Predicting Drug Addiction Using Ensemble Learning: A Majority Voting Classifier Approach

Ismail Idowu AKUJI[1] Mutiat. A OGUNRINDE[2] Taofik Abiodun AHMED[3] Sulaiman Olaniyi ABDULSALAM[4] Ronke Seyi BABATUNDE[5]

1,4,5 Department of Computer Science, Kwara State University, Malete, Nigeria.

2 Department of Mathematical and Computer Sciences, Fountain University, Osogbo, Nigeria.

3 Department of Computer Science, Kwara State College of Arabic and Islamic Legal Studies Ilorin, Nigeria.

KWASU Journal of Information, Communication and Technology (KJICT.) 2024, 3(1)

URL: http://kjict.ng/

Abstract

Globally, drug intake has claimed the lives of half a million people due to its complexity and devastating effects. Prompt intervention and prevention are crucial to curbing the fatalities and mortality caused by drug addiction. Leveraging machine learning for drug intake prediction and building on the existing research becomes imperative to reinforcing the effectiveness of machine learning utilisation, ultimately helping to curb the prevalence of drug intake. Therefore, this study proposes an ensemble learning approach, using a majority voting classifier to predict drug addiction from the analysis of facial images. The choice of the majority voting classifier over other ensemble methods, such as boosting, bagging, and stacking, is due to its simplicity, interpretability, and reduced potential for overfitting, as it averages out individual model predictions to produce more robust outcomes. A dataset containing 400 facial images was acquired from the Mendeley repository, comprising 240 addicted and 160 non-addicted images. Combining the predictions of multiple base models, namely ResNet, VGG, and Inception models, this study's approach ensures improved accuracy and robustness of the drug addiction prediction model. The Majority Voting classifier model was evaluated on the acquired dataset, achieving a 0.6573 loss and 69% accuracy rate on the training set and 2.8133 loss and 60% accuracy on the validation set. The discrepancy between training and validation performances indicates that the model is overfitting to the training data. As a result, further study can address the overfitting issue by considering data augmentation to ensure the model performs well in real-world scenarios.

Keywords: Drug addiction, Facial image analysis, Ensemble learning, Majority Voting classifier, Machine learning

1. Introduction

Drug addiction is a complex and chronic disorder that affects millions of people worldwide, causing significant harm to individuals, families, and society as a whole. The surge in illicit drug use, particularly among youth, highlights the urgent need for early identification and intervention to prevent the prevalence of drug addiction. Early detection and timely intervention are critical in preventing the progression of drug addiction and reducing its harmful consequences. Therefore, developing innovative and effective addiction detection and prevention methods is crucial.

In a bid to address the identified concern, several tools have been set aside and are in use. However, current methods (such as urine drug screening, blood tests, saliva tests, clinical interviews, physical examinations, and laboratory tests) rely heavily on subjective clinical evaluations and limited biomarkers. Recent advances in machine learning and data analytics offer a significant contribution to predicting drug addiction risk, enabling the development of

more targeted and effective prevention and treatment strategies. Moreover, various ensemble learning approaches have been utilised by previous studies, such as boosting (Faisal et al., 2019) and bagging (Choi et al., 2021; Kumara et al., 2023). The existing approaches have the potential to revolutionise the field of addiction research and treatment, although the methods are not optimally efficient to mitigate the devastating consequences of drug addiction.

Unlike the previous literature, which majorly focused on risk-based datasets, this study intends to employ a facial image dataset to contribute to the few existing studies, such as Kumara et al. (2023), Gnanasekar and Yanushkevich (2019) which also utilised facial image dataset but used algorithms that differed from the proposed algorithm in this study. While Kumara et al. (2023) used several algorithms, including CNN, Random Forest, and Decision Tree, Gnanasekar and Yanushkevich (2019) utilised pre-trained CNN (GoogleNet, ResNet50, and VGG16).

This study proposes a novel ensemble learning approach utilising a majority voting classifier to predict drug addiction risk. The method is chosen for its potential nature of simplicity, interpretability, and reduced tendency to overfitting (Ilyas, 2023). By combining the predictions of multiple base learners, namely ResNet, VGG, and Inception, this approach leverages the strengths of various algorithms to improve the accuracy and robustness of addiction risk prediction.

