Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 7, No. 2, May, Pp. 137-148

XGB-hybrid fingerprint classification model for virtual screening of meningitis drug compounds candidate

Mohammad Hamim Zajuli Al Faroby*[1], Helisyah Nur Fadhilah[2], Siti Amiroch[3], Rahmat Sigit Hidayat[4]
Department of Data Science, Faculty of Information Techonolgy and business, Institut Teknologi Telkom Surabaya, Indonesia[1,2,4]
Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Islam Darul 'ulum Lamongan, Indonesia[3]

Article Info
Keywords:
Drug Screening, Molecular Fingerprint, Extreme Gradient Boosting, Machine Learning, Meningitis

Article history:
Received: March 17, 2022
Accepted: April 18, 2022
Published: May 31, 2022

Cite:
M. H. Z. Al Faroby, H. N. Fadhilah, S. Amiroch, and R. S. Hidayat, "XGB-Hybrid Fingerprint Classification Model for Virtual Screening of Meningitis Drug Compounds Candidate", KINETIK, vol. 7, no. 2, May. 2022. https://doi.org/10.22219/kinetik.v7i2.1424

*Corresponding author.
Mohammad Hamim Zajuli Al Faroby
E-mail address:
alfaroby@ittelkom-sby.ac.id

Abstract
Meningitis is an infection of the lining of the brain caused by diffuse inflammation, and this condition is caused by viruses or bacteria that cause Meningitis. The mortality rate of untreated disease due to meningitis bacteria approached 100%, even though with special therapy the mortality ratio is only slightly reduced. The prevention of this disease is still strengthening antibodies with vaccines. The primary treatment for meningitis is with antibodies and anti-inflammatory drugs to relieve pain. However, drug candidates for inhibiting target protein still have not found optimal results in reducing mortality ration from meningitis. In a previous study by Yuan Nong et al, they got seven important proteins for Meningitis. We continue to investigate compounds associated with seven proteins that may be able to bind and inhibit them. We chose the in-silico process by utilizing data in an open database. We use several databases for the data collection process. After that, the compound data were extracted for bonding features and chemical elements using molecular fingerprints. We use two fingerprint methods, where both we combine with three types of combinations. The combined results produce three types of datasets with different matrix sizes. We establish the Extreme Gradient Boosting (XGB) method to form the classification model for the three datasets, select the best classification model, and compare it with other classification algorithms. The XGB model has better quality than the classification model of other algorithms. We used this model to predict and quantify compounds that strongly bind to seven vital meningitis proteins. The compound with the highest predictive score (we found more than 0.99) became a drug candidate to inhibit or neutralize Meningitis.

1. Introduction

Infection of the brain and spinal cord lining caused by diffuse inflammation is a severe disease. This disease attacks the lining of the host's brain, so patients experience clinical symptoms of headaches, nausea, and fever [1]. In medical terms, this disease is Meningitis. A viral or bacterial infection usually causes Meningitis. The development of Meningitis is quite fast, even untreated diseases cause death with a percentage approach to 100% [2]. In the acute stage, its development will progress in hours or days [3]. The prevention of this disease is still strengthening antibodies with vaccines. The primary treatment for meningitis is with antibodies and anti-inflammatory drugs to relieve pain. However, drug candidates for inhibiting target protein still have not found optimal results in reducing mortality ration from meningitis [4]. The pathogenetic causes of meningitis have not been fully elucidated so far. Therefore, the current anti-meningitis strategy needs to be explored further by finding potential active compounds that can bind and inhibit target proteins.

Potential active compounds require analysis of target proteins that are significant to the biological processes of viruses and bacteria that cause meningitis. From a previous study by Yuan Yong et al, EGFR, TNF, EGF, ATM, ESR1, CASP8, and NGF are vital pharmacological protein targets for meningitis that were analyzed using protein interactions (PPI) analysis [5]. The seven target proteins have several clusters of active compounds that can prevent the development of viral and bacterial meningitis (anti-meningitis). However, screening active compounds against the seven target proteins considered vital targets has not been thoroughly studied. Screening for active ingredients will take a long time when analyzing all possible compounds that can bind [6]. Thus, a more efficient data analysis method is needed to find possible compounds suitable for the seven target proteins.

The role of computing in data analysis can help trace some of these active compounds. The application of machine learning algorithms to analyze protein interaction networks [7] and searching for active compounds that can bind and inhibit proteins provides a lower cost than directly analyzing biological objects [8]. In this study, the model made from data on compounds that have the potential effect/efficacy to target the proteins is classified using a machine

learning model. The dataset features used to build the model, using the fingerprint feature of the chemical bond. There are approximately 881 chemical bonding features on the PubChem fingerprint [9] and approximately 4860 features on the Klekota-Roth fingerprint [10]. The two fingerprints are hybrid to enrich the data features.

The study aimed to obtain compounds that can be inhibitors of meningitis pathogens by analyzing seven vital protein targets. This research is needed because the need for meningitis drugs has not been fully efficient against this disease. Several vaccines are available but have not been able to handle these pathogens optimally. Because vaccines only provide temporary immunity to the human body against pathogens (bacteria or viruses) that have been previously recognized by vaccines [11]. However, the active compound that has the potential to be a drug can be an intermediary in healing meningitis. These compounds are capable of being inhibitors or affected/efficacy for the curative treatment of seven target proteins.

2. Research Method

In silico drug screening has two approaches with which several studies have been successful. The first approach is with molecular docking [12][13]; the analysis of ligand docking to the target protein can find compounds suitable for the protein receptor inhibition. The second approach is to create a predictive model using machine learning methods from past data. The second approach is to create a predictive model using machine learning methods from past data. This approach gives the characteristics of the biochemical compound data in the form of smiles code [14], which are related to the character of the target protein receptor. The approach is to design this drug using the Quantitative Structure-Activity Relationship (QSAR) method with classification prediction methods such as Artificial Neural Network (ANN) [15], Support Vector Machine (SVM) [16], Random Forest [17], to Deep Learning[18].

