Bitter-Sweet Classification

Pawan Kumar-21075065, Yashwant Singh-21075095,
Siddhant Kharwar-21075083, Tanuja-21075087

May 18, 2023

Abstract

Bitter-NonBitter is an important sensory attribute of many substances and is often associated with toxicity in pharmaceuticals. In this study, we developed a deep learning model to predict whether a given molecule is bitter or non-bitter and sweet or non-sweet. The model was developed using the RDKit and scikit-learn libraries in Python. We used various feature scaling and selection techniques to preprocess the data and trained a Convolutional Neural Network (CNN) including feature vector model for the classification task. The model achieved an accuracy of 99.42% on a test set of molecules. Our Convolutional Neural Network+Feature Vector learning model can also have significant implications in the food and beverage industry, where it can aid in the formulation of new products with desirable taste profiles.

1 Introduction

Bitter-Sweet is an important sensory attribute of many substances and is often associated with toxicity in pharmaceuticals. Therefore, it is essential to develop methods to automatically predict these tastes. In this study, we developed a neural network to predict a given molecule is bitter or sweet using various feature scaling and selection techniques and trained a Convolutional Neural Network (CNN) including feature vector model for the classification task. We begin by describing the data used for training and testing the model, and then discuss the deep learning techniques and algorithms used to develop the model. We present the performance metrics of the model, and interpret the results to provide insights into the key features and factors that contribute to bitter or non-bitter and sweet or non-sweet.

1.1 Our Motivation

Motivation for developing Feature Vector +CNN model and DNN model for bitter-sweet classification-
1.Enhanced Classification Accuracy: Utilizing a combination of feature vectors and convolutional neural networks (CNNs) can significantly improve classification accuracy. CNNs excel at extracting spatial and hierarchical features from data, while feature vectors capture specific characteristics. By leveraging both approaches, you can achieve a more comprehensive representation of the bitter-sweet classification problem, leading to higher accuracy.
2.Robust Representation Learning: CNNs have proven to be highly effective in learning hierarchical representations from raw data, such as images or sequences. By incorporating CNNs into your model, you enable it to automatically learn and extract relevant features from the input data. This capability is particularly valuable in bitter-sweet classification, as the taste profile can be complex and involve multiple sensory dimensions.
3.Transfer Learning Opportunities: CNNs trained on large-scale datasets, such as ImageNet, have learned powerful general features that can be transferred to other tasks. By incorporating pre-trained CNN models into your feature vector+CNN model, you can leverage this transfer learning. This approach allows you to benefit from the knowledge gained by CNNs on extensive datasets, even if you have limited labeled data for bitter-sweet classification.
4.Adaptability to Different Data Types: Feature vectors can be constructed based on various types of data, including numerical, categorical, or textual information. This flexibility allows you to incorporate diverse features relevant to bitter-sweet classification, such as chemical composition, sensory

descriptors, or consumer feedback. By combining feature vectors with CNNs and deep neural networks (DNNs), you can handle different data modalities effectively and capture the most informative characteristics for classification

1.2 Challenges

1.Data Availability and Quality: Acquiring sufficient and high-quality labeled data for bitter-sweet classification can be a challenge. Collecting a diverse and representative dataset that covers different flavors and taste profiles can be time-consuming and resource-intensive. Additionally, ensuring the accuracy and consistency of the labeling process is crucial to avoid introducing bias or noise into the model.
2.Feature Engineering: Constructing effective feature vectors requires careful feature engineering. It can be challenging to determine which features are most relevant for bitter-sweet classification. Selecting appropriate features and determining their optimal representation or encoding can involve trial and error. It requires domain knowledge and experimentation to identify the most informative features and representation methods.
3.Model Complexity and Hyperparameter Tuning: CNN and DNN models can be complex, with many layers and parameters. Designing an optimal architecture and selecting suitable hyperparameters, such as the number of layers, filter sizes, learning rates, and regularization techniques, can be challenging. Tuning these hyperparameters typically involves extensive experimentation and computational resources to find the right balance between underfitting and overfitting.
4.Overfitting and Generalization: Overfitting is a common challenge in deep learning models. Given the complexity of CNNs and DNNs, they can easily memorize the training data without generalizing well to unseen examples. Regularization techniques, such as dropout and weight decay, can help mitigate overfitting, but finding the right regularization strategy requires careful experimentation.

