9th International Conference on Computer Science and Computational Intelligence 2024 (ICCSCI 2024)

Leveraging Machine Learning to Analyze Social Media Use and Understanding the General Behavior in ADHD Communities

Tisha Jillian[a], Tara Nirmala Kusuma[a], Ghinaa Zain Nabiilah[a]*, Jurike V Moniaga[a]

[a]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract

Social media is an online platform that allows people to connect, share information, and create communities. It is a way for individuals, groups, and organizations to interact and engage with one another through the internet. ADHD (attention deficit hyperactivity disorder) also known as one of the most common neurodevelopmental disorders of childhood, mostly diagnosed in childhood and lasts into adulthood. People with ADHD have trouble paying attention, controlling their impulsive behavior, and being overly active. ADHD and social media often associate with one another because social media users are more likely to experience new ADHD symptoms. The symptoms are frequently seen in adults who often use social media and they tend to have short attention spans. This study interpreted a survey of active social media users on a likert scale involving the hours of social media usage and questions directed to ADHD symptoms. Classification method is used to elucidate how social media can be used to see how likely someone to have ADHD based on the survey. The result is the link between social media use and ADHD in individuals and which classifier would be the best to be used. Using Classification type of models as we are using supervised learning, we conclude that Support Vector Machines (SVM), Logistic Regression (LR), and Gaussian Naïve Bayes have the stable performance without overfitting or underfitting conditions compared to Random Forest (RF) that have the highest accuracy in training but then dropped on validation and testing leading to potential overfitting conditions.

Keywords: social media; ADHD; likert scale; classification; attention span; Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Gaussian Naïve Bayes, machine learning

* Corresponding author.
   E-mail address: ghinaa.nabiilah@binus.ac.id

1. Introduction

The relationship between social media use and Attention Deficit Hyperactivity Disorder (ADHD) has been a topic of growing interest in recent years. Social media use has become a major problem for someone who is diagnosed with ADHD implies (Panagiotidi & Overton, 2022) [16]. Short attention span can be seen when social media users eagerly scroll their smartphones to access other types of content, this behavior eventually leads to being overly active, trouble paying attention, and suffers from memory loss implies (Thorell et al., 2022) [21]. This behavior will slowly change people's life and their habits. Especially in 2020, at the beginning of the COVID-19 pandemic, social media and the internet was rapidly used. A study by (Bozzola et al., 2022) implies that social media health related problems are frequently found in this period of time, with depression, anxiety, and addiction as their main issues [6].

As found in the study done by (Núñez-Jaramillo et al., 2021), the development of ADHD appears to arise from complex gene-environment interactions influencing neurodevelopment, with certain genetic, environmental, and physiological risk factors converging primarily during sensitive prenatal and early life periods. Sleep disturbances may worsen inattention, impulsivity and hyperactivity as means to remain awake. The study denote that chronic sleep disorders may be one of the main causes of ADHD symptoms [15].

According to (Ríssola et al., 2020) there is some potential link between social media use and ADHD symptoms has been examined in several studies. The health problem regarding depression, anxiety, and addiction is primarily found in subjects whose adolescents have been exposed to social media use in their daily life [18]. The conditions regarding social media use have been found the same as someone who have an ADHD, (Thorell et al., 2022) implies there are significant associations between social media use and higher-level ADHD symptoms with the majority results of 74% from other studies, this symptoms level can directly and indirectly affect their daily life problems [21]. Another notable study by (Dekkers & van Hoorn, 2022) declared the association between ADHD and social media is not specific to cause ADHD symptoms, instead the higher intensity of social media use appears to be related to higher levels of ADHD, along with anxiety and fear of missing out. Hence, rather than causing ADHD as in a disorder, findings indicate that social media use potentially can cause enhanced symptoms of hyperactivity, impulsivity, and inattention [9].

Study conducts by (Dekkers & van Hoorn, 2022) implicate that adolescents with ADHD frequently struggle with offline social problems like rejection by peers and difficulties maintaining close friendships [9]. ADHD may use excessive social media to escape from unpleasant family situations or may project their negative feelings onto social media interactions implies (Dekkers & van Hoorn, 2022) [9].