The facial dataset available in the Mendeley repository was used in this study. The utilisation of a machine-learning approach has profound implications and is poised to revolutionise addiction research and treatment. By ensuring early detection of individuals at high risk of addiction, this approach can help prevent the escalation of drug dependence, thereby mitigating the substantial social, economic, and healthcare costs associated with this condition. This breakthrough has the potential to transform lives, families, and communities, ultimately contributing to a healthier and more resilient society. The remaining parts of this paper are as follows: Section 2 covers the literature review, followed by methodology (Section 3), results and discussion (Section 4), and finally, the conclusion, recommendations, and references (Section 5).

2. Literature Review

Predicting drug addiction is a crucial task that has garnered significant attention in recent years. Ensemble learning approaches have shown promising results in predicting drug addiction. Kumara et al. (2023) employed an Ensemble Learning approach, combining CNN, Random Forest Classifier, and Decision Tree Classifier, achieving high accuracy rates (91.09%–96.75%). Similarly, Choi et al. (2021) used Random Forest and LASSO algorithms to build prediction models for nicotine addiction among youth e-cigarette and hookah users, achieving high performance. Gu et al. (2021) proposed an Ensemble Learning approach using bi-modal signals (EEG and NIRS) and machine learning algorithms, achieving a 3-class accuracy score of 63.15%. Castro et al. (2021) used an Ensemble Learning approach to optimise 3D printed drug delivery systems, achieving a high accuracy score (93%) in predicting critical aspects of the 3DP formulation pipeline and in vitro dissolution properties.

The studies reviewed demonstrate the effectiveness of Ensemble Learning approaches in predicting drug addiction and related healthcare issues. Combining multiple predictive models and Ensemble Learning approaches can improve prediction accuracy and robustness, showing potential in addressing the complex issue of drug addiction. While these studies demonstrate the potential of ensemble learning in addressing drug addiction and related healthcare issues, the current study, unlike previous studies, focuses on utilising the majority voting classifier, generalising the effectiveness of ensemble learning in real-world scenarios. Like other ensemble learning classifiers, a majority voting classifier is also recognised as a practical approach to combining the predictions of multiple models to improve overall accuracy.

There have been many studies on predicting drug addiction using machine learning. However, quite a few of these studies employed facial images. Kumara et al. (2023) conducted "A Machine Learning Approach to Analyse Drug Addiction" by employing several algorithms, including CNN, Random Forest Classifier, and Decision Tree Classifier. The study achieved high accuracy rates (91.09%–96.75%), and the study's findings contribute to the development of precise, reliable, and personalised strategies for combating drug addiction. Kumari and Swetapadma (2023) proposed a machine-learning approach to detect heroin abuse using a random forest-based model. The study resulted in 94.697% accuracy score.

Ripperger et al. (2022) developed and validated

ensemble learning to predict opioid-related overdose risk using Tennessee statewide data from 2012 to 2017. The models, trained on 3 million patients and 71 million prescriptions, showed improved discrimination and calibration after ensembling (AUROC: 0.79-0.83). Abada and Bouramoul (2022) proposed a machine learning-based model to predict the risk of substance abuse using a Naïve Bayes algorithm and found an encouraging accuracy score of 91.4%, indicating that machine learning-based models can greatly assist addiction physicians in predicting individuals at risk for drug addiction.

Arif et al. (2021) also used logistic regression, support vector machines, and random forest to predict the risk of drug addiction. The model achieved an accuracy of 97.91% using logistic regression, outperforming other classifiers, and identifies significant factors for addiction, providing a valuable tool for prevention and intervention. Choi et al. (2021) presented "Machine Learning-Based Nicotine Addiction Prediction Models for Youth E-Cigarette and Waterpipe (Hookah) Users." The study developed predictive models for nicotine addiction among youth e-cigarette and hookah users using machine learning algorithms. The study adopted the National Youth Tobacco Survey (2019) data, containing 6511 instances of people who had used e-cigarettes or hookah. Random Forest and LASSO algorithms were used to build prediction models, with ReliefF and Davies-Bouldin clustering evaluation indexes identifying important predictor variables. The final analysis included 193 predictor variables, and the results showed high performance of the model in predicting nicotine addiction.

