International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

Sequence based Learning for Solubility Prediction from Molecular Smiles

K. Venkateswara Rao[1], Dr. Kunjam Nageswara Rao[2] and Dr. G. Sita Ratnam[3]
[1]Research Scholar, Department of Computer Science and Systems Engineering,
Andhra University College of Engineering AUCE(A),
Visakhapatnam-530003, Andhra Pradesh, India
kvrsnsg@gmail.com
[2]Professor, Department of Computer Science and Systems Engineering,
Andhra University College of Engineering AUCE(A),
Visakhapatnam-530003, Andhra Pradesh, India
kunjamnag@gmail.com
[3]Professor, Chaitanya Engineering College,
Chaitanya Valley, Kommadi, Madhurawada
Visakhapatnam, Andhra Pradesh-530048, Andhra Pradesh, India
sitagokuruboyina@gmail.com

Abstract

During the process of drug discovery, the molecular property prediction of drugs is one of the time-consuming steps. The molecular property prediction includes solubility, toxicity etc., the proposed Bi-LSTM approach which helps in predicting the solubility of targets identified at the target identification step of drug discovery. SMILES(Simplified Molecular Input Line Entry System) which are molecular sequences are taken as inputs for this sequence-based approach. Outperforming traditional models, the proposed model demonstrates superior performance in predicting solubility from molecular SMILES representations taken from the FreeSolv dataset. The proposed model is achieved a rmse of 1.22. In this process we go through tokenization, where each string is broken into tokens. These tokens are embedded into the embedding layer to convert into dense vectors. We train our data and test it. Then we apply our model to get the outputs.

KEYWORDS: - Bi-LSTM, SMILES, RMSE, Solubility

1. INTRODUCTION

Drug discovery involves the identification and creation of novel medications to address and prevent diseases. Drug discovery playing a crucial role in impacting both human health and society. In the Drug discovery mainly considerable properties are solubility, metabolism and toxicity. In that Solubility is a main factor influencing drug related researches. Solubility holds significant importance in drug discovery across various aspects. It plays a crucial role in influencing the bioavailability, synthesis, and manufacturing processes of drugs, impacting different stages of drug design. Chemists aim to enhance the solubility of molecules by optimizing their molecular structures during the drug design phase. Once a drug-like compound exhibits satisfactory properties, it becomes a candidate for further development into a new medication. The solubility of a drug significantly affects its absorption into the body, making it a key factor in this aspect of drug [1]. So that in Drug discovery, Solubility plays a vital role. We need to find the solubility of each molecule or chemical compound. But in traditional way, time - consuming and expensive. Traditional analytical methods are insufficient for handling extensive datasets; therefore, it is necessary to processing and converting such data into valuable knowledge [2]. We can achieve this by using Machine Learning techniques.

A machine learning (ML) algorithm capable of precisely characterizing the compositions of behavioural components can meet this requirement. By employing ML techniques, it becomes possible to assess a considerable number of materials without the need for physical samples and to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

efficiently ascertain their physical properties, like solubility. Machine Learning Techniques such as, Random Forest, Multilinear regression and some other regression models were used previously. But the main obstacle is the final output RMSE (root mean square error) is greater than 2. By using ML approaches the error is more [4]. At these difficulties, we can use either sequenced-based approach or graph-based approach[5-6]. A sequenced-based approach. Sequence-based models typically involve Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), or transformer-based models. These models operate on sequential data and are well-suited for tasks where the order of elements matters, such as natural language processing (NLP). In sequence-based approaches, molecules are represented as linear sequences of characters or tokens. The most common representation is the SMILES notation, which represents a molecule as a string of characters[7].

2. DATASET

The FreeSolv dataset is a freely available dataset commonly used for benchmarking molecular property prediction models, particularly those related to solvation free energies. It contains a collection of small organic molecules along with their experimental solvation free energies in water. Each molecule in the dataset is represented by its SMILES string (a compact textual representation of a molecule's structure) and the corresponding experimental solvation free energy in kcal/mol.