(Dekkers & van Hoorn, 2022) emphasize they may minimize the imbalance between reward and cognitive control systems, which could lead to a decrease in problematic social media habits, although their real impact on social media use have not been evaluated [9]. However, some patients may experience negative side effects from these medications stated by (Dekkers & van Hoorn, 2022) [9].

This interest has led many people to question and wonder if that is the case nowadays and starting to analyse the relation between ADHD and people's behavior leading to understanding how machine learning could help them to track and analyse those behaviors regarding ADHD symptoms. A study conducted by (Chen et al., 2021) tries to demonstrate and analyze the demographics information from ADHD patients related to predicting ADHD diagnosis. By using 5 types of machine learning method, the experiment evaluates the potential patient historical behavior might be related to the occurrence of ADHD with the average accuracy of ~74.15%, supported by highest accuracy from Decision Tree method that reach 85.507% accuracy and lowest accuracy from K-Nearest Neighbor that reach 59.420% [7]. A study by (Anitha & Thomas Geroge, 2021) investigating ADHD diagnosis from fMRI data using SoftMax regression and Support Vector Machines (SVM) to tune data feature extraction as a combined for ADHD identification process. The proposed Support Vector Machines (SVM) model yields an accuracy of 94% using 200 holdout data, concluding it is feasible to diagnose ADHD subjects [3]. Another study by (Dekkers & van Hoorn, 2022) summarized that adolescents with ADHD are likely to display a high intensity and problematic use of social media. Strengthening the correlation between ADHD and social media use by conducting a meta-analysis and consequently presenting an integrative framework and empirical work to show the most consistent evidence [9].

Most of the research that has been conducted doesn't emphasize the relation between social media use and triggering ADHD directly, and it's difficult to find the same topic of interest as our research to some degree. There's some inadequate solution regarding the study that has been conducted, the conclusion doesn't come out identifying

the relation between social media use and ADHD by using machine learning methods. Study conducted by (Chen et al., 2021) proving that even using 5 types of methods, the highest result accuracy they could get from experimental on tha main assessment data are 82.609% by using Decision Tree continued by Random Forest (RF) with accuracy of 81.159% [7]. Implying that none of these methods could reach 100% accuracy with those data. Another problem occurs in research conducted by (Dekkers & van Hoorn, 2022) discussing longitudinal studies and meta-analyses rather than machine learning methods [9]. These problems arise from the difference of the dataset and the difference between methods used to find the best accuracy for this topic.

One notable study by (Maniruzzaman et al., 2022) applied machine learning to predict children with ADHD using behavioral activity data. The data utilized for the study is a nationally representative survey based on child health and well-being. The authors used Logistic Regression (LR) as feature selection method and compared eight machine learning classifiers including Random Forest (RF) and found that Random Forest (RF) classifier achieved the highest accuracy of ~87% in predicting ADHD [14].

Future research may focus on interventions that specifically target problematic social media use in individuals with ADHD and adapting existing evidence-based ADHD treatments to address this issue. (Maniruzzaman et al., 2022) applied machine learning algorithms such as Logistic Regression (LR) and Random Forest (RF) to predict children with ADHD using behavioral activity data. In this research, our focus will be on applying machine learning algorithms, namely Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes, and Support Vector Machines (SVM) on social media behavior survey data, then comparing them based on their effectiveness [14].

2. Methodology

To find out how social media use affects an individual; this section outlines the process of analyzing social media use dataset by using a classification model to predict and find the patterns of the likelihood of an individual experiences ADHD symptoms. Furthermore, this section also provides more information about the dataset used and the detailed method.

2.1 Dataset

The dataset used in this research is collected through a questionnaire done originally in purpose of finding potential relationship between social media use and user psychological well-being (Ahmed, 2022) [1]. The questionnaire itself consists of 7 features related to individual's identity, and 12 Likert scale-based questions measuring the intensity of various aspects of mental health. In the scale of 1 to 5, higher scores equal worse mental health conditions and lower scores equal better mental health conditions.

