," vol. 25, no. 9, 2022.

[11] S. Hongjaisee, C. Nantasenamat, T.S. Carraway, W.J.C.B. Shoombuatong, HIVCoR: A sequence-based tool for predicting HIV-1 CRF01_AE coreceptor usage, Chemistry, " vol. 80 (2019) 419–432.

[12] P. Charoenkwan, W. Chotpatiwetchkul, V.S. Lee, C. Nantasenamat, and W.J.S.R. Shoombuatong, "A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides," vol. 11, no. 1, p. 23782, 2021.

[13] P. Charoenkwan, N. Schaduangrat, M.A. Moni, B. Manavalan, W. J. C. i. B. Shoombuatong, and Medicine, "SAPPHIRE: A stacking-based ensemble learning framework for accurate prediction of thermophilic proteins," vol. 146, p. 105704, 2022.

[14] S. Akbar, A. Ahmad, M. Hayat, A.U. Rehman, S. Khan, F. Ali, "iAtbP-Hyb-EnC: Prediction of Antitubercular peptides Via Heterogeneous Feature Representation and Genetic Algorithm based Ensemble Learning Model,", Comput. Biol. Med. (2021) 104778.

[15] Y. Huang, B. Niu, Y. Gao, L. Fu, W.J.B. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, no vol. 26 (5) (2010) 680–682.

[16] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, S. Gul, "Prediction of Antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy,", Chemom. Intell. Lab. Syst. vol. 230 (2022) 104682.

[17] S. Akbar *et al.*, "Prediction of Amyloid Proteins using Embedded Evolutionary & Ensemble Feature Selection based Descriptors with eXtreme Gradient Boosting Model," 2023.

[18] S. Akbar *et al.*, "Identifying Neuropeptides via Evolutionary and Sequential based Multi-perspective Descriptors by Incorporation with Ensemble Classification Strategy," 2023.

[19] R. Alsini *et al.*, "Deep-VEGF: deep stacked ensemble model for prediction of vascular endothelial growth factor by concatenating gated recurrent unit with two-dimensional convolutional neural network," pp. 1-11, 2024.

[20] O. Alghushairy *et al.*, "Machine learning-based model for accurate identification of druggable proteins using light extreme gradient boosting," pp. 1-12, 2023.

[21] M. Arif, S. Ahmad, F. Ali, G. Fang, M. Li, D.-J. Yu, TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree, J. Comput. -Aided Mol. Des. vol. 34 (8) (2020).

[22] F. Ali, S. Ahmed, Z.N.K. Swati, S. Akbar, DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information, J. Comput. -Aided Mol. Des. vol. 33 (7) (2019) 645–658.

[23] A. Khan *et al.*, "AFP-SPTS: An Accurate Prediction of Antifreeze Proteins Using Sequential and Pseudo-Tri-Slicing Evolutionary Features with an Extremely Randomized Tree," 2023.

[24] M. Kabir, M. Arif, F. Ali, S. Ahmad, Z.N.K. Swati, D.-J. Yu, Prediction of membrane protein types by exploring local discriminative information from evolutionary profiles, Anal. Biochem. vol. 564 (2019) 123–132.

[25] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, S. Gul, "iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach,", Chemom. Intell. Lab. Syst. vol. 204 (2020) 104103.

[26] F. Ali, H. Kumar, S. Patil, A. Ahmed, A. Banjar, A. Daud, "DBP-DeepCNN: Prediction of DNA-binding proteins using wavelet-based denoising and deep learning,", Chemom. Intell. Lab. Syst. (2022) 104639.

[27] F. Ali, H. Kumar, S. Patil, K. Kotecha, A. Banjar, A. Daud, "Target-DBPPred: An intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting,", Comput. Biol. Med. vol. 145 (2022) 105533.

[28] F. Ali, S. Akbar, G. Ali, Z.A. Maher, A. Unar, D.B. Talpur, "AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information,", Comput. Biol. Med. (2021) 105006.

[29] Z.U. Khan, F. Ali, I. Ahmad, M. Hayat, and D. Pi, "iPredCNC: computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection, Chemom. Intell. Lab. Syst. vol. 195 (2019) 103876.

[30] Z.U. Khan, F. Ali, I.A. Khan, Y. Hussain, D. Pi, iRSpot-SPI: Deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components, Chemom. Intell. Lab. Syst. vol. 189 (2019) 169–180.

[31] A. Banjar, F. Ali, O. Alghushairy, A. Daud, "iDBP-PBMD: A machine learning model for detection of DNA-binding proteins by extending compression techniques into evolutionary profile,", Chemom. Intell. Lab. Syst. (2022) 104697.

[32] A. Ghulam, R. Sikander, F. Ali, Z.N.K. Swati, A. Unar, D.B. Talpur, "Accurate prediction of immunoglobulin proteins using machine learning model,", Inform. Med. Unlocked (2022) 100885.

[33] B. Rao, L. Zhang, and G.J.I.A.. Zhang, "Acp-gcn: the identification of anticancer peptides based on graph convolution networks," vol. 8, pp. 176005-176011, 2020.