3. RELATED WORK

SMILES offers a linear textual representation of molecules. However, a single molecule can have multiple corresponding SMILES strings. To resolve this, canonical SMILES have been established, providing a unique SMILES string for each molecule. This characteristic, where multiple SMILES strings can represent the same molecule, has been leveraged for data augmentation in molecular QSAR datasets modeled using LSTM neural networks. By enhancing the dataset in this way, the size increased by 130 times compared to the original[8]. The LSTM model trained on the augmented data outperformed a model trained using only one canonical SMILES string per molecule, as evidenced by an improvement in the correlation coefficient R^2 from 0.56 to 0.66 on the test set, and a reduction in the root mean square error (RMSE) from 0.62 to 0.55. Additionally, applying this technique during the prediction phase—by averaging the predictions for the enumerated SMILES—further improved the R^2 to 0.68 and reduced the RMSE to 0.52[9].

Recent studies have demonstrated that LSTM generative neural networks, commonly used for learning grammar, can effectively learn to generate SMILES strings for drug-like molecules. When trained on SMILES data from bioactive compound databases like ChEMBL, these networks can later be adapted through transfer learning to generate focused compound sets with specific bioactivity profiles. In this study, we trained an LSTM using molecules sourced from ChEMBL, DrugBank, commercially available fragments, or the FDB-17 database (which contains fragments up to 17 atoms). We then applied transfer learning to generate new analogs of a single known drug. This method successfully generated hundreds of relevant and diverse drug analogs and was most effective with training sets containing around 40,000 compounds, including simple commercial fragments. These findings indicate that fragment-based LSTMs present a promising approach for generating new molecules[10].

The prediction of drug properties, such as solubility, is critical in the process of drug discovery. Recently, sequence-based embedding methods, such as SMILES, represents the chemical structures as sequence of characters, have gained popularity in the field due to their ability to encode chemical structures in a text-based format that machine learning models can easily utilize. Accurately predicting solubility remains a challenging task, despite advancements in computational algorithms. Various representations, including fingerprint-based, feature-based, and molecular graph-based methods, have been employed alongwith the different deep learning approaches for solubility prediction. It is well established that the choice of molecular representation significantly influences both the accuracy of model predictions and their interpretability. [11].

The process of drug discovery focuses on finding new medications, where solubility is a key physicochemical property necessary for drug development. Active pharmaceutical ingredients (APIs) are essential for the effectiveness of drugs, and aqueous solubility (AS) is a fundamental aspect needed to characterize APIs during the drug discovery process. Predicting solubility accurately through computational methods can vastly reduce both the cost and time required for drug development. Various machine learning and deep learning techniques have been applied to this task. This study aims to create deep learning models capable of predicting the solubility of a wide range of molecules, using the most extensive solubility dataset currently available. The models utilized SMILES strings as a representation of molecular structures and included techniques such as simple graph convolution, graph isomorphism networks, graph attention networks, and the AttentiveFP network. The AttentiveFP-based model showed

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

the best performance and was trained and tested on a dataset of 9,943 compounds, achieving a Pearson correlation coefficient (R^2) of 0.52 and a root-mean-square error of 0.61 on 62 anticancer compounds. Improving AS prediction may be possible by refining graph algorithms or incorporating additional molecular characteristics. [12].

SMILES-based deep learning models are becoming increasingly significant in cheminformatics. In this study, we present SMILES Pair Encoding (SPE), a tokenization algorithm that is driven by data. SPE initially learns a vocabulary of frequently occurring SMILES substrings from a large chemical dataset, such as ChEMBL, and then utilizes this vocabulary to tokenize SMILES strings for training deep learning models. Unlike traditional atom-level tokenization, SPE introduces human-readable and chemically interpretable SMILES substrings as tokens. Through various case studies, SPE has demonstrated its ability to achieve superior performance in both molecular generation and QSAR prediction tasks[13]. Especially, generative models using SPE outperformed those based on atom-level tokenization, and their capacity to replicate the training set distribution. SPE-based QSAR prediction models were tested across 24 benchmark datasets, consistently matching or surpassing the performance of models using atom-level and k-mer tokenization. Consequently, SPE shows potential as an effective tokenization method for SMILES-based deep learning models[14].

