A DEEP LEARNING MODEL IMPLEMENTATION OF TABNET FOR PREDICTING PEPTIDE-PROTEIN INTERACTION IN CANCER

Hanif Aditya Pradana

School of Computing, Faculty of Informatics, Telkom University
Bandung, West Java, Indonesia
hanif.aditya@gmail.com

Ahmad Ardra Damarjati

Department of Computer Science, IPB University
Bogor, West Java, Indonesia
ahmad.ardra30@gmail.com

Isman Kurniawan

School of Computing, Faculty of Informatics, Telkom University
Bandung, West Java, Indonesia
ismankrn@telkomuniversity.ac.id

Wisnu Ananta Kusuma*

Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University
Tropical Biopharmaca Research Center, IPB University
Bogor, West Java, Indonesia
ananta@apps.ipb.ac.id

Abstract

Cancer has become one of the deadliest diseases in the world, mainly caused by the accumulation of somatic and inherited mutations. However, this phenomenon can be traced back to the molecular level, specifically, to proteins. Proteins are molecules responsible for various bioprocesses in the human body through their interactions with other molecules. Abnormalities in these interactions can lead to various undesirable outcomes, including disease and cancer. Peptides have the potential to serve as molecules that can be used in protein interactions to treat cancer. However, identification of peptides corresponding to target proteins in the laboratory is time-consuming and expensive. Therefore, there is a need for computational methods to aid identification. TabNet, a deep learning-based computational method was used in this study. For comparison purposes, we selected techniques from ensemble learning, including Random Forest and Extreme Gradient Boosting, along with methods from deep learning such as Convolutional Neural Network and Stacked Autoencoder-Deep Neural Network. Predictions are performed on a multi-feature peptide-protein interaction dataset, and the features include position-specific scoring matrices, intrinsic disorder, amino acid sequence, and physicochemical properties. Among our selected metrics, we found that TabNet achieved a better score in AUC of 0.7 and lower false negatives compared to other models.

***Keywords*: Cancer, Peptide-Protein Interaction, TabNet**

1. Introduction

As one of the deadliest diseases and a major health concern, the estimated number of cancer-related cases and deaths worldwide has reached approximately 19.3 million and 10 million in 2020 [Ferlay et al. (2021)]. Together with cardiovascular diseases, cancer has become the leading cause of death in 127 countries [Bray et al. (2021)]. Cancer is the result of an accumulation of inherited and somatic mutations in oncogenes and tumor

suppressor genes, as stated by Jonsson and Bates (2006). In simple terms, cancer occurs when there are mutations that make cells multiply uncontrollably. Many factors play a role in cancer, but they can be traced down to molecules, such as proteins for example. One example is the well-known tumor suppressor protein p53, where mutations of this protein and the gene that produces it, the TP53 gene, have been linked to more than 50% of cancer cases [Duffy et al. (2022)]. There have also been other findings of cancer-related proteins in different studies. Some other studies have also found proteins that have links to cancer. A study on the CCN protein family found that the expression of CCN proteins may have a role in the regulation of cancer cell growth, and another study on the ErbB/HER protein kinase family showed that mutations in this family can lead to malignancy in some cancers [Perbal (2003); Roskoski (2014)].

Proteins, referred to as the building blocks of life, are complex organic molecules made up of chains of amino acids usually consisting of 50 or more [LaPelusa and Kaushik (2023)]. One of the many properties of proteins is that they like to form bonds with other molecules, either through physical or chemical means. This is often the case in the human body, where in human cells alone, there are more than 39,000 identified protein interactions (PIs) [Gonzalez and Kann (2012)]. These interactions allow proteins to achieve their functions in bioprocesses such as cell communication, signal transduction, metabolism, and immune system regulation [Jonsson and Bates (2006)]. Therefore, disruptions caused by internal or external factors to these essential functions can initiate or further develop a disease, including cancer [Kuzmanov and Emili (2013)]. Among protein interactomes, peptides have emerged as promising candidates for therapeutic agents due to their safety, good efficacy, high selectivity, ease of synthesis, and good biocompatibility [Matijass and Neundorf (2021); Fosgerau and Hoffmann (2015)]. The use of peptides today has become widespread, with the history of their use as drugs beginning in 1922 and growing in popularity. It is estimated that in 2019, they accounted for 5% of the global pharmaceutical market or approximately $50 billion, and within this market share, 17% was dedicated to oncology [Muttenthaler et al. (2021)]. Considering the various side effects and limited effectiveness of conventional treatments in this domain, attempts to use peptides as cancer treatments may be a viable option [Cavalcanti and Soares (2021)].