[34] Z.U. Khan, D. Pi, S. Yao, A. Nawaz, F. Ali, S. Ali, piEnPred: a bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm, Front. Comput. Sci. vol. 15 (6) (2021) 1–11.

[35] R. Sikander, A. Ghulam, and F. J. S. r. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," vol. 12, no. 1, p. 5505, 2022.

[36] A. Ghulam, Z.N.K. Swati, F. Ali, S. Tunio, N. Jabeen, and N. Iqbal, "DeepImmuno-PSSM: Identification of Immunoglobulin based on Deep learning and PSSM-Profiles," 2023.

[37] Y. Chu *et al.*, "MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph," vol. 22, no. 6, p. bbab165, 2021.

[38] O. Barukab, F. Ali, S.A. Khan, "DBP-GAPred: An intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning,", J. Bioinforma. Comput. Biol. (2021) 2150018.

[39] S. Rahu *et al.*, "UBI-XGB: Identification of ubiquitin proteins using machine learning model," vol. 8, pp. 14-26, 2022.

[40] A. Ghulam, R. Sikander, and F. Ali, "AI and Machine Learning-based practices in various domains: A Survey," 2022.

[41] O. Barukab, F. Ali, W. Alghamdi, Y. Bassam, S.A. Khan, "DBP-CNN: Deep Learning-based Prediction of DNA-binding Proteins by Coupling Discrete Cosine Transform with Two-dimensional Convolutional Neural Network,", Expert Syst. Appl. (2022) 116729.

[42] A. Ghulam *et al.*, "Identification of Novel Protein Sequencing SARS CoV-2 Coronavirus Using Machine Learning," p. 47-58, 2021.

[43] H.E. Manoochehri, A. Pillai, M. Nourani, Graph convolutional networks for predicting drug-protein interactions. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2019, pp. 1223–1225.

[44] F. Ali, H. Kumar, S. Patil, A. Ahmad, A. Babour, A. Daud, "Deep-GHBP: Improving prediction of Growth Hormone-binding proteins using deep learning model,", Biomed. Signal Process. Control vol. 78 (2022) 103856.

[45] D. Chen and L.J.C.P. Wei, "A useful tool for the identification of DNA-binding proteins using graph convolutional network," vol. 18, no. 5, pp. 661-668, 2021.

[46] M. Khalid *et al.*, "An ensemble computational model for prediction of clathrin protein by coupling machine learning with discrete cosine transform," pp. 1-9, 2024.

[47] A. Khan, J. Uddin, F. Ali, A. Banjar, A. Daud, "Comparative analysis of the existing methods for prediction of antifreeze proteins,", Chemom. Intell. Lab. Syst. (2022) 104729.

[48] I.A. Khan, et al., A privacy-conserving framework based intrusion detection method for detecting and recognizing malicious behaviours in cyber-physical power networks, Appl. Intell. (2021) 1–16.

[49] T.N. Kipf and M.J.A.P.A. Welling, "Semi-supervised classification with graph convolutional networks," 2016.

[50] F. Ali, W. Alghamdi, A.O. Almagrabi, O. Alghushairy, A. Banjar, and M.J.I.J. o B.M. Khalid, "Deep-AGP: Prediction of angiogenic protein by integrating two-dimensional convolutional neural network with discrete cosine transform," p. 125296, 2023.

[51] A. Ahmad, et al., "Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks,", Chemom. Intell. Lab. Syst. vol. 208 (2021) 104214.

[52] A. Ghulam, F. Ali, R. Sikander, A. Ahmad, A. Ahmed, S. Patil, "ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network,", Chemom. Intell. Lab. Syst. vol. 226 (2022) 104589.

[53] F. Ali, A. Almuhaimeed, M. Khalid, H. Alshanbari, A. Masmoudi, and R.J.M. Alsini, "DEEP-EP: Identification of epigenetic protein by ensemble residual convolutional neural network for drug discovery," vol. 226, pp. 49-53, 2024.

[54] F. Ali, M. Hayat, Classification of membrane protein types using Voting Feature Interval in combination with Chou′ s Pseudo Amino Acid Composition, J. Theor. Biol. vol. 384 (2015) 78–83.

[55] R. Sikander, A. Ghulam, F. Ali, XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set, Sci. Rep. vol. 12 (1) (2022) 1–9.

[56] Z.N.K. Swati, et al., Brain tumor classification for MR images using transfer learning and fine-tuning, Comput. Med. Imaging Graph. vol. 75 (2019) 34–46.

[57] J. Hu, X.-G. Zhou, Y.-H. Zhu, D.-J. Yu, G.-J. Zhang, "TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning,", IEEE/ACM Trans. Comput. Biol. Bioinforma. vol. 17 (4) (2019) 1419–1429.

[58] F. Ali, A. Ghulam, Z.A. Maher, M.A. Khan, S.A. Khan, W. Hongya, "Deep-PCL: A deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features,", Chemom. Intell. Lab. Syst. vol. 221 (2022) 104484.