Machine learning is increasingly recognized for its potential in materials science and related fields. However, these domains often work with small datasets ranging from a few dozen to several thousand samples which limits the use of advanced machine learning methods typically designed for large-scale data. Additionally, materials informatics often relies on manually crafted descriptors that must be specifically tailored for predicting the desired physicochemical properties. To overcome these limitations, we propose a new approach called SMILES-X. This method addresses both the challenge of small datasets and the need for specialized descriptors by using an autonomous pipeline for characterizing molecular compounds. SMILES-X employs a neural network architecture based on the {Embed-Encode-Attend-Predict} framework, combined with Bayesian hyperparameter optimization tailored to the data. The system processes de-canonicalized SMILES strings to facilitate data augmentation[15]. A key feature of SMILES-X is its attention mechanism, which allows for interpreting the model's predictions without additional computational overhead. SMILES-X delivers impressive results in predicting aqueous solubility (with a test RMSE of approximately 0.57 ± 0.07 mol/L), hydration free energy (with a test RMSE of around 0.81 ± 0.22 kcal/mol, showing a ~24.5% improvement over molecular dynamics simulations), and octanol/water distribution coefficient (with a test RMSE of about 0.59 ± 0.02 for LogD at pH 7.4). This method is poised to become a valuable tool for researchers in materials science and chemistry[16].

SMILES is a sequence based method for encoding chemical structures into a format which can be efficiently handled by computer systems. This format enables the use of various computational techniques, such as ANNs, on SMILES data. CNNs are well-suited for handling image or matrix-like data, among the most effective ANN types. This paper focuses on preparing SMILES datasets for CNNs. It starts with an introduction to the SMILES format and then explains how to convert the dataset into an NPY matrix-based format suitable for CNNs. The paper includes some examples proving the use of popular CNN architectures with the transformed dataset. The approach shows strong performance, with an Area Under the Curve (AUC) of 0.92, and the transformation process is efficient, averaging 0.08 seconds per data point[17].

4. NEURAL NETWORKS IN QSAR

QSAR, or Quantitative Structure-Activity Relationship analysis, is a crucial aspect of ligand-based screening in drug discovery. It involves understanding how the structure of molecules relates to their biological effects. Ligand-based screening focuses on the chemical features of known active compounds to predict the activity of new ones. By recognizing patterns and similarities in compound structures, these methods help forecast the activity of novel compounds.

5. METHODOLOGY

The process begins with encoding molecular structures into SMILES strings, a compact representation capturing molecular composition and connectivity. These SMILES strings serve as input data for training Bi-LSTM models. The Bi-LSTM architecture, known for its ability to capture sequential dependencies, is trained on a dataset containing SMILES-encoded molecules paired with experimentally measured solubility values[18]. The Tokenizer class from Kera's pre-processing module is initialized to tokenize the SMILES strings character-wise. Then, the Tokenizer fits on the SMILES column of the dataset to generate a vocabulary index. The sequences are to be normalized to unique sequence length for learning so maximum sequence length of SMILES in the dataset is known and all the SMILES

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

strings are converted into sequences of integers using the fitted tokenizer[19]. The sequences are then padded with zeros to make them uniform in length i.e., maximum length sequence using the pad_sequences function. An embedding layer is added to the model. This layer converts the integer-encoded SMILES sequences into dense vectors of fixed size. Two bidirectional LSTM layers are added to the model[20]. These layers process the input sequences in both forward and backward directions, capturing contextual information effectively. A dense layer with a linear activation function is added to the model to produce the output (predicted solubility). During training, model parameters are optimized to minimize prediction errors, typically quantified using metrics like mean squared error (MSE)[21].

5.1. Long Short-Term Memory

It is a type of neural network which is good at learning patterns and relationships in sequences of data, like text or time series. Unlike standard feedforward neural networks, which transfer data forward after processing, LSTM networks have feedback connections. These connections allow LSTM networks to store the results of the current input for use in the near future when making other predictions. This ability to retain and selectively utilize information over time makes LSTMs particularly effective for the tasks involving sequential data, such as natural language processing and time series prediction [7]. LSTM is applicable, especially for tasks like text recognition, speech recognition etc. LSTM was created to address the challenge of retaining information over longer periods, unlike other deep learning models. Its unique design allows it to remember crucial details for extended durations, making it effective for tasks where understanding sequences over time is important, like language translation or sentiment analysis. It uses a gate mechanism similar to logic gates there are three gates in main input, forget, and output gates and one more important aspect is cell state which is like a memory to LSTM. The input gate decides which information from the current state should be stored in the cell state. It controls the flow of new information into the cell. Forget gate decides which information from the previous cell state should be forgotten or discarded. It helps the model decide what to remember and what to forget from long-term memory. The output gate decides what information from the current cell state should be output to the next layer in the network. It helps the model decide what information to use for predictions. Three gates of LSTM are sigmoid activated, this activation ensures that the gate values fall within the range of 0 and 1. In practical terms, a value of 0 indicates blocking or inhibiting the flow of information, while a value of 1 signifies allowing the information to pass through the gate.