Uncovering possible PIs is key in revealing suitable drug targets [Rao et al. (2014)]. Several in vivo and in vitro approaches are commonly used for this purpose, although there are technical difficulties along with poor scalability, inefficiency, and time constraints that have plagued such approaches [Lee et al. (2019)]. Over the years, computational approaches, otherwise referred to as in silico approaches, have always been the first choice in assisting this important task. Recently, the rapid advancement of artificial intelligence (AI) has made a great impact in revolutionizing industries and various fields of study. In the field of bioinformatics, the trend of integrating deep learning methods has seen a significant increase over the past decade [Min et al. (2017)]. Compared with traditional machine learning, deep learning methods can extract a higher level of data representation from inputs with stacked processing layers [Ahmed et al. (2023)]. Following the trend, considerations regarding our data, and the lack of implementation for this method in bioinformatics, we propose the use of TabNet to provide peptide-protein interaction (PepPI) prediction in cancer. We also performed a comparison with common deep learning and machine learning methods that have been adapted previously for this task.

2. Material and Methods

2.1. *Data Preparation and Features*

There are several things to do to get complete data, and we followed the steps of Lei et al.'s 2021 study. It starts with downloading a few files from each database. Proteins and peptides were retrieved from the RCSB PDB (https://www.rcsb.org/docs/programmatic-access/file-download-services), and in addition, some peptides were added from Drugbank (https://go.drugbank.com/releases/latest). Supporting protein and peptide data were also downloaded from SIFTS (https://www.ebi.ac.uk/pdbe/docs/sifts/quick.html), Uniprot (https://www.uniprot.org/), and TCGA (https://portal.gdc.cancer.gov/). After collecting the necessary files, the data was then forwarded to the protein-ligand interaction profiler (PLIP) to determine the interaction between the peptide and protein [Adasme et al. (2021)]. Then, a FASTA sequence search was initiated for each interacting protein-peptide. Next, we filtered only cancer proteins with the available cancer protein data retrieved from TCGA. With a set of cancer peptide-protein interaction data retrieved, the next step can be done to extract more features from the data. Fig. 1 shows the steps taken to obtain the complete data.