Each question collected crucial data to be worked on in our research. Each individual identity remains anonymous, collected data is not personal. Questions regarding the individual identity are age, gender, relationship status, occupation status, organizations affiliated, commonly used social media, and time spent on social media per day. Table 1 provides the questions related to mental health in the dataset.

Table 1. Visualization of the mental health related questions on the dataset.

9. How often do you find yourself using Social media without a specific purpose?	10. How often do you get distracted by Social media when you are busy doing something?	11. Do you feel restless if you haven't used Social media in a while?	12. On a scale of 1 to 5, how easily distracted are you?	13. On a scale of 1 to 5, how much are you bothered by worries?	14. Do you find it difficult to concentrate on things?	15. On a scale of 1-5, how often do you compare yourself to other successful people through the use of social media?	16. Following the previous question, how do you feel about these comparisons, generally speaking?	17. How often do you look to seek validation from features of social media?	18. How often do you feel depressed or down?	19. On a scale of 1 to 5, how frequently does your interest in daily activities fluctuate?	20. On a scale of 1 to 5, how often do you face issues regarding sleep?

- Questions 9, 10, 12, 14 assess symptoms of ADHD such as purposeless social media use, being easily distracted, and difficulty concentrating.
- Questions 11 and 13 evaluate characteristics of anxiety by asking questions about restlessness when not using social media and being bothered by worries.
- Questions 15, 16, 17 explore concerns related to self-esteem issues, comparing themselves to others on social media, how those comparisons affect their feelings, and seeking validation through social media.
- Question 18, 19, 20 check for signs of depression, including feeling of depression, loss of interest, and trouble sleeping.

2.2 Outlined methods

The research flow done in this research is represented in Fig. 1. Begins with data preprocessing that is segmented into smaller steps, followed by model development, training, evaluation, testing, performance comparison and analysis.
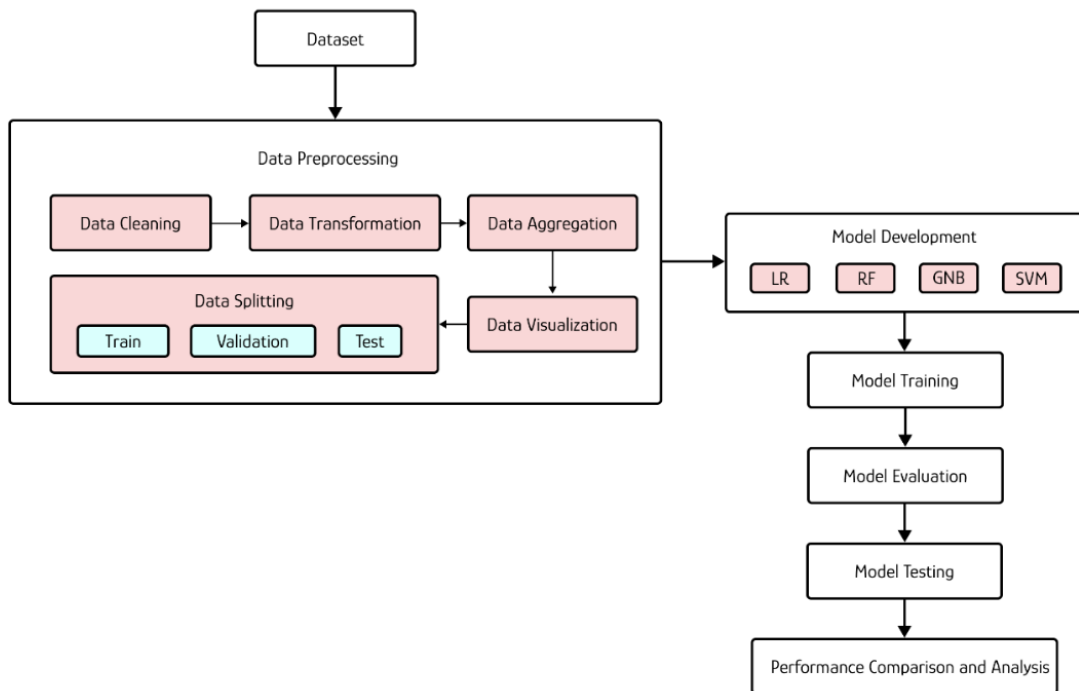

Fig. 1. The flowchart for this research.