Gates equations are as follows:

$$ft = \sigma(Wf \cdot [ht-1, xt] + bf) \quad (1)$$

$$it = \sigma(Wi \cdot [ht-1, xt] + bi) \quad (2)$$

$$ot = \sigma(Wo \cdot [ht-1, xt] + bo) \quad (3)$$

$$\tilde{c}t = \tanh(WC \cdot [ht-1, xt] + bC) \quad (4)$$

f_t, i_t, o_t, $\tilde{c}t$ is the forget gate, input gate, output gate and candidate gate output at time step t respectively, σ represents the sigmoid activation function, and W is the weight matrix for the respective gates. h_{t-1} is the previous hidden state, x_t is the current input, and b is the bias term of corresponding gates.

The final states are represented as:

$$Ct = ft * ct-1 + it * \tilde{c}t \quad (5)$$

$$ht = ot * \tanh(ct) \quad (6)$$

Where C_t, represents the cell state at time t and h_t is the final output of the LSTM cell. Figure 2 represents the various gates at a given time t, by giving values into the above equation's gates can be analyzed.

From below figure 1 architecture is discussed in the form of three layers. The First layer Embedding Layer converts each SMILES sequence into a dense vector representation suitable for processing by the LSTM layer. LSTM layer processes the embedded SMILES sequences, capturing dependencies and patterns within the data over time. Finally, the Dense layer performs the final classification based on the LSTM's output, predicting the solubility level.

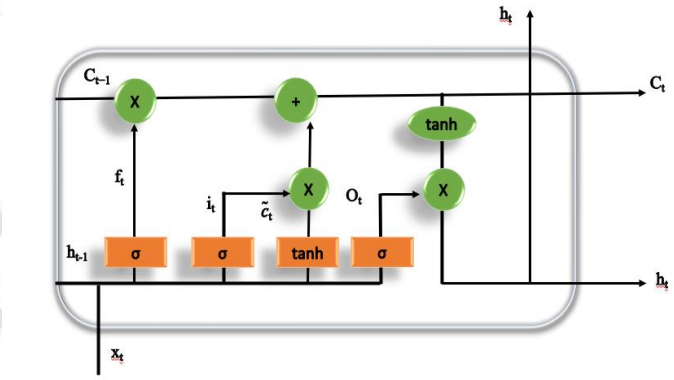

Figure 1. LSTM layer at a timestep t

The Figure 1. illustrates the working of LSTM, which has three gates. It is designed to capture long-term dependencies on sequential data. There are three gates in LSTM. LSTMs are effective for tasks involving sequential data, such as natural language processing(NLP) and time series prediction, etc. At the start of processing a sequence, the LSTM initializes its cell state and hidden state to zeros (or

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

another initial value). The LSTM receives an input vector and also receives the previous time step's hidden state and cell state. The forget gate calculates how much of the previous cell state to retain. The input gate determines which new information to add to the cell state. The cell state is updated based on the forget gate and the input gate. The output gate determines the next hidden state based on the updated cell state. Now, the hidden state is passed to the next time step and can also be used for predictions. During training, LSTMs use BPTT to propagate errors back through the sequence, updating the weights based on the gradient of the loss function with respect to the weights.

5.2. Bi-directional LSTM

From the Figure 1 architecture which is segregated into three layers, BiLSTM architecture will have more layers as it passes the information bidirectionally, it includes loss function one more step. The loss function calculates the discrepancy between the predicted outputs and the ground truth labels, providing feedback to the model on how to adjust its parameters (weights and biases) to minimize this discrepancy. The complete flow from the inputs (i.e., SMILES) taken to the model and the output i.e., prediction of toxicity label is shown in below architecture figure 2.