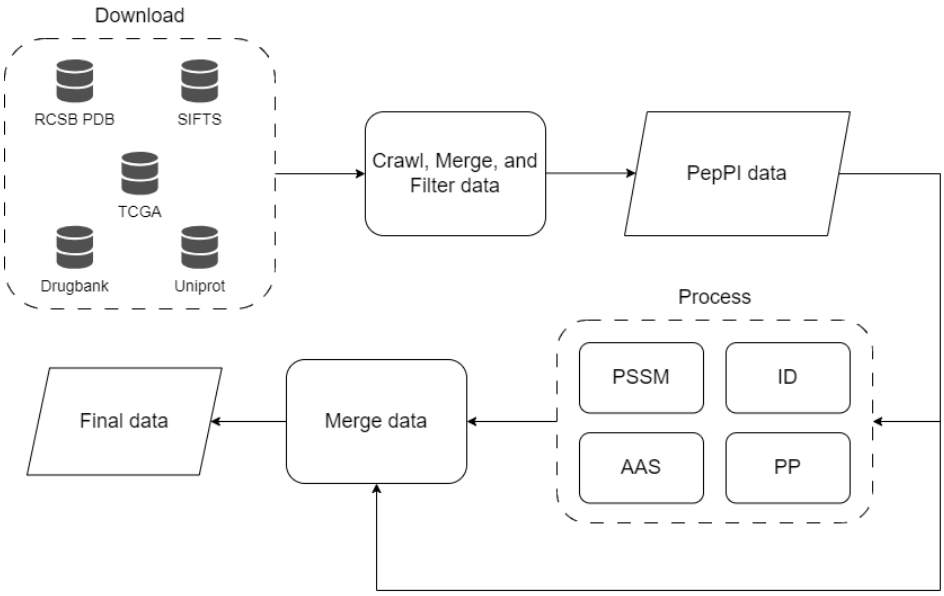

Fig 1. Steps in obtaining data.

For the features, we follow a preceding 2023 study done by Kusuma et al., and we chose 4 out of the 6 main features for proteins and peptides. Some of them are position-specific scoring matrix (PSSM), intrinsic disorder (ID), amino acid sequence (AAS), and physicochemical properties (PP). For the first feature, PSSM, describes the frequency of each amino acid at each position in the sequence served in a matrix representation. It was a popular feature to be used in protein research as it contributes to detecting homology between the sequences [Lei et al. (2021)]. We generated our PSSM profile using PSI-BLAST [Madeira et al. (2022)], and then transformed it using auto cross-covariance (ACC). However, PSSM itself is not enough as it suffers from the loss of some amino acid information [Khanh Le et al. (2019)]. Therefore, the other three features attempt to fill the gap in lost features. In the second feature, ID, refers to protein/peptide binding promiscuity due to its disordered/flexible region in the structure. We generated this feature using IUPred2A [Mészáros et al. (2018)]. In addition, the third feature, AAS, gives structural information through the protein/peptide sequence. Each letter of the sequence is encoded individually as integers. Finally, the last feature, PP, refers to its polarity and hydropathy of amino acids. The features were then merged with the original data to create the final data. We managed to obtain around 452 positive interactions and then continue to generate negative interactions through pairing random pairs of PepPI with one another that's not included in its pairing list. Due to our lack of overall data, we aim for 1 to 5 distribution of positive and negative interaction data, and we ended up with around 2260 negative interactions. This data can then be processed further by each method. More information about each feature is available in Table 1.

Features	**Encoding**
Position-Specific Scoring Matrix (PSSM)	Float, varies in number, can be negative
Intrinsic Disorder (ID)	Float of 0 to 1, where 0 represents complete order and 1 represents complete disorder
Amino Acid Sequence (AAS)	Integer of 0 to 21, where 0 serves as padding and 1 to 21 serves the available amino acids
Physicochemical Properties (PP)	1 = non-polar, positive hydropathy 2 = non-polar, negative hydropathy 3 = polar-uncharged, positive hydropathy 4 = polar-uncharged, negative hydropathy 5 = negatively charged, negative hydropathy 6 = positively charged, negative hydropathy 7 = unknown, unknown

Table 1. Description of Features

2.2. *TabNet*

TabNet is a deep learning architecture that uses sequential attention to choose which features to reason from at each decision step which allows it to focus on relevant information [Arık and Pfister (2021)]. Inspired by decision trees, TabNet combines its deep learning basis with a new level of interpretability. It was proposed by the researchers at Google and was motivated by the lack of deep learning methods for solving tabular data, despite it being said as one of the most common types of data used in the creation of AI [Chui et al. (2018)]. There are 3 main things that comprise TabNet, which are feature transformer, attentive transformer, and feature masking. More details can be seen in Fig. 2.

