CHAPTER 7

MOLECULAR PROPERTY PREDICTION OF MODIFIED GEDUNIN ($C_{26}H_{31}N_2O_6F$)USING MACHINE LEARNING

7.1. INTRODUCTION

Prediction of molecular characteristics in the realm of drug revelation is essential. Computer approaches for accurate forecasting can speed up the interaction between finding better applicants in a faster and cheaper manner. This is particularly convincing, considering that the usual cost of improving another drug is estimated at approximately 2,8 billion dollars **(Gao *et al*., 2019).** The usual silicon technique for predicting molecular characteristics depends mainly on removing fingerprints or handmade highlights, combined with AI (Artificial intelligence) computations. In conclusion, this kind of atomic portrayal is unilaterally presented by space professionals to gather the highlights essential for the task to be carried out**(Merkwirth and Lengauer,2005).** To advance beyond that type of tendency to a broader approach, several forms of IAs have been incorporated into the field of the prediction of molecular properties. In particular, the calculation of deep learning has resurged owing not only to the acceleration of computer power and the increase in accessibility of large information indices but also to its gigantic performance, for instance, in the field of characteristic language use**(Young et al., 2018)** and recognition **(Bhamare and Suryawanshi, 2018).** These networks are mechanized to study representations for a specific task and thus eliminate components' confusing design interaction **(Gilmer *et al*., 2017).** A suitable molecule portrait should be created to use deep-learning

calculations and cover the explicit component design region. As molecules may

be seen as diagrams, one technique involves using an atomic graph portrait – resulting in the improvement of graphic neural networks (GNNs), which have become increasingly commonplace and have been increasingly common in recent years **(Wu *et al.,* 2019) (Zhou *et al.,* 2020) (Mayr *et al.,* 2018).** Their achievements in the development of traditional AI techniques, in particular, with expectations of quantum mechanical properties **(Pappu *et al*., 2018) (Gao *et al.,* 2019)(Zemel *et al.,* 2019) (Shindo *et al.,* 2017**); physicochemical properties, Hydrophobicity**,(Yi *et al*., 2018) (Wei *et al.,* 2019) (Jaakkola *et al.,* 2020)** and predictions of toxicity, appear to give them perhaps the most encouraging profound learning strategies in explicit diagram tasks. Owing to the rapid growth of the distributions recognized with GNN (Graph neural network) expectations for molecular properties, it may be challenging to know the current situation in this sector.

Profound studies in synthetic scanning have become more critical at a faster rate, and the prediction of particle characteristics such as free solvent energy **(Guthrie *et al.,* 2014**) or ionizing energy **(Lilienfeld *et al.,* 2013)** is contending with or even outstripping existing abdominal muscle initio calculations. The range of potential AI (ML: Machine learning) calculation errors appears endless and can range from the grouping of bioactive mixtures into dynamic and non-dynamic HIV replication classifications **(Dubey *et al.,* 2018)** to retrograde assignments in which a particular value, such as lipophilicity, is predicted, depending on the information **(However *et al.,* 1998).** ML computations are not limited to class and quality forecasts. Depending on the input structures or desired communication goals, they can generate new compound buildings. Thus, they can maintain the plan of new molecules in materials and innovative chemistry **(Chung *et al.,* 2019) (Kim *et al.,***

2018) (Taylor *et al.*, 2005). Clustering further evaluates the partitioning of massive nuclear datasets into the most highly harmonic collections and can be utilized in therapeutic research to identify innovative lead structures and to speed up the selection of your PC **(Weininger *et al.*, 1988).** Clustering of K-Means is an unmonitored learning technique grouping the unmarked data into several clusters. Here, K determines the number of predefined clusters that must be generated; if K=2, two clusters are created, and for K=3, three are created.

