WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Ensemble Model for Predicting Cosmetic Product Toxicity by Analysing the Ingredients

Antony Arokiamary Sherene[1], Dr. Jerline Amutha. A[2]

[1]Student, PG Department of Computer Science and Technology, Women's Christian College, Chennai, Tamil Nadu, India
[2]Assistant Professor, PG Department of Computer Science and Technology, Women's Christian College, Chennai, Tamil Nadu, India

A B S T R A C T

Regular use of beauty products does contribute to our appearance, but most of them contain hazardous ingredients. Ignorance of these can lead to a variety of diseases, such as allergies, cancer, harm to a person's reproductive system in both genders, etc. The objective of this study is to create a dependable machine learning model that uses ingredient-level toxicity studies to predict the toxicity level of cosmetic items. The study suggests using an ensemble model, Support Vector Machine (SVM), Random Forest, and LightGBM Classifier to classify the toxicity of a specific ingredient and aggregate these data to determine the toxicity of the cosmetic product. The goals of this research are to minimize the harmful effects of toxins in beauty products on human health and to enhance consumer knowledge by developing an efficient method for identifying hazardous elements and producing comprehensive toxicity reports. This would enable the use of beauty products in a safer manner.

Keywords: Machine learning, Ensemble model, Cosmetic toxicity prediction.

1. INTRODUCTION

People use cosmetics to enhance their appearance all around the world. The producers utilize a lot of hazardous substances that are harmful to human health in order to improve the product's performance. Some cosmetic items have been found to include substances that are forbidden or restricted according to regulations in the USA (United States of America) and the EU (European Union), which raises concerns about their safety [10]. People will be vulnerable to dangerous illnesses as a result of their ignorance about these dangerous substances in cosmetics. Cosmetics can cause a wide range of diseases, including those that are carcinogenic (cause cancer), reprotoxic (cause difficulties with reproduction), immunotoxic (cause problems with the immune system), and more. Therefore, by analyzing the ingredients of cosmetic goods and giving consumers information about their level of toxicity, this research offers a machine learning model that helps people understand the safety of the items they use. Thus, protecting consumers from potentially harmful ingredients in cosmetics and helping them make cautious choices.

2. LITERATURE REVIEW

[1] Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis (Yunyi Wu & Guanyu Wang, 2018). It predicted toxicity using machine learning methods such as deep learning, random forests, and support vector machines based on integration of chemical structure information with human transcriptome analysis.

[2] Predicting toxicity properties through machine learning (Luz Adriana Borreroa et al, 2020). This research identified the most effective machine learning methods available right now for toxicity prediction as an ADME-Tox attribute. The decision tree significantly outperformed other models in terms of prediction accuracy and alignment with actual results, despite the fact that other models also produced impressive results.

[3] Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints (Huawei Feng et al, 2021). It predicted the reproductive toxicity of chemicals, ensemble learning models were constructed utilising support vector machine, random forest, and extreme gradient boosting.

[4] Cosmetic product selection using machine learning (Rubasri S et al, 2022). This paper proposes a simple cosmetic recommendation system that extracts effective cosmetic ingredients for each user attribute and develops a recommender system based on these ingredients using NLP techniques.

[5] Product ingredient analysis (Harshit Gautam et al, 2022). This study examines 1,472 cosmetic ingredients from Sephora and suggests a content-based recommendation system based on data science.

[6] hERG-toxicity prediction using traditional machine learning and advanced deep learning techniques (Erik Ylipaa et al, 2023). It predicted human ether-á-go-go related gene (hERG) toxicity based on traditional and advanced AI techniques. It discovered that older machine learning algorithms perform equally well as sophisticated strategies. The GNN model is notable for its better performance, low feature engineering, and automation benefits.

[7] Review of machine learning and deep learning models for toxicity prediction (Wenjing Guo et al, 2023). This paper determined the toxicity of chemicals using machine learning and deep learning. It found for Machine Learning models, Support Vector Machine, Random Forest, and Ensemble Learning are the most frequently used algorithms, respectively. For Deep Learning models, Multilayer Perceptron and Convolutional Neural Network are the widely used algorithms.

[8] Systematic approaches to machine learning models for predicting pesticide toxicity (Ganesan Anandhi & M. Iyapparaja, 2024). It found that the regression models such as SVM, k-NN, ANN, CNN, LDA, DQA, and RF are suitable for application in toxicity prediction.

[9] Review of toxic chemicals in cosmetics (Domina Petric, 2021). This article examined the literature on harmful substances found in cosmetics and stressed the value of toxicological research in the field of cosmetics.