Based on the dataset, we aim to classify by using machine learning model that is known and has been demonstrated by most of the researchers, (Sheykhmousa et al., 2020) implies it might be impossible to find the best classification accuracy by only performing a single classifier rather than building an ensemble of classifiers [19]. Naive Bayes, Logistic Regression (LR), Support Vector Machines (SVM), and RF are commonly used for supervised learning and uses a training set to teach models to yield the desired output; the dataset used are labeled data sets used in supervised learning.

Each of these machine learning models have different approaches, they play an essential role to identify and classify dataset to reach the best accuracy and avoid any overfitting and underfitting. Stated by (Ibrahim & Abdulazeez, 2021) many algorithms have good accuracy for predicting and classifying but those algorithms accuracy may change due to many aspects such as datasets, feature selection, and any other aspects [12].

2.2.1 Data preprocessing

Data preprocessing is crucial to be done to improve data quality for better model performance (Duong & Nguyen-Thi, 2021) [10]. The dataset used in this research is not yet pre-processed or can be named as raw. Hence, data preprocessing is necessary to be done to get a good result. As represented in the Fig. 1., steps involved inside data preprocessing in this research include:

- Data Cleaning covers features rearrangement, irrelevant features removal, missing values handling, remove duplication, and fixing inconsistencies.
- Data Transformation involves data normalization and perform feature scaling to convert features values to a similar scale.
- Data Aggregation creates new features from existing ones to improve model performance. The raw dataset contains questions other than ADHD, but for this research purpose, only ADHD questions are taken.
- Data Visualization uses plots to visualize features and their relationships.

Data Splitting splits the dataset into training, validation, and testing datasets done with "sklearn.model_selection" libraries "train_test_split", applying the random state of 42, with splitted dataset portions of 80% for training dataset, 10% for validation dataset, and 10% for test dataset. The models are going to predict less to no or higher to severe ADHD symptoms experience, based on features namely Age, Gender, and Time Spent per Day on Social Media. Hence, the target class is either 0 (less to no ADHD symptoms) and 1 (higher to severe ADHD symptoms). Table 2 presents the amount of each portion of splitted dataset.

Table 2. Dataset split distribution.