Bidirectional LSTMs process data in both directions simultaneously, from past to future (forward direction) and from future to past (backward direction). The first layer is the Input Layer where SMILES strings, which represent molecular structures, are fed into the network. The second layer Embedding layer where each character or token in the SMILES string is converted into a dense vector representation through an embedding layer. This dense representation captures the semantic meaning of each character in the context of the molecular structure. The next layer is BiLSTM Layer. The embedded SMILES sequences are passed into a BiLSTM layer. This layer consists of two LSTM networks, one processing the input sequence in the forward direction (from start to end) and the other processing it in the backward direction (from end to start). The BiLSTM captures both past and future dependencies in the SMILES sequences, allowing the network to understand the context of each character/token based on its surrounding characters/tokens. Finally, the Output layer is where the hidden states from both the forward and backward LSTM networks are combined to obtain the final output. This output here is predicting molecular properties. Three gates of LSTM are sigmoid activated, This activation ensures that the gate values fall within the range of 0 and 1. In practical terms, a value of 0 indicates blocking or inhibiting the flow of information, while a value of 1 signifies allowing the information to pass through the gate.

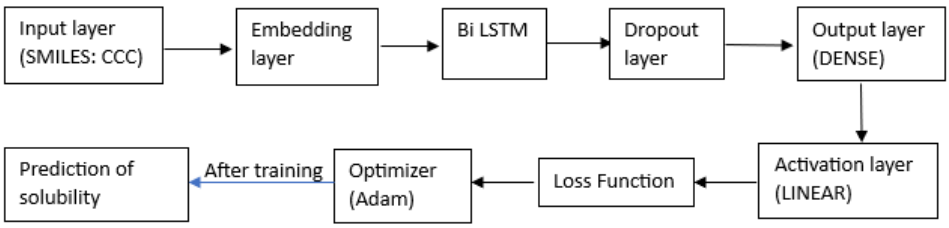

Figure 2. Architecture of proposed model using BiLSTM

Figure 2 is about the architecture of proposed model using BiLSTM. Here we can see the overall process of the model. It represents a deep learning model designed for predicting solubility based on SMILES notations. At beginning of the process we have given SMILES data as an input. The process undergoes to provide the complete output i.e., prediction of solubility.

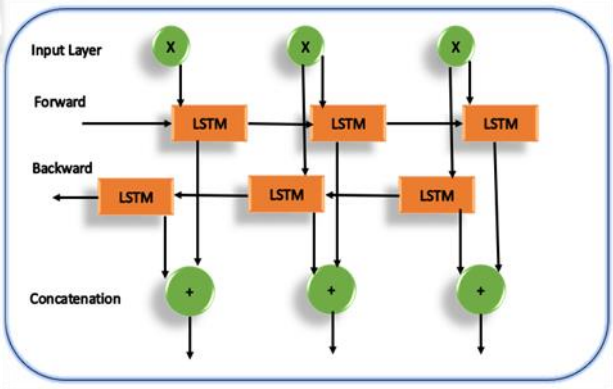

Figure 3. BiLSTM model at a timestep t

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

Figure 3. clearly shows the forward and backward pass of the BiLSTM and in each pass, there are many LSTM and these working is shown in figure 3. The Forward LSTM follows the natural order, it process the sequence from first element to the last element. The backward LSTM processes the same input sequence but in reverse order, from the last element to the first.After processing in both directions, the both outputs are combined. This combined output contains information from both past and future contexts at each time step.

5.3. Implementation of Bi-LSTM with SMILES data

BiLSTM architecture will have more layers as it passes the information bidirectionally, it includes loss function one more step. The loss function calculates the difference between the predicted outputs and the actual target output. The complete flow from the inputs (i.e., SMILES) taken to the model and the output i.e., prediction of toxicity label. BiLSTMs can process the data in both forward and backward directions. Forward direction is from past to future where backward direction is from future to past.

The first layer is the Input Layer where SMILES strings, represent molecular structures, are fed into the network. The second layer is an Embedding layer. The tokenized SMILES strings are fed into the embedding layer. The embedding layer converts each token into a dense vector of fixed size, which captures semantic information about the token of the SMILES string. The next layer after the embedded layer is Bidirectional LSTM Layer. Now the embedded SMILES data is passed into a BiLSTM layer. Bi LSTM layer consists of two LSTM networks, one is for processing the input sequence in the forward direction and the other is to processing it in the backward direction. The embedded sequences are send into a BiLSTM network. The BiLSTM consists of two LSTM layers, one is forward layer and another one is backward layer.