[10] Analysis of prohibited and restricted ingredients in cosmetics (Rimadani Pratiwi, 2022). This study examined analytical techniques that have been recently published for compounds that are banned in cosmetic products by the FDA and the EU.

3. METHODOLOGY

3.1 DESCRIPTION OF DATASET

The dataset was developed using toxicity data from EWG (Environmental Working Group), a nonprofit organisation that promotes environmental health and safer consumer products. The dataset includes ingredient names,cancer (level of causing marked as 1-low, 2-medium, 3-high), allergies (level of causing marked as 1-low, 2-medium, 3-high), immunotoxic (level of causing disease in the immune system marked as 1-low, 2-medium, 3-high), reprotoxic (level of causing disease in the reproductive system marked as 1-low, 2-medium, 3-high), developmental disorder (level marked as 1-low, 2-medium, 3-high), restrictions (level marked as 1-low, 2-medium, 3-high), source (natural-1, synthetic-2, both-3), status of the ingredients (banned – 1 or active - 0).Toxicological data for 400 compounds were obtained in total, out of which 200 were toxic ingredients and 200 were non-toxic ingredients.

3.2 DATA PRE-PROCESSING

The data was collected manually, thus duplicates (two identical ingredient information) and excessive white spaces in the component names were removed to prevent prediction errors. To convert categorical values to numerical values, label encoding was utilised. In the specified columns: cancer, allergens, immunotoxic, developmental, reprotoxic, restrictions, the labels 'low', 'medium', and 'high' are replaced with 1, 2, and 3, respectively and for source , the labels ‘natural’ ,’synthetic’,’both’ are replaced as 1,2,3 and for status, the labels ‘active’ ,’banned’ are replaced as 0,1.The "overall toxicity" score for each chemical was calculated by adding the values from six categories: cancer, allergies, immunotoxic, developmental, reprotoxic, and restrictions on usage. Equation 1 (Eq. 1) describes the formula below:

Overall_toxicity = cancer + allergies + immunotoxic + developmental + reprotoxic + restrictions ……..(Eq. 1)

Since each category (cancer, allergies, immunotoxic, developmental, reprotoxic, and restrictions on use) has a value of either 1, 2, or 3, the maximum possible score for each category is 3. Given there are 6 categories, the maximum possible value for the sum is calculated using the formula given below in equation 2 (Eq. 2):

Max Possible Overall Toxicity=6×3=18 ……..(Eq. 2)

To normalize the overall toxicity, the formula used is given in Fig. 1:

$$\text{normalized\_toxicity} = \frac{\text{overall\_toxicity}}{\text{Max Possible Overall Toxicity}}$$

Fig. 1 – Normalized toxicity formula

The ingredients whose normalized_toxicity value is greater than or equal to 0.5 are considered as toxic in this study. Finally, after preprocessing and calculations, the dataset was prepared with a total of 400 ingredient data points with 11 features. From 400 ingredient data points, 70% were used for training and 30% for testing. Sample dataset is given below in Fig. 2:

cosmetic_ingredients	cancer	allergies	immunotoxic	developmental	reprotoxic	se_restriction	source	status	overall_toxicity	normalized_toxicity
1,2-Hexanediol	1	1	1	1	1	1	2	0	6	0.333333333
2 Phenylphenol	2	2	2	1	1	2	2	1	10	0.555555556
4 Aminobiphenyl	3	1	1	1	1	3	2	0	10	0.555555556
4,4 Isobutylethylidenedip	1	1	1	2	2	3	2	1	10	0.555555556
Acetyl Glucosamine	1	1	1	1	1	1	1	0	6	0.333333333

Fig. 2 – Sample Dataset

3.3 PROPOSED METHOD

The proposed approach employs an ensemble learning model with Support Vector Machine (SVM), Random Forest, and LightGBM to assess the toxicity of a cosmetic product based on ingredient analysis. All three models are highly regarded for their unique specialities. By combining all three models into an ensemble, the model gains from LightGBM's accuracy and speed, as well as Random Forest's robustness and stability, resulting in more reliable and accurate predictions. SVM's classification capability can be increased by combining it with ensemble methods. A voting classifier is used, which incorporates SVM, LightGBM, and Random Forest predictions. The classifier is set to use soft voting, which means it predicts the class label by averaging the expected probability from each model. The ensemble takes advantage on the strengths of each model by combining their predictions, potentially boosting overall forecast accuracy and robustness**.** Workflow diagram of the proposed model is given in Fig. 3:

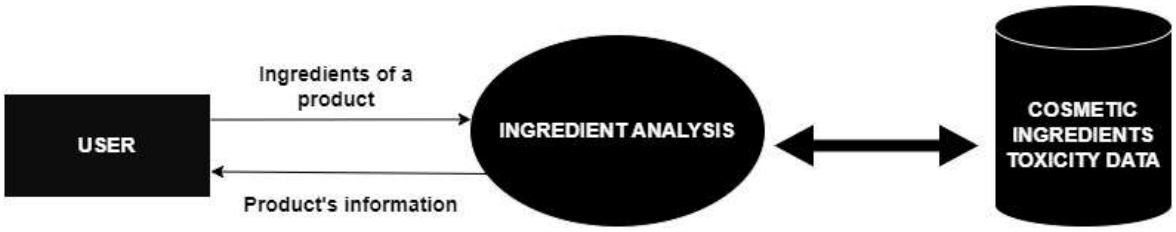

Fig. 3 – Workflow diagram of the proposed model

Ingrdient analysis is performed at two levels :

- Prediction of individual ingredient toxicity
- Prediction of product toxicity

3.3.1 PREDICTION OF INDIVIDUAL INGRDIENT TOXICITY

In this level, given the cosmetic ingredients, it will predict whether or not the ingredient is toxic based on the toxicity features selected. The features are chosen for training using recursive feature elimination (RFE). It is a feature selection approach that involves iteratively removing the least significant characteristics from a dataset and then creating a model with the remaining features. RFE feature selection was applied to all three models: SVM, Random Forest, and LightGBM, and four common features were selected from each model for the ensemble model's prediction. The following features were frequently employed in training: immunotoxic level, allergy level, restriction level, and reprotoxic level. These were the most important characteristics for identifying the toxicity of an ingredient. A comparison of feature selection using the three models is provided below in Table 1:

Table 1 - Comparison of features used in each model.

MACHINE LEARNING MODEL	FEATURES SELECTED USING RFE
SVM	developmental disorder level, immunotoxic level, reprotoxic level, restrictions level
Random forest	allergies level, immunotoxic level, reprotoxic level, restrictions level
LightGBM	allergies level, cancer level, restrictions level, source level
Ensemble (SVM, Random forest, LightGBM)	allergies level, immunotoxic level, reprotoxic level, restrictions level

3.3.2 PREDICTION OF PRODUCT TOXICITY

After assessing the toxicity of each ingredient, the overall product toxicity is projected by aggregating this information. It shows whether the product is toxic or not. If the product is toxic, or if it contains a banned ingredient (illegal in any country), it will notify the user along with the ingredient's name. In addition, the percentage of causing every illness is displayed along with the ingredients responsible for it. For example, the cancer-causing percentage is 20%. Ingredients causing cancer: Benzene, Tetrafluoropropene.

4. RESULTS AND DISCUSSIONS

This study analysed three models: Support Vector Machine(SVM), Random forest and LightGBM. All the 3 models performed well in predicting whether cosmetic items were hazardous or not based on ingredient toxicity data. Support Vector Machine (SVM) provided 97% accuracy, Random Forest supplied 98% accuracy, and LightGBM provided 98% accuracy. The ensemble model, which incorporated Support Vector Machine (SVM), LightGBM's accuracy and speed, and Random Forest's robustness and stability, obtained 98% accuracy. The Random Forest and LightGBM models performed well individually, predicting the product's toxicity with a 98% accuracy. In combining these three models, SVM, Random Forest, and LightGBM significantly improved SVM's classification capability, while its individual performance was slightly lower than the other two models. Several studies have identified Random Forest, Support Vector Machines (SVM), and gradient-boosting machine models as the most often utilised machine learning algorithms for predicting toxicity[1,3,6,7,8]. An ensemble model using Support Vector Machines (SVM), Random Forest, and LightGBM can accurately forecast the toxicity of a cosmetic product based on ingredient-level data. This concept aims to categorise risky ingredients and provide thorough toxicity reports to increase customer knowledge and lessen the negative impact of toxins in beauty products. Comparison of individual model accuracies with ensemble model is given below in Fig. 4:

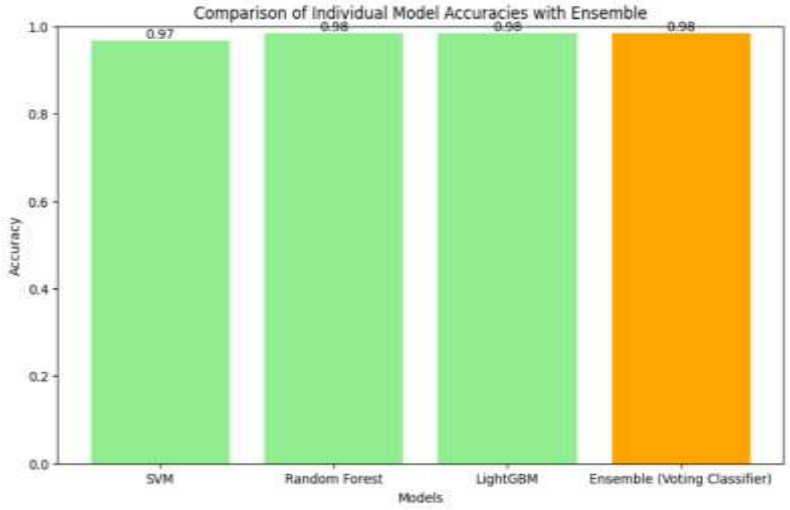

Fig. 4 – Comparison of individual model accuracies with ensemble

4.1 SCREENSHOTS

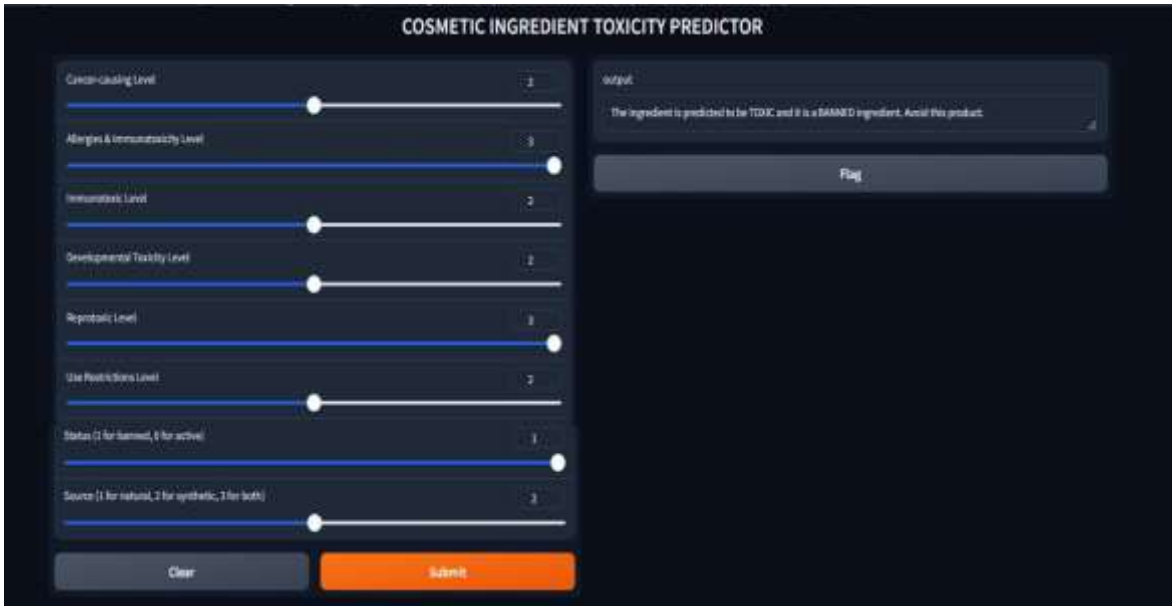

Fig. 5 – User interface for predicting single ingredient toxicity

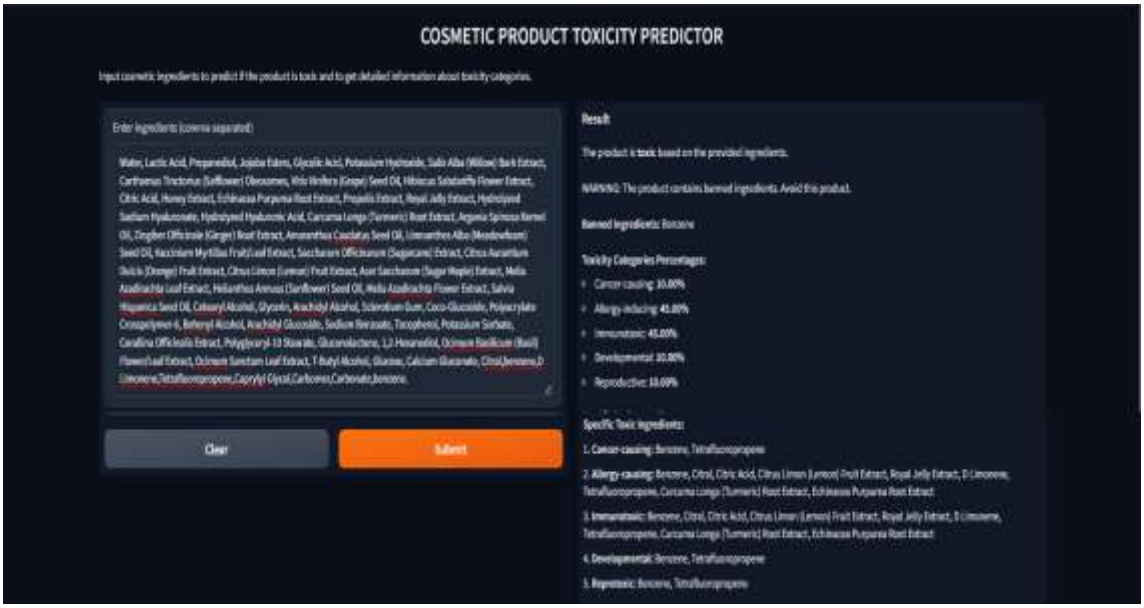

Fig. 6 – User interface for predicting product toxicity

5. CONCLUSION

The ensemble model, which incorporates Support Vector Machine (SVM), Random Forest, and LightGBM, works well on the dataset and predicts whether a cosmetic product is unsafe or not based on toxicity data for each individual ingredient. It has an accuracy of 98%. This model will develop an effective method for categorising hazardous ingredients and giving detailed toxicity reports in order to raise customer awareness and reduce the negative health consequences of toxins prevalent in beauty products. The research could be expanded to include more features like respiratory toxicity, neurotoxicity, and so on for predicting an ingredient's toxicity data. Increase the size of the dataset and experiment with different machine learning and deep learning models.

References

[1] Wu, Yunyi & Wang, Guanyu. (2018). Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. International Journal of Molecular Sciences. 19. 2358. 10.3390/ijms19082358.

[2] Borrero, Luz & Guette, Lilibeth & Lopez, Enrique & Pineda, Omar & Buelvas, Edgardo. (2020). Predicting Toxicity Properties through Machine Learning. Procedia Computer Science. 170. 1011-1016. 10.1016/j.procs.2020.03.093.

[3] Feng, Huawei & Zhang, Li & Li, Shimeng & Liu, Lili & Yang, Tianzhou & Yang, Pengyu & Zhao, Jian & Arkin, Isaiah & Liu, Hongsheng. (2021). Predicting the Reproductive Toxicity of Chemicals Using Ensemble Learning Methods and Molecular Fingerprints. Toxicology Letters. 340. 10.1016/j.toxlet.2021.01.002.