The use of strings is a simple method to address subatomic design. A string is a number, and letter progression is routinely used to enhance the determination of the subatomic information line transit (SMILES). In these strings, the SMILES string 'CCCO' would depict each particle with its edge iotas, neighbouring bonds, and an atom like n Propanol **(Tchekhovskoi *et al.*, 2015).** As the chirality was not adequately represented, the IUPAC offered an alternative string depiction called the International Chemical Identifier (InChI), which could not be easily interpreted by a chemist, apart from the focused chiral and the tautomerism**(Mitchell *et al.*, 2014).**

Subatomic descriptors are another method for addressing subatomic design. These descriptors may be directly available from the construction, highlights such as the number of iotas, the number of heteroatoms, or the subatomic load, to establish or reconstruct fundamental characteristics such as the second dipole and incomplete loading**(Pujadas *et al.*, 2015).** These characteristics are combined into a fixed-length vector in a described request. Additional features, such as subatomic fingerprints, are available and typically used. An atomic fingerprint is a longitudinal vector that addresses each component.

A recurrent neural network (RNN) is an artificial nerve network that employs sequence or time-series data. These deep learning methods have often been employed for ordinary or temporal issues, including language translation and processing natural languages (NLP), language recognition, and picture captions. Recent neural networks use training data to learn feedforward and convolutional neural networks (CNNs). They are characterized by their "memory" since they use previous inputs for information to impact the present input and output, whereas The artificial recurrent neural network (RNN) architecture utilized in deeper learning is the long-short memory (LSTM). In contrast to typical feedforward neural networks, LSTM contains feedback connections. Only single data points (such as pictures) and whole data sequences may be processed (such as speech or video). LSTM is appropriate for unsegmented, connected handwriting, voice recognition, and network traffic anomaly or intrusion detection systems (IDS). The visual geometry group VGGNet architecture is ILSVR2014's first classifying architecture, whereas GoogLeNet is the winner. This explains why VGGNet is built on top of the architecture using many contemporary model classifications.

This chapter focuses on molecular property prediction of modified gedunin ($C_{26}H_{31}N_2O_6F$) using machine learning through property prediction of this novel gedunin derivative($C_{26}H_{31}N_2O_6F$) using LSTM, VGGnet, and RNN algorithms using the CHEMBL and drug bank datasets.

7.2. Experimental

7.2.1. Datasets

All relapse businesses were conducted with the CHEMBL and Drug Bank datasets, and Python code under the Rkdit environment was used to transform the

SMILES texts into graphic pictures as part of data pre-processing **(Buhlmann *et al.*, 2020**).

7.2.2 Feature Extraction and clustering

VGGNet, a revolutionary neural network design with exhausting features, was proposed by **Andrew Zisserman and Karen Simonyan** at Oxford University in 2014. To cluster data in needed groupings based on a similarity index **(Liu *et al.*, 2018),** the k-means clustering method was employed. The cluster number was determined using the elbow technique.

7.2.3 Prediction

The properties of the modified gedunin ($C_{26}H_{31}N_2O_6F$) were anticipated based on the similarity index by predicting the fit of the cluster.

7.3 Results and Discussion

LSTM was used to convert the modified gedunin ($C_{26}H_{31}N_2O_6F$) smile into an image which was further classified in cluster 2 **(Figure 7.1.)** by k-means clustering out of the four clusters predicted by the elbow curve **(Figure 7.2.)** using features from a large dataset as the training and validation sets with model accuracy of 88.9, Model specificity of 90.4 and Model sensitivity of 80.5. Modified gedunin $C_{26}H_{31}N_2O_6F$ shares properties with the 933 compounds, as shown in **(Figure 7.3.).** These properties include interleukin-23 receptor inhibitor, adenosine kinase, mitotic spindle assembly protein MAD2B inhibitor, antiproliferative activity, cytotoxicity against human HCC366 cells, anticancer activity against human T47D cells, growth inhibition of human MCF7 cells, antiprogestational activity, anti-inflammatory activity, and anti-tumour activity.

```
result_df_pred.head(n = 10)
```

	Id	Category
0	C:\Users\PRIYA DAGAR\Desktop\all folder\ml pro...	2

```
Id,Category
C:\Users\PRIYA DAGAR\Desktop\all folder\ml project\data\Smile_Img_Gen\Smile_Img\inhibitor ,2
```

Figure 7.1. A molecule with the assigned cluster.