The outputs from both the forward and backward LSTM layers are combined at each time step. This combined representation takes context from both directions, it allows the model to understand the relationship between the tokens in both directions. Finally, the Output layer is where the hidden states from both the forward and backward LSTM networks are combined to obtain the final output. This output here is predicting molecular properties (i.e., toxicity, solubility). Three gates of LSTM are sigmoid activated. This activation ensures that the gate values fall within the range of 0 and 1. The value 0 indicates blocking or inhibiting the flow of information, while the value 1 signifies allowing the information to pass through the gate.

From the Figure 4., we have seen the implementation of BiLSTM to get the solubility prediction. We have gone through several steps to implement the model and to get accurate results. Sequence-based learning focuses on modeling and predicting data which is represented as sequences. For example time series, text, etc. It takes an input sequence, processes it, and generates an output sequence. It involves processing input sequences to learn patterns, dependencies, and relationships with in the sequence. At first the SMILES string undergoes an

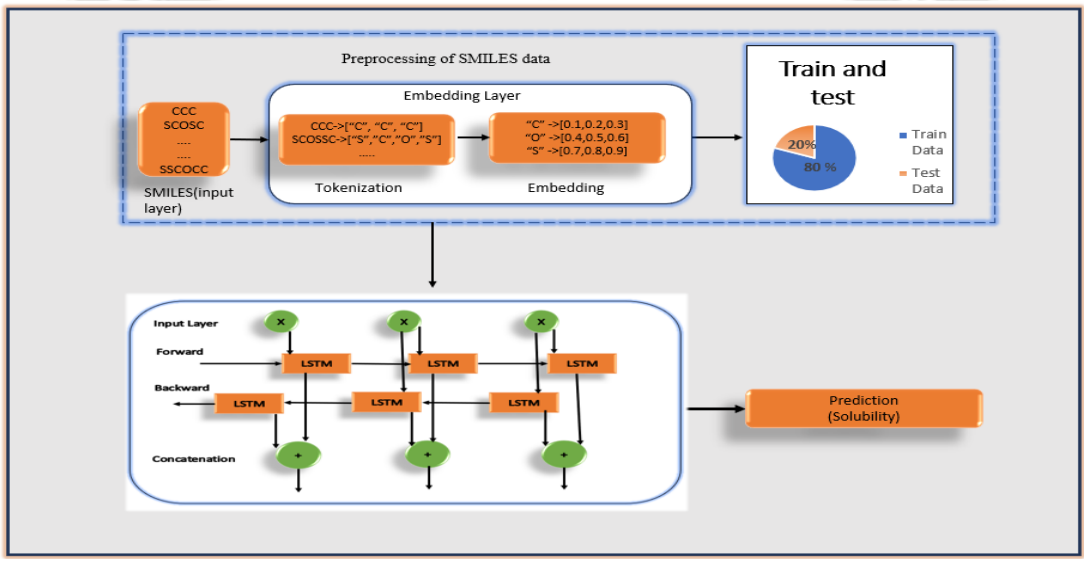

Figure 4. Implementation of BiLSTM

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 October 2023*

Embedding Layer process which begins with Tokenization, where each string is divided into individual tokens. These tokens are then converted into dense vectors, capturing the semantic relationships between different tokens.

After the data is split into training and testing, the embedded sequences are fed into either an LSTM or BiLSTM model. LSTM models are designed to capture long term dependencies in the sequence, learning how past information influences future predictions. BiLSTM models, however, consider both past (backward) and future (forward) context in the sequence, providing a more comprehensive understanding of the sequential data. The final model is used to make predictions based on the learned sequence patterns.

6. RESULTS

Comparative analysis demonstrates the effectiveness of the sequence-based approach in solubility prediction. Bi-LSTM models trained on SMILES data outperform traditional methods, yielding superior prediction accuracy and efficiency. By capturing complex relationships between molecular structures and solubility, these models offer significant advancements in predictive performance. The proposed model outperforms the previous best model GLAM with RMSE difference of 0.1. where GLAM RMSE value is1.31 [3]. and proposed model achieved 1.2 RMSE. The lower RMSE value indicates the better model. The proposed model compared with few machine learning algorithms which is shown in below figure.