Research conducted by Yuan Nong et al. at the end of 2020 stated that there were seven vital meningitis target proteins with Calycosin compounds, namely Epidermal Growth Factor Receptor (EGFR), Tumor Necrosis Factor (TNF), Epidermal Growth Factor (EGF), Ataxia Telangiectasia Mutated Protein (ATM), Estrogen Receptor Alpha (ESR1), Caspase-8 (CASP8), and Nerve Growth Factor (NGF) [5]. By using the QSAR method to analyze candidate active compounds, this study is expected to predict other active compounds besides Calycosin, which can be used as inhibitors of Meningitis, to produce an effective drug design.

2.1 Data Acquisition: Drug Compound

In this study, the target protein data used the results of previous studies in the form of seven vital proteins in the Meningitis [5]. To obtain data on compounds associated with these seven vital proteins, we acquired data from several online databases such as DUDE: Database[19], CAS A Division of the American Chemical Society, and PubChem. This study also enriched data on related ligand compounds in ChEMBL [20] and the Super Target and Matador database [21]. The process of acquiring this data obtained as many as 1146 compounds related to seven target proteins. Details of the compound's relationship with its target protein are given in Table 1.

Table 1. The number of ligands that we found in several databases that can bind to the seven vital meningitis proteins

Protein Targets	DUDE/CAS/PubChem	ChEMBL/Super Target
ATM	0	14
CASP-8	0	169
EGF	0	16
EGFR	542	0
ESR1	383	0
NGF	0	1
TNF	0	21

To form a classification model, we needed data on compounds not associated with the seven target proteins. These data are called data decoys. We get the decoy data from the DUDE database. We generate data from DUDE and generate 35050 data.

2.2 Molecular Fingerprint

Machine learning-based drug screening research uses a molecular fingerprint to extract features from a chemical compound. There are two standard molecular fingerprints found in RDKit, the first is PubChem which has 881 molecular features [22][23], and the second standard fingerprint is Klekota-Roth which has 4860 molecular features [24].

The feature extraction process records the structural bonds of molecules that are identified as having certain structural bonds. In the example in Figure 1, the molecular structure in Figure 1 has ring bonds (in the image above), then the substructure that is a feature will be coded with 1. If the compound data does not have a chemical bond in the substructure that is a crucial feature, then the feature of the substructure is coded 0.

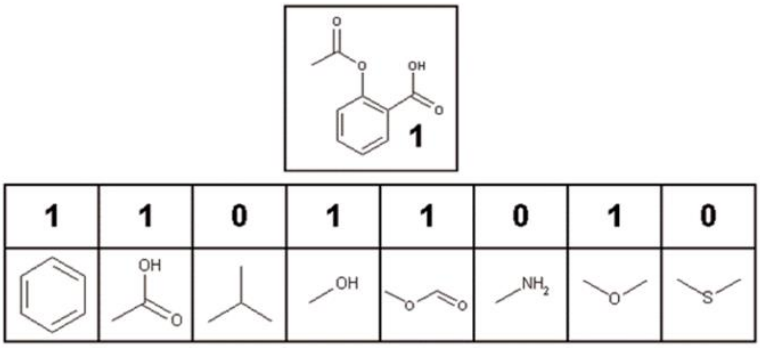


Figure 1. Illustration of Feature Extraction from a Compound using a Molecular Fingerprint

2.3 Extreme Gradient Boosting (XGB)

The Extreme Gradient Boosting (XGB) method is a reinforced ensemble tree. Each generated tree strengthens the classification model on the previous trees. This method is the sum of the lead weights on the generated ensemble tree. In Equation 1, suppose we have a dataset with n samples and feature dimensions m with $\psi = (x_i, y_i)$ for $\|\psi\| = n, x_i \in R^m$. An ensemble model formed from the dataset using the additive function K, where K is the number of trees generated to predict the result [25],

$$\hat{y}_i = \phi(x_i) = \sum_{i=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{1}$$

$\mathcal{F} = \{f(x) = v_q(x)\}$ is the domain space which represents the classification tree f_k, for $q: R^m \to T, v \in R^T$ and the structure of the classification tree is represented in q, which is a function of mapping the data to the corresponding leaf T. The f_k tree corresponds to independent structure q and leaf weight v. Each resurrected tree has a specific score on each leaf, representing the i leaf score. For example, we want to get the predictive value of data; the classification tree calculates the features corresponding to the leaf and calculates a predictive score (in terms of probability) by adding up all the corresponding leaves on the data in each generated tree.

Learning algorithms on data always have an objective function. The XGB classification model uses an objective function to minimize $\mathcal{L}$ [25],

$$\mathcal{L}(x) = \sum_{i=1} l(y_i, \hat{y}_i) + \sum_{K=1} \Omega(f_k) \tag{2}$$

for,

$$\Omega(x) = \gamma T + \frac{1}{2}\lambda\|v\|^2$$

l is a differentiable function that can distinguish between predicted $\hat{y}_i$ and target y_i. At the same time, Ω is a function that determines the risk taken from the complexity of the model. The loss function for the classification model uses the logistic regression formula called $log(likelihood)$. In contrast, it uses the mean squared for the regression. Equation (2 is generally a function of the parameters in the population area, so the optimization method in Euclid space cannot optimize it. To overcome these conditions, the model algorithm is trained additively. Let us say that the prediction of the ith event in iteration t is $\hat{y}_i(t)$, so it takes f_t to produce a minimal objective function in the next iteration,

$$\mathcal{L}^{(t)} = \sum_{i=1} (l(y_i, \hat{y}_i^{t-1}) + f_t(x_i)) + \Omega(f_t) \tag{3}$$

To estimate the Equation 4, a stochastic approach by utilizing the second-order Taylor expansion [26]. Equation (3 is extended to,

$$\mathcal{L}^{(t)} \cong \sum_{i=1}^{n} \left(l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_t) \tag{4}$$

Where, $g_i = \frac{\partial\, l(y_i, \hat{y}_i^{t-1})}{\partial, \hat{y}_i^{t-1}}$ and $h_i = \frac{\partial^2\, l(y_i, \hat{y}_i^{t-1})}{\partial, \hat{y}_i^{t-1^2}}$. g is the Gradient derived from the first derivative of the loss function. In contrast, h is called Hessian, the second derivative of the loss function used in the classification model. In the random forest algorithm, the loss function divides the features on the acquisition of essential information. It randomly combines the trees, and then the predicted value is based on the vote of each generated tree [27]. While XGB changes the loss function into a new function where each tree where each new tree reinforces the previous classification tree to choose the best threshold,

$$\mathcal{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\left(\sum_{i\in I_j} h_i + \lambda\right)} + \gamma T \tag{5}$$

$\mathcal{L}^{(t)}(q)$ is a scoring function to measure the quality of the tree structure q in the t iteration. In contrast, g and h are the first and second derivatives of the loss function, respectively. I_j is the instance set of a particular leaf node j. In this case, XGB may iteratively reduce loss and outperform other ensemble approaches.