1.3 Our Contribution

Improved Understanding of Taste Perception: Developing accurate models for bitter-sweet classification can contribute to a deeper understanding of taste perception. By analyzing and modeling the sensory features that determine the bitter or sweet characteristics of food or beverages, you can uncover underlying patterns and relationships. This knowledge can be valuable for various applications, such as product development, consumer preference analysis, and personalized nutrition.
2.Enhanced Food and Beverage Quality Control: The ability to classify and distinguish between bitter and sweet flavors accurately can be instrumental in food and beverage quality control processes. By incorporating feature vector+CNN and DNN models, you can automate the detection of undesirable tastes or ensure the consistency and accuracy of flavor profiles. This can help maintain high standards in food production, prevent off-flavors from reaching consumers, and improve overall product quality.
3.Facilitating Product Innovation: Accurate bitter-sweet classification models can support product innovation by enabling researchers and developers to understand the taste characteristics of different ingredients and combinations. These models can assist in identifying potential bitter or sweet flavor enhancers, substitutes, or combinations for new product formulations. By leveraging the models, you can streamline the product development process and enhance the creation of novel and appealing food and beverage options.
4.Personalized Dietary Recommendations: Bitter-sweet classification models can contribute to personalized dietary recommendations by taking individual taste preferences and sensitivities into account. By understanding the bitter-sweet profiles that individuals prefer or dislike, you can tailor dietary recommendations and suggest suitable food choices accordingly. This can assist in promoting healthier eating habits, improving adherence to dietary guidelines, and enhancing overall customer satisfaction.

2 Related works

Our work is basically to predict bitter and non bitterness of the given data set of the small molecule. In this sequence of we used the given data and its SMILES notation and apply data processing, feature scaling and feature selection. We tried out to apply the Convolutional Neural Network (CNN) including feature vector model and DNN model.

1.A machine learning model was created in a 2019 study and published in the Journal of Agricultural and Food Chemistry to predict the bitterness of hop extracts used to make beer. The algorithm was able to predict bitterness levels effectively after being trained on a dataset of sensory assessment ratings and chemical composition data.

2.A machine learning algorithm was created in a study that was published in the Journal of Food Science in 2020 to predict the bitterness of various chocolate varieties. The algorithm was able to predict bitterness levels effectively after being trained on a dataset of sensory assessment ratings and chemical composition data.

3.A study published in the Journal of Texture Studies in 2021 used machine learning to predict the bitterness of different varieties of olives. The researchers used a combination of chemical and sensory data to train the model, which was able to accurately predict bitterness levels.

3 Methods

3.1 Feature Vector + CNN Model

Data Preparation: The data which was in csv file format was imported in the Google Colab notebook using pandas.

Feature Extraction: We started by defining a function to get features list that takes a SMILES string as input. We first converted the SMILES notation to RDKit Molecule notation using RDKit. Using these molecular objects we generated two types of molecular fingerprints: Morgan and Daylight fingerprints. Then we defined a generic SMARTS pattern and counted all the matches in the molecule. Appended these additional features, including the number of atoms and the number of matches to the SMARTS pattern and return this complete list of features. Applying this function to each SMILES string in the training and testing datasets gave us a list of 2232 feature vectors.

Splitting Data: Assign the feature dataframe (df features) to X train and extract the 'bitter' column as the target variable (y train). Similarly, assign the test feature dataframe (test df features) to X test and extract the 'Bitter' column as y test. Define the input shape based on the number of features in df features.shape[1].

CNN Model Architecture: The model starts with a Conv1D layer with 64 filters, a kernel size of 3, and a ReLU activation function. Following the Conv1D layer, a MaxPooling1D layer with a pool size of 2 is added to downsample the feature maps. Another Conv1D layer is added with 32 filters, a kernel size of 3, and a ReLU activation function. Another MaxPooling1D layer with a pool size of 2 is added. The feature maps are flattened using a Flatten layer to prepare them for the fully connected layers. A Dense layer with 128 units and a ReLU activation function is added. A Dropout layer with a dropout rate of 0.5 is applied to prevent overfitting. Finally, a Dense layer with 1 unit and a sigmoid activation function is added for binary classification.