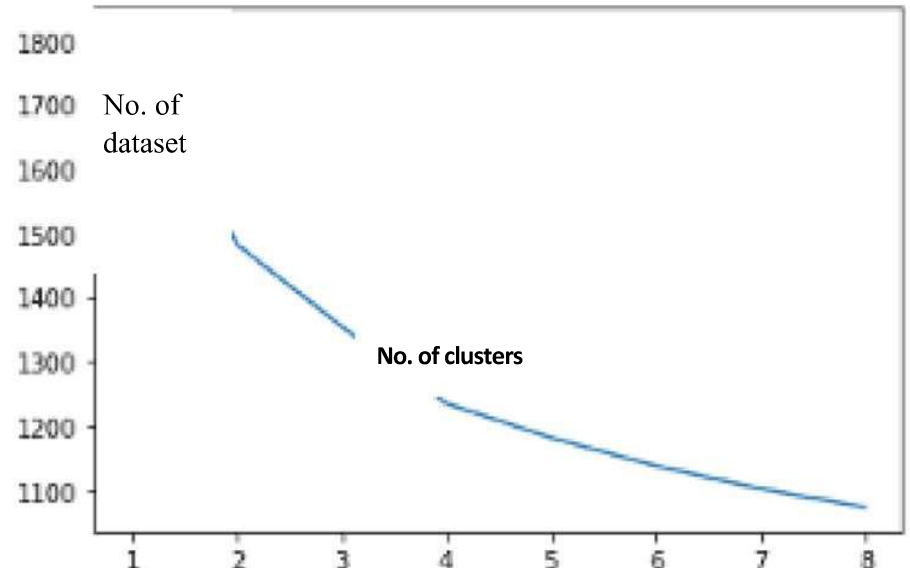
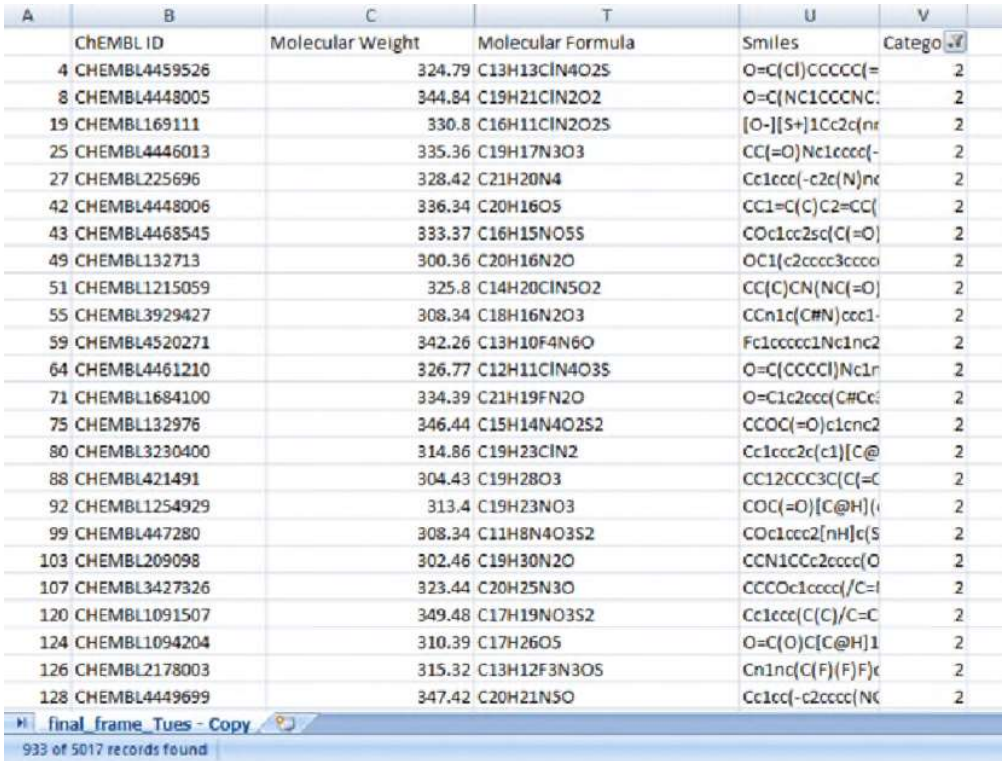