Gu et al. (2021) proposed "Application of Bi-Modal Signal in the Classification and Recognition of Drug Addiction Degree Based on Machine Learning." This study focused on an objective, quantified approach to evaluating drug addiction severity using bi-modal signals (EEG and NIRS) and machine learning algorithms. Data from 45 drug addicts (classified into mild, moderate, and severe groups) was utilised, along with convolutional neural networks (CNNs), to extract features from EEG and NIRS data. The features are then fused using decision fusion, achieving a 3-class accuracy of 63.15%. The results suggest that the approach can effectively measure drug addiction severity, providing a real-time, objective evaluation method. Castro et al. (2021) reviewed the "application of machine learning (ML) in optimising 3D printed drug delivery systems, overcoming the limitations of traditional trial-and-error approaches." The study used 968 data points formulated from 114 articles. The study found that ML models achieved high accuracy, up to 93%, in predicting critical aspects of the 3DP formulation pipeline and in vitro dissolution properties. Notably, an artificial neural network predicted drug release times with a mean error of ±24.29 minutes, identifying crucial variables for formulation development. This study showcases the potential of ML in revolutionising pharmaceutical 3DP, enabling rapid and accurate optimisation of personalised drug products.

Another empirical study titled "Applications of Graph Machine Learning in Drug Discovery and Development: A Multidisciplinary Review" was conducted by Gaudelet et al. (2021). The review chronologically traversed the drug development pipeline through a multidisciplinary lens, examining GML's role in target identification, small molecule and biologic design, and drug repurposing. While still emerging, GML has achieved notable milestones, including repurposed drugs advancing to in vivo studies. It indicates its potential to become a preferred modelling framework in biomedical machine learning, poised to transform drug discovery and development. Furthermore, Lekkas et al. (2021) used a consensus ensemble approach and social networking data (Instagram) to predict acute suicidal ideation (SI) in adolescents with a history of lifetime SI. The model achieved an accuracy of 0.702, outperforming previous efforts.

Adam et al. (2020) reviewed "the challenges and recent progress in using machine learning approaches to predict drug responses in cancer treatment." The review compares machine learning techniques for practical use by clinicians and non-experts, highlighting the potential of incorporating new data modalities like single-cell profiling and methods for rapid identification of effective drug combinations to enhance cancer care. The study concluded that by addressing the challenges and leveraging advances in machine learning, personalised drug response prediction may become a reality, improving patient outcomes and treatment success.

Deng et al. (2020) presented a multimodal deep learning framework, DDIMDL, for predicting drug-drug interaction (DDI) events. The framework outperformed state-of-the-art methods and baseline models, with chemical substructures emerging as

the most informative feature. The combination of substructures, targets, and enzymes yielded an accuracy of 0.8852 and an AUPR of 0.9208, demonstrating the potential of DDIMDL in advancing DDI prediction and drug development. The study realised this approach may improve drug safety and efficacy, enabling personalised medicine and more effective treatment strategies.

Faisal et al. (2019) proposed "A Supervised Machine Learning Approach to Predict Vulnerability to Drug Addiction." The study identified the features of individuals vulnerable to drug addiction based on their social, parental, and health factors. A dataset of 498 samples, each with 60 features, collected from rehabilitation centres, colleges, and universities in Dhaka, Bangladesh, was used and validated using Cronbach's alpha nominal test and analysed using ten supervised machine learning algorithms, including neural networks, random forests, and XGBooster. The results showed that XGBooster achieved the highest accuracy (95.20%) in predicting vulnerability to drug addiction. Feature selection techniques (mRMR, Chi-square, and PCA) identified vital features associated with drug addiction, enabling the development of a predictive model.

Gnanasekar and Yanushkevich (2019) presented "Face Attributes and Detection of Drug Addicts." The study proposed a two-stage approach for detecting drug addicts using "soft" face biometrics and machine learning algorithms. The strategy identifies "drug-affected facial attributes" and uses probabilistic reasoning to detect drug addicts. The study used pre-trained CNNs (GoogleNet, ResNet50, and VGG16) for face attribute classification, followed by dimensionality reduction and feature selection using PCA and Fisher's linear discriminant. The accuracy of the model was 90% using ResNet50. A Bayesian network was also created to classify subjects as drug addicts, achieving an accuracy of 84%. The result affirmed that the approach offers a novel method for detecting drug addicts using face attributes and machine learning algorithms.