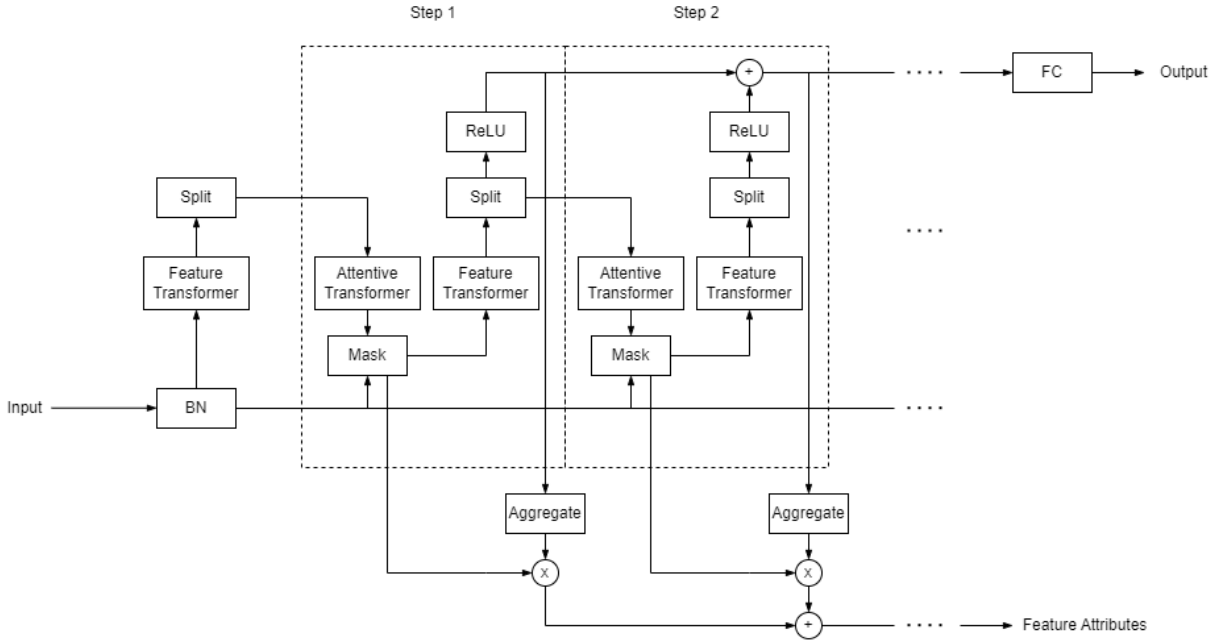

Fig 2. TabNet structure.

First, the data goes to a batch normalization (BN) layer and then passed on to the feature transformer. The feature transformer processed the normalized inputs and further extracted the features. Each feature transformer consists of 2 parts, where the first part shares all the parameters, and the second part did not share its parameters and trained separately with each step [Yan et al. (2021)]. Each of these parts consists of a fully connected (FC) layer, BN layer, and a gated linear unit (GLU) layer. The output from these parts were then normalized by $\sqrt{0.5}$ to help in stabilizing variance and the overall learning. The results taken from the feature transformer module were subsequently passed on to the split module where the feature representations get divided for attentive transformer module and the overall output (starting from the first step). At every single step, a feature matrix of $\mathbf{B} \times \mathbf{D}$ was used, where B is the batch size, and D is the feature dimension. Afterwards, an attentive transformer then selects the most salient features. Combined with the mask module, a mask can be generated with the following formula:

$$\mathrm{M}[i] = sparsemax(\mathrm{P}[i-1] \cdot h_i(a[i-1])) \qquad (1)$$

This is done by the work of several layers inside the attentive transformer, which began with an FC layer, BN layer, and then continued by trainable function h_i, prior scale $\mathbf{P}[i]$, and sparsemax. Prior scale is then used to give weight of the preceding step:

$$\mathbf{P}[i] = \prod_{j=1}^{i} (\gamma - \mathrm{M}[j]) \qquad (2)$$

In the original paper, sparsemax was used to promote sparsity by mapping the Euclidean projection onto the probabilistic simplex. But in this research, entmax was used since it offers better, smoother, and more differentiable curvature with the desired sparsity of sparsemax [Peters et al. (2020)]. Next, the mask continues to the feature transformer and the output gets split and aggregated for the next step. The process iterates through these steps until it reaches the final stage, where a summary of the steps is forwarded to an FC layer to generate the final output.