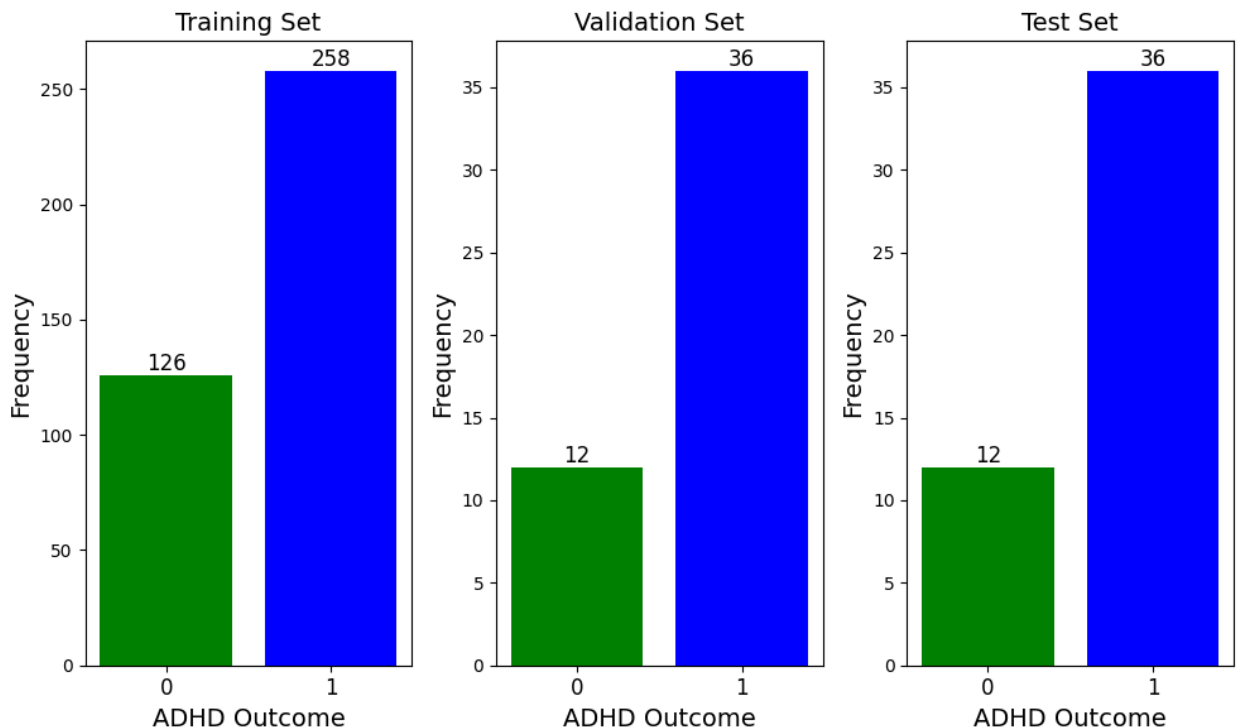| Train	Validation	Test
384	48	48

Fig. 2. Frequency of target class in each dataset split.

2.2.2 Model development

Likert scale dataset was used in this topic, this type of dataset for the most part uses classification as it is the most appropriate to apply implies (Janda & Endresen, 2017) [13]. There are a few models that have been used for this classification algorithm in machine learning, for instance there are Logistic Regression (LR), Naive Bayes, K-Nearest Neighbors, Random Forest (RF), and Support Vector Machines (SVM). These models have different types of approaches to maximize their accuracy and their training test, most of the models use scarce data. These types of models are commonly used for supervised learning algorithms as we use spam detection models by train databases to recognize patterns or anomalies in new data to organize spam and non-spam-related correspondences effectively. This research uses models and supporting tools imported from sklearn library.

2.2.2.1 Logistic Regression (LR)

Logistic Regression (LR) finds the correlation between features and the probability of a particular outcome by using classification algorithms. (Choi et al., 2020) stated that Logistic Regression (LR) uses sigmoidal curves to estimate probability and can be used as an analytic tool for testing hypotheses [8].

This research Logistic Regression (LR) model used lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) solver to find the optimal coefficients for Logistic Regression (LR), with maximum iterations of 100. The choice of lbfgs solver for Logistic Regression (LR) in this research is justified by its efficiency. The iterations limit choice is made 100 to manage computational cost and prevents the algorithm from running indefinitely.

2.2.2.2 Random Forest (RF)

A study conducts by (Sheykhmousa et al., 2020) developed by Brieman, Random Forest (RF) solves classification and regression problems and is known as an ensemble learning approach by integrating multiple models to boost accuracy. Boosting reform the previous error sequence from each model, also known as the process of building a sequence model while bagging is designed to improve the stability and accuracy while reducing variance [19].

The crucial hyperparameter "n_estimators" used for this research Random Forest (RF) model is 500. The choice of 500 trees is to improve accuracy and stability in predictions.

2.2.2.3 Gaussian Naïve Bayes

Gaussian Naïve Bayes is a simple method for creating classifiers based on probabilistic approaches that assume each class follows a normal distribution; naïve bayes can be trained effectively implies (Tan, 2021) [20].

2.2.2.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) uses classification and regression analysis to analyze data by using supervised learning models that are associated with learning algorithms. Even though Support Vector Machines (SVM) takes a lot of time for training, it has been reported that SVMs work better for classification stated by (Boateng et al., 2020) [5].

The chosen kernel function based on the dataset we used is Radial Basis Function (RBF) kernel because SVM is needed here for performing classification tasks, and the dataset used is nonlinearly separable.