Wang et al. (2018) investigated "Discrimination of Smoking Status by MRI Based on Deep Learning Method." The study utilised 3D-T1WI images from 127 subjects (61 smokers and 66 non-smokers) and developed two deep learning models: Conv3D and ConvLSTM. The results showed that Conv3D achieved an accuracy of 80.6%, a sensitivity of 80.0%, and a specificity of 81.3%, while ConvLSTM attained an accuracy of 93.5%, a sensitivity of 93.33%, and a specificity of 93.75%. These results outperformed the support vector machine (SVM) algorithm, which achieved less than 70% accuracy. The study concluded that deep learning-based MRI can accurately predict smoking status and suggests that larger sample sizes are needed to improve accuracy and predict nicotine dependence levels.

While the reviewed studies demonstrated high accuracy rates, machine learning has shown promising results in addressing drug addiction and related healthcare issues. Further research is needed to improve the accuracy and generalisation of machine learning models in real-world scenarios.

3. Material and Method

The section presents the set of tools and techniques used to develop the proposed model. The stages include data preparation, pretrained model description, facial analysis classification, feature selection, model development and evaluation. Each section is discussed as follows:

Data Acquisition

The dataset consists of 400 facial images, with 240 addicted images and 160 non-addicted images (https://data.mendeley.com/datasets/ywgv7mxzn4/1/files/38a3e93c-840d-4366-8b4d-a5483d2977c7). The images underwent pre-processing to enhance their quality and consistency. Facial landmark detection is employed for alignment, and ImageDataGenerator() is utilised for normalisation, resizing images to 224x224 pixels. This step is crucial in ensuring that the images are in a suitable format for the models to extract meaningful features and be trained by the images.

Pre-trained Majority Voting Classifier

This study leverages the capabilities of pre-trained models, specifically ResNet50, VGG16, and InceptionV3, to extract features from the facial images. The pre-trained models were already trained on large datasets and have learned to recognise general features that apply to various image classification tasks. Using pre-trained models, this study can tap into their knowledge and fine-tune them for our specific task of drug addiction classification. The proposed classifier predicts the class labels (addicted or non-addicted) based on the features extracted from the pre-trained models. This approach allows the combination of the predictions of multiple models and the realisation of a more informed decision.

Facial Analysis Classification

Each base learner is trained on the pre-processed dataset, and their predictions are combined using a majority voting classifier to determine the final class label. Facial analysis classification involves analysing facial features to identify patterns and correlations that can indicate drug addiction. Using an ensemble learning model, the proposed model can automatically extract relevant features from the facial images and classify them into either the addicted or non-addicted category. The facial analysis classification processes are represented as follows:

Equation 3.1

$$F = \{f1, f2, \ldots, fn\}$$

Where:

- F: set of facial features

- fi: individual facial feature (e.g. eye shape, nose size, etc.)

Each base model (ResNet50, VGG16, and InceptionV3) is trained to predict the class label (addicted or non-addicted) based on the facial features:

Equation 3.2

$$y1 = f1(F) = \text{argmax}(Ci \mid F)$$

Equation 3.3

$$y2 = f2(F) = \text{argmax}(Ci \mid F)$$

Equation 3.4

$$y3 = f3(F) = \text{argmax}(Ci \mid F)$$

Where:

- yi: predicted class label from each base model

- Ci: class label (addicted or non-addicted)

- f: facial analysis classification model

The predictions from each base model (Equations 3.2 to 3.4) are then combined using a majority voting classifier:

Equation 3.5

$$y = \text{mode}(y1, y2, y3)$$

Where the function mode() represents the majority vote.

The facial analysis classification process (Equation 3.1) is then represented as follows:

Equation 3.6

$$F \rightarrow [f1, f2, \ldots, fn] \rightarrow [y1, y2, y3] \rightarrow y$$

Where:

- F: input facial features

- [f1, f2, ...,fn]: extracted facial features

- [y1, y2, y3]: predictions from each base model

- y: final predicted class label

By using ensemble learning model for facial analysis classification, automatic extraction of relevant features from the facial images and classification into either the addicted or non-addicted category are feasible.

Feature Selection

The pre-trained models extract features from the facial images, which are then used for classification. Feature selection is an essential step in this process, as it allows identification of the most relevant features that contribute to accurate classification. By selecting the most informative features, dimensionality of the feature can be reduced thereby leading to an improved model's performance.