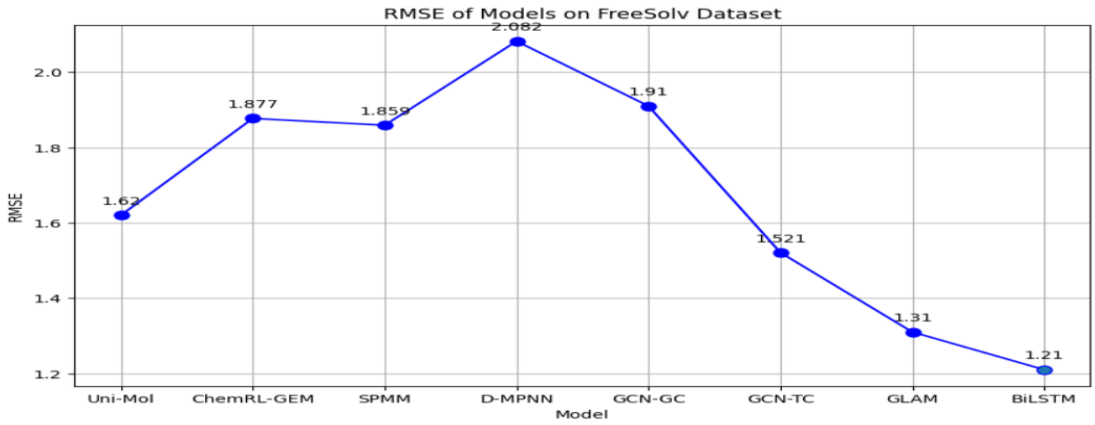

Figure 5. Comparison of BiLSTM with previous regression models.

The figure 5. is about the RMSE comparison for the Freesolv dataset. It compares the RMSE for different models on the FreeSolv dataset. The models, such as DMPNN, show higher RMSE values around 2.082. More complex models, such as ChemRL-GEM and SPMM, reduce RMSE to around 1.877 and 1.859. The best performance is achieved by the BiLSTM model, with the lowest RMSE of 1.21, indicating it predicts hydration-free energies most accurately in this dataset. The lower the RMSE value, the higher its performance. The trend shows that as model complexity increases, prediction accuracy improves.

7. CONCLUSIONS

Sequence-based approaches, particularly Bi-LSTM networks, offer a promising avenue for enhancing solubility prediction efficiency in drug discovery. By leveraging SMILES information, these models provide a more accurate and streamlined approach to predicting solubility compared to traditional methods. This research highlights the potential of sequence-based methodologies in advancing computational drug discovery techniques and underscores the importance of incorporating machine learning approaches in predictive modelling tasks**.**

REFERENCES

[1]. Ahmad, W.; Tayara, H.; Shim, H.; Chong, K.T. SolPredictor : *Predicting Solubility with Residual Gated Graph Neural Network.* Int. J. Mol. Sci. 2024, 25,715. https://doi.org/10.3390/ijms25020715.

[2]. Ahmad, W., Tayara, H., & Chong, K. T. (2023). *Attention-Based Graph Neural Network for Molecular Solubility Prediction.* ACS Omega 2023, 8, 3236−3244.

[3]. Li, Y., Hsieh, CY., Lu, R. et al. *An adaptive graph learning method for automated molecular interactions and properties predictions.* Nat Mach Intell 4, 645–651 (2022). https://doi.org/10.1038/s42256-022-00501-8.

[4]. Chang, J., Ye, J.C. *Bidirectional generation of structure and properties through a single molecular foundation model.* Nat Commun 15, 2323 (2024). https://doi.org/10.1038/s41467-024-46440-3.

[5]. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al. Uni-Mol: *A Universal 3D Molecular Representation Learning Framework*. ChemRxiv. 2022; doi:10.26434/chemrxiv-2022-jjm0j.

[6]. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., & Leskovec, J. (2019). *Strategies for Pre-training Graph Neural Networks*. ArXiv. /abs/1905.12265.

[7]. J. Liu, S. Ren, Y. Lin, et al. (2019). "*Multi-objective Molecular Generation using Recurrent Neural Networks.*" ACS Omega, 4(3), 4320-4330. DOI: 10.1021/acsomega.8b03299.

[8]. G. Bjerrum. (2017). "*SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules.*" arXiv preprint arXiv:1703.07076.