[59] F. Ali, M. Hayat, Machine learning approaches for discrimination of Extracellular Matrix proteins using hybrid feature space, J. Theor. Biol. vol. 403 (2016) 30–37.

[60] F. Ali, O. Barukab, A.B. Gadicha, S. Patil, O. Alghushairy, A.Y. Sarhan, DBP-iDWT: Improving DNA-Binding Proteins Prediction Using Multi-Perspective Evolutionary Profile and Discrete Wavelet Transform, Comput. Intell. Neurosci. vol. 2022 (2022).

[61] F. Ali, et al., DBPPred-PDSD: Machine learning approach for prediction of DNA-binding proteins using Discrete Wavelet Transform and optimized integrated features space, Chemom. Intell. Lab. Syst. vol. 182 (2018) 21–30.

[62] F. Ali, H. Kumar, W. Alghamdi, F.A. Kateb, and F. K. J. A. o. C. M. i. E. Alarfaj, "Recent Advances in Machine Learning-Based Models for Prediction of Antiviral Peptides," pp. 1-12, 2023.

[63] P. Charoenkwan, S. Kongsompong, N. Schaduangrat, P. Chumnanpuen, and W. J. B. b. Shoombuatong, "TIPred: a novel stacked ensemble approach for the

accelerated discovery of tyrosinase inhibitory peptides," vol. 24, no. 1, p. 356, 2023.

[64] P. Charoenkwan, S. Waramit, P. Chumnanpuen, N. Schaduangrat, and W.J.P. o. Shoombuatong, "TROLLOPE: A novel sequence-based stacked approach for the accelerated discovery of linear T-cell epitopes of hepatitis C virus," vol. 18, no. 8, p. e0290538, 2023.

[65] P. Charoenkwan, N. Schaduangrat, P. Lio, M.A. Moni, P. Chumnanpuen, and W. J. A. o. Shoombuatong, "iAMAP-SCM: a novel computational tool for large-scale identification of antimalarial peptides using estimated propensity scores of dipeptides," vol. 7, no. 45, pp. 41082-41095, 2022.

[66] P. Charoenkwan, S. Kanthawong, N. Schaduangrat, P. Li', M.A. Moni, and W.J.A.O. Shoombuatong, "SCMRSA: a new approach for identifying and Analyzing anti-MRSA peptides using estimated propensity scores of dipeptides," vol. 7, no. 36, pp. 32653-32664, 2022.

Farman Ali received his BS and MS degrees in Computer Science from University of Peshawar and Abdul Wali Khan University Mardan, Pakistan in 2009 and 2016, respectively. He secured Ph.D. degree in Computer Science and Technology with Machine Learning and Bioinformatics Group at School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China in 2021. He completed postdoctoral fellowship in computer science at University of Jeddah, Saudi Arabia in August, 2022. His major research areas are Bioinformatics, Machine Learning, and Deep Learning. Currently, he is working as Assistant Professor at Department of Computer Science, Bahria University Islamabad Campus, Pakistan.

Majdi Khalid received the Ph.D. degree from Colorado State University, Fort Collins, USA, in 2019. He is currently an Associate Professor with the Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia. His research interests include machine learning algorithms, computer vision, natural language processing, and other applications of Artificial Intelligence.

Abdullah Almuhaimeed is an Associate Research Professor in computer science at the Digital Health Institute at King Abdulaziz City for Science and Technology (KACST). He holds MSc and PhD degrees in computer science (2011 and 2016, respectively) from the University of Essex in the UK. Also, he has a bachelor's degree in computer science from Imam Muhammad Ibn Saud Islamic University (2007). His research interests include Semantic Web, Ontologies, Artificial intelligence, Machine learning, Data Science, Sentiment Analysis, Recommendation systems, Search Engines, Big Data, Neutral Language Processing, Deep Learning, Fuzzy Logic and Bioinformatics.

Atef Masmoudi completed his engineering diploma in informatics at the National School of Computer Science in 2003. He then pursued a master's in signal processing from the National Engineering School of Tunisia in 2005. In 2010, he obtained his PhD in computer science from the University of Montpellier II, France. He works as an Associate Professor in the College of Computer Science at King Khalid University. His primary research focus includes privacy, security, data hiding of multimedia, and machine learning.

Wajdi Alghamdi is PhD holder specialized in Computer Science (Data mining) from the Department of Computing, Goldsmiths College, University of London, London, United Kingdom. His is currently an associate professor at Information Technology Department, Computing and Information Technology College, King Abdul-Aziz University, Jeddah, Saudi Arabia. He is mostly interested in Knowledge Discovery in Databases, Data Mining and Statistical Computing. His main research is focusing on applying Machine Learning and Statistical Learning methods to genotype, phenotype and clinical data in-order to discover patterns of interest, including the identification of clinical and genetic predictors with respect to diseases.

Ayman Yafoz received the M.Sc. degree in web technology from the University of Southampton, U.K. in 2015, and the Ph. D. degree in computer science from the University of Regina, Canada, in 2021. He is currently an Assistant Professor with the Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University. His research interests include natural language processing and data science.