Predictive value of tree leaves based on the optimal weight of the loss function. To get the optimal value of an equation, you can find the critical point of an equation by finding $f'(x) = 0$, as a result, the predicted score is,

$$v_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{6}$$

We have Equations *(5* and Equations (*6*, which can be used as extraction functions to measure the quality of the tree structure q. Listing all the possible trees formed is impossible because of the large number of compositions. So, this method uses a greedy algorithm by starting at a single leaf (later as a root) and adding tree branches from the w_j^* calculation. After splitting, let I_L and I_R be the sample sets for the left and right sides. Given $I = I_L \cup I_R$, the root of the tree before adding children, the loss reduction Equation 7 after separation is as follows.

$$\mathcal{L}_{gain} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\left(\sum_{i\in I_L} h_i + \lambda\right)} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\left(\sum_{i\in I_R} h_i + \lambda\right)} + \frac{(\sum_{i\in I} g_i)^2}{(\sum_{i\in I} h_i + \lambda)}\right] \tag{7}$$

The $\mathcal{L}_{gain}$ equation is like the entropy of a decision tree but differs in the rules for separating it. Equation 7, $\mathcal{L}_{gain}$ can ignore the coefficient because it is considered a multiplier of $\mathcal{L}_{gain}$.

2.4 Methodology

The methodology covers the initial research process from how to obtain data to the results of obtaining a candidate for meningitis inhibitor compounds. The methodological stages in the outline include starting from data acquisition. After that, clean and organize the data to be processed or as supporting information. The following process extracts feature data in the form of compound smiles code with a molecular fingerprint. After the features are obtained, the two fingerprints used are combined to build a hybrid fingerprint dataset. After that, the dataset is used to train the XGB algorithm. An algorithm optimization process is needed to get a good classification model by getting the appropriate parameters from the data conditions. After that, test the classification model and get the test matrix value to determine how good the model is. The finished classification model is used to predict candidate drug compounds.

3. Results and Discussion

This section describes the research results on screening candidate drug compounds for meningitis. The sub-section discusses how to construct datasets, model results, and results of compounds with a significant predictive score that are candidates for meningitis drug compounds.

3.1 Dataset Cunstruction

This study used secondary data from protein and ligand databases from the DUDE: Database [28], PubChem Database [29], ChEMBL Database [30] and Super Target database[31]. The search results for seven vital target proteins we get several ligands associated with these target proteins, the total target protein results from online database searches are listed in Table 2 below.

Table 2. The Number of Active Compound (Ligans) Data On Vital Target Proteins

Protein Codes	Protein Names	Number of Ligands
ATM	Serine Protein Kinase ATM	14
CAPSE8	Truncated pro-caspase 8 (amino acids 213−496)	169
EGF	Epidermal Growth Factor	16
EGFR	Epidermal Growth Factor Receptor	542
ESR1	Estrogen Receptor 1	383
NGF	*Nerve growth factor*	1
TNF	Tumor necrosis factor	21

In addition to the 1146 ligand data, we initialized the dummy compound in the DUDE: Database for 35050 data. This data decoy as an inactive compound against all existing target proteins. However, not all data decoys are used to form the dataset. This study randomly selected decoy data to balance and enrich the data used [32]. The election results leave as many as 6608 decoy data which means about 5% of the total previous data. The sample size of the active compound and decoy looks unbalanced. So, to balance the ligand data and decoy data, this study applies the random oversampling method [33]. This method is helpful for randomly duplicating ligand data to balance the data decoy. In the end, the random over-sampling dataset resulted in 13216 data consisting of 6608 ligand data and 6608 decoy data.

3.2 Feature Extraction: Hybrid Molecular Fingerprint

The feature extraction algorithm reads smiles from compounds on the dataset. The feature extraction results using the PubChem method produce a data matrix (rows) with features (columns) measuring 13216×881. Meanwhile, the result of feature extraction using the Klekota-Roth method produces a matrix of 13216×4860 size. The dataset resulting from feature extraction with a molecular fingerprint is binary, where 1 means the data feature (chemical bond or chemical element) in the compound is detected. While 0 means not detected. An example of a complete dataset is displayed with the fingerprint feature in Figure 1.

Table 3. Hybrid Dataset Information

Hybrid Fingerprint	Description	Dataset Matrix
Full Combining	Combination of: PubChem features: 1-881 Klekota-Roth features: 1-4860	13216×5741
Initial Combining	Combination of: PubChem features: 1-441 Klekota-Roth features: 1-4860	13216×5301
Latter Combining	Combination of: PubChem features: 1-441 Klekota-Roth features: 2872-4860	13216×2871

After extracting the features of these compounds, the following process is to form a hybrid from the results of the two extraction methods used. The first hybrid is called Full Combining; this process combines all the features extracted from PubChem and Klekota-Roth. The result of combining all these features produces a matrix measuring 13216×5741. The second hybrid is called Initial Combining; This process combines features 1-441 of the PubChem extraction with all the features of the Klekota-Roth yield. Combining the front features from the PubChem results and all the Klekota-Roth features results in a matrix measuring 13216×5301. The last hybrid is called Latter Combining, where this process combines the first half of the PubChem feature extraction and the last half of the Klekota-Roth feature extraction. The results of this third combination form a dataset matrix measuring 13216×2871 [12]. Detailed information about the hybrid fingerprint dataset is in Table 3.