2.2.3 Model performance and analysis

Carrying out training, evaluation, and testing for each model. Developing Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes, and Support Vector Machines (SVM), those are models that are going to be trained using the training dataset. After fitting the model, performance accuracy deployed to measures the correct predictions for further comparisons. Evaluate the Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes, and Support Vector Machines (SVM) trained models using the validation dataset. Evaluation is done to explore how well the

models perform and identify weaknesses. Steps after the models predicted the validation dataset are to evaluate the results including accuracy, precision, recall, and F1-score of the models' performance. Test the trained Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes, and Support Vector Machines (SVM) models on the testing dataset to ensure the correctness, accuracy, and capability of trained models to measure the required success parameters, utilizing classification report from sklearn library is needed to visualize complete accuracy, precision, recall, and F1-score table. From the obtained results, a comparison to the performance of Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes, and Support Vector Machines (SVM) models. Further result analysis and assessing the best model for the dataset based on the learned evaluation metrics.

3. Results and Discussion

The process of this research was done in Google Colab, using python programming language. It is built on top of Jupyter Notebooks, which are interactive computing environments that allow users to create and share documents containing live code and visualizations make it an ideal environment for developing and experimenting with machine learning models like Logistic Regression (LR), Random Forest (RF), Gaussian Naïve Bayes, and Support Vector Machines (SVM). The models are processed based on training, validation, and testing dataset that contains 3 features namely Age, Gender, and Time Spent per Day on Social Media, and target class of 0 (less to no ADHD symptoms) or 1 (higher or severe ADHD symptoms). Table 3 represents accuracy score of each model on training, validation, and testing. Table 4 represents precision, recall, and F1 score of each model on training, validation, and testing.

Table 3. Accuracy Score of Each Model

Classifier Model	Accuracy		
	Train	Validation	Test
Logistic Regression	**0.750000**	**0.812500**	**0.791667**
Random Forest	0.856771	0.687500	0.666667
Gaussian Naïve Bayes	0.736979	0.770833	0.770833
SVM	**0.718750**	**0.791667**	**0.791667**

Based on Table 3, conclude that Logistic Regression (LR) shows an accuracy of 0.75000 on the training set, 0.81250 on the validation set, and 0.791667 on the test set. This shows that the model generalizes well from training data to unseen data. Random Forest (RF) has a higher accuracy on the training set at 0.85671, that drops to 0.68750 on validation set, then drops again to 0.666667 on the test set. The notable drop from training to validation to test suggests that overfitting potentially happened. Gaussian Naïve Bayes has an accuracy of 0.736979 on the training set, 0.770833 on the validation set, and 0.770833 on the test set. The model shows consistent accuracy from validation to test, indicating good generalization. Support Vector Machines (SVM) has an accuracy of 0.718750 on the training set, which improves to 0.791667 on both the validation and test sets. The improvement suggests that the model may be more conservative and generalizes well. Support Vector Machines (SVM) models can improve their accuracy and it can be seen on our accuracy sets, both validation and test sets have been enhanced compared to the training test. This improvement may happen because of a few factors such as sample size, data distribution, feature compatibility, and preprocessing data.

Table 4. Evaluation Report of Each Model

Classifier Model	Precision			Recall			F1 Score		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Logistic Regression	0.766447	0.885714	0.825000	0.903101	0.861111	0.916667	0.829181	0.873239	0.868421
Random Forest	0.848797	0.800000	0.763158	0.957364	0.777778	0.805556	0.899818	0.788732	0.783784
Gaussian Naïve Bayes	0.766102	0.857143	0.837838	0.875969	0.833333	0.861111	0.817360	0.845070	0.849315
SVM	0.727273	0.825000	0.782609	0.930233	0.916667	1.000000	0.816327	0.868421	0.878049

The evaluation report of each model in Table 4, indicates that Logistic Regression (LR) model is quite balanced between precision and recall, especially on unseen data. The Random Forest (RF) model shows good performance on

the training set but the performance drops on the validation and test sets, which may indicate overfitting. Gaussian Naïve Bayes performs relatively consistent across all datasets. Support Vector Machines (SVM) model shows a strong recall, especially on the test set, and a good balance between precision on recall as shown in the F1 score.