Model Development

The Majority Voting classifier predicts the class labels (addicted or non-addicted) based on the features extracted from the pre-trained models. The proposed model's ability to classify individuals into either the addicted or non-addicted category can provide timely interventions. The drug classification task can be represented using the following equation:

Equation 3.7

$$y = f(X) = \text{argmax}(Ci \mid X)$$

Where:

- y: predicted class label (addicted or non-addicted)

- X: input features extracted from the pre-trained models

- Ci: class label (addicted or non-addicted)

- f: Majority Voting classifier

The Majority Voting classifier combines the predictions from multiple base models (ResNet50, VGG16, and InceptionV3) to make a final prediction.

Model Evaluation

The performance of each base model (ResNet50, VGG16, and InceptionV3) is evaluated using loss (categorical cross-entropy) and accuracy (percentage of correctly classified images). These evaluation metrics provide insights into the models' performance and facilitate comparison. The dataset is split into training (80%) and testing (20%) sets to evaluate the performance of each base model and the Majority Voting classifier. The models are trained on the training set, and their performances are assessed on the testing set. The results are reported in terms of accuracy and loss, enabling a comprehensive evaluation of the proposed approach.

Equations:

- Loss (categorical cross-entropy):

Equation 3.8

$$L(y, y') = -\sum(y * \log(y') + (1 - y) * \log(1 - y'))$$

- Accuracy:

Equation 3.9

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- y: true label

- y': predicted label

- TP: true positives

- TN: true negatives

- FP: false positives

- FN: false negatives

4. Result and Discussion

Table 4.1: Model Result

Dataset Split	Loss	Accuracy Rate
Training Set	0.6573	69%
Validation Set	2.8133	60%

Table 4.1 reveals that the model's result on the training set is reasonable, with an accuracy rate of 69%. However, the relatively high loss value of 0.6573 demonstrates that the model may have struggled to capture the underlying patterns in the data.

On the validation set, the model's performance is less impressive, with an accuracy rate of 60% and a significantly higher loss value of 2.8133. This indicates that the model may have overfitted to the training data, failing to generalise well to unseen data.

Despite the model's performance being unsatisfactory, this study's findings may still provide a useful foundation for future research. However, further improvements are necessary to enhance the model's generalisation abilities and accuracy.

Models Comparison

The aforementioned results were compared with existing studies as shown in Table 4.2.

Table 4.2: Comparison of the results of this study with the previous studies

Study	Methodology	Accuracy Rate
Kumara et al. (2023)	Stacking (CNN, RF, DT)	96.75%
Gu et al. (2021)	Ensemble Learning (bi-modal signals)	63.15%
Castro et al. (2021)	Ensemble Learning (3D printed drug delivery systems)	93%
Current Study	Majority Voting Classifier (facial images)	69% (training set), 60% (validation set)

Table 4.2 provides a concise comparison of the accuracy rates achieved by various existing studies with the present study, highlighting the differences in methodologies and performance.

Given the previous studies, a 60% validation accuracy can be considered suboptimal because it is lower than the accuracy rates reported for

the existing studies, which range from 63.15% to 96.75%. Moreover, a 2.8133 loss is relatively high, demonstrating the need to further optimise the model.

Model Testing

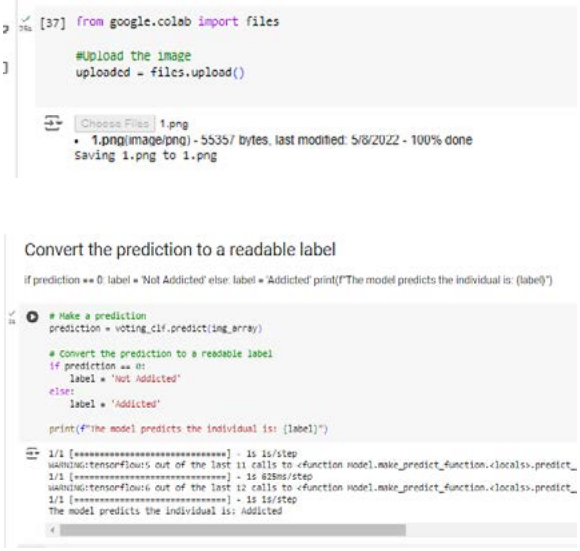

Figure 4.1: Model Testing

Figure 4.1 illustrates the Google Colab interface, where the model is developed, tested, and validated using novel input data.