3.3 Classification and Optimization Model

We prepared three datasets to form a classification model of meningitis inhibitory compounds. Before training the data against the XGB algorithm, we divided the data into training data of 70% of the total data of each dataset and 30% as test data. This division aims to be a data builder model and a data testing model. Each dataset forms one classification model, thus forming three XGB models. The settings for the XGB parameters are same, with a learning rate of 0.15, the number of trees generated is 99, the gamma is 0.4, and the maximum depth of each tree is 16. The XGB parameters were obtained based on trial and error on the model classification. The purpose of trial and error is to optimize the model conditions so that the cost of the computational process (both in the form of time and memory speed)

is not too high. Figure 2 shows the loss graph that the model generates each time it builds a new tree. The graph shows the loss conditions generated by the model every time it generates a smaller booster tree, and the graph shows constant conditions. The differences in circumstances between the losses in the training and test processes demonstrate that the classification model is not underfitting or overfitting [34]. Conditions like this show the classification model in training and predicting the data is running well.

In addition to looking at the condition of the model losses, looking at the condition of the classification error graph can also be considered to optimize the model. We present the classification error graph in Figure 3; the classification error is calculated based on data that does not match the class [35]. The graph shows the classification error that occurs in each booster tree that is generated. The comparison of the conditions of the former model shows that the ROC value for the XGB-Initial Combination model is better than the XGB model with other Hybrid data. The ROC score for the XGB-Initial Combination model is 0.9958, higher than the ROC score for the XGB-Full Combination and XGB-Latter Combination models, which have scores of 0.9937 and 0.9932, respectively. From this comparison, it can be concluded that the model used to predict candidate meningitis inhibitor compounds is the XGB-Initial Combination model.

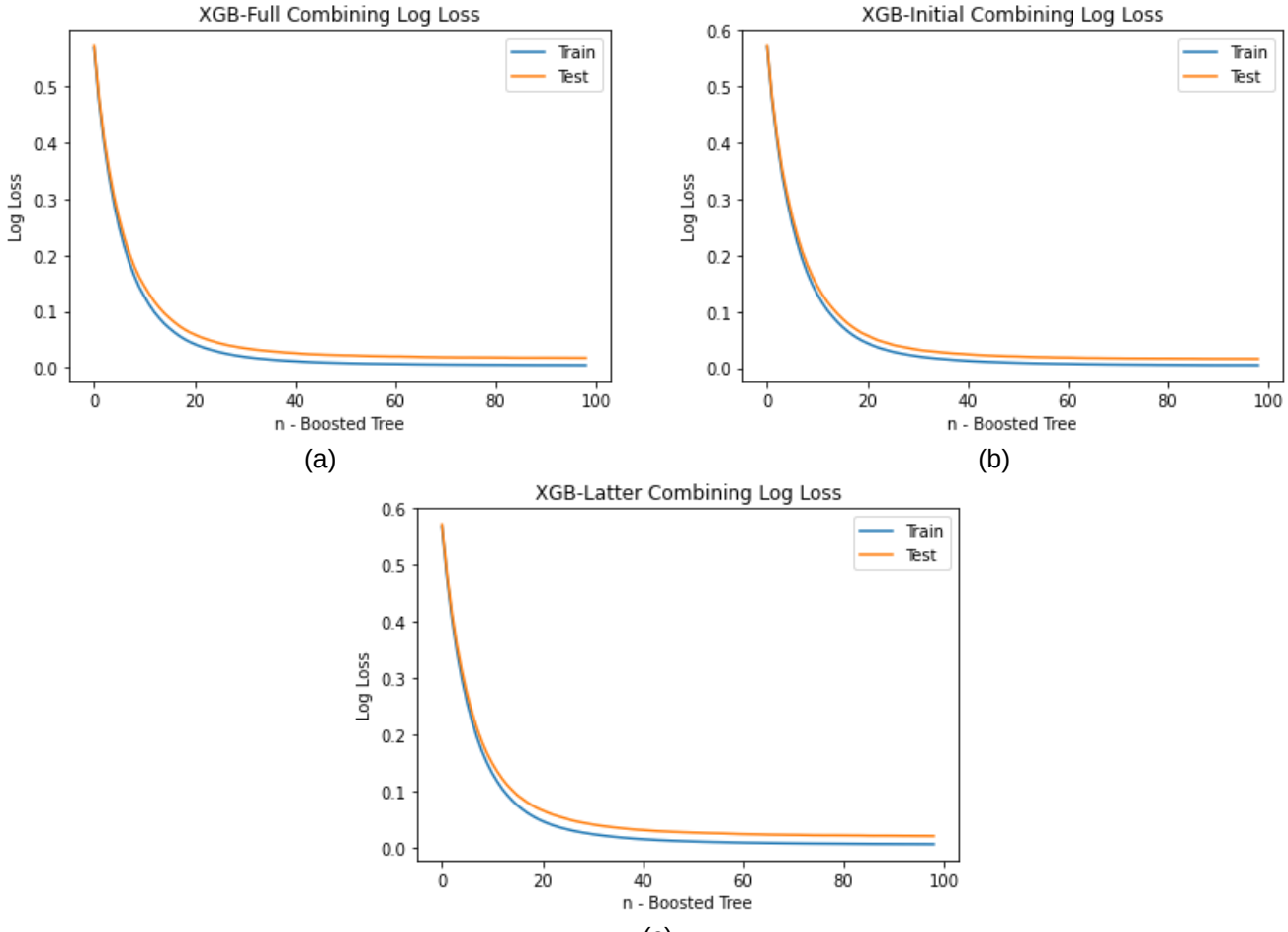


Figure 2. Comparison Graph of Model Performance to Optimize the Model Based on Losses Generated by the Model from the Objective Function for Each Generated Booster Tree. (a) Shows a Graph of the Loss Function of the Full Combination dataset, (b) Shows a Graph of the Loss Function of the Initial Combination Dataset, and (c) Shows a Graph of the Loss Function of the Latter Combination Dataset