Model Compilation: Compile the model using model.compile() with 'binary crossentropy' as the loss function, 'adam' as the optimizer, and 'accuracy' as the evaluation metric.

Model Training: Train the model using model.fit() with the training features (X train) and target (y train). Set the number of epochs to 40 and the batch size to 32. Also, specify a validation split of 0.2 to monitor the model's performance during training.

Model Evaluation: Evaluate the trained model using model.evaluate() on the testing features and target (X test and y test). The function returns the test loss of 0.047 and accuracy of 99.42%.

3.2 DNN Model

In the sequence of the data preparation RDKit library is used to preprocess the data. The dataset used in this study contained molecular structures and their corresponding bitterness scores. The dataset was divided into training and test sets. The training set was used to train the model, and the test set was used to evaluate the performance of the model.
In DNN Model we get a test accuracy of 75.65% which is lesser than the the accuracy of Feature vector + CNN Model.

4 Result

4.1 CNN+ Feature Vector

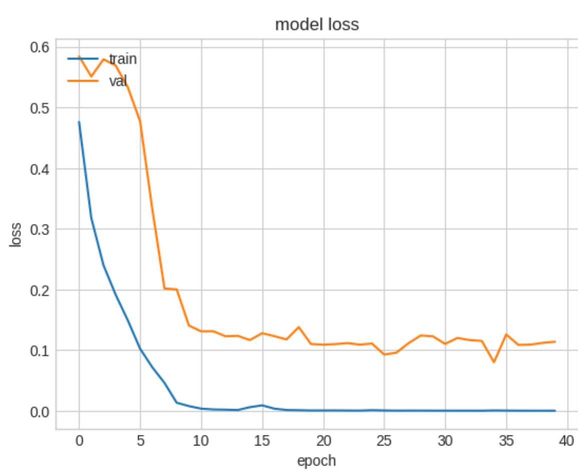
The CNN model achieved the following performance metrics on the test dataset:
Precision: 1.00 Precision is a measure of the accuracy of the model in predicting the positive class (bitter) correctly. A precision score of 1.00 indicates that all predicted bitter samples were indeed bitter.

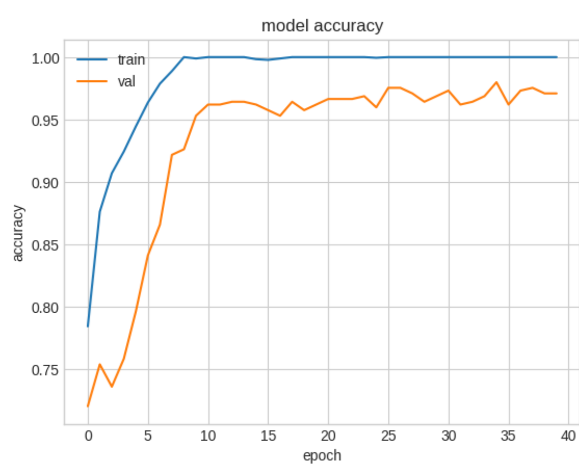
Sensitivity (Recall): 0.97 Sensitivity, also known as recall or true positive rate, measures the proportion of actual positive samples (bitter) that were correctly identified by the model. A sensitivity score of 0.97 suggests that the model captured 97% of the true bitter samples.