2.3. *Evaluation Metrics*

We compared our proposed method with some of the methods from ensemble and deep learning. Ensemble methods are chosen due to their common use in bioinformatics, as they can support high-dimensional data better than common methods [B.Meshram and M. Shinde (2015)]. As for deep learning methods, it served as a benchmark for TabNet. Representing ensemble methods, we chose Random Forest (RF) and Xtreme Gradient Boosting (XGBoost), and deep learning methods, we chose Convolutional Neural Network (CNN) and Stacked Autoencoder-Deep Neural Network (SAE-DNN) from Kusuma et al., 2023 research. We split our data with the ratio of training and testing of 80 to 20. We also utilized stratified k-fold with k value of 5 for every method and performed hyperparameter tuning using a hyperparameter optimization framework Optuna to ensure a better result [Akiba et al. (2019)]. Explanation of the tuned hyperparameters can be seen on Table 2.

Model	Tuned Hyperparameter
Random Forest (RF)	{'n_estimators': 720, 'max_depth': 24}
Xtreme Gradient Boosting (XGBoost)	{'n_estimators': 513, 'max_depth': 64, 'objective': 'binary:logistic', 'learning_rate': 0.3760674530242582, 'subsample': 0.8253520839713084}
Convolutional Neural Network (CNN)	{'n_layers': 1, 'hl_node': 856, 'lr': 0.006428422621179661, 'do': 0.4313130008029596, 'epochs': 31, 'batch_size': 106}
Stacked Autoencoder-Deep Neural Network (SAE-DNN)	{'epochs': 72, 'batch_size': 34, 'hl_node': 272, 'lr': 0.005353369339157353, 'do': 0.11720777275661616}
TabNet	{'mask_type': 'entmax', 'n_da': 40, 'n_steps': 3, 'gamma': 1.1, 'n_shared': 2, 'lambda_sparse': 2.387592998912264e-06, 'patienceScheduler': 5, 'epochs': 45, 'v_batch_size': 128}

Table 2. Model Hyperparameter.

For measuring how each model performs, we chose evaluation metrics consisting of accuracy, recall, precision, f-measure, AUC score, and confusion matrix. Accuracy measures how correct the prediction is against test data. Recall, or sensitivity, measures capability on correctly identifying positive value of the data. Precision measures the rate of correctly predicted positive value. F-measure is a balance of precision and recall, where it mostly measures model's performance on an imbalanced dataset. AUC score measures the area under the receiver operating characteristic (ROC) curve, where this measures model's performance in differentiating between positive and negative value. Finally, confusion matrix can support other metrics in judging a model's performance through visualization of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) score of the model.

3. Results and Discussion

Model	Accuracy	Recall	Precision	AUC Score	F-measure
RF	0.85324±0.006	0.61424±0.051	0.33187±0.020	0.64470±0.009	0.42970±0.018
XGBoost	0.85176±0.008	0.59538±0.052	0.35834±0.027	0.65439±0.012	0.44601±0.024
CNN	0.75404±0.026	0.19039±0.057	0.13719±0.028	0.50731±0.024	0.15818±0.037
SAE-DNN	0.81379±0.005	0.33607±0.060	0.42307±0.027	0.62268±0.027	0.37332±0.046
TabNet	0.77175±0.012	0.48220±0.069	0.36119±0.021	0.65592±0.027	0.41148±0.035

Table 3. Model results after 5-fold.

The results for all five models were summarized in Table 3. In terms of accuracy, it was observed that ensemble methods were better than deep learning methods, except for SAE-DNN, where it reached around ≥ 80% accuracy. TabNet excelled over other deep learning methods in recall, AUC score, and f-measure, but it fell short against ensemble methods in terms of accuracy, recall, and f-measure. This might be attributed to the nature of ensemble methods, where it exhibited better adaptability to small size data in contrast to deep learning methods. Among the available metrics, all models were not good at precision, suggesting the failure in predicting actual positive interactions. This issue might be caused by less positive interaction in the data than the negative one.