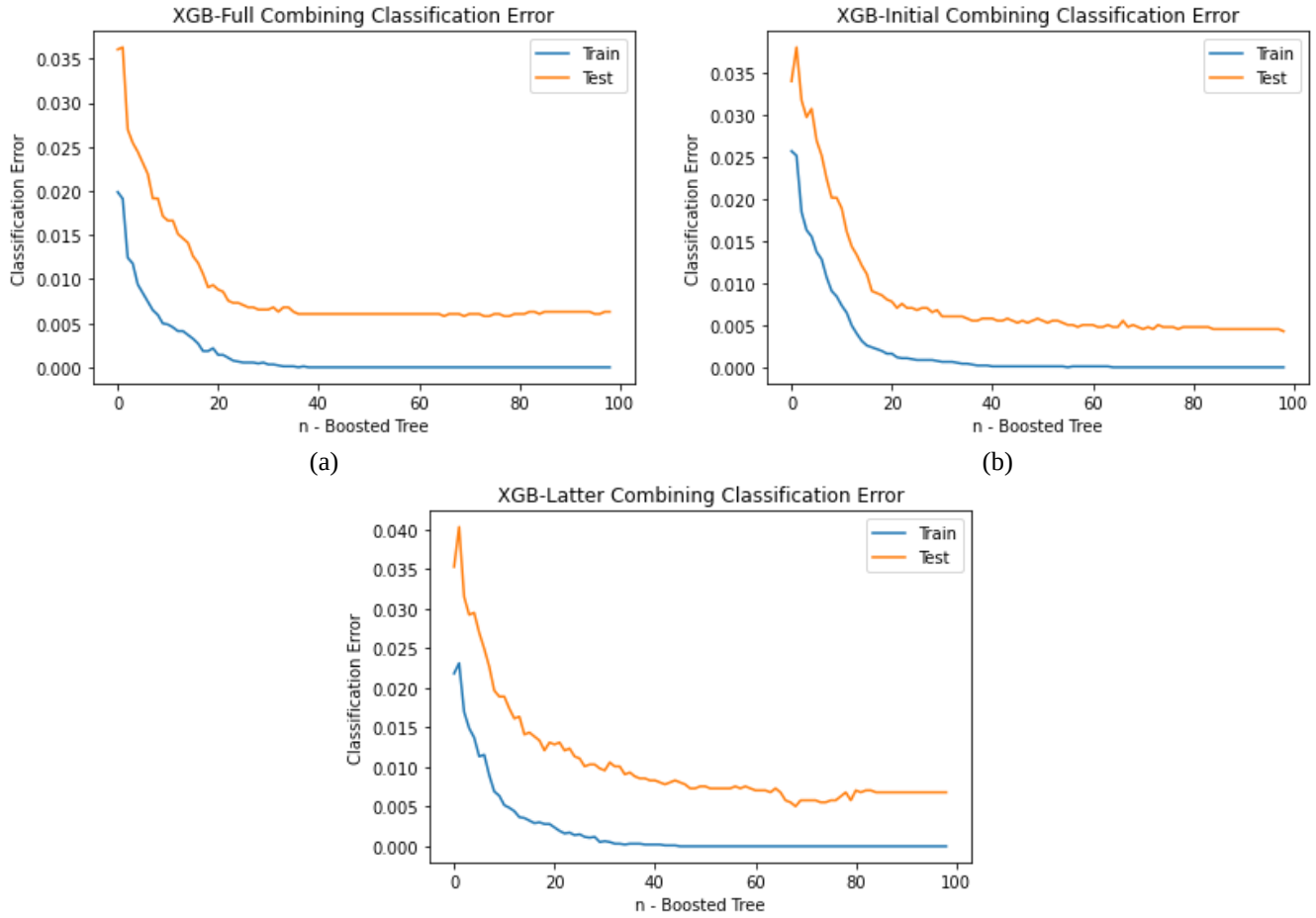


Figure 3. Comparison Graph of Model Performance to Optimize the Model Based on Model Misclassification for Each Generated Booster Tree. (a) Shows a Graph of the Full Combination Classification Error, (b) Shows a Graph of the Initial Combination Classification Error, and (c) Shows a Graph of the Latter Combination Classification Error

3.4 Comparison of Other Classification Models

The effectiveness of the classification model greatly affects the final day of the benefits of the built model. To add value to its effectiveness, we compared several classification models built with the Initial Combination dataset. This dataset was selected based on the previous results, which had better model performance than the other two datasets. We compared the XGB-Initial Combination model with several other classification methods. We compared it against other classification methods such as Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and Multilayer Perceptron. In addition, we also compared the results with other ensemble classification methods such as Random Forest and Gradient Boosting. All the results of the comparison of the quality of the model are shown in Table 4, which contains the ratio of accuracy, sensitivity, recall, and ROC.

Table 4. Comparison of Several Machine Learning Model Measurements to the Initial Combination Dataset Based on the Matrix of Accuracy, Sensitivity, Recall, and ROC

Classification Models	Accuracy	Sensitivity	Recall	ROC Score
Logistic Regression	0.8545	0.8711	0.8290	0.8542
K-nearest Neighbours	0.9624	0.9319	0.9969	0.9627
Support Vector Machine	0.9838	0.9822	0.9852	0.9838
Multilayer Perceptron	0.9838	0.9822	**1.0000**	0.9838
Random Forest	0.9089	0.9063	0.9104	0.9089
Gradient Boosting	0.9581	0.9549	0.9608	0.9582
Extreme Gradient Boosting	**0.9954**	**0.9909**	**1.0000**	**0.9955**

The XGB-Initial combination model has higher accuracy and ROC matrix than the other classification models. The accuracy matrix score of XGB-Initial Combination is 0.9954 and the ROC ratio is 0.9955. From the comparison in Table 4, we know that the measure of the quality of the XGB-Initial Combination model is on average above the other models for the four measurement matrices. As shown in Figure 4, the ROC score also shows that the model from the Extreme Gradient Boosting (XGB) method is superior to the model with other methods. The ROC score of the XGB-Initial Combination model is 0.9955. If we look further, the model produced by the Logistic regression method is the worst model, with a ROC score of 0.8542. This condition makes the Logistic Regression method less suitable if used to model binary data shapes with very large data feature dimensions. In contrast, the Support Vector Machine method has advantages in datasets with large feature dimensions. These reasons allow the SVM algorithm to produce a reasonably good model compared to other models except XGB.