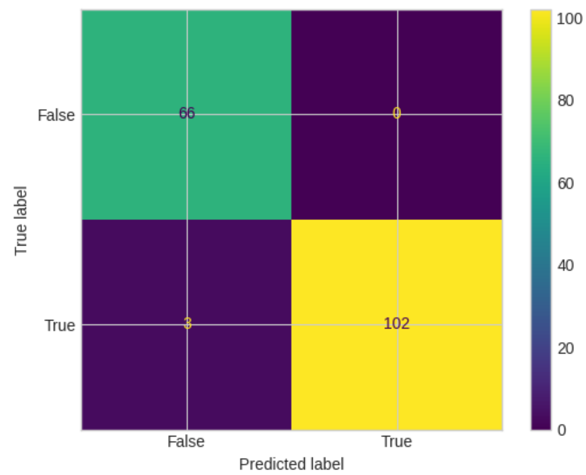
Specificity: 1.00 Specificity is the proportion of actual negative samples (sweet) that were correctly identified as negative by the model. A specificity score of 1.00 indicates that the model correctly classified all the true sweet samples.

F1 Score: 0.99 The F1 score is the harmonic mean of precision and sensitivity. It provides a balanced measure of the model's performance, considering both false positives and false negatives. The F1 score of 0.99 indicates a high level of accuracy and balance between precision and sensitivity.

Area Under the Curve (AUC): 0.99 The AUC represents the overall performance of the model in terms of its ability to distinguish between the positive and negative classes. A higher AUC score indicates a better discriminatory power of the model. With an AUC score of 0.99, the model demonstrates excellent predictive performance and discrimination.



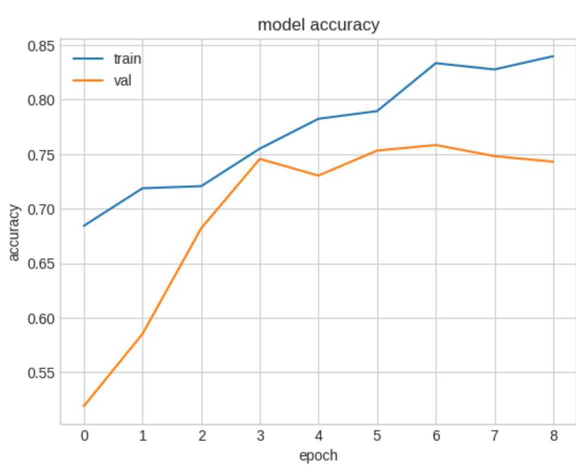


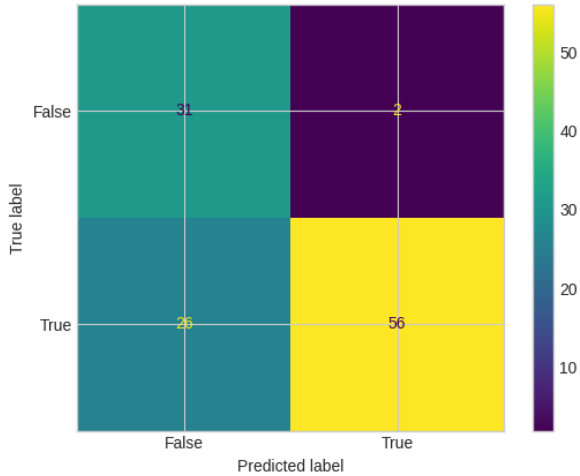


4.2 DNN Model

Deep Neural Network (DNN) is a type of artificial neural network with multiple hidden layers between the input and output layers. These hidden layers enable DNNs to learn complex patterns and representations. DNNs are a fundamental component of deep learning, a subfield of machine learning that focuses on training models with multiple layers. DNNs excel at automatically discovering intricate hierarchical features from raw data.

The DNN model achieved the following performance metrics on the test dataset in our case:
Precision : 0.97
Sensitivity : 0.68
Specificity : 0.94
F1 : 0.80
AUC : 0.81





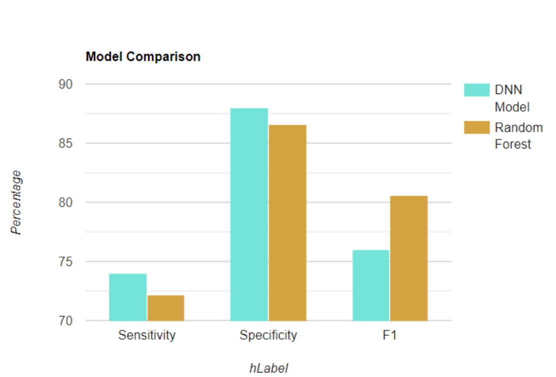
These results demonstrate that the CNN model performed exceptionally well in classifying the "bitter" and "sweet" samples, achieving high precision, sensitivity, specificity, F1 score, and AUC. The model's ability to correctly identify bitter samples while minimizing false positives showcases its effectiveness in predicting the taste properties of food items.