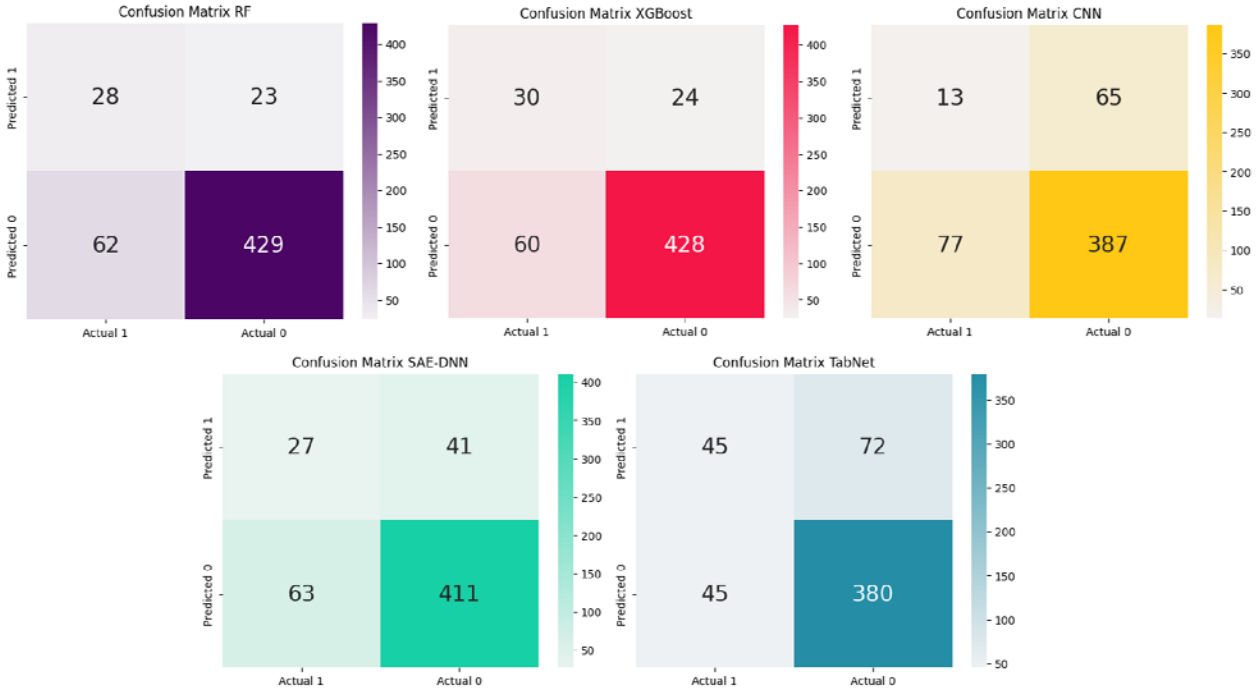

Fig 3. Confusion matrix of each model.

Fig. 3 shows the confusion matrix of each model. TabNet demonstrated a low false negative in comparison to other models, however it also exhibited a high false positive as a tradeoff. Meanwhile, ensemble methods gave the opposite of TabNet, with low value of false positive and high value of false negative. SAE-DNN performed as average in this metric, and CNN performed lower than average, where both showing no superiority in either false positive or false negative. In the case of PepPI prediction, a low false negative means that actual positive interaction has a lower chance to be labeled as negative, and this leads to less likelihood in overlooking potentially important PepPI. On the other hand, a low false positive has the benefits in saving time and resources for laboratory investigation of most prominent interaction, and there is less of a necessity to revalidate every result.