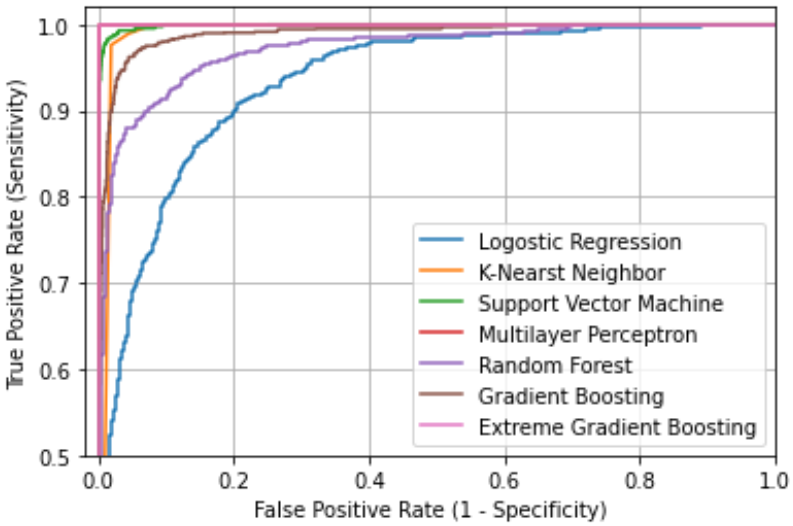
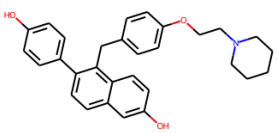
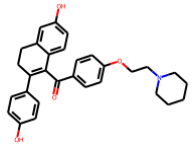


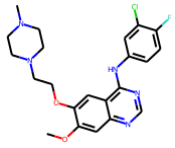
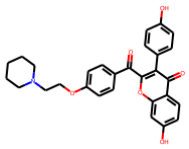
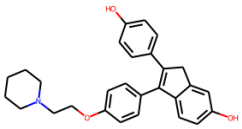
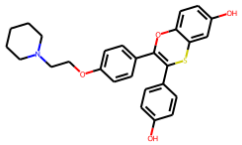
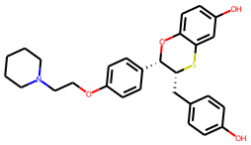
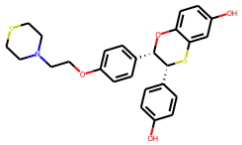
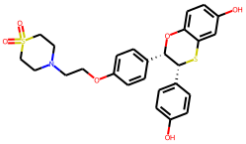
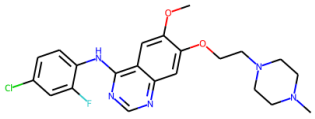
Figure 4. ROC Curve Showing Model Quality with Whole Combining Hybrid Dataset with XGB Compared to other machine Learning Models

3.5 Significant Compounds as Meningitis Inhibitors

After comparing the classification models from several methods to XGB-Initial Combination, we believe that the XGB-Initial Combination model is suitable for predicting ligand compounds with a large effect and can be used as meningitis inhibitors. We rank the ligand compounds and extract the features of these ligands according to the feature conditions used in the Initial Combination dataset. The ranking results of these ligands are presented in Table 5, where only the top ten rankings are presented. We present the ten ligands in 2D bonds and Smiles codes. These compounds are candidates for meningitis inhibitors based on data mining analysis in open databases. The top ten ligands have predictive scores on the classification model above 0.99. This value is quite good and means the compound is suitable for the seven target proteins that bind.

Table 5. Ten potential ligands to be Meningitis inhibitors according to XGB-Initial combination model.

Compound/ ChEMBL Codes	Smiles Codes	Compound 2D Structures
CHEMBL539849	Oc5ccc(c2ccc1cc(O)ccc1c2Cc4ccc(OCCN3CCCCC3)cc4)cc5	
CHEMBL176509	Oc5ccc(C4=C(C(=O)c2ccc(OCCN1CCCCC1)cc2)c3ccc(O)cc3CC4)cc5	

CHEMBL300791	COc3cc2ncnc(Nc1ccc(F)c(Cl)c1)c2cc3OCCN4CCN(C)CC4	
CHEMBL285483	Oc5ccc(c4c(C(=O)c2ccc(OCCN1CCCCC1)cc2)oc3cc(O)ccc3c4=O)cc5	
CHEMBL381645	Oc5ccc(C4=C(c2ccc(OCCN1CCCCC1)cc2)c3ccc(O)cc3C4)cc5	
CHEMBL418327	Oc5ccc(c2sc1cc(O)ccc1oc2c4ccc(OCCN3CCCCC3)cc4)cc5	
CHEMBL313941	Oc5ccc(C[C@H]2Sc1cc(O)ccc1O[C@H]2c4ccc(OCCN3CCCCC3)cc4)cc5	
CHEMBL362718	Oc5ccc([C@H]2Sc1cc(O)ccc1O[C@H]2c4ccc(OCCN3CCSCC3)cc4)cc5	
CHEMBL185383	Oc5ccc([C@H]2Sc1cc(O)ccc1O[C@H]2c4ccc(OCCN3CCS(=O)(=O)CC3)cc4)cc5	
CHEMBL283088	COc3cc2c(Nc1ccc(Cl)cc1F)ncnc2cc3OCCN4CCN(C)CC4	

It should be noted that the compounds above are still candidates and are included in the small molecule type. So, further research is needed on bonding these compounds to the protein-protein interactions that occur. Further research can continue molecular dynamic docking. Based on the reference classification model to predict these compounds, this condition has a very significant similarity to the compounds that become the data to train the model.

3.6 Discussion

The ligand compounds candidates for Meningitis inhibitors are random in an open database. Thus, these compounds have been verified based on research related to these compounds. However, determining which compound is more effective against Meningitis takes many times. Compounds consisting of various kinds of elemental bonds are very numerous; to examine one by one requires a very high cost. This classification model makes selecting compounds prioritized for in vitro tests easier. The selection of compounds based on the classification model is quite objective because the model built is derived from compounds that have been shown to affect seven vital meningitis pharmacological protein targets. Those ten candidates of drug compounds also consist of Polycyclic Aromatic Hydrocarbon (PAH), making them more lipophilic. It is important for drug delivery to the brain that needs to surpass the blood-brain barrier, which is a very lipophilic [36].