5. Conclusion and Recommendation

This study develops a facial image-based addiction detection system using an ensemble learning technique called the Majority Voting Classifier. The model's potential drawbacks, owing to overfitting, highlight the need for further refinement. Hence, future work can explore techniques like hyperparameter tuning, data augmentation, and feature engineering to enhance the model's performance. Above all, this research contributes to the growing field of facial image analysis for healthcare applications, demonstrating the potential of machine learning to drive innovation in this area.

References

Abada, W., & Bouramoul, A. (2022). Using Machine Learning Techniques to Predict People At-Risk for Drug Addiction: A Bayesian-Based Model. *In 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)* (pp. 1-7). IEEE.

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, *4*(1), 19.

Arif, M. A. I., Sany, S. I., Sharmin, F., Rahman, M. S., & Habib, M. T. (2021). Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population. *International Journal of Electrical and Computer Engineering*, *11*(5), 4471.

Castro, B. M., Elbadawi, M., Ong, J. J., Pollard, T., Song, Z., Gaisford, S., & Goyanes, A. (2021). Machine learning predicts the 3D printing performance of over 900 drug delivery systems. *Journal of Controlled Release*, *337*, 530-545.

Choi, J., Jung, H. T., Ferrell, A., Woo, S., & Haddad, L. (2021). Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (hookah) users—*Journal of Clinical Medicine*, *10*(5), 972.

Deng, Y., Xu, X., Qiu, Y., Xia, J., Zhang, W., & Liu, S. (2020). A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics*, *36*(15), 4316-4322.

Faisal, F., Shahriar, A., Mahmud, S. U., & Shuvo, R. A. (2019). *A supervised machine learning approach to predict vulnerability to drug addiction* (Doctoral dissertation, Brac University).

Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Re-

gep, C., Liu, G., & Taylor-King, J. P. (2021). Utilising graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, *22*(6), bbab159.

Gnanasekar, S. T., & Yanushkevich, S. (2019). Face Attributes and Detection of Drug Addicts. In *2019 Eighth International Conference on Emerging Security Technologies (EST)* (pp. 1-6). IEEE.

Gu, X., Yang, B., Gao, S., Yan, L. F., Xu, D., & Wang, W. (2021). Application of bi-modal signal in the classification and recognition of drug addiction degree based on machine learning. *Math. Biosci. Eng*, *18*(5), 6926-6940.

Ilyas, A. (2023). What is Hard and Soft Voting in Machine Learning? *Medium.* Available at *https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32*

Kumara, U. G. H. T., Siriwardana, S. S. A., Weerasinghe, L., Shavindi, R. A. K. I., Chiranjeewa, H. P. R. C., & Siriwardana, S. (2023). A Machine Learning Approach to Analyse the Drug Addiction. In *2023 5th International Conference on Advancements in Computing (ICAC)* (pp. 113-118). IEEE.

Kumari, D., & Swetapadma, A. (2023). Building a Heroin Abuse Prediction Model: A Generalized Machine Learning Approach. In *International Conference on Recent Trends in Image Processing and Pattern Recognition* (pp. 3-13). Cham: Springer Nature Switzerland.

Lekkas, D., Klein, R. J., & Jacobson, N. C. (2021). Predicting acute suicidal ideation on Instagram using ensemble machine learning models. *Internet interventions*, *25*, 100424.

Ripperger, M., Lotspeich, S. C., Wilimitis, D., Fry, C. E., Roberts, A., Lenert, M., Cherry, C., Latham, S., Robinson, K., Chen, Q., McPheeters, M. L., Tyndall, B., & Walsh, C. G. (2022). Ensemble learning to predict opioid-related overdose using statewide prescription drug monitoring program and hospital discharge data in the state of Tennessee. *Journal of the American Medical Informatics Association*, *29*(1), 22-32.

Wang, S., Zhang, R., Deng, Y., Chen, K., Xiao, D., Peng, P., & Jiang, T. (2018). Discrimination of smoking status by MRI based on deep learning method. *Quantitative Imaging in Medicine and Surgery*, *8*(11), 1113.