[4] S, Rubasri & S, Hemavathi & Jayasakthi, K. & Aranganathan, Sangeerani & .K, Latha & Nithyanandam, Gopinath. (2022). Cosmetic Product Selection Using Machine Learning. 1-6. 10.1109/IC3IOT53935.2022.9767972.

[5] Gautam, Harshit & Singh, Vishal & Kumar, Deepak & Kaur, Sukhpreet. (2022). Product Ingredient Analysis. International Journal for Research in Applied Science and Engineering Technology. 10. 618-620. 10.22214/ijraset.2022.43745.

[6] Ylipää, Erik & Chavan, Swapnil & Bånkestad, Maria & Broberg, Johan & Glinghammar, Björn & Norinder, Ulf & Cotgreave, Ian. (2023). hERG-Toxicity Prediction using Traditional Machine Learning and Advanced Deep Learning Techniques. Current Research in Toxicology. 5. 100121. 10.1016/j.crtox.2023.100121.

[7] Guo, Wenjing & Liu, Jie & Dong, Fan & Song, Meng & Li, Zoe & Khan, Md & Patterson, Tucker & Hong, Huixiao. (2023). Review of machine learning and deep learning models for toxicity prediction. Experimental biology and medicine (Maywood, N.J.). 248. 15353702231209421. 10.1177/15353702231209421.

[8] Anandhi, Ganesan & Meenakshisundaram, Iyapparaja. (2024). Systematic approaches to machine learning models for predicting pesticide toxicity. Heliyon. 10. e28752. 10.1016/j.heliyon.2024.e28752.

[9] Petric, Domina. (2021). Review of Toxic Chemicals in Cosmetics. 10.14293/S2199-1006.1.SOR-.PPKO7OD.v1.

[10] Pratiwi, Rimadani & As, Nisa & Yusar, Rani & Shofwan, Adnan. (2022). Analysis of Prohibited and Restricted Ingredients in Cosmetics. Cosmetics. 9. 87. 10.3390/cosmetics9040087