Figure 7.2. Curve Elbow to determine the number of clusters where the y-axis is no. of the dataset and the x-axis is no. of clusters.



	ChEMBL ID	Molecular Weight	Molecular Formula	Smiles	Catego
4	CHEMBL4459526	324.79	C13H13ClN4O2S	O=C(Cl)CCCCC(=	2
8	CHEMBL4448005	344.84	C19H21ClN2O2	O=C(NC1CCCNC:	2
19	CHEMBL169111	330.8	C16H11ClN2O2S	[O-][S+]1Cc2c(n	2
25	CHEMBL4446013	335.36	C19H17N3O3	CC(=O)Nc1cccc(-	2
27	CHEMBL225696	328.42	C21H20N4	Cc1ccc(-c2c(N)n	2
42	CHEMBL4448006	336.34	C20H16O5	CC1=C(C)C2=CC(	2
43	CHEMBL4468545	333.37	C16H15NO5S	COc1cc2sc(C(=O)	2
49	CHEMBL132713	300.36	C20H16N2O	OC1(c2cccc3cccc	2
51	CHEMBL1215059	325.8	C14H20ClN5O2	CC(C)CN(NC(=O)	2
55	CHEMBL3929427	308.34	C18H16N2O3	CCn1c(C#N)ccc1-	2
59	CHEMBL4520271	342.26	C13H10F4N6O	Fc1ccccc1Nc1nc2	2
64	CHEMBL4461210	326.77	C12H11ClN4O3S	O=C(CCCCl)Nc1n	2
71	CHEMBL1684100	334.39	C21H19FN2O	O=C1c2ccc(C#Cc:	2
75	CHEMBL132976	346.44	C15H14N4O2S2	CCOC(=O)c1cnc2	2
80	CHEMBL3230400	314.86	C19H23ClN2	Cc1ccc2c(c1)[C@	2
88	CHEMBL421491	304.43	C19H28O3	CC12CCC3C(C(=C	2
92	CHEMBL1254929	313.4	C19H23NO3	COC(=O)[C@H]((	2
99	CHEMBL447280	308.34	C11H8N4O3S2	COc1ccc2[nH]c(S	2
103	CHEMBL209098	302.46	C19H30N2O	CCN1CCc2cccc(O	2
107	CHEMBL3427326	323.44	C20H25N3O	CCCOc1cccc(/C=I	2
120	CHEMBL1091507	349.48	C17H19NO3S2	Cc1ccc(C(C)/C=C	2
124	CHEMBL1094204	310.39	C17H26O5	O=C(O)C[C@H]1	2
126	CHEMBL2178003	315.32	C13H12F3N3OS	Cn1nc(C(F)(F)F)c	2
128	CHEMBL4449699	347.42	C20H21N5O	Cc1cc(-c2cccc(NC	2

final_frame_Tues - Copy

933 of 5017 records found

Figure 7.3. Similar property compounds in cluster 2 with reference to ($C_{26}H_{31}N_2O_6F$)

7.3.1 CNN

CNNs are fundamental neural organizations, which are incredibly layered, and the large majority feature comparable essential capacities, including convolution layers, pooling, and grouping layers **(Figure 7.4.).** CNN contrast by introducing and grouping these essential layers and developing an organizing plan. First, the information images were standard-specifically preprocessed **(He *et al.*, 2020).** The input stream is then transferred to many convolutionary layers with pooling layers, in which extraction and recurrence are highlighted. The basic features were built continuously in a constructive manner. Each highlight was then partially consolidated, and the following highlights signified the design aspect of the class.