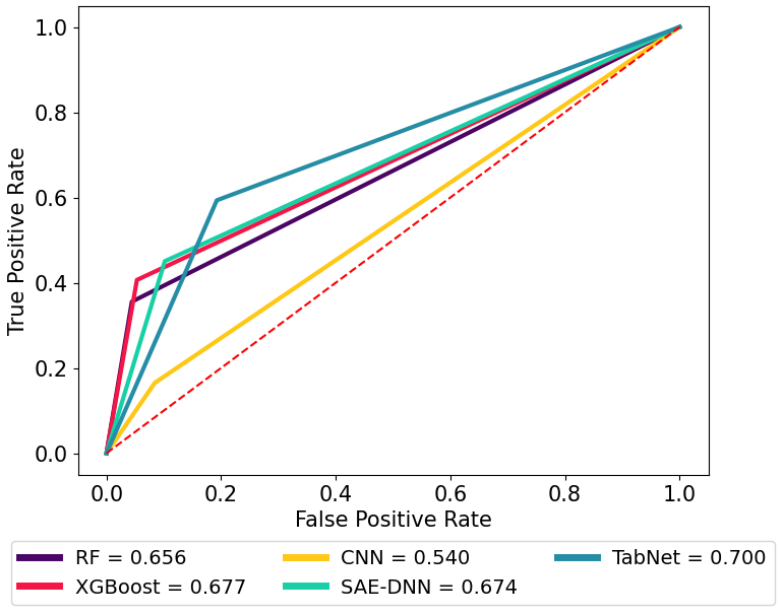

Fig 4. Comparison of AUC for each model.

Figure 4 shows the ROC curve and the highest achievable AUC score for each model across 5-fold validation. This graph describes the performance of a model in distinguishing two different classes. Notably, every model performed about the same as one another, while CNN was significantly worse than other models. The performance of SAE-DNN closely resembled that of XGBoost, with RF trailing not too far behind. TabNet however, achieved better scores than the other models, with the best AUC score in 5-fold of 0.7. In the context of PepPI prediction, TabNet provides better understanding in determining factors that affect positive and negative interaction.

4. Conclusion

In this research, we conducted an implementation of TabNet deep learning model in predicting interactions between peptides and cancer proteins. Each model produced comparable results in terms of performance in each metric, but they were not as impressive as those of other research. This was mainly contributed by limited data and the imbalance distribution of positive and negative interaction as shown by each model's confusion matrix. Additional efforts are required to acquire more positive interaction data for optimal results. As for TabNet, it managed to achieve satisfactory results in some metrics when compared to other ensemble and deep learning methods. More study can be done to further explore the potential use of TabNet.

Acknowledgments

This work was fully supported by Tropical Biopharmaca Research Center. Therefore, we are very grateful for the funding and support of this research.

Conflict of Interest

The authors have no conflicts of interest to declare in this study.