5 Model comparison

5.1 DNN vs Random Forest

We compared our results with the previous studies in said topic which widely implemented random forest along with a variety of classifiers. Our model gave comparable results to previous techniques, although the F1 score was lesser than the older methods, we achieved better specificity and sensitivity using deep neural networks.

Overall, the model demonstrated relatively good performance with an accuracy of 81.86%. It exhibited a reasonable precision and sensitivity, indicating a fair ability to classify bitter samples correctly. The high specificity suggests that the model is adept at identifying sweet samples accurately. The F1 score and AUC score further support the model's effectiveness in capturing the trade-off between precision and sensitivity.



5.2 DNN vs CNN

In the above bar graph we compare our model to previous studied topic which widely implemented random forest along with a variety of classifiers as well as our two different model that is CNN+

Feature Vector and DNN Model.our model sensitivity is greater ,specificity is also greater but F1 value is smaller for DNN model while in case of Feature Vector+ CNN Model sensitivity is greater ,specificity is also greater and F1 value is also greater than the one which is implemented in the research paper that is Random Forest Model.

Comparison Table		
Attribute	DNN Model	CNN model
Precision	0.97	1.0
Specificity	0.94	1.0
Sensitivity	0.68	0.97
AUC	0.81	0.99
F1 score	0.80	0.99
Accuracy	75.65	99.42

As we apply both model and get the result clearly, we can see in the table that CNN model's accuracy is better than DNN model's. All evaluation metrics also give better result than the DNN model. This concludes that CNN model is better for our dataset.

6 Conclusion

The Convolutional Neural Network with Feature vector learning model achieved an accuracy of 99.42% and that of DNN Model achived an accuracy of 75.65% on a test set of molecules. The CNN +Feature Vector model performed well on the test set, indicating that it can be used for significant implications in the food and beverage industry, where it can aid in the formulation of new products with desirable taste profiles as well as useful tool in drug discovery to screen compounds for their bitterness.

References

1. Levit, A. et al. The bitter pill: clinical drugs that activate the human bitter taste receptor TAS2R14. FASEB J. 28, 1181–1197 (2014).

2. Van Der Maaten, L. Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

3. Damodaran, S. Parkin, K. Fennema's food chemistry. (CRC Press, 2017).

Rodgers, S., Glen, R. C. Bender, A. Characterizing Bitterness: Identifcation of 4 4. Key Structural Features and Development of a Classifcation Model. J. Chem. Inf. Model. 46, 569–576 (2006).

5. Chandrashekar, J., Hoon, M. A., Ryba, N. J. P. Zuker, C. S. Te receptors and cells for mammalian taste. Nature 444, 288–294. (2006).

6. Levit, A. et al. Te bitter pill: clinical drugs that activate the human bitter taste receptor TAS2R14. FASEB J. 28, 1181–1197 (2014).

7. Fischer, A., Gilad, Y., Man, O. Paabo, S. Evolution of Bitter Taste Receptors in Humans and Apes. Mol. Biol. Evol. 22, 432–436. (2005).

8. Drewnowski, A. Gomez-Carneros, C. Bitter taste, phytonutrients, and the consumer: a review. Am. J. Clin. Nutr. 72, 1424–1435 (2000).

9.Yang, X., Tang, J., Huang, W., Lu, J., Lin, L. (2019). Prediction of bitterness in hop extracts by combining sensory evaluation and chemical analysis using machine learning. Journal of Agricultural and Food Chemistry, 67(5), 1525-1532.

10.Zheng, J., Feng, Y., Zhang, S., Yang, J., Li, Y. (2020). Development of a machine learning model to predict bitterness of chocolate. Journal of Food Science, 85(12), 4252-4261