In the long term, these highlights are transferred to the related layer, which measures the order **(Shangwei *et al.,* 2021).**

Pre-processing Layer

Several information images must be loaded from the information layer. Before this, a few preprocessing activities are necessary, including measurement and standardization **(Wang *et al.,* 2021).** However, CNNs require far fewer preprocessing tasks than other neural networks. Another critical factor in removing uninteresting contrasts is the basic preprocessing layer.

Convolutional Layer

Eliminating interesting highlights, a convolutionary layer consolidates the contribution, including maps due to different component locators, as illustrated previously. In the first convolutionary layer, the neurons scrub straight into the edges. In the accompanying convolutionary layers, neurons can collect data to obtain a larger image of the image, thereby making a high-demand distinction between them **(Bazgir *et al.,* 2021)**. Every turning bite is equipped to highlight the removal across the information plane, although neurons are allocated to diverse parts of the photographic information plane to provide part maps with the exact size of the responsive field.

Convolution

Each convolutionary layer has limitations such as information size, bit size, guide stack depth, zero coiling, and step **(Ghosh *et al.,* 2021).**

Activation

Actuation should occur after the weighted whole and the propensity in all cases. Apart from pure perceptrons, a simple direct mix of information is predicted to

break apart and make it viable for a neuronal organization to become the general approximator of non-stop capabilities in a Euclidean environment. However, Jarrett et al. eventually brought the corrected linear units (ReLUs) **(Dittmer *et al.,* 2020)** to CNNs to enhance the display. For some time now, Xavier Glorot et al. pointed out that the very strong nonlinearity, differentiability, and insufficient components of ReLU must be acknowledged. Ultimately, ReLUs are generally recognized for initiating convolutional yields.

Pooling Layer

Pooling or sub-sampling includes several activities, such as broad pooling and pooling, which usually interfere within a few convolutionary levels. The most frequently used pooling techniques in CNNs are max pooling and average. Pooling offers several benefits to CNNs. In particular, grouping goals to prevent overcrowding by completing information near a window decreases the dimensionality of the information. Dimensional information further decreases assistance in reducing the estimates **(Hartmann *et al.,* 2021).** In addition, pooling achieves invariance, including interpretation, pivot, and scale, because a few distinctions or scaling are not distinct after the appropriate pooling.

Classification Layer

The group layer is the organization's top tier, where the final enmeshed element is collected, and a segment vector is returned, where each column is directed to a class. Each component of the yield vector deals with the probability assessment for each class and the number of components **(Sporea *et al.,* 2021).** While convolutions and information regarding the crude map in the element space are linked, the arrangement layer results in an example space presentation that offers

a remarkable exhibition of order. Unique CNN's generally yield a fully related layer. The entirely linked characterization layer is regarded as a heritage that goes out to artificial intelligence with 'maintain extraction and grouping **(Surekcigil *et al.*, 2021).** All the high requirements are joined and weighed by complete connections of neuronal contributions to achieving spatial change.

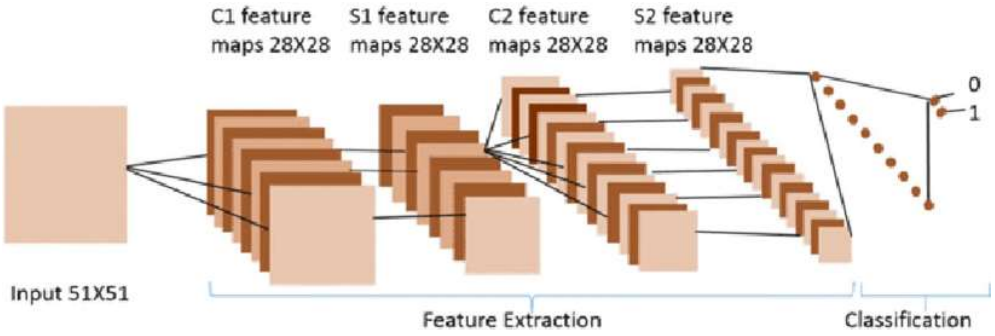


Figure 7.4. It depicts the architecture of CNN layers for feature extraction and classification **(Khoshdeli *et al.*, 2017).**