References

[1] Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., and Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, *149*(4).
[2] Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16).
[3] Jonsson, P. F., and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, *22*(18).
[4] Duffy, M. J., Synnott, N. C., O'Grady, S., and Crown, J. (2022). Targeting p53 for the treatment of cancer. In *Seminars in Cancer Biology* (Vol. 79).
[5] Perbal, B. (2003). The CCN3 (NOV) cell growth regulator: A new tool for molecular medicine. In *Expert Review of Molecular Diagnostics* (Vol. 3, Issue 5).
[6] Roskoski, R. (2014). The ErbB/HER family of protein-tyrosine kinases and cancer. In *Pharmacological Research* (Vol. 79).
[7] LaPelusa, A., and Kaushik, R. (2023). *Physiology, Proteins*.
[8] Gonzalez, M. W., and Kann, M. G. (2012). Chapter 4: Protein Interactions and Disease. *PLoS Computational Biology*, *8*(12).
[9] Kuzmanov, U., and Emili, A. (2013). Protein-protein interaction networks: Probing disease mechanisms using model systems. In *Genome Medicine* (Vol. 5, Issue 4).
[10] Matijass, M., and Neundorf, I. (2021). Cell-penetrating peptides as part of therapeutics used in cancer research. *Medicine in Drug Discovery*, *10*.
[11] Fosgerau, K., and Hoffmann, T. (2015). Peptide therapeutics: Current status and future directions. In *Drug Discovery Today* (Vol. 20, Issue 1).
[12] Muttenthaler, M., King, G. F., Adams, D. J., and Alewood, P. F. (2021). Trends in peptide drug discovery. In *Nature Reviews Drug Discovery* (Vol. 20, Issue 4).
[13] Cavalcanti, I. D. L., and Soares, J. C. S. (2021). Conventional Cancer Treatment. In *Advances in Cancer Treatment*.
[14] Rao, V. S., Srinivas, K., Sujini, G. N., and Kumar, G. N. S. (2014). Protein-Protein Interaction Detection: Methods and Analysis. *International Journal of Proteomics*, *2014*.
[15] Lee, A. C. L., Harris, J. L., Khanna, K. K., and Hong, J. H. (2019). A comprehensive review on current advances in peptide drug development and design. In *International Journal of Molecular Sciences* (Vol. 20, Issue 10).
[16] Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. In *Briefings in bioinformatics* (Vol. 18, Issue 5).
[17] Ahmed, S. F., Alam, M. S. Bin, Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., Mofijur, M., Shawkat Ali, A. B. M., and Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*.
[18] Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D., and Zeng, J. (2021). A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature Communications*, *12*(1).
[19] Adasme, M. F., Linnemann, K. L., Bolz, S. N., Kaiser, F., Salentin, S., Haupt, V. J., and Schroeder, M. (2021). PLIP 2021: Expanding the scope of the protein-ligand interaction profiler to DNA and RNA. *Nucleic Acids Research*, *49*(W1).
[20] Kusuma, W. A., Fadli, A., Fatriani, R., Sofyantoro, F., Yudha, D. S., Lischer, K., Nuringtyas, T. R., Putri, W. A., Purwestri, Y. A., and Swasono, R. T. (2023). Prediction of the interaction between Calloselasma rhodostoma venom-derived peptides and cancer-associated hub proteins: A computational study. *Heliyon*, *9*(11), pp. e21149.
[21] Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research*, *50*(W1).
[22] Khanh Le, N. Q., Nguyen, Q. H., Chen, X., Rahardja, S., and Nguyen, B. P. (2019). Classification of adaptor proteins using recurrent neural networks and PSSM profiles. *BMC Genomics*, *20*.
[23] Mészáros, B., Erdös, G., and Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, *46*(W1).

[24] Arık, S., and Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, *8A*.
[25] Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., and Malhotra, S. (2018). Notes From the Ai Frontier Insights From Hundreds of Use Cases. *McKinsey Global Institute*.
[26] Yan, J., Xu, T., Yu, Y., and Xu, H. (2021). Rainfall forecast model based on the tabnet model. *Water (Switzerland)*, *13*(9).
[27] Peters, B., Niculae, V., and Martins, A. F. T. (2020). Sparse sequence-to-sequence models. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.
[28] B.Meshram, S., and M. Shinde, S. (2015). A Survey on Ensemble Methods for High Dimensional Data Classification in Biomedicine Field. *International Journal of Computer Applications*, *111*(11).
[29] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Authors Profile

Hanif Aditya Pradana is a final year informatics student at Telkom University. He completed an internship at the Tropical Biopharmaca Research Center with the research for this paper. He has a passion in data science, machine learning, and is now finding a new interest in bioinformatics.

Ardra Damarjati is a final-year Computer Science major at IPB University, and he is bringing a wealth of experience from his bioinformatics internship at the Tropical Biopharmaca Research Center. Passionate about machine learning, Ardra is dedicated to applying it for data-driven research, aiming to solve real-world problems.

Isman Kurniawan is a lecturer in the School Computing of Telkom University. He has an expertise in the fields of cheminformatics, machine learning, artificial intelligence, and modeling and simulation.

Wisnu Ananta Kusuma received his bachelor's and master's degrees from Bandung Institute of Technology, as well as his Ph.D. from Tokyo Institute of Technology, in 2012. He is currently an Associate Professor at the Department of Computer Science, IPB University. He also serves as Executive Secretary of Institute for International Research on Advanced Technology, IPB University, coordinator of Bioinformatics Working Group, Faculty of Mathematics and Natural Sciences, IPB University, and coordinator of Bioinformatics and High Performance Computing Research Group, Advanced Research Laboratory, IPB University. He has been the author of more than 60 articles and has reviewed for international journals. His current research interests include machine learning, high-performance computing, and bioinformatics research.