This research needs further activities to confirm the behavior of the compound towards the target protein and its protein-protein interactions. Suppose we take the example of the compound CHEMBL539849. This compound in the open database from previous research has not yet been named, and the behavior of the compound against the target protein is unknown. So, it is necessary to proceed to the simulation of molecular dynamic docking of the ligand to the

target protein. From molecular dynamic docking results, we can find out more clearly about the behavior of the ligand to its suitability to the target protein. Supposedly with features that are very similar to the ligands in the training data, the possibility of a match on the target protein is very large because the fingerprint's characteristics are the elements and bonds contained in the compound.

4. Conclusion

Initial Combination is a good hybrid method for building XGB datasets and Classification models. This model has an accuracy score of 0.9957 and a ROC of 0.9958 over models with other hybrid datasets. In addition, for binary data with high dimensions, the XGB method can model well compared to other classification methods, including SVM, which has advantages in high-dimensional datasets. The method with the worst quality results is the model produced by Logistic Regression-Initial Combination. This comparison adds to the confidence that the model built with XGB-Initial Combination has good quality for predicting meningitis-related compounds.

The results of the prediction of related compounds show that some compounds have very good prediction scores. This condition is proven by the prediction score in the top ten rankings above 0.99. The compound with code CHEMBL539849 was the most significant candidate for meningitis inhibitor. We also found that the ten compounds have a Polycyclic Aromatic Hydrocarbon structure. It is essential for drug delivery to the brain that needs to surpass the blood-brain barrier, which is very lipophilic. However, to find out the behavior of these compounds, which are recommendations for meningitis inhibitor candidates, we need to do further research by molecular dynamic docking these ligands to meningitis vital target proteins using 3D protein computations. After that, it is necessary to carry out an In Vitro test; if the results are as expected, the candidate compounds can be the latest drugs to prevent meningitis.

Acknowledgment

The research received support from the Institute for Research and Community Service, Telkom Institute of Technology, Surabaya, with Decree No. REK. 094/PNLT1/REK/III/2021. However, the source of funds in this study was obtained from research funds.

References

[1] A. Kohil, S. Jemmieh, M. K. Smatti, and H. M. Yassine, “Viral meningitis: an overview,” *Arch. Virol.*, vol. 166, no. 2, pp. 335–345, Jan. 2021. https://doi.org/10.1007/s00705-020-04891-1

[2] M. W. Tenforde *et al.*, “Mortality in adult patients with culture-positive and culture-negative meningitis in the Botswana national meningitis survey: a prevalent cohort study,” *Lancet Infect. Dis.*, vol. 19, no. 7, pp. 740–749, Jul. 2019. https://doi.org/10.1016/S1473-3099(19)30066-0

[3] T. A. Erickson *et al.*, “The Epidemiology of Meningitis in Infants under 90 Days of Age in a Large Pediatric Hospital,” *Microorganisms*, vol. 9, no. 3, p. 526, Mar. 2021, doi: 10.3390/MICROORGANISMS9030526.

[4] D. van de Beek *et al.*, “ESCMID guideline: diagnosis and treatment of acute bacterial meningitis,” *Clin. Microbiol. Infect.*, vol. 22 Suppl 3, pp. S37–S62, May 2016. https://doi.org/10.1016/j.cmi.2016.01.007

[5] Y. Nong, Y. Liang, X. Liang, Y. Li, and B. Yang, “Pharmacological targets and mechanisms of calycosin against meningitis,” *Aging (Albany. NY).*, vol. 12, no. 19, pp. 19468–19476, 2020. https://doi.org/10.18632/aging.103886

[6] T. Rogers *et al.*, “Impact of Antibiotic Therapy in the Microbiological Yield of Healthcare-Associated Ventriculitis and Meningitis,” *Open forum Infect. Dis.*, vol. 6, no. 3, Mar. 2019. https://doi.org/10.1093/ofid/ofz050.

[7] M. H. Z. Al Faroby, M. I. Irawan, and N. N. T. Puspaningsih, “Prediction insulin-protein interactions associated based on ontology genes using extreme gradient boosting and centrality method,” *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Contr*, vol. 4, no. 5, pp. 253–262, 2020. https://doi.org/10.22219/kinetik.v5i4.107

[8] T. B. Kimber, Y. Chen, and A. Volkamer, “Deep Learning in Virtual Screening: Recent Applications and Developments,” *Int. J. Mol. Sci.*, vol. 22, no. 9, p. 4435, Apr. 2021. https://doi.org/10.3390/ijms22094435

[9] Y. Liu *et al.*, “Machine Learning Models for the Classification of CK2 Natural Products Inhibitors with Molecular Fingerprint Descriptors,” *Processes*, vol. 9, no. 11, p. 2074, Nov. 2021. https://doi.org/10.3390/pr9112074

[10] N. R. Das, S. P. Mishra, and P. G. R. Achary, “Evaluation of molecular structure based descriptors for the prediction of pEC50(M) for the selective adenosine A2A Receptor,” *J. Mol. Struct.*, vol. 1232, p. 130080, May 2021. https://doi.org/10.1016/j.molstruc.2021.130080

[11] N. Principi and S. Esposito, “Bacterial meningitis: new treatment options to reduce the risk of brain damage,” *Expert Opin. Pharmacother.*, vol. 21, no. 1, pp. 97–105, Jan. 2019. https://doi.org/10.1080/14656566.2019.1685497

[12] S. Das, S. Sarmah, S. Lyndem, and A. Singha Roy, “An investigation into the identification of potential inhibitors of SARS-CoV-2 main protease using molecular docking study,” *J. Biomol. Struct. Dyn.*, vol. 39, no. 9, pp. 3347–3357, 2021. https://doi.org/10.1080/07391102.2020.1763201

[13] F. Fernando, M. I. Irawan, and A. Fadlan, “Bat Algorithm for Solving Molecular Docking of Alkaloid Compound SA2014 Towards Cyclin D1 Protein in Cancer,” *J. Phys. Conf. Ser.*, vol. 1366, no. 1, 2019. https://doi.org/10.1088/1742-6596/1366/1/012089