7.3.2 Transfer Learning

A deep CNN model was developed using transfer learning to increase the accuracy of the CNN expectation. The Chembl dataset was used for the pre-workout of the model. The model was pretrained on 5021 compounds, and particular atomic characteristics were removed **(Cai *et al.*, 2020).** After the preparation of the scenes, a flattened layer trapped through a dense layer, a dropout layer, and a single-node dense cap was used to shorten and subtract the pre-trained models.

7.3.3 K-means Clustering

The K-means partition algorithm is a bunch-like method proposed by J. B. MacQueen. This solo algorithm is typically used in mining information, such as recognition. This algorithm is defined by a target-group execution file, square error, and error. This method finds K divisions to show a certain standard for determining the optimization outcome. **(Roover *et al.,* 2013).** Right off the bat, select a few spots to address the central points of the underlying group (usually, the primary K pay points are selected in a case for the underlying bunch point of convergence); furthermore, assemble the remaining example dabs in the middle of the central focus. Based on the lower distance, the underlying arrangement will be obtained **(Timmerman *et al.,* 2013).** The K-means isolation-dependent algorithm is a bunch algorithm that benefits from speed, skill, and speed. However, this technique depends significantly on introductive specks and the difference in starting to select instances which always yields different results. In addition, this track-dependent algorithm constantly uses the slope method to obtain high values **(Pei *et al.,* 2020).** The search is constantly performed using the angle technique along the direction where energy is diminished, which results in the lower point of the neighbourhood when the underlying convergence point is not legitimate. Four clusters were chosen to divide the molecules into clusters based on comparable functions or characteristics based on the silhouette score **(Figure 7.2.).** This molecule was assigned to a cluster **(Figure 7.1.).**

7.3.4 Long short-term memory (LSTM)

It is a falsified repetitive neural organization (RNN), a deep-learning engineering company. LSTM does not have critical connections, such as conventional feed for

neural organizations. It can cycle single-focused information (such as photos) and entire information groups (such as discourse or video) **(Jia *et al.*, 2021).** For example, LSTM is suitable for messages such as unsegmented penalty recognition, speech identification, and detecting anomalies in network traffic or interruption location systems (IDSs). LSTM was used to produce the SMILE of modified gedunin $C_{26}H_{31}N_2O_6F$ **(Figure 7.5.)**

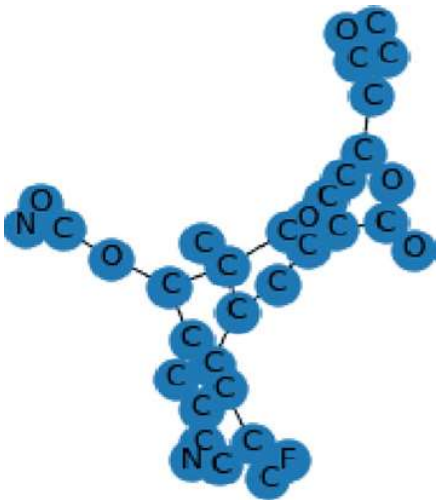


Figure 7.5. Molecular image of newly generated SMILES using LSTM.

7.4 Conclusion

According to the analyses mentioned earlier, it is obvious to forecast the molecular characteristics of newly produced compounds with smiles as input data by applying deep learning (CNN and transfer learning) and the K-MEANS algorithm. More than 5000 molecules were utilized for training, and vggnet was employed for feature extraction. Based on the resemblance of characteristics

based on LST/RNN, clusters were developed with pre-exercise data and the newly generated smile of modified gedunin ($C_{26}H_{31}N_2O_6F$).