Based on evaluation metrics presented in Table 3 and Table 4, (Pandria et al., n.d.) implies that each classifier optimizes all the validation scores by reporting the performance metrics of accuracy, precision, recall, and F1 scores [17]. Logistic Regression (LR) and Support Vector Machines (SVM) models show a good balance between precision and recall, as presented by their F1 scores, and they generalize well to unseen data based on their accuracy score. Meanwhile, Random Forest (RF) shows potential overfitting, as indicated by the high accuracy and F1 score on the training set but lower scores on the validation and test sets. Gaussian Naïve Bayes shows consistent performance across all datasets but does not achieve the highest accuracy or F1 score compared to the other models. Other than Random Forest (RF), all models train and validation data test indicates that they have steady result concluding that they do not undergo any underfitting, (Guleria & Sood, 2023) explain it as models perform poorly on training data or overfitting as it learns the irrelevant data hence the performance on training data increases [11].

Table 5. Accuracy Score Comparison

Classifier Model	Our Model Accuracy	(Bhatnagar et al., 2022) Model Accuracy
Gaussian Naïve Bayes	77,08%	71,05%
SVM	79,16%	75.55%

In previous study by (Bhatnagar et al., 2022) stated that machine learning algorithms such as Decision Trees, Naïve Bayes, Random Forest (RF), and Support Vector Machines (SVM) almost perform equally good, with Support Vector Machines (SVM) by being the second best-performance models with accuracy of 75.55% compared to ours with accuracy of 79.16% [4] however, based on our evaluation report indicates that Logistic Regression (LR) and Support Vector Machines (SVM) have the best performance.

Another study stated by (Alsharif et al., 2024) declares that Support Vector Machines (SVM) based classifiers attained the best-performing models as it was able to recognize many instances and identify risk factors associated with symptoms of adult ADHD [2]. This statement could support the result of our accuracy test that emphasizes Support Vector Machines (SVM) have the best performing models for classifying.

4. Conclusion

The complete process of this research indicates that an individual's time spent per day on social media tends to link to an individual's ADHD symptoms severity, the more of an individual spent time per day on social media, the higher the chance of them to experience ADHD symptoms. Those ADHD symptoms include hyperactivity, short attention-span, impulsivity, inattention, and anxiety, known as general or common behavior for someone with ADHD. Someone with ADHD usually has a hard time doing their daily life activity as they are more likely to get distracted, not only could affect their mental health but their capability of doing daily activities as well. The age and gender of an individual does not lessen chance of an individual to experience ADHD symptoms, various age and gender can all experience ADHD symptoms.

According to the predictive modelling performance evaluation report, Logistic Regression (LR) and Support Vector Machines (SVM) seem to appear to be the best-performing models in terms of generalization and balanced metrics. Random Forest (RF) might require further adjustment to address overfitting, and for Gaussian Naïve Bayes, although it is reliable, it may gain increasing accuracy from feature engineering or model adjustments.

This research not only explores the importance of machine learning in daily life basis, but also to help ADHD community to stand firm and help them to get better. This paper may help other potential researchers to improve their study and give new ideas for a new topic research. Suggestion for the next paper, utilizing machine learning for community that struggle with their mental health by using other models beside Support Vector Machines (SVM) and Gaussian Naïve Bayes.

References

[1] Ahmed, S. (2022, April). *Social Media and Mental Health: Correlation between Social Media use and General Mental Well-being*. https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health/data/version/1

[2] Alsharif, N., Al-Adhaileh, M. H., & Al-Yaari, M. (2024). Accurate Identification of Attention-deficit/Hyperactivity Disorder Using Machine Learning Approaches. *Journal of Disability Research*, *3*(1). https://doi.org/10.57197/jdr-2023-0053

[3] Anitha, S., & Thomas Geroge, S. (2021). Adhd Classification from FMRI Data Using Fine Tunining in SVM. *Journal of Physics: Conference Series*, *1937*(1). https://doi.org/10.1088/1742-6596/1937/1/012014

[4] Bhatnagar, S., Agarwal, J., & Sharma, O. R. (2022). Detection and classification of anxiety in university students through the application of machine learning. *Procedia Computer Science*, *218*, 1542–1550. https://doi.org/10.1016/j.procs.2023.01.132