[14] S. Lim and Y. O. Lee, “Predicting chemical properties using self-attention multi-task learning based on SMILES representation,” in *Proceedings - International Conference on Pattern Recognition*, 2020, pp. 3146–3153. https://doi.org/10.1109/ICPR48806.2021.9412555

[15] L. Gentiluomo *et al.*, “Application of interpretable artificial neural networks to early monoclonal antibodies development,” *Eur. J. Pharm. Biopharm.*, vol. 141, pp. 81–89, Aug. 2019. https://doi.org/10.1016/j.ejpb.2019.05.017

[16] J. W. Liang, M. Y. Wang, S. Wang, S. L. Li, W. Q. Li, and F. H. Meng, “Identification of novel CDK2 inhibitors by a multistage virtual screening method based on SVM, pharmacophore and docking model,” *J. Enzyme Inhib. Med. Chem.*, vol. 35, no. 1, pp. 235–244, Jan. 2020. https://doi.org/10.1080/14756366.2019.1693702

[17] Y. Zhou *et al.*, “Quantitative Structure-Activity Relationship (QSAR) Model for the Severity Prediction of Drug-Induced Rhabdomyolysis by Using Random Forest,” *Chem. Res. Toxicol.*, vol. 34, no. 2, pp. 514–521, Feb. 2021. https://doi.org/10.1021/acs.chemrestox.0c00347

[18] C. Schneider, A. Buchanan, B. Taddese, and C. M. Deane, “DLAB: deep learning methods for structure-based virtual screening of antibodies,” *Bioinformatics*, vol. 38, no. 2, pp. 377–383, Jan. 2022. https://doi.org/10.1093/bioinformatics/btab660

[19] S. Pokhrel *et al.*, "Spike protein recognizer receptor ACE2 targeted identification of potential natural antiviral drug candidates against SARS-CoV-2," *Int. J. Biol. Macromol.*, vol. 191, pp. 1114–1125, Nov. 2021. https://doi.org/10.1016/j.ijbiomac.2021.09.146
[20] F. M. I. Hunter, A. P. Bento, N. Bosc, A. Gaulton, A. Hersey, and A. R. Leach, "Drug Safety Data Curation and Modeling in ChEMBL: Boxed Warnings and Withdrawn Drugs," *Chem. Res. Toxicol.*, vol. 34, no. 2, pp. 385–395, Feb. 2021. https://doi.org/10.1021/acs.chemrestox.0c00296
[21] K. Nandhini and G. V Sriramakrishnan, "A Review of Drug Target Interaction Prognostication Using Artificial Intelligence," Ann. Rom. Soc. Cell Biol., vol. 25, pp. 832–838, May 2021.
[22] M. D. M. Fernández-Arjona, J. M. Grondona, P. Fernández-Llebrez, and M. D. López-Ávalos, "Microglial activation by microbial neuraminidase through TLR2 and TLR4 receptors," *J. Neuroinflammation*, vol. 16, no. 1, 2019. https://doi.org/10.1186/s12974-019-1643-9
[23] A. Capecchi, M. Awale, D. Probst, and J. Reymond, "PubChem and ChEMBL beyond Lipinski," *Mol. Inform.*, vol. 38, no. 5, p. 1900016, May 2019. https://doi.org/10.1002/minf.201900016
[24] Y. Hua, Y. Shi, X. Cui, and X. Li, "In silico prediction of chemical-induced hematotoxicity with machine learning and deep learning methods," *Mol. Divers.*, vol. 25, no. 3, pp. 1585–1596, Aug. 2021. https://doi.org/10.3389%2Ffphar.2021.793332
[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-Augu, pp. 785–794. https://doi.org/10.1145/2939672.2939785
[26] X. Su and M. Bai, "Stochastic gradient boosting frequency-severity model of insurance claims," *PLoS One*, vol. 15, no. 8, p. e0238000, Aug. 2020. https://doi.org/10.1371/journal.pone.0238000
[27] S. Kabiraj *et al.*, "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm," Jul. 2020. https://doi.org/10.1109/ICCCNT49239.2020.9225451
[28] H. Kuswanto, R. Y. Nurhidayah, and H. Ohwada, "Comparison of Feature Selection Methods to Classify Inhibitors in DUD-E Database," in *Procedia Computer Science*, Jan. 2018, vol. 144, pp. 194–202. https://doi.org/10.1016/j.procs.2018.10.519
[29] S. Kim *et al.*, "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1388–D1395, Jan. 2021. https://doi.org/10.1093/nar/gkaa971
[30] A. Capecchi, D. Probst, and J. L. Reymond, "One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome," *J. Cheminform.*, vol. 12, no. 1, pp. 1–15, Jun. 2020. https://doi.org/10.1186/s13321-020-00445-4
[31] N. Hecker *et al.*, "SuperTarget goes quantitative: Update on drug-target interactions," *Nucleic Acids Res.*, vol. 40, no. D1, Jan. 2012. https://doi.org/10.1093/nar/gkr912
[32] T. Mancini, I. Melatti, and E. Tronci, "Any-horizon uniform random sampling and enumeration of constrained scenarios for simulation-based formal verification," *IEEE Trans. Softw. Eng.*, 2021. https://doi.org/10.1109/TSE.2021.3109842
[33] A. Salazar, L. Vergara, and G. Safont, "Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets," *Expert Syst. Appl.*, vol. 163, p. 113819, Jan. 2021. https://doi.org/10.1016/j.eswa.2020.113819
[34] Y. Peng and M. H. Nagata, "An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data," *Chaos, Solitons & Fractals*, vol. 139, p. 110055, Oct. 2020. https://doi.org/10.1016/j.chaos.2020.110055
[35] M. Rahman, Y. Cao, X. Sun, B. Li, and Y. Hao, "Deep pre-trained networks as a feature extractor with XGBoost to detect tuberculosis from chest X-ray," *Comput. Electr. Eng.*, vol. 93, p. 107252, Jul. 2021. https://doi.org/10.1016/j.compeleceng.2021.107252
[36] M. A. Mallah *et al.*, "Polycyclic aromatic hydrocarbon and its effects on human health: An overeview," *Chemosphere*, vol. 296, p. 133948, Jun. 2022. https://doi.org/10.1016/j.chemosphere.2022.133948