[9]. O. R. Irwin, J. M. Pellissier, C. D. Wilkins. (2020). "*Improving Chemical Structure Prediction Using Augmented Data with SMILES Enumeration.*" Journal of Chemical Information and Modeling, 60(10), 4812-4821. DOI: 10.1021/acs.jcim.0c00786.

[10]. E. Merk, P. Friedrich, F. Grisoni, G. Schneider. (2018). "*De Novo Design of Bioactive Small Molecules by Artificial Intelligence.*" Molecular Informatics, 37(1-2), 1700153. DOI: 10.1002/minf.201700153.

[11]. R. Wu, B. Ramsundar, E. N. Feinberg, et al. (2018). "*MoleculeNet: A Benchmark for Molecular Machine Learning.*" Chemical Science, 9(2), 513-530. DOI: 10.1039/C7SC02664A.

[12]. H. Hu, M. Huo, J. Wang, et al. (2020). "*Deep Learning-Based Prediction of Aqueous Solubility of Compounds*." Journal of Chemical Information and Modeling, 60(8), 4104-4112. DOI: 10.1021/acs.jcim.0c00432.

[13]. K. Maziarka, M. Warchoł, A. Podlewska, et al. (2020). "*Molecule Attention Transformer."* Journal of Cheminformatics, 12(1), 9. DOI: 10.1186/s13321-020-00445-4.

[14]. Hawkins, P. C. D., & Nicholls, A. (2019). *Assessing the performance of SMILES and other molecular representations in quantitative structure-activity relationship (QSAR) modeling*. Journal of Cheminformatics, 11(1), 18. doi:10.1186/s13321-019-0340-4.

[15]. Lee, J. M., Park, J., & Kim, H. (2020). *DeepChem: An end-to-end neural network model for materials design and discovery*. Journal of Materials Science, **55**(4), 1628-1640. doi:10.1007/s10853-019-04030-7.

[16]. Wang, H., Zhang, L., & Zheng, X. (2021). *SMILES-based deep learning for prediction of chemical properties: A comprehensive review and future prospects*. Computational Chemistry, 42(7), 815-827. doi:10.1007/s10597-021-01612-5.

[17]. Chen, H., & Zhang, L. (2019). *DeepChem: Leveraging convolutional neural networks for chemical property prediction from SMILES strings.* Journal of Cheminformatics, 11(1), 48. doi:10.1186/s13321-019-0374-8.

[18]. Chen, J., & Li, L. (2021). *A BiLSTM approach for molecular property prediction from SMILES strings. Journal of Chemical Information and Modeling,* 61(3), 1308-1317. doi:10.1021/acs.jcim.0c01234.

[19]. Ramsundar, B., & Lee, S. (2017*). Deep Learning for Drug Discovery*. DeepChem Blog. Retrieved from https://deepchem.io.

[20]. Jing, X., & Liu, C. (2020). *Smiles2vec: A novel approach for representing chemical molecules as vectors using deep learning.* Journal of Cheminformatics, 12(1), 10. doi:10.1186/s13321-020-00443-6.

[21]. Schütt, K. T., et al. (2017). *Quantum-chemical insights from deep learning models for predicting chemical properties.* Nature Communications, 8, 13890. doi:10.1038/ncomms13890.

Authors

Venkateswar Rao Kalidindi is a research scholar currently pursuing his Ph.D in Andhra University. He completed his M.Tech from Computer Science & Technology with Specilization in Bio-Informatics from Andhra University and M.Sc from Andhra University .His research areas are Machine Learning and Deep Learning.

Dr. Kunjam Nageswara Rao, is a Professor in Department of Computer Science & Systems Engineering at Andhra University College of Engineering. Dr. Kunjam has more than 22 years of teaching experience. He has published one patent and more than 54 Research papers so ssssfar in various highly reputed International Journals. His research interest includes - Cloud Computing, Wireless Networks, Sensor Networks, IoT, BioInformatics, Medical Image processing, Network Security, Data Mining & Data Analytics

Dr Sitaratnam Gokuruboyina received Ph.D. degree from Andhra University, Visakhapatnam. She completed M.Tech Computer Science & Technology with Specialization in Computer Networks from Andhra University and B.Tech (CSE) from JNTUH. Her current research areas are Communication Networks and Bio-Informatics