[5] Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machines, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, *08*(04), 341–357. https://doi.org/10.4236/jdaip.2020.84020

[6] Bozzola, E., Spina, G., Agostiniani, R., Barni, S., Russo, R., Scarpato, E., Di Mauro, A., Di Stefano, A. V., Caruso, C., Corsello, G., & Staiano, A. (2022). The Use of Social Media in Children and Adolescents: Scoping Review on the Potential Risks. *International Journal of Environmental Research and Public Health*, *19*(16). https://doi.org/10.3390/ijerph19169960

[7] Chen, T., Antoniou, G., Adamou, M., Tachmazidis, I., & Su, P. (2021). Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Machine Learning. *Applied Artificial Intelligence*, *35*(9), 657–669. https://doi.org/10.1080/08839514.2021.1933761

[8] Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2). https://doi.org/10.1167/tvst.9.2.14

[9] Dekkers, T. J., & van Hoorn, J. (2022). Understanding Problematic Social Media Use in Adolescents with Attention-Deficit/Hyperactivity Disorder (ADHD): A Narrative Review and Clinical Recommendations. In *Brain Sciences* (Vol. 12, Issue 12). MDPI. https://doi.org/10.3390/brainsci12121625

[10] Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1). https://doi.org/10.1186/s40649-020-00080-x

[11] Guleria, P., & Sood, M. (2023). Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies*, *28*(1), 1081–1116. https://doi.org/10.1007/s10639-022-11221-2

[12] Ibrahim, I., & Abdulazeez, A. (2021). The Role of Machine Learning Algorithms for Diagnosing Diseases. *Journal of Applied Science and Technology Trends*, *2*(01), 10–19. https://doi.org/10.38094/jastt20179

[13] Janda, L. A., & Endresen, A. (2017). Five statistical models for Likert-type experimental data on acceptability judgments. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *3*(2), 217–250. https://doi.org/10.1558/jrds.30822

[14] Maniruzzaman, M., Shin, J., & Hasan, M. A. M. (2022). Predicting Children with ADHD Using Behavioral Activity: A Machine Learning Analysis. *Applied Sciences (Switzerland)*, *12*(5). https://doi.org/10.3390/app12052737

[15] Núñez-Jaramillo, L., Herrera-Solís, A., & Herrera-Morales, W. V. (2021). Adhd: Reviewing the causes and evaluating solutions. In *Journal of Personalized Medicine* (Vol. 11, Issue 3, pp. 1–25). MDPI AG. https://doi.org/10.3390/jpm11030166

[16] Panagiotidi, M., & Overton, P. (2022). Attention deficit hyperactivity symptoms predict problematic mobile phone use. *Current Psychology*, *41*(5), 2765–2771. https://doi.org/10.1007/s12144-020-00785-2

[17] Pandria, N., Petronikolou, V., Lazaridis, A., Karapiperis, C., Kouloumpris, E., Spachos, D., Fachantidis, A., Vasileiou, D., Vlahavas, I., & Bamidis, P. (n.d.). *An Information System for Symptom Diagnosis and Improvement of Attention Deficit Hyperactivity Disorder: The ADHD360 Project*. https://doi.org/10.2196/preprints.40189

[18] Ríssola, E. A., Aliannejadi, M., & Crestani, F. (2020). Beyond Modelling: Understanding Mental Disorders in Online Social Media. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12035 LNCS*, 296–310. https://doi.org/10.1007/978-3-030-45439-5_20

[19] Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support Vector Machines Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (Vol. 13, pp. 6308–6325). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/JSTARS.2020.3026724

[20] Tan, H. (2021). Machine Learning Algorithm for Classification. *Journal of Physics: Conference Series*, *1994*(1). https://doi.org/10.1088/1742-6596/1994/1/012016

[21] Thorell, L. B., Burén, J., Ström Wiman, J., Sandberg, D., & Nutley, S. B. (2022). Longitudinal associations between digital media use and ADHD symptoms in children and adolescents: a systematic literature review. In *European Child and Adolescent Psychiatry*. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00787-022-02130-3